Building Human-Like Machine Intelligence: Advancing Attention by Modeling, Alignment, and Explainability

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Xianyu Chen

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Catherine Qi Zhao

November, 2025

Acknowledgements

Pursuing a doctoral degree has been a demanding yet profoundly rewarding journey that I will never forget. During this journey, there are a great number of people who support me, and therefore I would like to express my deepest gratitude to all of you.

First, I want to thank my advisor, Professor Catherine Qi Zhao. Your recognition of my potential in computer vision and your offer to join your research group provide an opportunity to start pursuing my scientific career. Without your influence, I might have continued on a different path as an engineer with a more stable but less interesting life. It is a pleasure to work under your supervision, which has greatly shaped my technical skills and my aspirations for a research career.

Moreover, I would like to thank my doctoral committee: Professors Pen-Chung Yew, Yao-Yi Chiang, and Changhyun Choi. Your valuable insights and constructive feedback during my preliminary exam, thesis proposal, and defense have been essential in refining the quality and direction of my work.

This dissertation would not have been possible without the collaboration and support of many colleagues. I am grateful to Dr. Ming Jiang, Dr. Shi Chen, Dr. Yan Luo, Yifeng Zhang, James Yang, and Hunmin Lee in the Vision & Image Processing Lab from the University of Minnesota for their help and encouragement. I want to thank Brian Lim and Dr. Zhi Yang at Fasikl Incorporated for their productive discussions and excellent collaboration.

Some of my most valued experiences over the past few years have come from the time spent with my friends. I am deeply grateful to Kuan Wu, Jianrui Xie, and Shaoming Xu, listed here in alphabetical order, for the many thoughtful and stimulating conversations we shared. Although we were not academic collaborators, your encouragement and support contributed meaningfully to my progress throughout this Ph.D. journey. I also

would like to express my sincere appreciation to Nuo Chen, Hanlin Ding, Xinchen Fan, Biao Kang, Jiawei Luo, Tianci Song, Xinze Wan, Ruichen Wang, Haoyu Yang, Ruipeng Zhao, and Zifeng Zhao, listed alphabetically. Living near campus together provided countless memorable experiences that I will always cherish.

And of course, I'm most thankful to my family for their endless love and support. To my parents, Xiling Chen and Ruoling Zhou-thank you for giving me a warm home, teaching me invaluable life lessons, and shaping me into who I am today.

Dedication

To those who held me up over the years.

Abstract

A long-standing ambition of Artificial Intelligence (AI) is to build human-like machine intelligence capable of solving real-world tasks in general and complex circumstances. In recent years, progress in machine learning-made possible by the availability of large datasets and the increasing power of Graphics Processing Units (GPUs)-has brought major changes to fields such as computer science, neuroscience, and healthcare. In certain benchmark tasks and real-world applications-such as the game of Go and AlphaFold for protein structure prediction-intelligent systems have even surpassed human performance. However, discrepancies between human and machine intelligence remain. Much of this gap is driven by the following fundamental challenges. (1) The mechanisms by which the human brain processes visual information and produces intelligence remain largely unknown, and the application of neuroscience discoveries to advance machine intelligence is still poorly understood; (2) Human intelligence, developed through evolution and experience, possesses innate capabilities that progress from perceiving and understanding to active learning, reasoning, and planning, and ultimately to creativity, critical thinking, and problem-solving. In contrast, current machine intelligence is limited to narrow and well-structured tasks, with no evidence suggesting it can replace human intelligence in open-ended and free-form real-world applications; (3) Evidence suggests that current machine intelligence relies heavily on statistical priors rather than genuinely reasoning from observations, functioning as a black box. This lack of transparency and interpretability limits its applicability in scenarios that require reasoning and justification.

With the overarching goal of developing machine intelligence with human-like capabilities, we first leverage the visual attention as an interface to investigate how humans and machines prioritize their focus when addressing general and complex problems. Furthermore, we aim to uncover the reasons behind how humans allocate attention to visual stimuli when performing tasks, shedding light on the underlying psychological and cognitive mechanisms to transform black-box models into white-box systems with interpretable and transparent decision-making. Our works focus on the following three aspects of visual attention, *i.e.*, modeling, alignment, and explainability, all of which are

designed to bridge the gap between human attention and machine intelligence. More specifically: (1) **Modeling:** First, we aim to understand task-driven spatio-temporal patterns of eye fixations (a.k.a. scanpaths) and individual attention patterns, which is still a crucial yet underexplored area of research in artificial intelligence and psychology; (2) **Alignment:** We investigate how to align human attention and machine attention to understand real-world problems and hence provide step-by-step solutions, which aims to enable machine intelligence with human-like capabilities; (3) **Explainability:** Finally, while attention mechanisms shape human eye fixations during visual exploration, the reason behind these fixations is still not fully understood. Closing this gap is vital for making human intelligence more explainable and for applying such understanding to improve machine intelligence. Building on these efforts, this dissertation presents observations that underscore the importance of human attention and its relevance to creating human-like computational models. By drawing on insights from human vision, it contributes to the AI systems that are interpretable, transparent, and trustworthy, and that can address real-world challenges.

Contents

Acknowledgements **i**

Dedication **iii**

Abstract **iv**

Contents **vi**

List of Tables **x**

List of Figures **xiii**

1 Introduction **1**

1.1 The Gap between Machine and Human Intelligence 3

1.2 Our Approaches 4

1.2.1 Modeling 5

1.2.2 Alignment 5

1.2.3 Explainability 6

1.3 Contributions 7

1.4 Publications 9

2 Predicting Human Scanpaths in Visual Question Answering **10**

2.1 Introduction 11

2.2 Related Work 13

2.3 Method 15

2.3.1 Network Architecture 15

2.3.2 Objective . . . 17
2.4 Experiments . . . 19
2.4.1 Experiment Settings . . . 19
2.4.2 Are the predicted scanpaths plausible? . . . 21
2.4.3 What contributes to the model's performance? . . . 24
2.4.4 What do the predicted scanpaths fixate? . . . 24
2.4.5 Which VQA model is the most effective? . . . 25
2.4.6 Does the proposed method generalize? . . . 26
2.5 Conclusion . . . 27

3 Beyond Average: Individualized Visual Scanpath Prediction 28
3.1 Introduction . . . 29
3.2 Related Works . . . 31
3.2.1 Eye-Tracking Datasets . . . 32
3.2.2 Visual Scanpath Prediction . . . 33
3.3 Methodology . . . 33
3.3.1 Observer Encoding . . . 33
3.3.2 Observer-Centric Feature Integration . . . 35
3.3.3 Adaptive Fixation Prioritization . . . 36
3.4 Experiments . . . 37
3.4.1 Experiment Settings . . . 37
3.4.2 Quantitative Results . . . 40
3.4.3 Ablation Study . . . 41
3.4.4 Qualitative Examples . . . 41
3.4.5 From Scanpaths to Saliency Maps . . . 43
3.4.6 Semantic Analyses . . . 43
3.4.7 Application . . . 45
3.5 Conclusion . . . 46

4 VisualHow: Multimodal Problem Solving 48
4.1 Introduction . . . 49
4.2 Related Work . . . 51
4.2.1 Visual Captioning and Storytelling . . . 51

4.2.2 Visual Question Answering and Dialog . . . 52
4.2.3 Multimodal Instructions . . . 53
4.2.4 Multimodal Representation Learning . . . 53
4.3 The VisualHow Dataset . . . 54
4.3.1 Problems and Solutions . . . 54
4.3.2 Data Annotation . . . 55
4.3.3 Data Statistics . . . 57
4.4 Experiments . . . 61
4.4.1 Baseline Models . . . 61
4.4.2 Task 1: Solution Steps Prediction . . . 61
4.4.3 Task 2: Solution Graph Prediction . . . 63
4.4.4 Task 3: Problem Description Generation . . . 65
4.4.5 Task 4: Solution Captions Generation . . . 66
4.5 Conclusion . . . 67

5 Every Problem, Every Step, All In Focus: Learning to Solve Vision-Language Problems with Integrated Attention 68
5.1 Introduction . . . 70
5.2 Related Works . . . 72
5.2.1 Problem Solving Methods . . . 72
5.2.2 Attention in Vision-Language Tasks . . . 73
5.2.3 Supervised Learning of Attention . . . 73
5.3 Problem Statement . . . 74
5.4 Method . . . 75
5.4.1 Solution Graph Attention Network . . . 76
5.4.2 Integrated Attention Mechanism . . . 78
5.4.3 Learning Objectives . . . 80
5.5 Experiments . . . 83
5.5.1 Experimental Setup . . . 83
5.5.2 Quantitative Results . . . 87
5.5.3 Qualitative Results . . . 88
5.5.4 Performance Analyses . . . 90

5.6 Conclusion 101

6 GazeXplain: Learning to Predict Natural Language Explanations of Visual Scanpaths 103

6.1 Introduction 104
6.2 Related Work 106
6.3 GazeXplain 107
6.3.1 Data 107
6.3.2 Model 110
6.4 Experiments 114
6.4.1 Experimental Setup 114
6.4.2 Scanpath Prediction Results 115
6.4.3 Ablation Study for Scanpath Prediction and Explanation 115
6.4.4 Scanpath Explanation Results 117
6.4.5 Qualitative Analysis 119
6.5 Conclusion 120

7 Conclusion and Discussion 122

8 List of All Publications 124

References 126

List of Tables

2.1 Scanpath prediction results on the AiR dataset (VQA). In each panel, the first row indicates the correct scanpaths and the second row indicates the incorrect scanpaths. The best results are highlighted in bold. Underlines indicate scores above human performance. 21

2.2 Ablation study of TG, SCST and CDL on the AiR dataset. In each panel, the first row indicates the correct scanpaths and the second row indicates the incorrect scanpaths. The best results are highlighted in bold. 23

2.3 Percentage of fixations in ROI, non-ROI, and background. In each panel, the first row indicates the correct scanpaths and the second row indicates the incorrect scanpaths. 25

2.4 Performances on the OSIE dataset (free-viewing). The best results are highlighted in bold. Underlines indicate scores above human performance. 26

2.5 Performances on the COCO-Search18 dataset (visual search). The best results are highlighted in bold. Underlines indicate scores above human performance. 26

3.1 Comparison of value-based evaluation results for models' ability to predict the scanpaths of individual observers. 37

3.2 Comparison of ranking-based evaluation results for models' ability to distinguish different observers. 39

3.3 Ablation study for the proposed technical components: observer encoder (OE), observer-centric feature integration (FI), and adaptive fixation prioritization (FP). 39

3.4 Spearmans' correlation coefficients of fixation proportions in 3 semantic ROIs (*i.e.*, social, nonsocial, and background) between the ground truth and predictions. Bold numbers indicate significant positive correlations ($p < 0.05$). 44
4.1 Comparison between VisualHow and related datasets. 57
4.2 Most common tokens in the caption among 1) nouns; 2) verbs; 3) other parts of speech (POS). 58
4.3 Most common tokens in the annotated phrases among 1) nouns; 2) verbs; 3) other parts of speech (POS). 60
4.4 Quantitative results of Task 1: solution steps prediction. 60
4.5 Quantitative results of Task 1: solution steps prediction (with attention supervision). 63
4.6 Quantitative results of Task 2: solution graph prediction. 65
4.7 Quantitative results of Task 3: problem description generation. 66
4.8 Quantitative results of Task 4: solution captions generation. 67
5.1 Solution graph prediction results from retrieval and dependency aspects. In each panel, the first row (I) indicates the image modality and the second row (C) indicates the caption modality. The best results are highlighted in bold. 87
5.2 Solution graph prediction results for sequential and non-sequential solutions. The best results are highlighted in bold. 91
5.3 Intra-step attention evaluation results. The best results are highlighted in bold. 92
5.4 Inter-step attention evaluation results. The best results are highlighted in bold. 93
5.5 Pearson's r between attention evaluation score and our proposed SGAN model's performance. Bold numbers indicate significant positive correlations ($p < 0.05$). 94
5.6 Ablation study of the number of integrated attention layers. The best results are highlighted in bold. 94

5.7 Evaluations on Intra-step attention, Inter-step attention, and solution graph prediction results across layers. The best results are highlighted in bold. 95
5.8 Ablation study of the proportion of fine-grained data annotations. The best results are highlighted in bold. 95
5.9 Solution graph prediction results with different sources of attention annotations. In each panel, the first row (I) indicates the image modality and the second row (C) indicates the caption modality. The best results are highlighted in bold. 98
5.10 Comparison of multimodal procedure planning models. The best results are highlighted in bold. 99
5.11 Ablation study of similarity functions used in the proposed evaluation metrics. 99
5.12 Ablation study on different combinations of dependency threshold λ_d and retrieval threshold λ_r. The best results are highlighted in bold. 100
6.1 Statistics of the eye-tracking datasets with annotated explanations. . . . 109
6.2 Scanpath prediction results. The best results are highlighted in bold. . . 116
6.3 Ablation study on AiR-D [1] for the proposed technical components: language decoder (EXP), semantic alignment (ALN), and cross-dataset cotraining (CT). The best results are highlighted in bold. 116
6.4 Explanation prediction results and diversity analysis. The best results are highlighted in bold. 118

List of Figures

2.1 Visual scanpaths of humans can reveal their decision-making strategies and explain their performance. Those who pay attention to relevant visual cues can achieve high levels of task performance. This example compares the scanpaths of people who succeed or fail to answer a question, where the dots represent fixations. The number and radius indicate the fixation order and duration, respectively. The blue and red dots indicate the beginning and the end of the scanpath, respectively. 11
2.2 Overview of the proposed scanpath prediction network. 13
2.3 Examples of the predicted scanpaths. Each column compares the prediction results and human scanpaths given specific answer correctness. The number and radius indicate the fixation order and duration, respectively. The blue and red dots indicate the beginning and the end of the scanpath, respectively. 22
2.4 Comparison of VQA models' answer accuracy, scanpath accuracy, and machine attention accuracy (bubble size). 25
3.1 Understanding and predicting the distinct eye movements of each observer is the key objective of individualized scanpath prediction. These examples reveal the variations in the scanpaths of different observers, showing their distinct attention preferences in (a) faces, (b) objects, and (c) background. Each dot represents a fixation, with the number and radius indicating its order and duration, respectively. The blue and red dots indicate the beginning and the end of the scanpath, respectively. . 29

3.2 Our proposed method incorporates an observer encoder for characterizing individualized attention traits, followed by observer-centric feature integration for holistic processing, and adaptive fixation prioritization for refined predictions. 32
3.3 Qualitative examples of scanpaths predicted by ChenLSTM-FT, ChenLSTM-ISP, and ground truth. Each row compares the model predictions and the ground truth scanpath of one observer. These observers show different gaze patterns, including (a) focusing on the image center, (b) exploring different people and objects, (c) exploring broadly in the scene, and (d) focusing on a particular region. The blue and red dots indicate the beginning and the end of the scanpath, respectively. 42
3.4 Saliency evaluation results of the baselines, fine-tuned (FT) models, and ISP models. Error bars indicate the standard error of the mean. 43
3.5 Statistical comparison between the predicted fixations for the ASD and Control groups [2]. Error bars indicate the standard error of the mean. Asterisks indicate significant differences (unpaired t-test, $p < 0.05$). . . . 45
3.6 Visualization of features extracted from ISP models (numbers indicate observer identities) and results of ASD classification using the features. 46
4.1 VisualHow is a vision-language task aiming to infer the solution to a real-life problem. The solution consists of multiple steps each described with an image and a caption. 49
4.2 An overview of the VisualHow dataset. We provide a hierarchical structure that organizes our data into categories, sub-categories, problems, solution graphs, steps with image-caption pairs, and multimodal attention. Example steps are highlighted in the solution graph. Steps without a dependency are connected to an empty node. 52
4.3 Crowdsourcing interface of the VisualHow task, which contains 1) an overview of the wikiHow Article, 2) annotation of the multimodal attention, and 3) annotation of the solution graph. 56
4.4 Number of problems in each category and the three types of solution graphs. 58
4.5 Distribution of solution steps and attention annotations. 59

4.6 Qualitative results for attention supervision. Important regions and keywords are highlighted with red and black colors. 64
5.1 Problem-solving tasks such as "how to decorate the tables for a vintage-themed wedding" often follow a non-sequential procedure. For example, steps 1, 3, and 4 can be completed in no particular order, as long as step 1 takes place before step 2, step 4 happens before step 5, and all of them take place before step 6. Our method represents such problem-solving procedures in a graph structure. Steps are represented as nodes, and dependent steps are directly connected by edges indicating ordering constraints. In this way, our approach can handle various types of step dependencies in free-formed procedures. Attention is optimized end-to-end over the full graph-based solution structure. 69
5.2 Overview of the proposed SGAN architecture. The input consists of node representations $\boldsymbol{h}^{(0)}$, features for images/caption candidates $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N$. The network leverages an integrated attention mechanism that progressively processes the input features and predicts the output intra-step attention $\boldsymbol{\alpha}^{(L)}$ for capturing salient information from the input images or captions, the inter-step attention $\boldsymbol{P}^{(L)}$ characterizing the probabilities of dependencies across different steps, and the final updated node representations $\boldsymbol{h}^{(L)}$. 75
5.3 Qualitative comparison of the predicted solution graphs and intra-step attention maps. The green edges in the graphs indicate correct predictions, while the red ones indicate wrong predictions. 89
6.1 This example reveals how observers strategically investigate a scene to find out if the person is walking on the sidewalk. Fixations (circles) start centrally, locating a driving car, then shifting to the sidewalk to find the person, and finally looking down to confirm if the person is walking. By annotating observers' scanpaths with detailed explanations, we enable a deeper understanding of the "what" and "why" behind fixations, providing insights into human decision-making and task performance. . 104

6.2 LLaVA generates the ground-truth explanation for each fixation using an input image with a red circle marking the fixation. The model's response provides information within the marked area, serving as a basis for further refinement. 108
6.3 GazeXplain's architecture consists of a general vision-language encoder and a novel attention-language decoder. The decoder outputs an explanation for each fixation in the predicted scanpath, with a semantic alignment mechanism facilitating the semantic consistency between fixations and explanations. The model is developed on three public datasets using a cross-dataset co-training technique. 110
6.4 ROC analysis of fixations and explanations. 118
6.5 Quantitative examples from GazeXplain compared to Gazeformer and the ground truth. Each row shows scanpaths and explanations of two key fixations. 120

Chapter 1

Introduction

Recent years have witnessed impressive progress and remarkable successes, bringing the world's attention to artificial intelligence (AI). In particular, the rapid development and maturation of deep learning since the beginning of this century have enabled modern machine intelligence systems to perform a broad range of tasks in academia and industry, fostering mutual benefits through their interaction. Convolutional neural networks (CNNs) [3, 4, 5, 6] have achieved state-of-the-art performance in computer vision, achieving human-level or even superhuman performance in a variety of vision challenges, including image recognition [7], object detection [8, 9], object segmentation [10, 11], the game of Go [12, 13], AlphaStar [14] for StarCraft II, and AlphaFold [15] for protein folding. The Transformer [16], proposed eight years ago, continues to demonstrate its capabilities in language understanding and vision representation. It also serves as the foundation with great compatibility for the revolution in large language models (LLMs) [17, 18, 19, 20, 21]/large vision language models (LVLM) [22, 23, 24] and Stable Diffusion (SD) [25, 26, 27]. Its strong compatibility with pretraining and the alignment of vision and language has drawn significant attention from both the computer vision and natural language processing communities, driving efforts to develop foundation models powered by billion-scale datasets [28, 10, 11, 29, 30, 31]. Beyond academic achievements, the development of AI provides industrial companies with unprecedented opportunities to advance their next-generation applications. OpenAI GPT-5 [32, 33] advances conversational AI by integrating methods that let it switch between rapid replies

and deeper reasoning as the context requires. This design enables it to deliver expert-level answers, demonstrate reasoning capabilities similar to a Ph.D. level researcher, and process multimodal inputs in a manner that more closely resembles human perception. Autonomous driving [34, 35] represents a major milestone in AI advancement, as self-driving vehicle technologies are being extensively developed and tested by various companies throughout the United States. Innovations like Neuro-AI therapy [36, 37] are currently transforming healthcare by enabling personalized treatments for different neurological conditions. These significant achievements suggest that the era of artificial general intelligence-an AI capable of solving any intellectual task like a human-may finally become a reality in the foreseeable future.

Despite these impressive advances, many studies reveal critical limitations in current machine intelligence and indicate that achieving general artificial intelligence remains a distant goal. At first, machine intelligence may not fully match the cognitive capabilities of humans, since humans' sensory and motor systems have been refined over hundreds of thousands of years of evolution, interacting with the natural world. In contrast, machine intelligence has limited multi-modality input to observe the world. As mentioned in feature-integration theory [38], humans focus attention on certain features of the environment to bind them together for coherent perception and reasoning, while machines that learn on data with specific statistics may fail to generalize to data with different statistics or novel environment. Furthermore, although machines can outperform their human counterparts in narrow and structured tasks, such as image recognition and visual question answering, they still struggle in free-form and open-ended general real-world tasks. Finally, unlike many machine systems that lean on superficial statistical patterns, people tend to draw on conscious reasoning, intuition, and an ability to adjust to the situation at hand. When thinking is visible in this way, it helps build trust, keeps decision-makers accountable, and supports ethical choices-especially when things get complicated.

The observations above emphasize the essential differences in how machines and humans perceive and interpret the visual world. Machines excel at learning from large datasets and solving specific tasks by identifying underlying correlations within the data of limited modalities. However, the learned knowledge is often task-specific, dataset-dependent, and implicit, limiting its ability to generalize to free-form and open-ended

scenarios. Furthermore, humans have a highly sophisticated ability to select and process information effectively. We believe that developing generalizable and human-like intelligence is an important component for the next generation of AI. In this dissertation, we explore our efforts to advance this goal through improvements in attention modeling, alignment, and explainability, to understand and interact with the environment in a way that mimics human cognitive abilities.

1.1 The Gap between Machine and Human Intelligence

Human intelligence draws on a layered set of innate abilities. It starts with self-awareness, then builds through learning, creativity, and planning, and reaches further into common sense and problem-solving, ending with the capacity for rational skepticism. These abilities have taken shape over the long course of evolution and have been honed through each person's life experiences, giving people the agility to deal with complex situations and to adapt when circumstances change. In contrast, while machine intelligence draws inspiration from the biological structures inside the human visual system, it fundamentally differs in terms of cognitive qualities and abilities [39, 40]. Although machines have surpassed human performance in certain tasks [7, 12, 14, 15], a significant gap remains between machine and human intelligence in the following several critical aspects:

Reasoning Ability. One of the key distinctions between machines and humans lies in their ability to reason. Humans are naturally able to handle complex reasoning across many areas-whether deductive, inductive, abductive, or causal [41, 42, 43, 44, 45]. This skill makes it possible to work with incomplete or uncertain information, draw sound conclusions, and adapt our thinking quickly to unfamiliar situations. Human reasoning is deeply rooted in biological evolution, experiential learning, and an understanding of abstract relationships, enabling us to solve problems and make decisions in real-world, dynamic contexts [46]. By contrast, computational models-though they can outperform in certain specific, well-defined tasks [7, 12, 13]-still fall far behind when it comes to reasoning. Current AI systems heavily rely on pattern recognition and statistical correlations within large datasets, rather than understanding causal relationships or engaging in genuine reasoning. Furthermore, they struggle with reasoning under uncertainty or

transferring reasoning skills across domains to address real-world challenges, which are hallmarks of human intelligence.

Generalizability. Humans naturally generalize their knowledge to unfamiliar environments and diverse tasks, enabling them to adapt effectively. This inherent generalization capability, rooted in both biological evolution and lifelong learning, enables them to integrate knowledge from various contexts to solve complex real-world problems. In contrast, despite being trained on millions of structured samples, computational models struggle significantly with generalization [47, 48, 49], as real-world scenarios frequently present unseen tasks. This pronounced disparity in learning efficiency reflects fundamental differences in how machines and humans process visual information. More critically, it underscores that successes made in narrow, well-structured tasks do not necessarily translate into genuine competence in solving real-world problems.

Interpretability. Humans are not only proficient at conducting comprehensive reasoning and generalizing to complex environments but are also adept at justifying the rationales behind their decisions. Explaining decisions is a unique human capability, requiring communication, understanding, and reasoning. However, existing machine intelligence systems typically rely on black-box deep neural networks to model statistical correlations within data and fail to effectively demonstrate their decision-making processes. The absence of interpretability makes it more difficult to understand the mechanisms driving machine intelligence and limits its use in areas where transparency is important, such as healthcare, finance, autonomous systems, and law.

1.2 Our Approaches

With the overarching goal of building machine intelligence that can understand and solve real-world tasks, we begin by modeling human attentional traits and aligning attention mechanisms across humans and machines. We treat visual attention as an interface between perception and reasoning. Because attention is intrinsically interpretable and localizes task-relevant sensory input, it offers a direct window into how both humans and models allocate focus during task execution. To provide a comprehensive understanding of attention's role in decision-making, we advance our research step by step, from modeling and alignment to explainability.

1.2.1 Modeling

In Chapter 2 and Chapter 3, we focus on studying how human visual attention operates step-by-step to understand and perform general tasks from the perspective of reasoning processes. Specifically, we design computational models with a novel attention integration mechanism and develop a new scanpath (temporal sequence of eye fixations) prediction model to predict human attention during task completion.

We first utilize task-driven visual scanpaths as a new benchmark to explore the effectiveness of model attention in modeling human attention (Chapter 2). The goal of this study is to understand how humans deploy their attention during visual reasoning in naturalistic environments and to develop a model that examines how scanpaths affect task performance. Unlike previous studies that predict attention by estimating a static probabilistic saliency map under free-viewing conditions, we conduct the first study to understand and predict the temporal sequences of eye fixations during general tasks and investigate how scanpaths influence task outcomes. Our experiments and analyses highlight the advantages of our model for human attention modeling and provide valuable insights into the effectiveness of different attention mechanisms and their generalizability across tasks.

Next, we concentrate on understanding how attention varies across individuals, which has significant scientific and societal impacts (Chapter 3). This study seeks to examine how individuals allocate their attention across a variety of visual tasks and to construct a model that enables a comprehensive analysis of individualized scanpath prediction. Unlike existing models of saccadic eye movement that predominantly focus on modeling generic gaze patterns manifested as observer-agnostic scanpaths, we study the underexplored task of individualized scanpath prediction, focusing on modeling how an observer's unique traits affect their eye movement.

1.2.2 Alignment

Alignment between model attention and human attention is fundamental for advancing toward human-like intelligence. In Chapter 4, we aim to demonstrate the superiority of aligning model attention across multi-modalities human attention, in understanding real-life problems and providing step-by-step guidance to solve them. To achieve this

goal, we first propose a free-form and open-ended research framework that focuses on understanding real-life problems and deriving their evolution by incorporating key components across multiple modalities. We also offer two types of annotations absent from existing studies: solution graphs that describe dependencies between different steps and multimodal attention that highlights and associates important keywords and regions of interest. Through extensive experiments on intermediate model attention and the collection of multimodal human attention, we reveal the discrepancies between the visual behaviors of machines and humans during the understanding of real-life problems. More importantly, our results shed light on the fact that achieving the goal of providing step-by-step guidance to solve real-life problems requires the ability to understand the relationships between the problem and the solution, the relationships between different steps of the solution, and the relationships between visual and textual information. We also show that, by encouraging the model to align with human multimodal attention throughout the problem-solving process, it is possible to enhance reasoning abilities across various problem-solving tasks.

Next, in Chapter 5, we further develop a novel graph-based approach that leverages a novel integrated attention mechanism, which jointly considers the importance of features within each step as well as across multiple steps. Moreover, this attention mechanism can be progressively learned to predict sequential and non-sequential solution graphs depending on the characterization of the problem-solving procedure. We also introduce a set of quantitative metrics that consider attention propagation across the entire graph of solution steps to jointly and progressively supervise the integrated attention for free-form and open-ended tasks.

1.2.3 Explainability

Explainability is an important component of human communication, reasoning, and decision-making. Humans are not only good at identifying and reasoning with relevant information but also at justifying the rationale behind their decisions. Current research achieves state-of-the-art results in tracking visual exploration trajectories and predicting "when" and "where" people shift their attention; however, little attention is paid to explicitly explaining the "what" and "why" – the insights behind each fixation. Our work, discussed in Chapter 6, advances existing studies by providing a venue for weaving

a narrative thread that connects fixations to their underlying meaning and a novel model for generating explanations that faithfully reflect the fixated visual information.

To enable a comprehensive understanding of human visual attention, we first propose a novel task that predicts visual scanpaths along with the rationales behind them. Furthermore, our approach goes beyond predicting the trajectory of visual fixations to demonstrate the underlying rationale behind these gaze shifts. We find that aligning visual and textual information is key to improving prediction performance and deepening cognitive insight in modeling human visual attention.

1.3 Contributions

This dissertation introduces our efforts to close the gap between human intelligence and machine intelligence. It focuses on modeling human attention, aligning different attention mechanisms, and enhancing explainability to enable machine intelligence with reasoning ability, generalizability, and interpretability. In summary, it makes the following contributions:

Chapter 2: Predicting Human Scanpaths in Visual Question Answering

- Introducing a deep reinforcement learning model for predicting scanpaths in task-driven contexts like visual question answering (VQA), incorporating task performance into scanpath prediction for the first time.
- Developing a method to integrate task-specific attention maps from deep neural networks, enabling task-relevant information encoding and interpretability analysis by comparing human and model attention.
- Proposing a self-critical sequence training method to optimize non-differentiable evaluation metrics, addressing the gap between training and testing, along with a novel loss function to differentiate correct and incorrect scanpaths.
- Achieving human-level performance across VQA, free-viewing, and visual search tasks, outperforming state-of-the-art methods and demonstrating the approach's generalizability.

Chapter 3: Beyond Average: Individualized Visual Scanpath Prediction

- Investigating the underexplored task of individualized scanpath prediction, focusing on how unique attention traits influence eye movements.
- Proposing an individualization method with three novel components, including an observer encoder for observer-centric feature integration and adaptive fixation prioritization, enabling accurate and personalized predictions.
- Demonstrating the effectiveness and generalizability of the method across multiple datasets, architectures, and tasks using value-based and ranking-based metrics.

Chapter 4: VisualHow: Multimodal Problem Solving

- Proposing a new VisualHow framework to advance vision-language methods and multimodal understanding of real-world problems and solutions.
- Curating a dataset with diverse problem categories, multimodal solution descriptions, and fine-grained annotations.
- Conducting comprehensive experiments on tasks addressing different aspects of the VisualHow problem and performing extensive analyses of baseline models.

Chapter 5: Every Problem, Every Step, All In Focus: Learning to Solve Vision-Language Problems with Integrated Attention

- Presenting a graph neural network-based approach to represent procedural solutions as graphs, learning complex step dependencies for enhanced problem-solving.
- Designing an integrated attention mechanism to assess multimodal feature importance within and across steps, maintaining both step-level focus and a holistic view of the solution.
- Introducing quantitative attention metrics to optimize attention propagation across the solution graph, improving supervised learning for complex tasks.

Chapter 6: GazeXplain: Learning to Predict Natural Language Explanations of Visual Scanpaths

- Proposing a novel task to jointly predict visual scanpaths and generate natural language explanations, providing deeper semantic understanding of visual attention.

- Annotating ground-truth explanations on three public eye-tracking datasets, offering fixation-level insights for studying human attention and reasoning.
- Developing a general model architecture featuring an attention-language decoder with semantic alignment for consistent fixation-explanation pairing and cross-dataset co-training to enhance generalizability.
- Demonstrating through extensive experiments that the model generates accurate scanpaths and explanations, highlighting the importance of explanation prediction, semantic alignment, and cross-dataset co-training.

1.4 Publications

The relevant publication list for this dissertation is as follows:

- (Chapter 2) Predicting Human Scanpaths in Visual Question Answering [50]. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- (Chapter 3) Beyond Average: Individualized Visual Scanpath Prediction [51]. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- (Chapter 4) VisualHow: Multimodal Problem Solving [52]. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- (Chapter 5) Every Problem, Every Step, All In Focus: Learning to Solve Real-World Problems with Integrated Attention [53]. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024.
- (Chapter 6) GazeXplain: Learning to Predict Natural Language Explanations of Visual Scanpaths [54]. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. (Oral Paper, 2.3% acceptance rate)

Chapter 2

Predicting Human Scanpaths in Visual Question Answering

Attention has been an important mechanism for both humans and computer vision systems. While state-of-the-art models to predict attention focus on estimating a static probabilistic saliency map with free-viewing behavior, real-life scenarios are filled with tasks of varying types and complexities, and visual exploration is a temporal process that contributes to task performance. To bridge the gap, we conduct a first study to understand and predict the temporal sequences of eye fixations (a.k.a. scanpaths) during performing general tasks, and examine how scanpaths affect task performance. We present a new deep reinforcement learning method to predict scanpaths leading to different performances in visual question answering. Conditioned on a task guidance map, the proposed model learns question-specific attention patterns to generate scanpaths. It addresses the exposure bias in scanpath prediction with self-critical sequence training and designs a Consistency-Divergence loss to generate distinguishable scanpaths between correct and incorrect answers. The proposed model not only accurately predicts the spatio-temporal patterns of human behavior in visual question answering, such as fixation position, duration, and order, but also generalizes to free-viewing and visual search tasks, achieving human-level performance in all tasks and significantly outperforming the state of the art.

2.1 Introduction

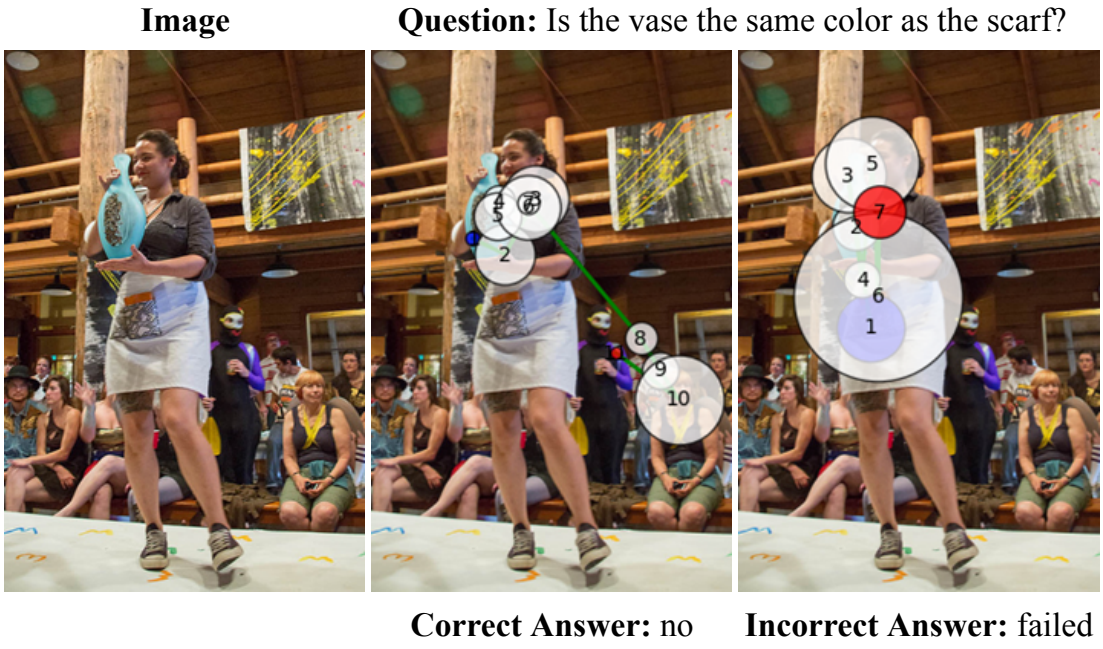

Figure 2.1: Visual scanpaths of humans can reveal their decision-making strategies and explain their performance. Those who pay attention to relevant visual cues can achieve high levels of task performance. This example compares the scanpaths of people who succeed or fail to answer a question, where the dots represent fixations. The number and radius indicate the fixation order and duration, respectively. The blue and red dots indicate the beginning and the end of the scanpath, respectively.

Visual attention plays an essential role in everyday tasks. While existing works focus on stimulus-driven attention with free-viewing behavior, underlying daily tasks is another form of attention, *i.e.*, task-driven attention, that selects task-relevant information to make a decision or to accomplish a task. Besides, beyond the static saliency map that highlights the relative importance of a visual input, temporal sequences of eye fixations encode a more comprehensive and natural representation of attention. Understanding and predicting visual scanpaths in general tasks will not only shed light on the decision-making process but also be a useful tool for a variety of computer vision applications.

Task-driven visual scanpaths reflect the visual exploration to accomplish the task, which also strongly correlates with task performance. As an example (figure 2.1), to answer the question "Is the vase the same color as the scarf?" while exploring the scene, humans need to actively explore the scene and search for the vase and the scarf. While looking at the right places at the right time would usually lead to correct answers (figure 2.1, middle), failing to do so may result in incorrect answers (figure 2.1, right).

As a step toward understanding and modeling general task-driven attention, we propose a novel deep reinforcement learning method leveraging task guidance as an important modality to predict the visual exploration behavior of humans performing general tasks. We first introduce a task guidance map to specify task-relevant image regions. The map is designed and demonstrated to generalize across tasks. To address the exposure bias that arises between training- and test-time contexts, we introduce a reinforcement learning method that directly optimizes non-differentiable test-time evaluation metrics [55]. To differentiate eye-movement patterns that lead to different performances, we further introduce a novel loss function to account for the consistency and divergence between correct and incorrect scanpaths.

Our work has three distinctions from previous scanpath prediction studies: (1) While state-of-the-art scanpath prediction studies focus on free-viewing [56, 57, 58, 59] or well-structured tasks such as visual search [60], this paper for the first time studies the complex scanpath patterns in general decision-making tasks, and investigates the correlation of scanpaths and performances in this context. (2) Scanpath prediction has not been as popular (compared with saliency prediction) or achieved excellent performance (compared with humans), partly due to the exposure bias – the discrepancy between training-time and test-time contexts. Here we close the gap using self-critical sequence training in the reinforcement learning method, leading to significantly boosted performance that is better than humans. (3) We go beyond a single task and design a new mechanism to encode general task-relevant information that is easily adaptable to other tasks with varying nature and levels of complexity. The proposed method has been demonstrated by three tasks with human-level performance.

In sum, this work makes the following contributions:

1. We develop a deep reinforcement learning model to understand and predict scanpaths in the general task-driven context with visual question answering (VQA).

Task performance is for the first time taken into account to predict scanpaths.

2. We propose to explicitly integrate attention maps from task-specific deep neural network models, allowing the encoding of task-relevant information as well as providing an alternative to measure the interpretability of task-specific models through analyzing model *vs*. human attention.

3. To address the discrepancy between training and testing that may have limited the development of scanpath prediction methods, we apply self-critical sequence training to directly optimize non-differentiable evaluation metrics. We further introduce a novel loss function to learn discriminative features and differentiate correct and incorrect scanpaths.

4. The proposed method significantly outperforms the state-of-the-art and shows human-level performance on three tasks: VQA, free-viewing, and visual search, demonstrating the generalizability of the method.

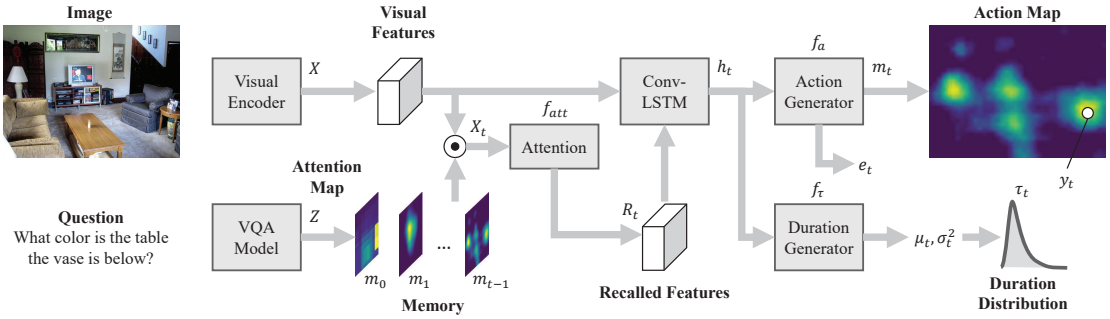

Figure 2.2: Overview of the proposed scanpath prediction network.

2.2 Related Work

Scanpath prediction. To precisely predict where humans look is not trivial, as eye movements are governed by several confounding factors [61]. Existing attention models either generate a saliency map where fixations can be sampled based on probability distribution and a winner-take-all strategy [62, 63, 64, 65, 66], or predict a sequence

of fixations by modeling their spatio-temporal complexity [56, 57, 67, 58, 68, 69, 70, 59, 71, 72, 73, 74, 75]. Our work is mostly related to the recent studies of task-driven attention [60]. Instead of studying structured vision tasks such as visual search [60], we aim to address a broader scope of general tasks. We use VQA as an example due to its generality and complexity, while further demonstrating the generalizability and flexibility of our method by adapting it for other tasks with various levels of complexity. To the best of our knowledge, our method is the first scanpath prediction method that successfully predicts human eye-movement behavior in the VQA task, and we further take the correctness of answers into account. Our model not only approaches human-level accuracy in the VQA task but is also highly generalizable across different tasks and datasets.

Human and machine attention in VQA. A unique characteristic of our work is the explicit integration of machine attention in the prediction of human scanpaths. With the rapid development of deep neural networks, the attention mechanism has become an essential component for improving the performance and explainability of VQA models [1, 76, 77]. However, due to their intrinsic differences, machine attention disagrees with human attention in many cases [77]. To study the relationship between human attention and machine attention, Chen *et al*. [1] and Jiang *et al*. [76] have developed datasets and computational methods to measure, model, and comparatively analyze the attention maps of humans and VQA models. While these analyses focus on the spatial difference of attention between correct and incorrect answers, our method generates individual fixations to study how people *maintain* and *shift* their attention which also encodes temporal information such as durations and orders. With the explicit incorporation of machine attention, our method also provides an alternative to measure the interpretability of VQA models based on their effectiveness in guiding scanpath prediction.

Reinforcement learning in attention prediction. A plausible approach to human attention prediction is reinforcement learning [78, 79, 80]. Early studies consider selective attention as a Markov decision process [81, 82] that can be optimized using policy iteration and a predefined reward function [78, 79, 80]. Recent scanpath prediction methods [83, 84, 60] adopt inverse reinforcement learning [85, 86] to automatically learn the unknown reward function from humans' eye-movement behavior. Although these

methods are promising, there is still a significant performance gap between scanpath prediction models and humans. We hypothesize that the performance gap is mainly caused by the exposure bias that commonly exists in sequence prediction tasks [87]. Exposure bias indicates the contextual discrepancy between the training and test settings. In scanpath prediction studies, many evaluation metrics are based on non-differentiable sequence comparison algorithms. Thus most computational methods are only able to use conventional cross-entropy or saliency evaluation metrics for training, leading to the discrepancy between training-time and test-time contexts. In this work, we adopt self-critical sequence training (SCST) [87] to address this bias by directly optimizing the non-differentiable test-time metrics. Leveraging the effectiveness of SCST, we further introduce a Consistency-Divergence loss to learn the differences between correct and incorrect scanpaths.

2.3 Method

We develop a deep reinforcement learning model to study and predict complex scanpath patterns in general decision-making tasks, while taking the task performance into account. This section presents the architecture of the proposed network and the machine learning methods to train the network with correct and incorrect scanpaths. Key technical novelties include the creation of a task guidance map to dynamically guide the prediction of fixation positions and durations, a reinforcement learning method with self-critical sequence training to address the exposure bias, and a novel Consistency-Divergence loss to learn the differences between correct and incorrect scanpaths.

2.3.1 Network Architecture

Where humans look during the VQA task is largely dependent on the input question. Existing task-driven attention models use a one-hot vector [60] or language embeddings [76] to encode the task input. These encoding methods provide semantic guidance to the model, to generate task-dependent outputs, but do not spatially align the task semantics with the visual contents. Differently, we compute a general task guidance map to highlight task-relevant image regions. This task guidance map is designed to be easily adaptable for other tasks. For example, it can be an all-zero matrix for predicting

scanpaths in the free-viewing task, or object detection masks can be used to provide task guidance in visual search. In this section, we summarize our method with the general VQA task.

As shown in figure 2.2, we design a neural network model to dynamically generate a sequence of fixation positions and durations. A memory module and an attention mechanism are developed to selectively memorize and recall task-relevant visual information. Specifically, given an image and a question, our goal is to generate a sequence of fixations positions $y = \{y_1, y_2, \cdots, y_T\}$ and durations $\mathcal{T} = \{\tau_1, \tau_2, \cdots, \tau_T\}$. At each step t, the fixation position y_t is sampled from a predicted action map m_t, and the fixation duration τ_t is sampled from a log-normal distribution with two predicted parameters (μ_t, σ_t^2). Besides, a scalar output e_t indicates the end of the scanpath. The specific network design is as follows:

Inputs and task guidance. On the input side, we adopt a CNN-based visual encoder [6] to extract visual features X from the image. The influence of the question is represented as a task guidance map highlighting task-relevant image regions. Trained on large VQA datasets, machine attention can better bridge the task semantics and visual contents by highlighting task-relevant spatial regions that are important for answering the question. Therefore, we guide the prediction of eye fixations using the machine attention of an externally trained VQA model [88, 1, 89, 90]. We preprocess the VQA model's attention into a 2D task guidance map Z with its values normalized within the range of $[0, 1]$.

Memory and attention. Answering complex questions requires dynamically updated memory and attention mechanisms to trace the reasoning process over time [1, 76]. The memory is denoted as $M_t = \{m_0, m_1, \cdots, m_{t-1}\}$, which explicitly maintains all previously computed action maps, as well as the task guidance map $m_0 = Z$. This memory as a whole can be seen as a spatio-temporal attention volume. By applying it to the visual features X, we can obtain the memorized features $X_t = M_t \circ X$, where $\circ$ indicates the Hadamard product. The attention module recalls the most relevant information from the memory, denoted as

$$R_t = f_{att}(X_t; \theta_{att}), \tag{2.1}$$

where the θ_{att} indicates learnable parameters. It computes a temporal attention vector indicating the dynamic importance of each historical time step [76], to determine what to recall from the memory for the prediction of the current fixation.

ConvLSTM and outputs. We design a ConvLSTM network to simultaneously predict the distributions of fixation positions and durations. The image features X and the recalled features R_t are fed into a ConvLSTM layer to encode the spatio-temporal patterns of scanpaths. With its current hidden state h_t, the outputs are computed as

$$p_t^a(a_t|a_{1:t-1}) = \text{softmax}(f_a(h_t;\theta_a)), \tag{2.2}$$

$$[\mu_t, \sigma_t^2] = f_\tau(h_t;\theta_\tau), \tag{2.3}$$

where f_a and f_τ indicate the output layers and θ_a and θ_τ are learnable parameters. We use $[m_t, e_t] = p_t^a(a_t|a_{1:t-1})$ to represent the distribution of the actions including the action maps m_t and the end-of-scanpath indicator e_t. Finally, we sample the fixation point y_t following the discrete probability values in the action map m_t, and sample the fixation duration following the parametric function $\tau_t \sim p_t^\tau(\tau|\mu_t, \sigma_t^2)$. We model the duration distribution p_t^τ as a log-normal function following previous experimental studies [91, 92].

2.3.2 Objective

Scanpath prediction is a typical sequential learning task. To address the discrepancy between training and testing contexts in sequential learning, we propose to apply self-critical sequence training (SCST) [87] to directly optimize the non-differentiable evaluation metrics. We further introduce a novel loss function to help differentiate correct and incorrect scanpaths.

Supervised learning. It is widely used in sequential learning to minimize a maximum-likelihood loss at each step. In our context, the objective is to jointly optimize the fixation action a_t and the duration τ_t:

$$L(\theta) = -\sum_{t=1}^{T+1} \log p_t^a(a_t^*|a_{1:t-1}^*;\theta) - \lambda \sum_{t=1}^{T} \log p_t^\tau(\tau_t^*|\mu_t, \sigma_t^2), \tag{2.4}$$

where T is the length of the ground-truth fixations, a_t^* and τ_t^* are the ground-truth

action (one-hot vector indicating the fixation position or end of the scanpath) and fixation duration, respectively. The hyperparameter λ balances the contributions of the two loss terms. With this loss function, we simultaneously train two networks with the correct and incorrect scanpaths. They share most of their parameters, except for the memory and output layers.

However, this objective function does not always produce the best results on the non-differentiable metrics for scanpath evaluation. This discrepancy between training and testing contexts has been observed in similar sequence generation tasks [88, 87]. To address this issue, we propose to use SCST in scanpath prediction and optimize the network based on test-time evaluation metrics.

Reinforcement learning with SCST. Specifically, in the context of scanpath prediction, the objective is to minimize the negative expected reward:

$$L_r(\theta) = -\mathbb{E}_{y,\mathcal{T}}[r(y, \mathcal{T})], \tag{2.5}$$

where $r(\cdot, \cdot)$ is a reward function (*i.e.* ScanMatch [55]), while y and $\mathcal{T}$ indicate the sampled fixation positions and durations, respectively. The main idea of SCST is to baseline the REINFORCE algorithm with the reward achieved by the current model under the corresponding evaluation metric used at the test time [87]. To reduce the variance of the gradient estimate and accelerate the training, for each network, we compute the average rewards of k scanpaths and use their mean reward as the corresponding baseline. We denote their corresponding loss functions as $L_r^+(\theta)$ and $L_r^-(\theta)$, respectively. Without loss of generality, in this paper, we use the superscripts $+$ and $-$ to distinguish the notations for correct and incorrect scanpaths, respectively.

Consistency-Divergence loss. The level of difference between correct and incorrect scanpaths is image-specific, so it is difficult to distinguish them by directly learning from the data. We combine the SCST objective with a novel Consistency-Divergence loss (CDL) to explicitly quantify the consistency and divergence of human scanpaths and force the model predictions to resemble such statistics. Specifically, given the correct and incorrect ground-truth scanpaths, we first compute their within-group similarity r^{*+}_{within}, r^{*-}_{within}, and the between-group similarity r^{*}_{between}, by averaging the pair-wise evaluation scores within and between the correct and incorrect groups. The differences $\Delta r^{*+} =$

$r^{*+}_{\text{within}} - r^{*}_{\text{between}}$ and $\Delta r^{*-} = r^{*-}_{\text{within}} - r^{*}_{\text{between}}$ measure the consistency of scanpaths within each group compared with the diversity between the two groups. Intuitively, high within-group similarity and low between-group similarity suggest that the differences between correct and incorrect scanpaths are more distinguishable. Similarly, we can evaluate the predicted scanpaths in the same way to obtain $\Delta r^{+}(y, \mathcal{T})$ and $\Delta r^{-}(y, \mathcal{T})$. The objective of the proposed CDL is to let $\Delta r^{+}(y, \mathcal{T})$ approximate Δr^{*+} and $\Delta r^{-}(y, \mathcal{T})$ approximate Δr^{*-}, so that the differences between the predicted scanpaths are similar to those of the ground-truth. Therefore, the CDL is computed as

$$\begin{aligned} L_{\text{CD}}(\theta) = \mathbb{E}_{y,\mathcal{T}}\big[|\triangle r^{+}(y, \mathcal{T}) - \triangle r^{*+}|\big] \\ + \mathbb{E}_{y,\mathcal{T}}\big[|\Delta r^{-}(y, \mathcal{T}) - \Delta r^{*-}|\big], \end{aligned} \tag{2.6}$$

Finally, we define the total loss as a linear combination of the negative expected reward and the CDL (2.6):

$$L'(\theta) = L^{+}_{r}(\theta) + L^{-}_{r}(\theta) + \gamma L_{\text{CD}}(\theta). \tag{2.7}$$

The hyperparameter γ balances the contribution of the loss terms in the policy gradient update stage [93].

2.4 Experiments

We evaluate the proposed method with extensive experiments. Our quantitative and qualitative results demonstrate the performance and generalizability of the proposed method, sheding light on some interesting research questions about scanpath prediction.

2.4.1 Experiment Settings

Dataset. We conduct our experiments mainly on the AiR dataset [1]. It consists of images and questions selected from the balanced validation set of GQA [94] and provides the eye-tracking data collected from 20 participants who answer the questions. Each question is answered by 10 different participants, and their eye-tracking data are associated with their answers. The numbers of fixations in the recorded scanpaths are similar between the correct answers (10.12 ± 0.99) and the incorrect answers ($10.27 \pm$

1.54). Their spatial priors are also highly similar. These similarities ensure that models do not differentiate between correct scanpaths and incorrect scanpaths based on their prior distributions. We randomly split this dataset into a training set of 1137 questions, a validation set of 142 questions, and a test set of 143 questions. The proportion of correct answers are balanced among these subsets.

Evaluation metrics. To evaluate the models, we generate 10 correct/incorrect scanpaths with each model and compare them with the corresponding ground-truth scanpaths using a combination of four evaluation metrics: The *ScanMatch* [55, 95] measures scanpath similarity based on the Needleman-Wunsch algorithm [96]. It has been commonly used to evaluate scanpath prediction models due to its robustness to the substantial noise inherent in the scanpaths. The *MultiMatch* [97] is a multidimensional evaluation metric, composed of five similarity measures regarding shape, direction, length, position, and duration. The *String-Edit Distance (SED)* [98, 99] is a dissimilarity measure that converts scanpaths into strings by associating each image region with a character. The *Scaled Time-Delay Embedding (STDE)* [73] measures the average of the minimum Euclidean distances of each sub-sequence of the compared scanpaths. For SED and STDE, we report the mean and best evaluation scores. While the mean scores are the averages of all subjects, the best scores are computed based on the most similar human scanpath [100]. These complementary evaluation metrics provide a comprehensive view of the prediction results.

Implementation details. We use ResNet-50 [6] to encode the visual features and use AiR [1] VQA model to compute the task guidance maps. The object-based attention weights are converted to spatial maps by computing a weighted average of their bounding box masks [1]. The resolution of the input image is 240×320. We discrete the fixation position into a 30×40 action map. In supervised learning, we train our model using the Adam [101] optimizer with learning rate 10^{-4} and weight decay 5×10^{-5}. To avoid the divergence of loss, we also adopt the warmup strategy [102] followed by a linear decay of the learning rates. In reinforcement learning, we also use the Adam [101] optimizer with linearly decayed learning rates starting at 5×10^{-5} and weight decay 5×10^{-5}. In SCST, we sample $k = 5$ different scanpaths for the correct and incorrect answers, respectively. The reward function is defined as the harmonic average of the two ScanMatch scores, one with duration and the other without. Our implementation of the ScanMatch metric

Method	ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
Human	0.421	0.391	0.945	0.747	0.938	0.879	0.522	7.836	4.804	0.867	0.918
	0.375	0.358	0.938	0.734	0.929	0.851	0.526	8.611	6.406	0.841	0.895
SaltiNet [57]	0.112	0.130	0.933	0.676	0.930	0.696	0.504	10.620	9.264	0.729	0.765
	0.120	0.138	0.930	0.676	0.926	0.696	0.506	10.650	9.750	0.734	0.754
PathGAN [56]	0.210	0.212	0.940	0.637	**0.937**	0.806	0.589	8.658	6.535	0.832	0.862
	0.221	0.218	0.937	0.637	0.927	0.821	0.612	9.071	7.750	0.844	0.861
IOR-ROI [59]	0.171	0.202	0.918	**0.724**	0.908	0.782	0.570	9.210	7.332	0.791	0.818
	0.198	0.216	0.917	**0.737**	0.905	0.793	0.590	9.177	7.945	0.801	0.817
Ours	**0.394**	**0.391**	**0.950**	0.717	0.933	**0.879**	**0.615**	**7.523**	**5.701**	**0.869**	**0.893**
	0.365	**0.368**	**0.946**	0.705	**0.930**	**0.864**	**0.632**	**7.955**	**6.772**	**0.856**	**0.877**

Table 2.1: Scanpath prediction results on the AiR dataset (VQA). In each panel, the first row indicates the correct scanpaths and the second row indicates the incorrect scanpaths. The best results are highlighted in bold. Underlines indicate scores above human performance.

in training and evaluation follows [55, 95]. The hyperparameters λ and γ are empirically set to 1.0 and 2.0, respectively, based on the validation set performance.

2.4.2 Are the predicted scanpaths plausible?

We first evaluate how well the predicted scanpaths simulate human behavior. Since we are the first to predict scanpaths in the VQA task, for a fair comparison, we customize the most relevant deep-learning-based scanpath prediction models (*i.e.* SaltiNet [57], PathGAN [56], and IOR-ROI [59]), by combining the BERT embedding [103] of the question with the visual features and jointly predicting the correct and incorrect scanpaths. Following [59, 60], we measure human performance by computing the inter-observer agreements within the correct and incorrect groups, respectively. For each image, we measure the similarity of every pair of human scanpaths from the same group and compute their mean values.

table 2.1 reports the quantitative results of the compared methods. Our method significantly improves the prediction of both fixation positions and durations. It outperforms the other methods on 9.5/11 metrics by a substantial margin. For example, its ScanMatch scores are over 84% (correct) and 69% (incorrect) higher than the state-of-the-art methods. It even outperforms humans on 6.5/11 metrics.

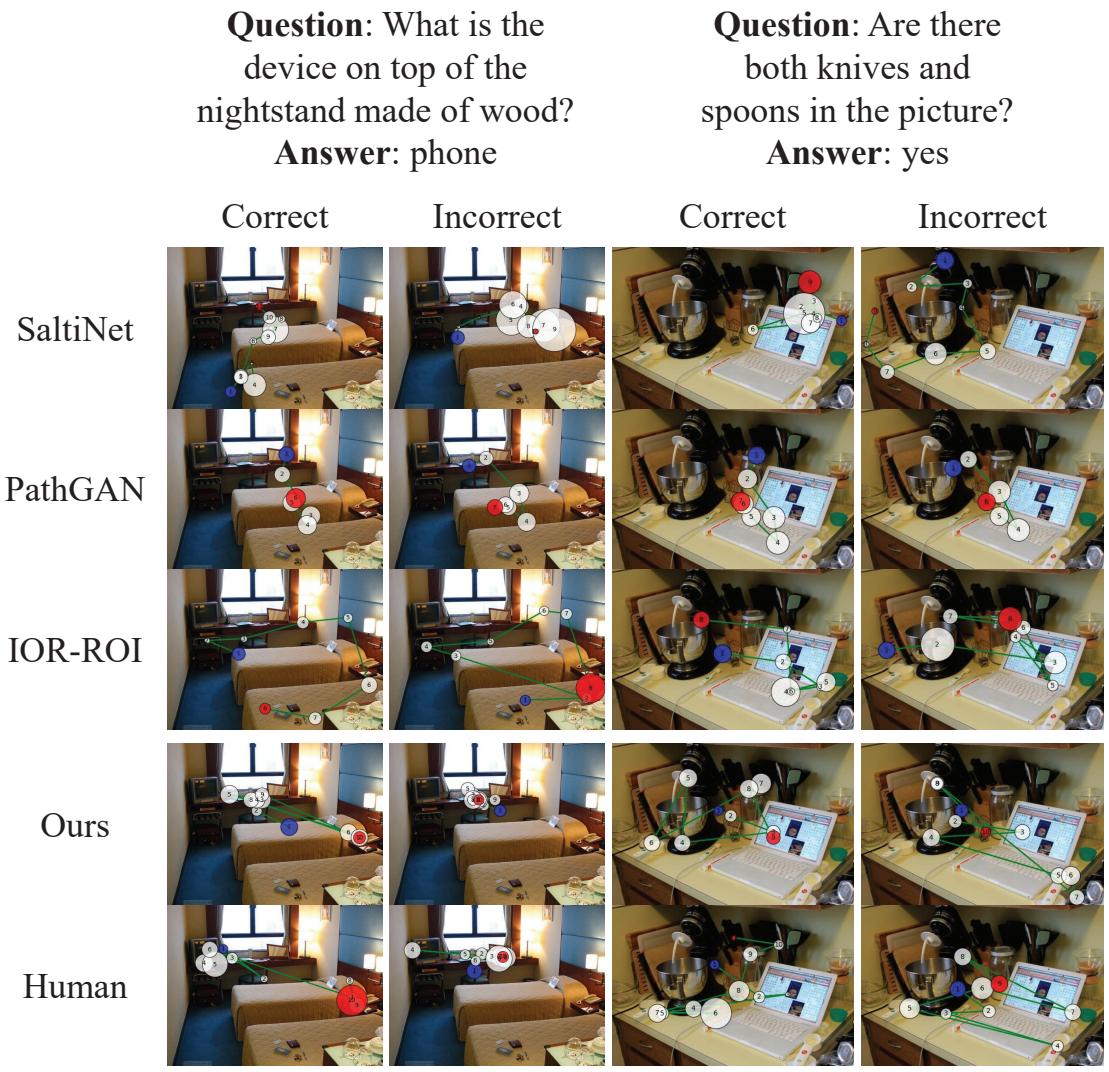

Figure 2.3: Examples of the predicted scanpaths. Each column compares the prediction results and human scanpaths given specific answer correctness. The number and radius indicate the fixation order and duration, respectively. The blue and red dots indicate the beginning and the end of the scanpath, respectively.

Method			ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
TG	SCST	CDL	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
			0.290	0.323	0.927	**0.719**	0.914	0.845	0.537	8.539	6.829	0.838	0.858
			0.280	0.310	0.920	0.713	0.909	0.831	0.544	8.797	7.667	0.827	0.845
✓			0.296	0.329	0.927	**0.719**	0.914	0.849	0.533	8.438	6.733	0.841	0.862
			0.288	0.317	0.922	**0.717**	0.910	0.837	0.546	8.749	7.682	0.831	0.850
	✓		0.360	0.363	0.948	0.705	0.930	0.865	0.612	7.752	5.961	0.860	0.885
			0.350	0.350	0.943	0.704	0.925	0.852	0.627	8.013	6.818	0.850	0.871
	✓	✓	0.369	0.370	0.949	0.713	**0.933**	0.869	0.605	7.741	5.982	0.860	0.883
			0.350	0.352	0.944	0.716	0.927	0.856	0.616	8.066	6.946	0.849	0.870
✓	✓		0.385	0.383	0.949	0.714	0.932	0.876	0.614	7.569	5.736	0.867	0.891
			0.348	0.354	0.945	0.703	0.928	0.855	0.620	8.011	6.796	0.849	0.873
✓	✓	✓	**0.394**	**0.391**	**0.950**	0.717	**0.933**	**0.879**	**0.615**	**7.523**	**5.701**	**0.869**	**0.893**
			0.365	**0.368**	**0.946**	0.705	**0.930**	**0.864**	**0.632**	**7.955**	**6.772**	**0.856**	**0.877**

Table 2.2: Ablation study of TG, SCST and CDL on the AiR dataset. In each panel, the first row indicates the correct scanpaths and the second row indicates the incorrect scanpaths. The best results are highlighted in bold.

figure 2.3 presents qualitative examples of the predicted scanpaths. While the state-of-the-art models look at salient objects in general, our predicted scanpaths align better with task-related objects and the human eye-movement behavior regarding fixation positions, durations, and orders. Note that subtle differences of scanpaths can determine the correctness of answers: the incorrect scanpaths consistently miss important objects (*i.e.* phone and knives).

Note that besides our significant performance boost in predicting correct scanpaths, our method is also effective in predicting scanpaths that lead to incorrect answers thus to be avoided. We find that incorrect scanpaths are less consistent compared with correct ones (also corroborated with Human scores), possibly due to the variety of factors that may lead to an incorrect decision. Yet with the task guidance and the novel CDL loss, our method can capture the subtle differences between the correct and incorrect scanpaths, and learn discriminative features relevant to answer correctness to successfully predict both correct and incorrect scanpaths.

2.4.3 What contributes to the model's performance?

Our proposed method has three major technical contributions: VQA model attention as the task guidance (TG), SCST to address the exposure gap, and the novel Consistency-Divergence loss (CDL). To demonstrate the contribution of each component, we incrementally apply them to a baseline (*i.e.* a task-ignorant supervised-learning variant of our method). As shown in table 2.2, each component helps predict both correct and incorrect scanpaths. In particular, though TG results in relatively minor improvements by itself (under supervised learning), it plays a more important role in reinforcement learning with SCST. This observation suggests that SCST can help the model to make better use of the task input to fixate task-relevant regions. Finally, using the new CDL loss together with SCST optimizes the within-group and between-group consistencies of the correct and incorrect scanpaths, thus further increasing the model performance.

2.4.4 What do the predicted scanpaths fixate?

To investigate how the predicted scanpaths fixate different objects, we align the fixation positions with the ground-truth object annotations provided by the GQA dataset [94]. We segment each image into three regions: 1) Region of Interest (ROI) is composed of all the objects in the questions and answers; 2) Non-ROI is composed of the other annotated objects that are not included in the ROI; 3) Background is the empty regions without object annotations. For each compared model, we compute the percentage of fixations in each region. As shown in table 2.3, in general, higher-performance models generate more fixations in the ROI. Our proposed techniques (*i.e.* TG, SCST, CDL) improve the accuracy of fixating task-relevant objects, allowing our method to perform significantly better than the state-of-the-art methods [56, 57, 59]. The percentage of fixations to ROI of our full model is similar to that of humans. Besides, humans' correct scanpaths fixate the ROI more frequently than the incorrect ones, showing the correlation between their attention allocation and task performance. Our method replicates this correlation, while the compared methods fail to do so. The proposed techniques allow our model to learn more discriminative features and better distinguish correct and incorrect scanpaths.

Method	Fixations Position %		
	ROI ↑	Non-ROI ↓	Background ↓
Human	26.43	67.48	6.09
	21.60	71.92	6.48
SaltiNet [57]	4.17	77.88	17.95
	3.96	78.49	17.55
PathGAN [56]	7.82	84.34	7.83
	7.17	86.10	6.73
IOR-ROI [59]	9.14	82.99	7.87
	9.79	82.53	7.67
Ours	25.04	69.70	5.26
	22.33	72.27	5.40

Table 2.3: Percentage of fixations in ROI, non-ROI, and background. In each panel, the first row indicates the correct scanpaths and the second row indicates the incorrect scanpaths.

Figure 2.4: Comparison of VQA models' answer accuracy, scanpath accuracy, and machine attention accuracy (bubble size).

2.4.5 Which VQA model is the most effective?

The explicit use of VQA models in our method allows us to evaluate and visualize VQA models from a human attention's perspective, which has not been explored before. We evaluate the effectiveness of four VQA models: AiR [1], UpDown [88], HAN [90] and MLB [89]. figure 2.4 compares their VQA accuracy on the GQA (test-dev) dataset, machine attention accuracy (AiR-E [1]), and the scanpath prediction performance (ScanMatch w/ duration). As can be seen, both the machine attention accuracy and VQA accuracy are positively correlated with the scanpath prediction performance. Object-based attention maps tend to be more accurate and provide better task guidance: AiR [1] achieves the best performance, thanks to its explicit attention supervision with the ground-truth object annotations. UpDown [88] computes implicitly supervised object-based attention, achieving lower performances in scanpath prediction. HAN [90] relies on attention ground-truth from a specific group of questions [104], which leads to lower

Method	ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
Human	0.390	0.386	0.941	0.695	0.931	0.851	0.621	7.486	5.001	0.844	0.906
Itti *et al.* [65]	0.211	0.088	0.824	0.653	0.763	0.685	0.415	8.701	6.529	0.714	0.757
SGC [71]	0.211	–	0.906	0.658	0.870	0.717	–	8.422	6.194	0.771	0.837
Wang *et al.* [73]	0.151	–	0.857	0.641	0.801	0.625	–	9.051	7.129	0.682	0.739
Le Meur *et al.* [70]	0.228	–	0.864	0.657	0.831	0.701	–	8.573	6.536	0.739	0.788
STAR-FC [75]	0.204	–	0.920	0.662	0.900	0.668	–	8.393	6.314	0.751	0.828
SaltiNet [57]	0.169	0.142	0.868	0.647	0.840	0.655	0.566	8.948	7.001	0.706	0.763
PathGAN [56]	0.077	0.079	0.919	0.572	0.905	0.511	0.678	9.414	7.677	0.611	0.691
IOR-ROI [59]	0.267	0.265	0.891	**0.709**	0.860	0.759	0.634	8.180	6.003	0.789	0.844
Ours	**0.383**	**0.377**	**0.943**	0.651	**0.924**	**0.847**	**0.684**	**7.155**	**4.579**	**0.852**	**0.905**

Table 2.4: Performances on the OSIE dataset (free-viewing). The best results are highlighted in bold. Underlines indicate scores above human performance.

Method	ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
Human	0.526	0.490	0.944	0.755	0.934	0.913	0.685	2.181	0.359	0.920	0.974
SaltiNet [57]	0.199	0.127	0.909	0.546	0.907	0.740	0.551	4.037	2.742	0.759	0.829
PathGAN [56]	0.277	0.198	0.930	0.561	0.926	0.839	0.604	2.820	1.694	0.847	0.901
IOR-ROI [59]	0.316	0.274	0.919	0.665	0.907	0.834	0.586	4.384	2.595	0.846	0.896
IRL [60]	0.403	–	0.904	0.630	0.887	0.825	–	2.734	1.002	0.898	0.952
Ours	**0.554**	**0.510**	**0.941**	**0.706**	**0.927**	**0.914**	**0.721**	**1.852**	**0.484**	**0.923**	**0.965**

Table 2.5: Performances on the COCO-Search18 dataset (visual search). The best results are highlighted in bold. Underlines indicate scores above human performance.

performances and difficulties to generalize. MLB [89] is based on image features, so its spatial attention maps may not highlight objects, leading to the lowest performances. In sum, our method suggests that a well-designed machine attention mechanism not only improves the performance of VQA models but also benefits human attention prediction. It also enables further correlational studies between human and machine attention mechanisms.

2.4.6 Does the proposed method generalize?

Our method can generalize across tasks with different complexities. Similar to what we observe in the VQA task, results in the free-viewing and visual search tasks also show a significant performance boost, achieving a human-level performance. First, for the free-viewing task (*i.e.* task guidance and CDL are not applicable), we conduct experiments

on the OSIE dataset [105] following the settings of Sun *et al*. [59]. table 2.4 shows that our method significantly outperforms the state-of-the-art methods [56, 57, 65, 70, 59, 71, 73, 75] on 10/11 metrics with over 42% higher ScanMatch scores. Next, we conduct experiments on COCO-Search18, a visual search dataset [60], using a CenterNet [106] detector to detect the search targets and generate the task guidance maps. As shown in table 2.5, our method outperforms the state-of-the-art approaches [56, 57, 59, 60] by a large margin and reaches human-level performance on 6/11 metrics. Particularly, our ScanMatch scores are over 37% better than the state-of-the-art [60] and over 5.3% better than humans. These overwhelming performances demonstrate the robustness and generalizability of our method in different task settings.

2.5 Conclusion

We propose the first model for predicting human scanpaths during visual question answering. By explicitly integrating a task guidance map, the model learns to predict a sequence of task-driven scanpaths that lead to correct or incorrect answers. To address the exposure bias, we propose an SCST approach that optimizes the model based on scanpath evaluation metrics and a Consistency-Divergence loss to distinguish between correct and incorrect scanpaths. Our method significantly outperforms the state-of-the-art methods on multiple datasets and tasks. Our experiments suggest that our model can predict human-like scanpaths and reveal the critical fixation patterns that determine the task performance. The improved performance of human scanpath prediction will push forward the research on task-driven attention and advance a wide range of applications in the development of intelligent robots, automatic design and advertising systems, human-computer interaction systems, and diagnostic tools for mental healthcare.

Chapter 3

Beyond Average: Individualized Visual Scanpath Prediction

Understanding how attention varies across individuals has significant scientific and societal impacts. However, existing visual scanpath models treat attention uniformly, neglecting individual differences. To bridge this gap, this paper focuses on individualized scanpath prediction (ISP), a new attention modeling task that aims to accurately predict how different individuals shift their attention in diverse visual tasks. It proposes an ISP method featuring three novel technical components: (1) an observer encoder to characterize and integrate an observer's unique attention traits, (2) an observer-centric feature integration approach that holistically combines visual features, task guidance, and observer-specific characteristics, and (3) an adaptive fixation prioritization mechanism that refines scanpath predictions by dynamically prioritizing semantic feature maps based on individual observers' attention traits. These novel components allow scanpath models to effectively address the attention variations across different observers. Our method is generally applicable to different datasets, model architectures, and visual tasks, offering a comprehensive tool for transforming general scanpath models into individualized ones. Comprehensive evaluations using value-based and ranking-based metrics verify the method's effectiveness and generalizability.

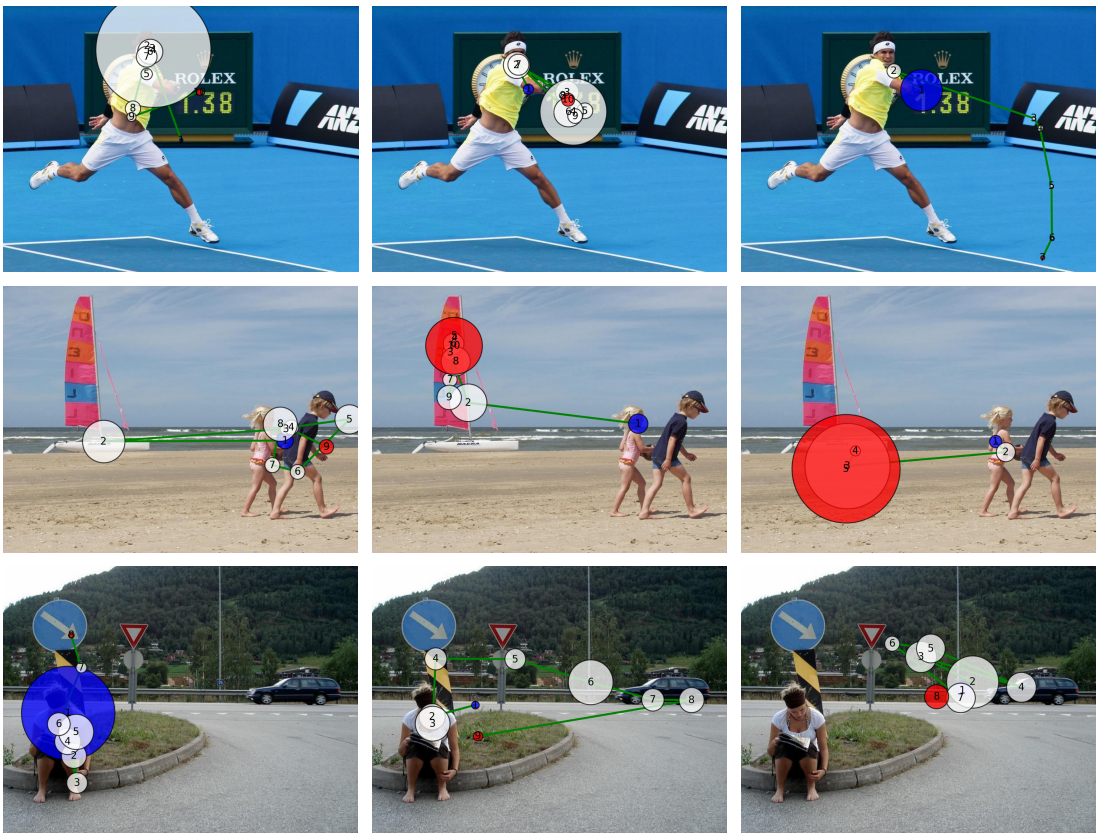

Figure 3.1: Understanding and predicting the distinct eye movements of each observer is the key objective of individualized scanpath prediction. These examples reveal the variations in the scanpaths of different observers, showing their distinct attention preferences in (a) faces, (b) objects, and (c) background. Each dot represents a fixation, with the number and radius indicating its order and duration, respectively. The blue and red dots indicate the beginning and the end of the scanpath, respectively.

3.1 Introduction

Saccadic eye movements, such as fixations and saccades, enable individuals to shift their attention quickly and redirect their focus to different points in the visual field. Studying various factors driving people's eye movements is important for understanding

human attention and developing human-like attention systems. Computational models predicting eye movements have broad impacts across various domains, such as assessing image and video quality [107, 108, 109], developing intuitive human-computer interaction systems [110, 111, 112, 113, 114], creating immersive virtual reality experiences [115, 116, 117], enhancing the safety and efficiency of autonomous vehicles [118, 119, 120], and diagnosing neurodevelopmental conditions [121, 122, 123].

While existing models of saccadic eye movements predominantly focus on modeling generic gaze patterns manifested as observer-agnostic scanpaths (*i.e.*, a spatiotemporal sequence of fixations), this work seeks to model the individual variations in eye movements. As shown in Figure 3.1, there exists significant inter-observer variations in visual scanpaths. Such variations can be attributed to a multitude of individual characteristics, such as gender, age, and neurodevelopmental conditions [124, 125]. For instance, females show more explorative gaze patterns than males [126, 127, 128], older adults prefer faces [129] and objects with high color visibility [130], individuals with neurodevelopmental disorders, such as autism spectrum disorder (ASD), may show a preference for repetitive patterns while avoiding social cues [2, 131, 132]. Therefore, developing tailored models that cater to the uniqueness of each observer is an essential step toward more precise and adaptive attention modeling.

Existing research efforts have failed to address the divergence between the personalized nature of human attention and the collective nature of current scanpath models. This is due to the lack of standardized methods for quantifying and representing individual attention traits, as well as the absence of comprehensive frameworks that can accommodate the diverse range of observer characteristics. In this paper, we resolve this significant challenge with a novel individualized scanpath prediction (ISP) method comprising three novel components: (1) The observer encoder is a key component for personalized scanpath modeling. It efficiently captures an observer's unique attention traits by introducing an observer-specific identifier as an additional input, forming the basis for individualized scanpath predictions. (2) The observer-centric feature integration module adopts a comprehensive approach, fusing visual features, task guidance, and observer-specific attention traits spatially and channel-wise. This ensures consideration of diverse bottom-up and top-down cues, simplifying subsequent processing and enhancing the efficient prediction of individualized scanpaths. (3) The adaptive fixation

prioritization module enhances scanpath precision by dynamically assigning priorities to the output features, generating a probability map for each fixation. This adaptability ensures refined predictions of individualized scanpaths.

Our method has three distinctions from previous visual scanpath studies: (1) We go beyond prior work focusing on general scanpath modeling and propose the first comprehensive investigation of individualized scanpath prediction. (2) We emphasize the tight integration of observer features into the scanpath prediction process, distinct from trivial individualization techniques such as fine-tuning with single-observer data. (3) Our method is generally applicable to various model architectures and visual tasks, broadening its usability in real-world applications.

The main contributions of this work are as follows:

1. We study the underexplored task of individualized scanpath prediction, focusing on modeling how an observer's unique attention traits affect their eye movements.

2. We propose an individualization method featuring three novel technical components: The observer encoder is an important addition to scanpath models, which enables observer-centric feature integration and adaptive fixation prioritization. These components enable the model to adapt to individual observers, yielding accurate and individualized predictions.

3. We comprehensively evaluate scanpaths from individual observers' perspectives, using both value-based and ranking-based metrics. Experimental results on multiple eye-tracking datasets, with different model architectures and visual tasks, prove our method's effectiveness and generalizability for predicting individualized scanpaths.

3.2 Related Works

Our work is related to prior studies on eye-tracking datasets and visual scanpath prediction methods.

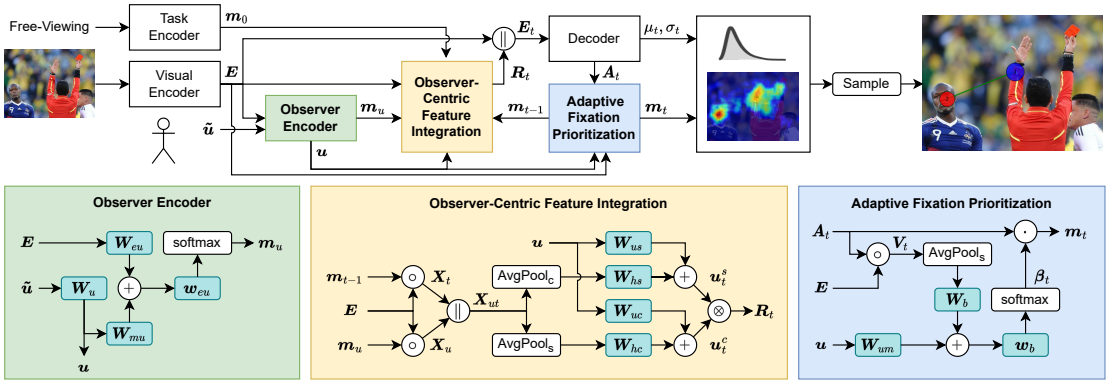

Figure 3.2: Our proposed method incorporates an observer encoder for characterizing individualized attention traits, followed by observer-centric feature integration for holistic processing, and adaptive fixation prioritization for refined predictions.

3.2.1 Eye-Tracking Datasets

The foundation for attention modeling relies on diverse, thoughtfully curated eye-tracking datasets spanning various stimuli, tasks, and observers [1, 2, 105, 60, 123]. These datasets, from those dedicated to free-viewing [2, 105, 123] to those capturing goal-directed behaviors [1, 60], serve as invaluable resources for training and evaluating attention models. Specifically, several well-recognized eye-tracking datasets have provided benchmarks to quantify the performance of saliency models [133, 134, 105, 135] and scanpath models [2, 105, 123]. Subsequent studies have developed datasets of goal-directed behaviors to characterize how observers search for an object in an image [60] or answer image-related questions [1]. These efforts facilitate the development of static saliency models [136, 137, 138, 139, 140, 141, 142, 143] as well as dynamic scanpath models [50, 144, 59, 60, 145, 146, 147, 148, 149]. Our work sets itself apart from individualized saliency models [150, 151, 152, 142, 153, 154] by predicting dynamic scanpaths rather than static saliency maps. It utilizes datasets from various visual tasks and observer groups to expand scanpath modeling, with emphasis on the distinct attention traits of each observer.

3.2.2 Visual Scanpath Prediction

Scanpath prediction has been an underexplored topic in the field of attention modeling. Early studies generate scanpaths by sampling fixations from saliency maps using the inhibition-of-return mechanism [73, 70, 75, 65, 74, 153]. Recent studies have developed computational models directly predicting the sequence of fixations and saccades [50, 144, 59, 60, 145, 146, 147, 148]. Several scanpath models harness the power of deep neural networks [50, 144, 59, 60, 145, 146, 147, 148, 155, 114], reinforcement learning techniques [50, 60, 146], and transformer-based models [148, 144], ultimately improving the accuracy of scanpath prediction to the human level. These developments have significantly deepened our understanding of the temporal dynamics of human attention. However, existing models focus on predicting general scanpaths rather than taking individual variations into account. Differently, our method places particular emphasis on characterizing individual attention traits and integrating them into a general scanpath model, thus enabling tailored predictions that align with each observer's gaze behavior. This unique approach extends the horizon of attention modeling, underlining the importance of individual differences within the broader context of human attention.

3.3 Methodology

The core challenge in individualized scanpath modeling is the need to predict unique gaze patterns for different observers. This arises due to the inherent variations in attention traits. Figure 3.2 presents an overview of our method. It offers a threefold solution: (1) an observer encoder, (2) an observer-centric feature integration module, and (3) an adaptive fixation prioritization module. These components are designed to be flexible as they can be applied on a general scanpath model based on the encoder-decoders, (*e.g.*, with a visual encoder and task encoder, and an LSTM [156] or Transformer [103] decoder) to provide robust and precise predictions tailored to each observer.

3.3.1 Observer Encoding

At the core of our proposed method is an *Observer Encoder*, a key component designed to enable the novel task of individualized scanpath prediction. It takes as input an

observer-specific identifier $\tilde{\boldsymbol{u}}$ (*e.g.*, a one-hot vector) and efficiently computes an observer feature $\boldsymbol{u}$. This feature represents the unique characteristics and preferences of each observer. Our approach utilizes a linear embedding operation to derive the observer feature:

$$\boldsymbol{u} = \boldsymbol{W}_u \tilde{\boldsymbol{u}}, \tag{3.1}$$

where $\boldsymbol{W}_u$ indicates learnable parameters. The linear embedding operation provides a straightforward mapping that retains important characteristics, offering a practical and computationally efficient solution for capturing unique attention traits.

This observer encoder can be seamlessly integrated into an existing scanpath model. As shown in Figure 3.2, a typical *Visual Encoder* is used to transform the input image into multi-channel feature maps $\boldsymbol{E}$ characterizing the bottom-up attention. To model the interaction between the visual feature $\boldsymbol{E}$ and the observer feature $\boldsymbol{u}$, an observer guidance map can be computed through a linear combination:

$$\boldsymbol{m}_u = \mathrm{softmax}\big(\boldsymbol{w}_{eu}^T \tanh(\boldsymbol{W}_{eu}\boldsymbol{E} + \boldsymbol{W}_{mu}\boldsymbol{u})\big), \tag{3.2}$$

where $\boldsymbol{w}_{eu}$, $\boldsymbol{W}_{eu}$, $\boldsymbol{W}_{mu}$ are learnable parameters. This observer guidance localizes salient image regions of specific interest to the observer.

Some scanpath models use a *Task Encoder* to process task-relevant information guiding the gaze behavior, such as a search target or a general question to answer. Such top-down guidance can be represented as a spatial attention map $\boldsymbol{m}_0$ prioritizing task-relevant regions. These bottom-up and top-down features are typically processed with a decoder (*e.g.*, LSTM or Transformer) to predict a sequence of probability maps $\{\boldsymbol{m}_t \,|\, t = 1, 2, \ldots, T\}$ and distribution parameters $\{(\mu_t, \sigma_t^2) \,|\, t = 1, 2, \ldots, T\}$ for sampling fixation positions and durations, respectively, where T is the number of fixations.

In Sections 3.3.2 and 3.3.3, we present specific modules that leverage the observer feature $\boldsymbol{u}$ to individualize the scanpath model. For clarity, our method description focuses on its integration with an LSTM model [50]. Please refer to Section 3.4.1 for details about its adaptation to a Transformer network [144].

3.3.2 Observer-Centric Feature Integration

With the encoded observer features characterizing each observer's distinct attention traits, we design observer-centric feature integration to address the critical need to fuse various inputs, including visual features, task relevance, and observer-specific characteristics, into a unified representation. The motivation behind this integration is to create a comprehensive understanding of individualized attention patterns. This integration process results in a sequence of observer-centric feature maps $\{\boldsymbol{R}_t \,|\, t = 1, 2, \ldots, T\}$ representing spatiotemporal fixation patterns, thus enabling the model to track individualized attention dynamics over time [50, 1, 76].

Specifically, to guide the prediction at each step, we leverage the predicted fixation distribution from the previous step (*i.e.*, $\boldsymbol{m}_{t-1}$) as a soft attention map, applying it to the visual features to derive the previously fixated visual features $\boldsymbol{X}_t = \boldsymbol{E} \circ \boldsymbol{m}_{t-1}$, where the symbol $\circ$ denotes the Hadamard product. It is noteworthy that the task guidance map $\boldsymbol{m}_0$ is used initially to guide the first fixation, mimicking the cognitive process that initially directs eye movements based on the visual task. Similarly, the observer guidance map $\boldsymbol{m}_u$ is used as the attention weights to obtain observer-centric visual features $\boldsymbol{X}_u = \boldsymbol{E} \circ \boldsymbol{m}_u$.

To seamlessly integrate the fixated visual features and observer-centric visual features, we concatenate the two types of feature maps

$$\boldsymbol{X}_{ut} = \boldsymbol{X}_t \parallel \boldsymbol{X}_u, \tag{3.3}$$

and perform spatial and channel-wise feature fusion, which are achieved by average-pooling the feature maps along the channel (AvgPool_c) and spatial dimensions (AvgPool_s), respectively, followed by linear layer processing and the addition of encoded observer features:

$$\boldsymbol{u}_t^s = \text{ReLU}(\boldsymbol{W}_{hs}\text{AvgPool}_c(\boldsymbol{X}_{ut}) + \boldsymbol{b}_{hs}) + \boldsymbol{W}_{us}\boldsymbol{u}, \tag{3.4}$$

$$\boldsymbol{u}_t^c = \text{ReLU}(\boldsymbol{W}_{hc}\text{AvgPool}_s(\boldsymbol{X}_{ut}) + \boldsymbol{b}_{hc}) + \boldsymbol{W}_{uc}\boldsymbol{u}, \tag{3.5}$$

where $\boldsymbol{W}_{hs}$, $\boldsymbol{W}_{hc}$, $\boldsymbol{W}_{us}$, $\boldsymbol{W}_{uc}$, $\boldsymbol{b}_{hs}$, and $\boldsymbol{b}_{hc}$ are learnable parameters. Ultimately,

combining $\boldsymbol{u}_t^s$ and $\boldsymbol{u}_t^c$ yields the final observer-centric feature maps

$$\boldsymbol{R}_t = \boldsymbol{u}_t^s \otimes \boldsymbol{u}_t^c, \tag{3.6}$$

where $\otimes$ is the outer product. It represents the dynamic importance of individual attention traits in the prediction of the current fixation, providing a more profound understanding of individualized visual behavior.

3.3.3 Adaptive Fixation Prioritization

While the observer-centric feature integration focuses on the fusion of input features, the adaptive fixation prioritization module addresses the variations of gaze behavior at the output end of the decoder. To achieve this, instead of directly predicting fixation positions, our approach, aimed at individualizing fixation predictions, takes a distinct path. We start by extracting semantic feature maps, denoted as $\boldsymbol{A}_t$, from the decoder. These feature maps are subsequently prioritized using attention weights specific to each observer, providing a pragmatic means of refining fixation outputs based on their unique attention traits.

To elaborate on the process, we begin by element-wise multiplication of the semantic feature maps $\boldsymbol{A}_t$ with the input visual features $\boldsymbol{E}$, and then perform average-pooling along the spatial dimensions, resulting in a feature vector that characterizes the observer's attention distribution across different semantic feature channels, defined as

$$\boldsymbol{V}_t = \mathrm{AvgPool}_s(\boldsymbol{E} \circ \boldsymbol{A}_t). \tag{3.7}$$

Considering that the visual preferences of various semantic features may vary for different observers, we introduce normalized attention weights $\boldsymbol{\beta}$ that prioritize the different feature channels, taking into account the observer feature:

$$\boldsymbol{\beta}_t = \mathrm{softmax}(\boldsymbol{w}_b^T \tanh(\boldsymbol{W}_b \boldsymbol{V}_t + \boldsymbol{W}_{um}\boldsymbol{u})), \tag{3.8}$$

where $\boldsymbol{W}_b$, $\boldsymbol{W}_{um}$ and $\boldsymbol{w}_b$ are learnable parameters. Finally, the attention weights $\boldsymbol{\beta}_t$

	OSIE [105]			OSIE-ASD [2]			COCO-Search18 [60]			AiR-D [1]		
Method	SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓
Human	0.386	0.808	7.486	0.370	0.783	7.720	0.458	0.809	1.777	0.405	0.801	7.966
SaltiNet [57]	0.151	0.739	8.790	0.137	0.735	8.688	0.127	0.712	3.821	0.116	0.747	10.661
PathGAN [56]	0.056	0.744	9.393	0.042	0.732	9.342	0.231	0.714	2.454	0.072	0.739	9.888
IOR-ROI [59]	0.294	0.791	7.966	0.301	0.788	7.655	0.197	0.787	7.087	0.239	0.791	8.584
ChenLSTM [50]	0.373	0.804	7.309	0.341	0.791	7.602	0.454	0.799	1.932	0.356	0.808	7.845
Gazeformer [144]	0.372	0.809	7.298	0.388	0.792	7.081	0.432	0.796	2.023	0.349	0.810	8.004
ChenLSTM-FT	0.378	0.808	7.344	0.394	0.796	7.067	0.454	0.804	1.936	0.341	0.806	8.282
Gazeformer-FT	0.373	0.810	7.319	0.387	0.795	7.083	0.432	0.796	2.026	0.350	0.812	8.068
ChenLSTM-ISP	0.377	0.810	7.284	0.401	**0.798**	**6.599**	**0.480**	**0.811**	**1.862**	**0.371**	0.813	**7.651**
Gazeformer-ISP	**0.390**	**0.813**	**7.163**	**0.406**	0.797	6.823	0.455	0.806	1.997	0.362	**0.814**	7.911

Table 3.1: Comparison of value-based evaluation results for models' ability to predict the scanpaths of individual observers.

are applied to the corresponding semantic feature maps $\boldsymbol{A}_t$ to compute the output

$$\boldsymbol{m}_t = \boldsymbol{\beta}_t^T \boldsymbol{A}_t. \tag{3.9}$$

This mechanism reshapes the scanpath prediction process into a weighted combination of multi-channel feature maps, allowing for the adaptive integration of these maps into the output fixation map. This approach allows the models to refine the fixation positions, providing a precise prediction of an individual's unique scanpath.

3.4 Experiments

This section reports comprehensive experimental results and analyses, demonstrating the effectiveness and generalizability of our method across various datasets, model architectures, and visual tasks.

3.4.1 Experiment Settings

Tasks and Datasets. We conduct experiments on four eye-tracking datasets featuring a variety of visual tasks, including free-viewing, visual search, and visual question answering: *OSIE* [105] comprising 700 images with free-viewing gaze data from 15 undergraduate and graduate students aged 18–30, *OSIE-ASD* [2] with free-viewing gaze data from 20 individuals with ASD and 19 controls, spanning ages 21 to 60, including 33 males and 6 females, *COCO-Search18* [60] (target-present subset) featuring 6202

images with gaze data from 6 males and 4 females aged 18 to 30, collected under a visual search task, and *AiR-D* [1] offering images and questions from the GQA dataset [94] with gaze and question-answering data from 16 males and 4 females aged 18 to 38. Dataset splits follow ChenLSTM [50] for the OSIE, OSIE-ASD, and AiR-D datasets, and the Gazeformer [144] for the COCO-Search18.

Evaluation Metrics. We conduct individualized scanpath prediction evaluation using two complementary sets of metrics: value-based metrics and ranking-based metrics. The **value-based** metrics measure the similarity or dissimilarity between the prediction and ground-truth scanpaths of the same observer. Different from existing studies [50] that compare a generic prediction with all observers' ground-truth scanpaths, we evaluate each individualized prediction against the corresponding observer's ground truth. Specifically, *ScanMatch (SM)* [55, 95] measures the similarity of fixation position and duration using the Needleman-Wunsch algorithm [96]; *MultiMatch (MM)* [97] measures scanpath similarity regarding shape, direction, length, position, and duration; *String-Edit Distance (SED)* [100, 98, 99] converts scanpaths into strings by associating each image region with a character. To evaluate how well the model predicts distinctly different scanpaths for different observers, we also employ **ranking-based** metrics. For each predicted scanpath, we rank the ground-truth scanpaths based on their ScanMatch similarity. *Recall at K (R@K)* [157, 158] quantifies whether the correct scanpath (*i.e.*, that from the same observer) is within the top-K most similar scanpaths. *Mean Reciprocal Rank (MRR)* [159, 160, 52] measures the quality of the ranking by calculating the reciprocal of the rank of the correct scanpath. Thus, the combination of value-based metrics focusing on the specific observer and ranking-based metrics considering all observers offers a comprehensive and robust performance evaluation.

Compared Models. We implement two individualized scanpath prediction models representing typical autoregressive and non-autoregressive sequential processing paradigms, respectively: *ChenLSTM-ISP* adapts the ChenLSTM [50] model, incorporating the observer encoder and the observer-centric feature integration for input processing. The model's LSTM decoder outputs are further modified for the proposed adaptive fixation prioritization. Similarly, we implement the *Gazeformer-ISP* model upon the Gazeformer [144] architecture. It replaces the original visual-semantic joint embedding with our observer-centric feature integration and changes the Transformer decoder outputs

	OSIE [105]			OSIE-ASD [2]			COCO-Search18 [60]			AiR-D [1]		
Method	MRR ↑	R@1 ↑	R@5 ↑	MRR ↑	R@1 ↑	R@5 ↑	MRR ↑	R@1 ↑	R@5 ↑	MRR ↑	R@1 ↑	R@5 ↑
SaltiNet [57]	0.213	5.619	32.286	0.107	2.454	12.454	0.293	10.114	49.804	0.295	10.210	49.930
PathGAN [56]	0.221	6.667	33.048	0.110	2.601	12.894	0.294	10.082	50.245	0.293	10.000	50.629
IOR-ROI [59]	0.218	6.762	31.524	0.109	2.784	12.454	0.292	9.673	50.507	0.291	9.814	48.567
ChenLSTM [50]	0.222	7.048	32.952	0.108	2.418	13.114	0.296	10.199	50.719	0.297	9.957	51.433
Gazeformer [144]	0.223	7.048	32.476	0.107	2.564	11.758	0.292	9.873	50.114	0.299	10.459	51.361
ChenLSTM-FT	0.225	6.667	34.381	0.113	2.711	12.637	0.298	10.641	49.820	0.294	10.118	50.262
Gazeformer-FT	0.217	6.000	32.857	0.108	2.528	13.223	0.293	10.183	50.000	0.300	9.599	51.863
ChenLSTM-ISP	**0.291**	**12.667**	**44.095**	**0.147**	**4.835**	**19.194**	**0.369**	**16.639**	**61.769**	**0.338**	**13.610**	57.235
Gazeformer-ISP	0.268	10.095	41.905	0.141	4.286	18.571	0.353	15.299	60.020	0.334	13.539	**57.450**

Table 3.2: Comparison of ranking-based evaluation results for models' ability to distinguish different observers.

Modules			ChenLSTM						Gazeformer					
OE	FI	FP	SM ↑	MM ↑	SED ↓	MRR ↑	R@1 ↑	R@5 ↑	SM ↑	MM ↑	SED ↓	MRR ↑	R@1 ↑	R@5 ↑
			0.341	0.791	7.602	0.108	2.418	13.114	0.388	0.792	7.081	0.107	2.564	11.758
✓			0.377	0.791	7.112	0.110	2.601	13.000	0.397	0.796	7.079	0.122	3.017	15.092
✓	✓		0.389	0.795	7.064	0.122	3.150	15.238	0.398	0.796	6.982	0.134	3.810	17.509
✓		✓	0.389	0.795	7.063	0.112	2.784	13.150	0.397	0.797	7.073	0.120	3.077	15.165
✓	✓	✓	**0.401**	**0.798**	**6.599**	**0.147**	**4.835**	**19.194**	**0.406**	**0.797**	**6.823**	**0.141**	**4.286**	**18.571**

Table 3.3: Ablation study for the proposed technical components: observer encoder (OE), observer-centric feature integration (FI), and adaptive fixation prioritization (FP).

from fixation coordinates to feature maps. We compare these ISP models with their general counterparts and other general scanpath prediction models, including SaltiNet [57], PathGAN [56], and IOR-ROI [59]. In addition, we fine-tune the general models on individual observer data (*i.e.*, ChenLSTM-FT, Gazeformer-FT) to provide a baseline for assessing the impact of explicitly incorporating observer-specific characteristics.

Implementation Details. We implement ChenLSTM [50] and Gazeformer [144] following the original methods, such as using the same visual encoder (*i.e.*, ResNet-50 [6]) and task encoder (*i.e.*, RoBERTa [161] or AiR-M [1] or CenterNet [106] object detector). For both models, the number of output feature channels for $\boldsymbol{A}_t$ is empirically set to 4. Specifically, for ChenLSTM [50] and Gazeformer [144], we adopt supervised learning for 15 epochs and self-critical sequence training (SCST) [50, 87] for the remaining 10 epochs. In supervised learning, we train our model using the Adam [101] optimizer with learning rate 10^{-4} and weight decay 5×10^{-5}, while in the SCST, we linearly decayed learning rates starting at 10^{-5}. To improve the learning of discriminative features across observers, each training batch includes different scanpaths for the

same image.

3.4.2 Quantitative Results

We present value- and ranking-based evaluation results to assess the effectiveness of our ISP models in capturing the unique attention traits of individual observers.

Table 3.1 presents the **value-based** evaluation results revealing how model predictions resemble the ground truth scanpath of each observer. While fine-tuning leads to minor improvements in some cases (*e.g.*, OSIE and OSIE-ASD), it struggles on datasets with less distinct inter-observer differences (*e.g.*, COCO-Search18 and AiR-D). In contrast, the ISP models consistently outperform the general methods and fine-tuning, indicating their ability to adapt to the unique attention traits of observers. This is particularly evident in the improved performance (*e.g.*, Gazeformer-ISP, SM=0.406) on the OSIE-ASD dataset with a diverse range of observer demographics. These results suggest that our method, by directly targeting the modeling of observer-specific attention patterns, offers more robust and effective individualization.

Table 3.2 presents **ranking-based** evaluation comparing models' ability to distinguish ground-truth scanpaths. General models, which are observer-agnostic, cannot differentiate the ground-truth scanpaths from similar ones (*e.g.*, ChenLSTM, R@1=2.4% on OSIE-ASD, lower than random). Even after fine-tuning with individual eye-tracking data, their performance improvements are marginal (*e.g.*, ChenLSTM-FT, R@1=2.7% on OSIE-ASD), because independently tuning parameters cannot effectively learn features that distinguish each observer from the others. Differently, the individualized models achieve promising results across all metrics and datasets. From ChenLSTM to ChenLSTM-ISP, R@1 is significantly improved to 4.8% on the OSIE-ASD dataset, doubling the probability of finding the correct scanpath. It suggests that the ISP models can predict scanpaths that align closely with an observer's unique attention traits. Between network architectures, ChenLSTM-ISP consistently outperforms Gazeformer-ISP when ranking scanpaths. This performance gain may be attributed to LSTM's autoregressive nature which is more effective than Transformer's parallel approach in learning fine-grained spatiotemporal differences.

3.4.3 Ablation Study

To evaluate the significance of the three technical components: observer encoder (OE), observer-centric feature integration (FI), and adaptive fixation prioritization (FP), we conduct an ablation study on the OSIE-ASD dataset [105] by applying them incrementally to the ChenLSTM and Gazeformer models. Table 3.3 shows that a fundamental module OE results in a significant improvement in the value-based evaluation and highlights its role of encoding attention traits of observers. Furthermore, based on OE, both FI and FP have notable impacts on the model performance. First, both components achieve similar performance improvements in SM, MM, and SED, demonstrating their ability to improve the overall accuracy of scanpath predictions. Further, regarding the MRR, R@1, and R@5 metrics, FI results in more significant improvements than FP, suggesting that the seamless integration of various input features is more substantial than FP's ability to prioritize where to look at the output end. We also notice that combining both modules leads to the most significant overall performance improvements, indicating that FI and FP offer complementary enhancements.

3.4.4 Qualitative Examples

To understand how the predicted scanpaths align with observer-specific gaze patterns, we present a qualitative comparison in Figure 3.3. Figure 3.3a and Figure 3.3b compare the scanpaths between an observer with autistic traits and a non-autistic observer. It can be seen that observer (a) focused on the center of the image while avoiding direct gaze at people, while observer (b) looked at people more frequently. Figure 3.3c and Figure 3.3d compare the scanpaths of two observers responding to the question 'What is the device on top of the nightstand made of wood?' with different answers. Observer (c) successfully found the correct answer 'phone' by searching broadly within the image, but observer (d) responded with an incorrect answer 'television' because the fixations were mostly distributed around the television. Notably, while the fine-tuning approach (column 1) falls short in capturing observer-specific gaze patterns, the ISP models' predictions (column 2) better align with the scanpaths of the human observers (column 3). This capability of ISP models opens up new avenues for understanding and interpreting individual differences in visual perception and decision-making processes.

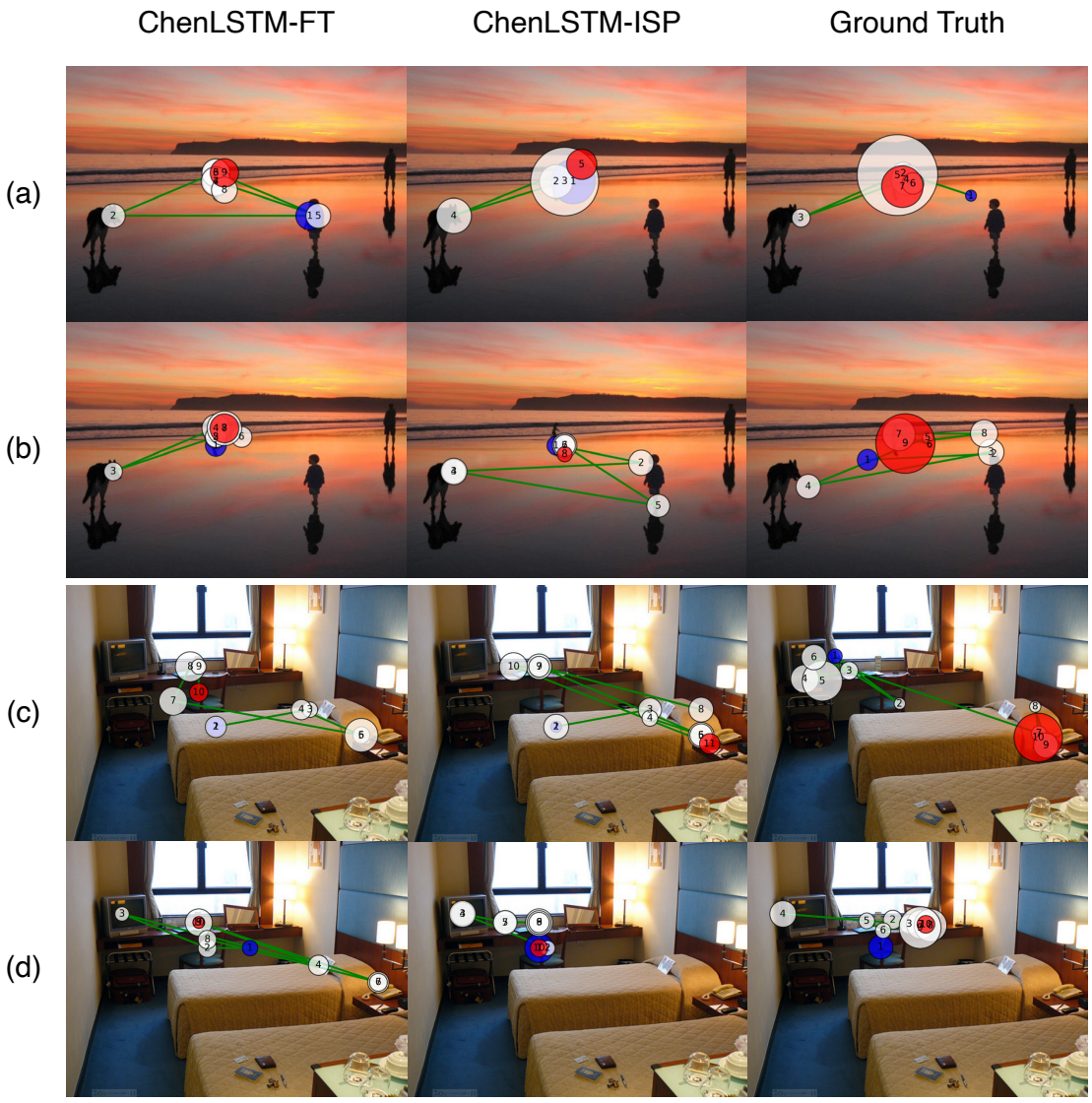

Figure 3.3: Qualitative examples of scanpaths predicted by ChenLSTM-FT, ChenLSTM-ISP, and ground truth. Each row compares the model predictions and the ground truth scanpath of one observer. These observers show different gaze patterns, including (a) focusing on the image center, (b) exploring different people and objects, (c) exploring broadly in the scene, and (d) focusing on a particular region. The blue and red dots indicate the beginning and the end of the scanpath, respectively.

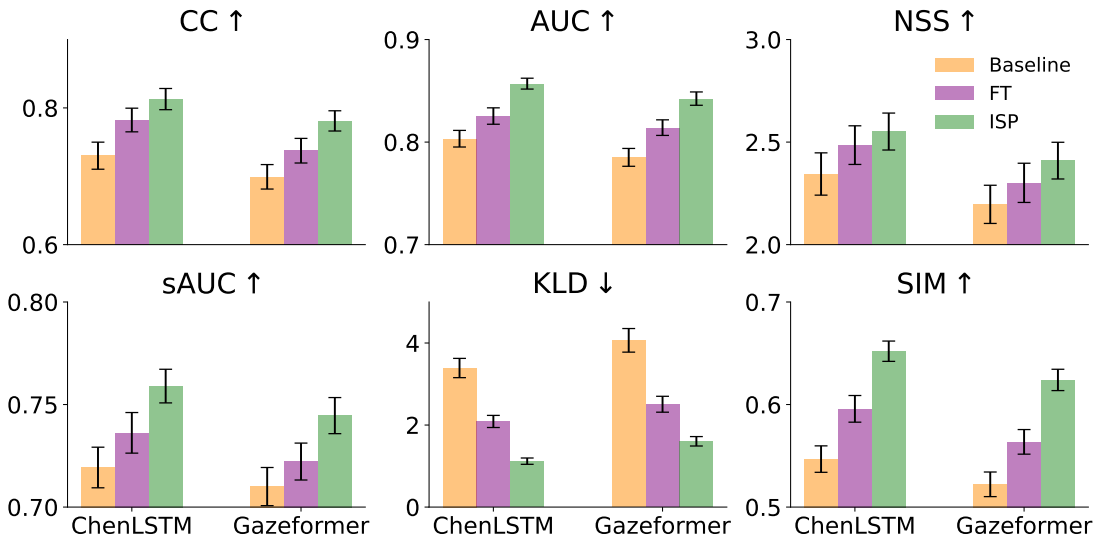

Figure 3.4: Saliency evaluation results of the baselines, fine-tuned (FT) models, and ISP models. Error bars indicate the standard error of the mean.

3.4.5 From Scanpaths to Saliency Maps

To further confirm the effectiveness of our ISP method, we assess the spatial accuracy of the predicted fixations using established saliency evaluation metrics [133, 138], including Linear Correlation Coefficient (CC), Area Under the ROC curve (AUC), Normalized Scanpath Saliency (NSS), shuffled AUC (sAUC), Kullback-Leibler divergence (KLD), and similarity metric (SIM). Saliency maps are generated by aggregating predicted fixations from all observers and applying a Gaussian kernel smoothing to all fixation points. Figure 3.4 shows the substantial improvement of the ISP models over the baselines and fine-tuned models when applied to the OSIE-ASD [2] dataset. This improvement shows that our method not only accurately predicts individual observers' fixations but also enhances the overall prediction of fixation distributions for the population.

3.4.6 Semantic Analyses

Moving forward, we conduct statistical analyses on the OSIE-ASD dataset to test ISP models' ability to learn the attention differences across observers and populations. While the evaluations above focus on fixation positions and durations, this analysis considers

how the predicted fixations align with the ground truth regarding their semantic-level statistics. Specifically, we group fixations into three categories based on the region of interest (ROI) annotations provided by OSIE [105], which are social regions (directly relating to humans, including faces, emotion, touched, gazed), nonsocial regions (*e.g.*, implied motion, relating to nonvisual senses, designed to attract attention, and other objects), and background. Each observer has a unique fixation distribution over the three categories (*i.e.*, social, nonsocial, and background), which enables the following individual-level and population-level analyses.

Method	Social	Nonsocial	Background
ChenLSTM [50]	0.181	-0.159	0.067
Gazeformer [144]	-0.141	-0.253	-0.211
ChenLSTM-FT	0.137	0.040	-0.166
Gazeformer-FT	0.045	0.164	0.051
ChenLSTM-ISP	**0.621**	**0.655**	**0.720**
Gazeformer-ISP	**0.692**	**0.572**	**0.699**

Table 3.4: Spearmans' correlation coefficients of fixation proportions in 3 semantic ROIs (*i.e.*, social, nonsocial, and background) between the ground truth and predictions. Bold numbers indicate significant positive correlations ($p < 0.05$).

Individual Level. To evaluate how the predicted scanpaths resemble human fixation statistics, we rank observers by their proportion of fixations in each category. The fixations can be obtained from the model predictions or the ground truth. Table 3.4 presents Spearman's rank correlation coefficient [162] to compare the observer rankings between the predictions and the ground-truth fixations. While fine-tuning is less effective, showing low correlations across all categories, ISP models consistently achieve significant and high positive correlations, suggesting their ability to resemble each human observer's unique fixation patterns.

Population Level. Beyond individual characterization, ISP models also effectively capture and reproduce distinctive attention traits observed at the population level. For example, individuals with ASD exhibit lower proportions, higher latency, and shorter duration of fixations to both social and nonsocial cues [2, 131, 132]. Figure 3.5 shows that fixations predicted by the ISP models achieve similar statistics. The statistical agreement between the model predictions and the ground-truth scanpaths demonstrates our

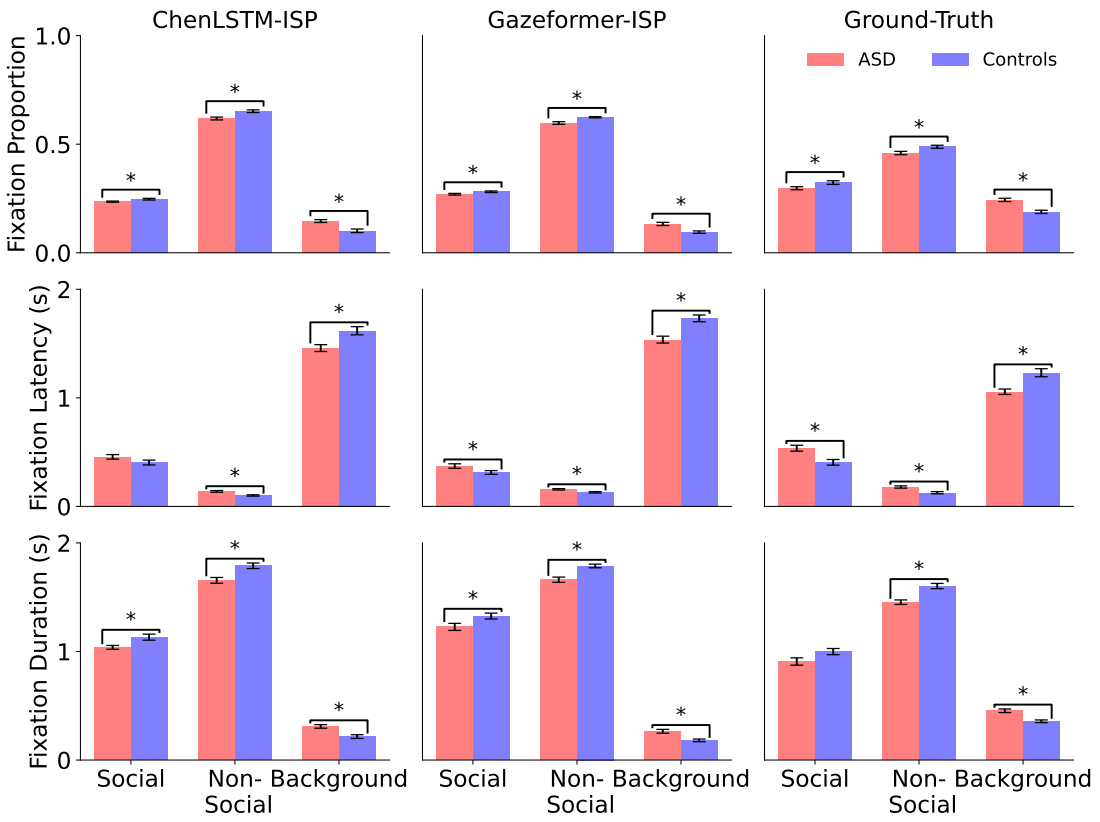

Figure 3.5: Statistical comparison between the predicted fixations for the ASD and Control groups [2]. Error bars indicate the standard error of the mean. Asterisks indicate significant differences (unpaired t-test, $p < 0.05$).

method's ability to generalize and represent population-level characteristics, reinforcing its potential utility in a variety of applications.

3.4.7 Application

To showcase the potential applicability of ISP models in the diagnosis of neurodevelopmental conditions, we visualize ISP model features and use these features to classify people with ASD. First, the individualization ability of our method is highlighted through t-distributed stochastic neighbor embedding (t-SNE) visualization. By concatenating all observer-specific features from Equations (3.2), (3.4), (3.5), and (3.8),

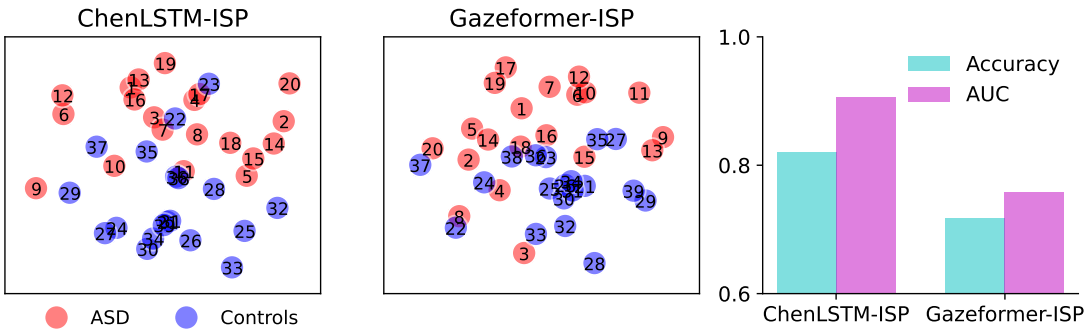

Figure 3.6: Visualization of features extracted from ISP models (numbers indicate observer identities) and results of ASD classification using the features.

into $\boldsymbol{v} = [\boldsymbol{W}_{mu}\boldsymbol{u} \parallel \boldsymbol{W}_{us}\boldsymbol{u} \parallel \boldsymbol{W}_{uc}\boldsymbol{u} \parallel \boldsymbol{W}_{um}\boldsymbol{u}]$, where $\parallel$ represents the vector concatenation, Figure 3.6 shows that the ISP model features can clearly distinguish people with ASD from the controls. It is noteworthy that such features are learned in an unsupervised manner without knowing each observer's class label, suggesting the strong learning power of the ISP models. Further, based on a leave-one-out cross-validation, we train a two-layer perceptron to classify people with ASD using the extracted feature $\boldsymbol{v}$. ChenLSTM-ISP and Gazeformer-IPS achieve 82.1% and 71.8% classification accuracy, respectively, similar to clinical gold standards [163, 164]. These results demonstrate ISP models' potential in real-world healthcare applications.

3.5 Conclusion

We have introduced a novel approach to predicting individualized human visual scanpaths. Our approach features three novel components: observer encoder, observer-centric feature integration, and adaptive fixation prioritization. Through extensive experiments across multiple datasets, network architectures, and visual tasks, our method consistently outperforms state-of-the-art scanpath prediction methods and individualization based on observer-specific fine-tuning. The results demonstrate the method's ability to generate human-like scanpaths and account for individual observers' gaze

patterns. By providing a better understanding of how individuals process visual information, our study has significant implications for tailored, user-centric solutions, such as improving the design of interfaces, products, and services across a wide range of application domains.

Chapter 4

VisualHow: Multimodal Problem Solving

Recent progress in the interdisciplinary studies of computer vision (CV) and natural language processing (NLP) has enabled the development of intelligent systems that can describe what they see and answer questions accordingly. However, despite showing usefulness in performing these vision-language tasks, existing methods still struggle in understanding real-life problems (*i.e.*, how to do something) and suggesting step-by-step guidance to solve them. With an overarching goal of developing intelligent systems to assist humans in various daily activities, we propose VisualHow, a free-form and open-ended research that focuses on understanding a real-life problem and deriving its solution by incorporating key components across multiple modalities. We develop a new dataset with 20,028 real-life problems and 102,933 steps that constitute their solutions, where each step consists of both a visual illustration and a textual description that guide the problem solving. To establish better understanding of problems and solutions, we also provide annotations of multimodal attention that localizes important components across modalities and solution graphs that encapsulate different steps in structured representations. These data and annotations enable a family of new vision-language tasks that solve real-life problems. Through extensive experiments with representative models, we demonstrate their effectiveness on training and testing models for the new tasks, and there is significant scope for improvement by learning effective attention

mechanisms.

Figure 4.1: VisualHow is a vision-language task aiming to infer the solution to a real-life problem. The solution consists of multiple steps each described with an image and a caption.

4.1 Introduction

The remarkable progress in vision-language studies has developed visual systems with the ability to understand and generate natural language information. Existing vision-language models mainly focus on the understanding of visual input in task-free (*i.e.*, Image Captioning [88, 165, 166] and Visual Storytelling [167]) or question-directed (*i.e.*,

Visual Question Answering [168, 169] and Visual Dialog [159]) settings. In other words, their aim is to develop visual systems that can "look and tell", by describing or answering questions about what is observed in a scene. On large-scale vision-language datasets [170, 166, 167, 168, 169, 171, 159, 172, 173, 174], state-of-the-art models have obtained promising achievements in understanding and predicting visual and textual information. Although achieving significant progress, these methods only perform well on standardized vision-language inference benchmarks and do not generalize to real-life situations to solve problems, which makes their scope of application relatively limited.

We believe that the next generation of visual intelligence systems will need to develop the ability to help humans solve real-life problems more directly. Achieving the goal requires them to provide step-by-step solutions with both textual descriptions and visual illustration. Applications of such systems may include: 1. Teaching people everyday and/or domain-specific skills, such as to tie a tie, to make a sandwich, or to change a bicycle tire. 2. Helping people decompose an abstract goal into actionable items, such as to improve social skills, to sleep better, or to become a soccer player. To this end, we introduce a novel research problem – VisualHow – along with a large-scale dataset and a systematic evaluation of various modeling approaches. The main objective of VisualHow is to generate a step-by-step vision-language description of how to solve a problem, where a step will be described using an image and a caption. An example of VisualHow data is shown in figure 4.1. To "involve a pet in Christmas", one may need to take a series of different actions. While people may still find it difficult to understand how to involve a pet in Christmas by only reading the textual descriptions, looking at the visual illustrations will offer great help in the process. Therefore, given the description of the problem and the previous steps, the specific goal of VisualHow is to predict a pair of well-matched and complementary image and caption to describe what to do next. Achieving the goal requires the ability to understand three types of relationships: the relationship between the problem and the solution, the relationships between different steps of the solution, and the relationships between the visual and textual information.

Our goal is to enable the development of intelligent systems for tackling various real-life problems. Compared to conventional vision-language tasks, our proposed VisualHow task has the following differentiating factors: 1. **Real-life problems and**

multimodal solutions. Rather than focusing on specific vision-language tasks [167, 159, 170, 175, 168, 169], our dataset contains 18 categories and 317 subcategories of real-life problems. Solutions to these problems are described in multiple steps, each with an image-caption pair, enabling the understanding of the decision-making process in problem solving. 2. **Fine-grained annotations.** Our VisualHow dataset offers two types of annotations that are absent from existing studies: the solution graphs describing dependencies between different steps, and multimodal attention that highlights and associates important keywords and regions of interest. They play an essential role in developing a structured understanding of the problem-solving procedure and closing the semantic gaps between vision and language. 3. **New vision-language tasks.** Our dataset enables several new vision-language tasks for various aspects of problem solving. Our experiments lead to several interesting observations and suggestions on improving the model performance.

To summarize, the contributions of this work are:

1. A new VisualHow study aiming to provide the foundation for developing novel vision-language methods and pushing the boundaries of multimodal understanding of real-life problems and solutions;

2. A new dataset that consists of diverse categories of problems, multimodal descriptions of solutions, and fine-grained annotations;

3. Experiments on multiple new tasks on different aspects of the VisualHow problem and extensive analyses of various baseline models.

4.2 Related Work

This paper is related to a series of studies including visual captioning and storytelling, visual question answering and dialog, multimodal instructions and multimodal representation learning.

4.2.1 Visual Captioning and Storytelling

There is a large body of research centering around generating textual descriptions of visual inputs. For example, the image captioning task [170, 175, 176, 177, 178] focuses on describing a single image with natural language, while visual storytelling [167] aims

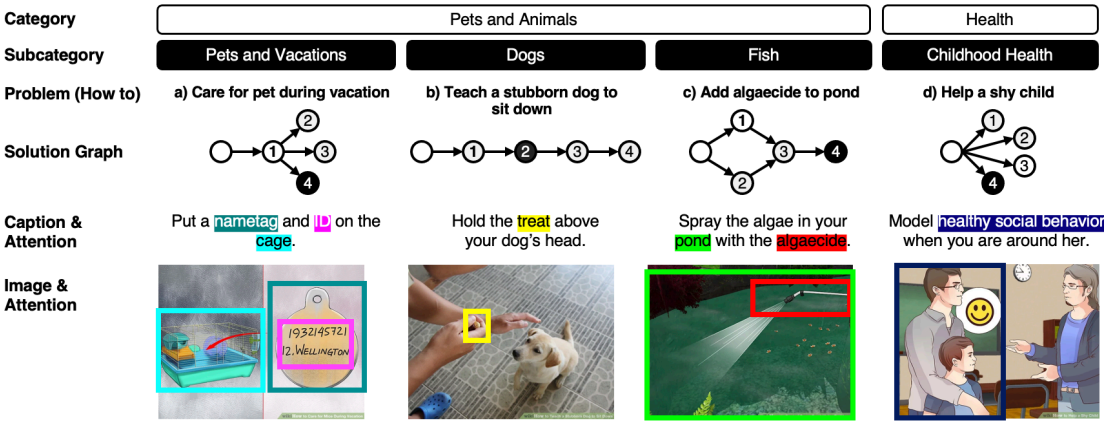

Figure 4.2: An overview of the VisualHow dataset. We provide a hierarchical structure that organizes our data into categories, sub-categories, problems, solution graphs, steps with image-caption pairs, and multimodal attention. Example steps are highlighted in the solution graph. Steps without a dependency are connected to an empty node.

to generate a narrative with a sequence of sentences about multiple images. The shared goal of these studies is to develop methods to effectively encode the input images into representative features and transform them into a sequence of words that naturally and fluently describes the images. Therefore, in their standard configuration, image captioning and visual storytelling are image-to-sequence prediction tasks whose inputs are pixels and outputs are a sequence of words decoded according to a given vocabulary. While they focus on passively describing visual inputs without being directed by a specific purpose, the VisualHow task is different: First, it jointly predicts the images and captions that complement each other for the description of a solution, and second, the prediction is conditioned on the problem to solve. These differences make VisualHow a distinct and challenging research problem.

4.2.2 Visual Question Answering and Dialog

Previous studies have attempted to solve simple problems. For example, visual question answering [168, 169] and visual dialog [159, 179, 180] aim to answer questions about visual information based on the understanding of multimodal inputs. A number of recent studies have proposed large-scale datasets [168, 169, 94, 159] and neural network

models [181, 182, 183, 184, 185] for free-form and open-ended VQA and visual dialog. However, these studies typically have restricted categories of questions, and their answers are in simplified forms (*i.e.*, categories or short phrases) [159]. On the contrary, the goal of our VisualHow task is to provide step-by-step description of the solution for various types of real-life problems. It not only requires the ability to understand both visual and textual information, but also involves constructing a reasonable structure of solutions to represent the relationship between different steps.

4.2.3 Multimodal Instructions

Our work is also related to existing studies on multimodal instructions. Datasets of instructional images [186, 187] and videos [188, 189, 190, 191, 192] provide step-by-step instructions about specific tasks. These datasets either focus on specific tasks or do not consider complex attention or structure in solutions. However, understanding the textual description of problems and providing step-by-step solutions each with a pair of well-matched captions and images have not been considered. Our work is different by considering diverse contents, multimodal attention, and solution structures, where the captions and images jointly describe the solution rather than each other. It contributes a large dataset with diverse and challenging problems, multimodal attention annotations, and non-sequential solutions.

4.2.4 Multimodal Representation Learning

There has been a long line of studies aiming to learn vision-language representations [157, 158, 193, 194, 195, 196]. They improve the representation learning using advanced attention mechanisms [197], better multimodal fusion methods [198, 199], multistep reasoning [200, 1], incorporation of object relations [195, 201, 196] and compositional reasoning models [202, 203]. Our study is most related with visual semantic embedding (VSE) [158, 193, 157, 204, 205, 206, 207], a typical category of approaches that learn a joint embedding space for visual and language representations. With VSE, compatibility score of visual features and language features can be computed as a simple inner-product. Specifically, DeViSE [204] learns to match the visual embeddings and semantic embeddings for zero-shot image recognition[208]. LSTM-SCNLM [205] encodes

the sentence as the semantic embedding via bi-directional LSTMs. VSE++ [193] is a fundamental VSE method that uses average pooling as the feature aggregator with online hard-negative mining. VSRN [206] captures key objects and semantic concepts of a scene to generate visual representations. Global pooling operation (GPO) [157] learns to automatically adapt itself to the best pooling strategy for different features while staying effective and efficient. These studies have provided suitable baselines for the proposed VisualHow task, and inspired the development of computational models for problem-solving in real-life scenarios.

4.3 The VisualHow Dataset

The goal of this work is to introduce a new benchmark with a focus on real-life problems and high-quality annotations to the community of vision-language understanding. It consists of 18 categories of real-life problems and step-by-step solutions described with images and captions. The diversity and generality of problems and solutions also make VisualHow a more challenging dataset. In addition to the image-caption pairs, VisualHow provides annotations for solution graph and multimodal attention, which are essential for the understanding of problem-solution relationships and aligning the semantics between vision and language. An example of VisualHow data is shown in figure 4.2. In this section, we describe the data collection method, annotations, and the data statistics.

4.3.1 Problems and Solutions

Building a general problem-solving dataset brings a series of unprecedented challenges. First, with the diversity and generality of real-life problems, manually defining and categorizing problems is impractical. Next, since many of the problems require domain expertise (*e.g.*, those about health or finance), general online contents or non-expert workers can hardly generate high-quality solutions. To address these challenges, we collect real-life problems and solutions from the wikiHow [209, 210, 211] knowledge base, which is known for its high-quality instructional articles. The wikiHow articles are generated by a pool of well-qualified experts with the help of a rigorous quality screening process. All articles come with detailed step-by-step descriptions and very

relevant visual illustrations in high resolution. Specifically, each problem consists of a language description (*e.g.*, a question starting with "How to") and is provided with a step-by-step solution that describes a method to solve it. The solution is composed of multiple steps described with an image and a caption. To control the data quality, VisualHow focuses on the proportion of wikiHow data with higher user ratings and popularity. A distinction from other wikiHow-based datasets is that for VisualHow we only select contents created by domain experts and with more than 50% of the users who voted and find it helpful, which ensures the quality of VisualHow contents. For problems with multiple solution methods, we consider each method a unique sample with the method title appended to the problem description.

4.3.2 Data Annotation

Learning to solve general problems is a challenging task, which requires knowledge to be learned from a variety of visual and textual information and organized in a structured representation. To tackle these challenges and benefit the development of future vision-language understanding methods, VisualHow offers fine-grained annotations on the solutions. As shown in table 4.1, as distinguished from related studies, we collect these annotations with crowdsourcing and implement an effective quality control mechanism.

Crowdsourcing. The annotations are conducted in Amazon Mechanical Turk (AMT) with a custom annotation paradigm and a user interface (see figure 4.3). First, an overview of the problem and solution (*i.e.*, the wikiHow article) is presented to the workers. Next, they browse through all steps one at a time. In each step, they select the important phrases from the caption and annotate the corresponding image regions, which reflects their attention towards the multimodal information when performing different actions. Finally, they are asked to annotate the dependency between each pair of steps, which will formulate a directed solution graph to provide a structured representation of the problem-solving process. This research does not collect personal data from crowd workers and is exempt from IRB review.

Quality control. Our dataset requires an effective mechanism for quality control, so crowd workers can generate high quality annotations. Collecting high-quality multimodal annotations is challenging. For example, determining what are important and

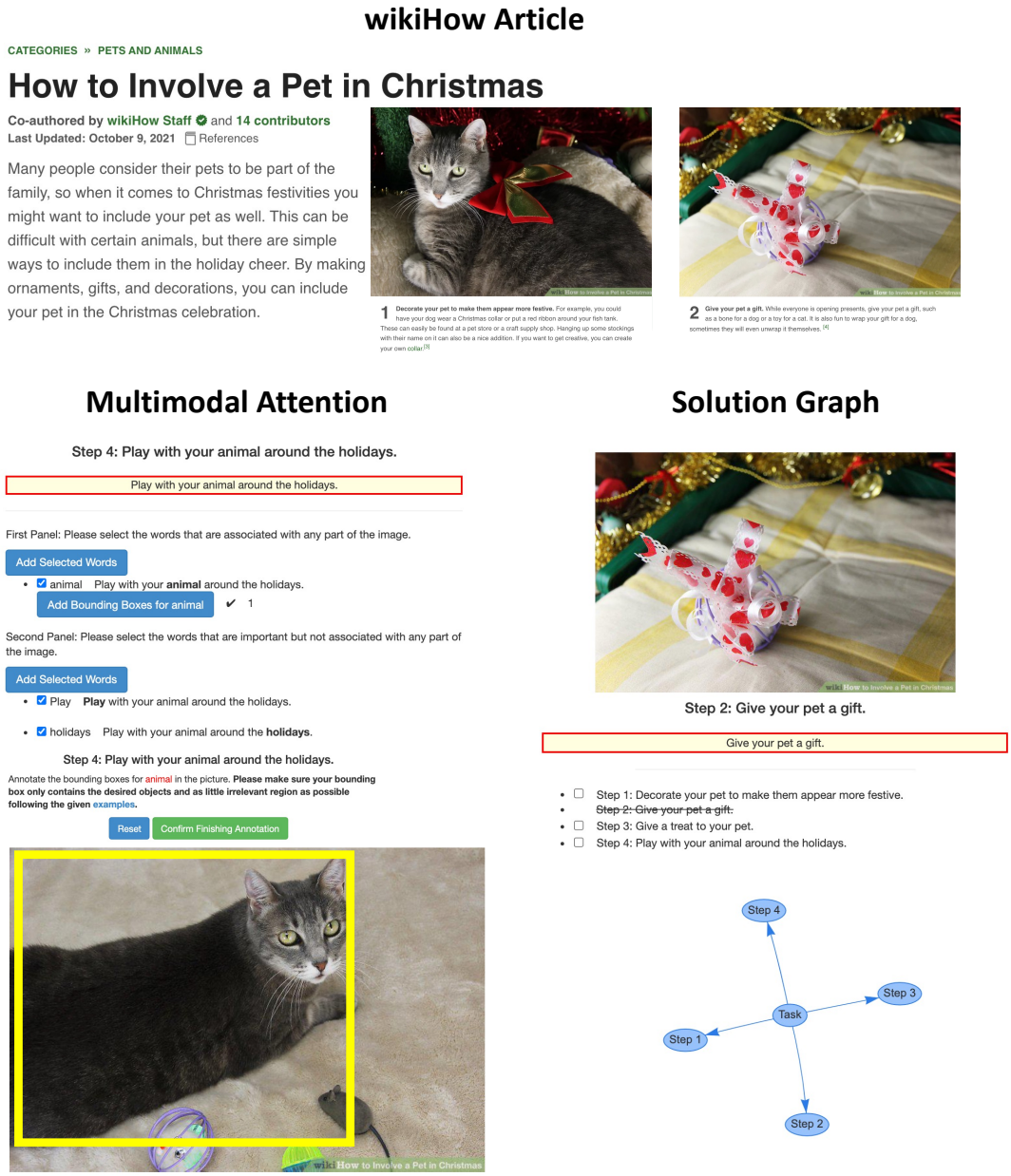

Figure 4.3: Crowdsourcing interface of the VisualHow task, which contains 1) an overview of the wikiHow Article, 2) annotation of the multimodal attention, and 3) annotation of the solution graph.

	VisualHow	ViPT [187]	COIN [191, 192]
Data Source	wikiHow	Snapguide; instructables	YouTube
Multimodal Attention	Yes	No	No
Solution Graph	Yes	No	No
Solution Types	Various	Procedure	Procedure

Table 4.1: Comparison between VisualHow and related datasets.

need to be annotated can be subjective. To control the data quality, objectiveness, and consistency, we implement a series of quality control methods including qualification, correction, and expert review. We first compose a list of specific guidelines and require each worker to complete two qualification Human Intelligence Tasks (HITs), and examine the quality and consistency of their annotations to make sure that both the multimodal attention and the solution graphs are reasonably correct. At the end of each HIT, the workers are asked to review their annotations and correct what they find problematic. We review the HITs with automatic anomaly detection and manual examination, and problematic HITs will be sent back to the workers for correction. Through these steps, we ensure all workers follow the same quality standard.

4.3.3 Data Statistics

The VisualHow dataset consists of 20,028 real-life problems and their solutions that vary in number of steps and fine-grained annotations. In this section, we conduct in-depth analyses and report key statistics of the dataset.

Problems. VisualHow contains 20,028 problems grouped in a hierarchy of 18 categories and 317 subcategories. Some of our major categories, such as Family Life, Computers and Electronics, Health, Finance and Business, have been rarely explored in previous vision-language studies. As shown in figure 4.4, the number of problems in each category ranges from 405 to 2,952, reflecting a naturally skewed distribution of wikiHow data. Despite that, VisualHow is still much more diverse than related datasets such as ViPT [187] and COIN [191], where a vast majority of samples are cooking or other household problems.

Solutions. As shown in figure 4.5, each solution consists of 3 to 10 steps described

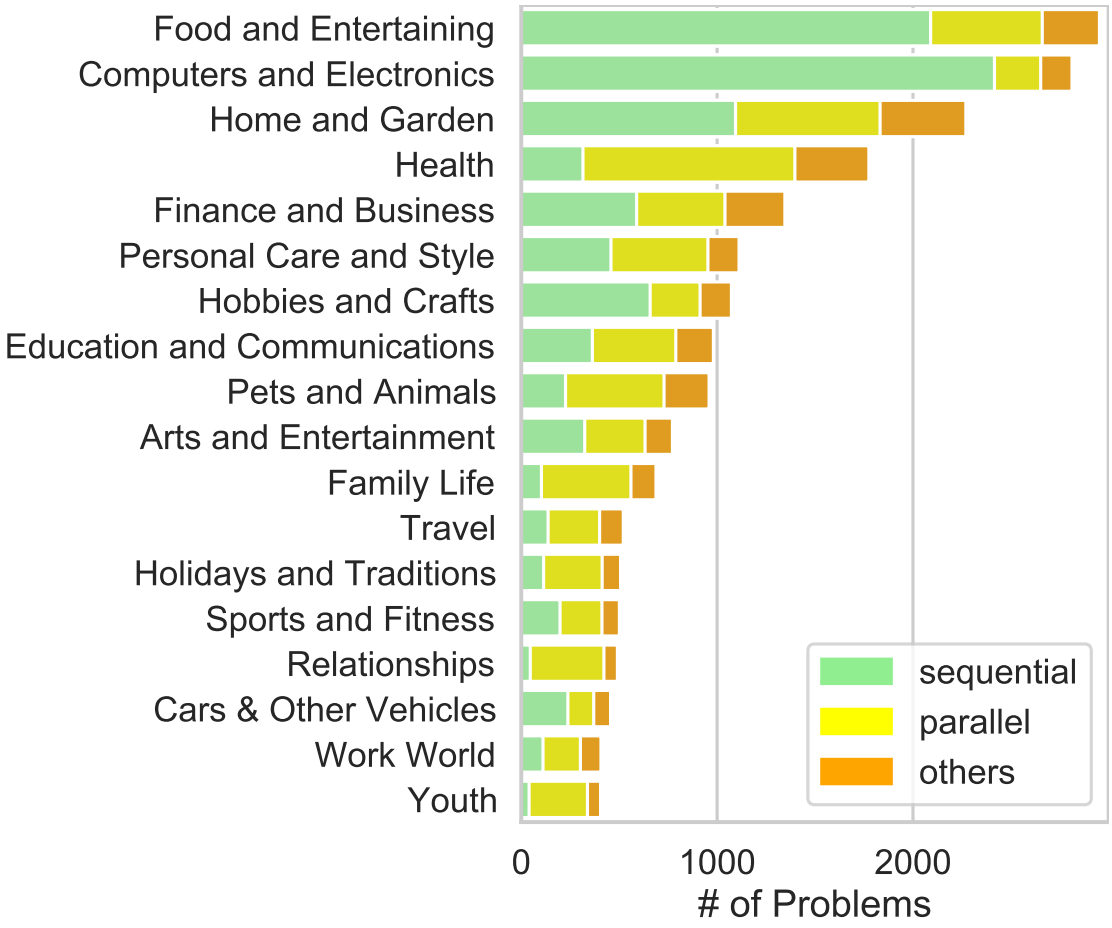

Figure 4.4: Number of problems in each category and the three types of solution graphs.

Nouns		Verbs		Other POS	
water	icon	click	want	new	right
minutes	account	tap	open	around	first
time	hair	use	take	one	small
button	area	make	choose	sure	away
app	oil	add	remove	next	dry

Table 4.2: Most common tokens in the caption among 1) nouns; 2) verbs; 3) other parts of speech (POS).

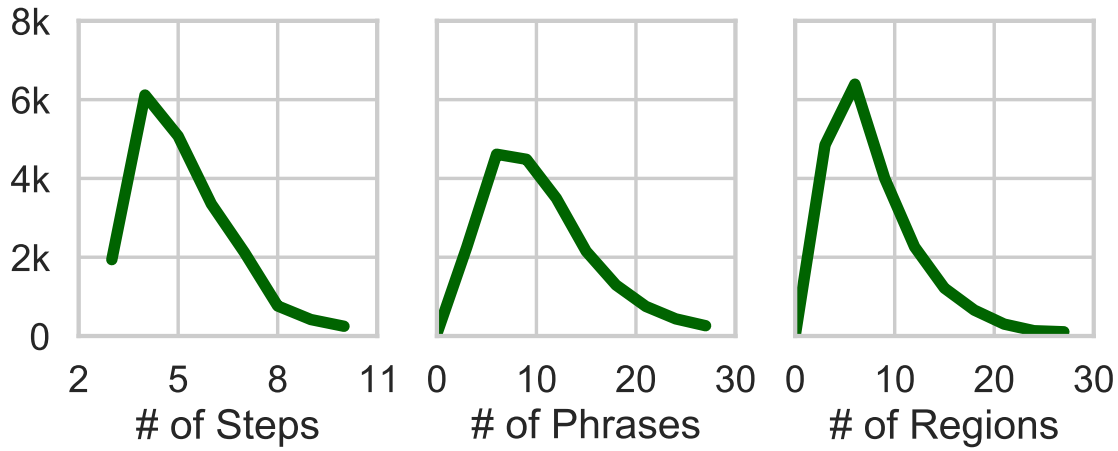

Figure 4.5: Distribution of solution steps and attention annotations.

with images and captions. On average, each solution consists of 5.14 steps. The images and captions are more diverse than existing datasets, thanks to the wide variety of wikiHow data. Of all the images, 36.5% are realistic photos, and 58.6% are abstract images such as cartoons, drawings, handwritings, charts, *etc*. The rest 4.9% are mixed with both realistic and abstract contents. The captions also have a vocabulary of 30k tokens. table 4.2 shows the most common nouns (52.3%), verbs (30.0%), and other parts of speech (17.7%) in the captions.

Multimodal attention. We have collected abundant multimodal annotations about important image regions and phrases, which enables fine-grained learning of visual semantic alignment. In figure 4.5, on average, 9.13 image regions and 11.69 phrases are annotated for each solution. Over 98% of all steps have at least one instance of multimodal attention in both the image and caption, and around 99.5% of all steps have at least one annotated phrase. For each step of a solution, an average of 1.56 instances of multimodal attention are annotated in both the image and the caption. In addition, each step has an average of 0.72 important phrases annotated without specific image regions, and 0.13 image regions are annotated without their corresponding phrases. In table 4.3, the tokens in the annotated phrases include nouns (60.8%), verbs (31.6%), and other parts of speech (7.6%). Compared to their distributions in the captions (see table 4.2), the annotations contain more nouns, corresponding to various object instances in the

Nouns		Verbs		Other POS	
water	oven	click	remove	online	overnight
doctor	bowl	tap	select	ok	comfortable
hair	child	open	choose	together	inside
settings	oil	add	check	regularly	daily
ingredients	food	use	make	outside	warm

Table 4.3: Most common tokens in the annotated phrases among 1) nouns; 2) verbs; 3) other parts of speech (POS).

	Image					Caption				
Method	MRR	R@1	R@3	R@5	Mean	MRR	R@1	R@3	R@5	Mean
GAP	0.495	28.583	62.313	77.882	3.758	0.535	34.449	64.785	79.284	3.558
GPO	0.501	29.441	62.249	78.676	3.695	0.549	35.695	67.165	81.203	3.392
ATT	0.505	29.589	63.420	79.579	3.649	0.572	38.563	69.240	83.213	3.186
ATT+CE	**0.507**	**29.865**	**63.688**	**79.644**	**3.620**	**0.580**	**39.402**	**69.950**	**83.592**	**3.133**

Table 4.4: Quantitative results of Task 1: solution steps prediction.

images. The abundance of verbs and others allow to infer a variety of semantics in both modalities. On average, each annotated image region takes about 36.0% of the image size, while an image region without textual correspondence takes about 30.7%. For incomplete annotations, we assign an empty placeholder for their counterparts that allow them to be used in model training or completed in later versions of the dataset.

Solution graphs. The solution graphs are diverse, complicated, and important for characterizing the relationships between solution steps. They broadly fall into three types based on their structures, including sequential (figure 4.2b, all steps are performed in a sequential order), parallel (figure 4.2d, steps can be performed independently in any order), and others (figure 4.2a and figure 4.2c, some of the steps depend on another). As shown in figure 4.4, the distribution of the three types of solution graphs varies across categories. For some categories (*e.g.*, Food and Entertaining, Computers and Electronics, Hobbies and Crafts), a majority of the graphs are sequential because they require to follow a certain procedure. For other categories (*e.g.*, Health, Pets and Animals, Relationships), the solutions often contain multiple steps that address different aspects of the problem (*i.e.* parallel) or have complex dependencies between steps.

These data and annotations enable fine-grained studies of understanding multimodal information in problem solving.

4.4 Experiments

Our VisualHow dataset enables new developments of intelligent problem-solving models that understand and generate solutions to real-life problems. In this section, we systematically analyze a series of baseline models that address new vision-language tasks based on the VisualHow dataset: 1) predicting the solution steps of paired images and captions, 2) predicting the dependencies of different solution steps, 3) describing the problem based on a given solution, and 4) generating captions of images in solution steps. These experiments demonstrate the success of benchmarking baseline models on the proposed VisualHow dataset. They also provide interesting analyses and observations and shed light on new research areas in multimodal understanding and real-life problem solving.

4.4.1 Baseline Models

In our experiments, we adopt state-of-the-art pretrained models to extract features from the visual and language modalities. In particular, the visual features are extracted from a ResNeXT-101 (32×8d) [6] pretrained on Instagram (WSL) [212], while the language features are obtained with a pretrained BERT model [213]. We explore three baseline methods to transform these features for downstream tasks: 1) GAP – a global average pooling method that independently processes features from different regions and words without considering their importance, 2) GPO – a generalized pooling operator [157] that aligns visual and language features and jointly considers them during feature aggregation, and 3) ATT – an attention mechanism to highlight the important semantic region of each modality and then aggregate them by the learned weights. Based on these methods, we develop baseline models for each of our four experiment tasks.

4.4.2 Task 1: Solution Steps Prediction

The main research objective of our work is to enable the learning of intelligent models that can predict step-by-step solutions to real-life problems with both visual illustrations and language descriptions. The joint prediction of multimodal descriptions has not been fully explored by existing vision-language studies. We achieve the goal by carrying out demonstrative experiments on the proposed VisualHow dataset with baseline models

that simultaneously generate the multimodal solutions.

Implementation. Specifically, given the problem description and the previous solution steps, the models are asked to predict the image and caption of the next solution step by sorting two sets of candidate images and captions. We encode the problem, images, and captions using three encoders. The encoded features are dynamically integrated with a bidirectional GRU [214, 215]. To predict the next step of the solution, we develop a triplet network [193, 157] to maximize the cosine similarity between the features of a positive candidate and the GRU features integrated from all previous steps, and to minimize that of a negative candidate.

Evaluation. At evaluation time, the candidates are sampled from the validation set following [159], which includes three sets of correct or incorrect solution steps: 1) the correct next step of the ground-truth solution, 2) 'hard negative' steps from solutions to the 10 most similar problems, 3) random solution steps from the same problem category. To capture this, all questions are embedded into a vector space by concatenating the averaged GloVe [216] embeddings of all words in the problem description. To generate 20 candidates, we first find the union of the correct and hard negative steps, and include other random steps until a unique set of 20 is found. The model is evaluated with three metrics: 1) mean reciprocal rank (MRR) of the correct step, 2) Recall@K, *i.e.*, existence of the correct step in top-K ranked steps, and 3) mean rank of the correct step.

Results. table 4.4 shows the evaluation results of this task. First, we observe that conventional vision-language methods such as GPO [157] achieve mediocre performance, although better than the GAP baseline, suggesting that solving a real-life problem is more challenging than existing vision-language tasks. Further, the results show that attention mechanisms (ATT) can effectively improve the model performance even without explicit supervision, suggesting the importance of focused attention for understanding and solving real-life problems. Finally, it is noteworthy that the performance rankings across all evaluation metrics are consistent, which suggests that our dataset offers a fair benchmark for evaluating solution step prediction models.

Analyses of attention. The rich multimodal attention annotations of VisualHow dataset may act as a guidance for semantic alignment between the two modalities, which allows us to learn more accurate attention with explicit supervision and improve the prediction of solutions. To demonstrate this, we introduce auxiliary cross entropy

	Image			Caption		
Method	MRR	SIM	KLD	MRR	SIM	KLD
ATT	0.505	0.293	1.937	0.572	0.371	1.586
ATT+CE	**0.507**	**0.520**	**0.852**	**0.580**	**0.665**	**0.543**

Table 4.5: Quantitative results of Task 1: solution steps prediction (with attention supervision).

(CE) losses to supervise the visual attention and language attention of models, and analyze the improvement of attention accuracy as well as task performance. We measure the attention accuracy with two popular evaluation metrics, Similarity (SIM) and KL-Divergence (KLD) [217].

table 4.5 shows the quantitative results of models learned with (*i.e.*, ATT+CE) or without (*i.e.*, ATT) attention supervision. Consistent with past observations [1], we find that explicit attention supervision during training may help models focus on important visual and language features, resulting in improved SIM and KLD scores. It also improves their image and caption prediction performance (*i.e.*, MRR). figure 4.6 further compares the attention output of the two models learned with and without explicit supervision. They show that explicit attention supervision not only helps the model locate important regions and words in the multimodal solutions, but also plays an essential role in correlating key components across the two modalities (*e.g.*, fish oil, steak) and deriving more accurate solutions. These observations highlight the important role of multimodal attention for deriving comprehensive solutions to real-life problems.

4.4.3 Task 2: Solution Graph Prediction

Next, given the problem and solution descriptions, we develop models to predict the solution graph. This experiment aims to demonstrate the solution graph as a fine-grained annotation for developing a better understanding about the order and dependency of different solution steps.

Implementation. To capture the relationships between different steps, we concatenate the features extracted from images, captions and the problem description, and learn a single linear layer with a sigmoid activation function to predict the dependency matrix that indicates the dependencies between every two steps.

Evaluation. Evaluation of solution graph prediction is an open problem. In this work, we calculate the intersection over union (IoU) [218, 219] given specific thresholds to compare the similarity between the predicted probability matrix and the ground-truth solution graph. Specifically, we apply a threshold (*e.g.*, 0.25, 0.5, 0.75) to the model output to determine the graph edges and count the edges for the intersection and union between the graph and the ground truth to compute the IoU score.

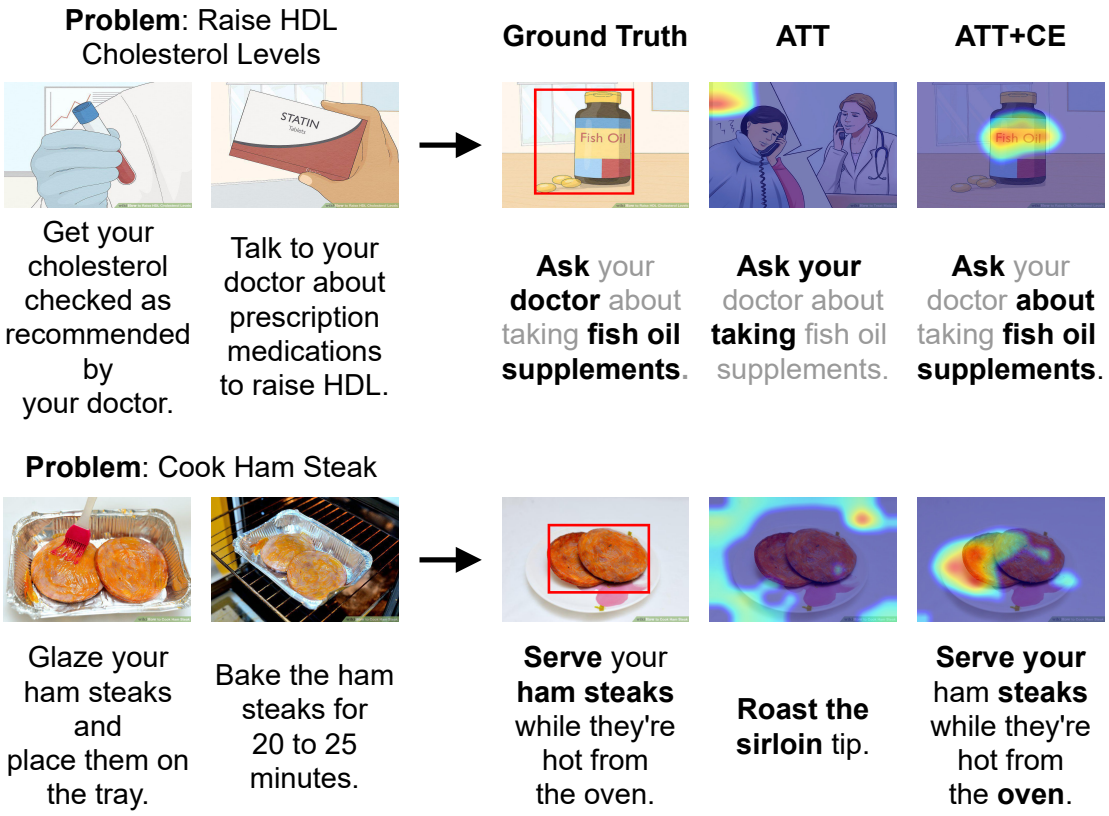

Figure 4.6: Qualitative results for attention supervision. Important regions and keywords are highlighted with red and black colors.

Results. As shown in table 4.6, understanding and predicting the dependencies between solution steps is a challenging task for the baseline models, while the ranks of different models remain similar to Task 1. Similarly, the IoU performance can be improved with the attention mechanism and explicit supervision. These results demonstrate the

Method	IoU@0.25	IoU@0.5	IoU@0.75
GAP	0.484	0.377	0.268
GPO	0.468	0.380	0.302
ATT	0.473	0.389	0.319
ATT+CE	**0.494**	**0.434**	**0.376**

Table 4.6: Quantitative results of Task 2: solution graph prediction.

potential of learning fine-grained solution structures based on the understanding of vision and language descriptions.

4.4.4 Task 3: Problem Description Generation

To further demonstrate the usage of our VisualHow as a general vision-language benchmark, we present a demonstrative experiment for the generation of problem description based on the visual and textual descriptions of a solution. This experiment resembles those for the conventional vision-language tasks (*e.g.*, image captioning and visual question answering) and focus on estimating the models' capability of understanding multimodal contents and performing language generation.

Implementation. For this task, the visual and language features are directly concatenated across all steps, and a BUTD captioning model [88] is adapted to generate the problem description. We implement the attention-based methods (*i.e.*, ATT and ATT+CE) using different inputs: images only (I), captions only (C), and both (I+C).

Evaluation. To evaluate problem description models, we adopt four automatic metrics that are widely used for captioning evaluation, including BLEU [220], METEOR [221], ROUGE-L [222], and CIDEr [223].

Results. table 4.7 presents the results of problem description generation. We observe that leveraging both images and captions (I+C) leads to a clear improvement over single-modality models (*i.e.*, image-only (I) and caption-only (C)). Furthermore, the results show that attention supervision (ATT+CE) has a positive impact on the task performance. Notably, the improvement is bigger with both modalities compared with single modality, suggesting the usefulness of the attention data and supervision methods that highlight multimodal attention alignment.

Method	B-1	B-2	B-3	B-4	M.	R.	C.
ATT (I)	16.7	8.5	4.7	2.9	6.7	16.5	22.9
ATT (C)	22.1	11.4	6.3	3.9	9.8	22.1	44.5
ATT (I+C)	22.7	12.0	6.8	4.4	9.9	22.4	46.7
ATT+CE (I)	16.9	9.5	5.3	3.7	7.3	18.5	24.6
ATT+CE (C)	22.8	11.7	6.3	3.8	9.9	22.3	47.0
ATT+CE (I+C)	**24.1**	**13.1**	**7.7**	**4.8**	**10.7**	**23.2**	**50.8**

Table 4.7: Quantitative results of Task 3: problem description generation.

4.4.5 Task 4: Solution Captions Generation

Our proposed VisualHow dataset can also serve as a useful testbed for evaluating the models' capability of jointly considering multiple images and generating fluent stories. For the final task in our experiments, we consider generating the solution captions based on the input problem description and solution images. It can be considered as a visual storytelling task, but with additional emphasis on the contextual relationship between the goals of problems and the different steps for achieving them.

Implementation. We adapt the AREL [214] model that achieves the state-of-the-art performance on the ViST [167] dataset. We feed the solution images and a BERT embedding of the problem description to the model, to obtain a sequence of captions corresponding to the images [167, 214].

Evaluation. The training and evaluation of models follow the standard visual storytelling paradigm. BLEU [220], METEOR [221], ROUGE-L [222], and CIDEr [223] are used as evaluation metrics to compare the generated captions with the ground truth.

Results. Quantitative results of this task are demonstrated in table 4.8. From the results, we observe that the generation of captions is less challenging than the prediction of problem descriptions, as suggested by the higher BLEU [220], METEOR [221] and ROUGE-L [222] scores. However, the CIDEr [223] scores are significantly lower than those of Task 3. It is because the length of solution captions is much longer than that of the problem description and the models are prone to predict the common words that are discounted by CIDEr. Comparing the different models, we observe that ATT+CE obtains the best performance, while the ATT and GPO fall slightly behind, and the GAP achieves the lowest performance. This suggests that learning to focus on important features can help with the understanding of solution images and generating

Method	B-1	B-2	B-3	B-4	M.	R.	C.
GAP	28.2	13.0	7.3	4.5	23.2	24.1	12.7
GPO	33.0	15.7	7.4	5.6	27.0	26.4	23.0
ATT	33.6	16.4	7.4	5.8	27.2	27.1	23.4
ATT+CE	**33.8**	**17.0**	**9.9**	**6.2**	**28.1**	**28.2**	**24.3**

Table 4.8: Quantitative results of Task 4: solution captions generation.

their corresponding captions. These observations suggest that VisualHow is a challenging benchmark for visual storytelling models, and accurate attention is important for generating fluent descriptions.

4.5 Conclusion

The ability to solve real-world problems is an important step toward human-like intelligence. In this paper, we have introduced VisualHow, a large-scale dataset for solving real-life problems. Utilizing expert-generated internet contents and crowdsourcing, we collected and annotated 20,028 problems and solutions. Dataset statistics demonstrate that the problems, solutions, and annotations contain rich multimodal solutions for a variety of problems in real-life scenarios. Understanding and predicting solutions to real-life problems is an inherently challenging problem. These data and annotations enable a family of new vision-language tasks and computational methods for understanding and solving problems. Our results indicate that there is significant scope for improvement. We hope that this work will facilitate future research to better understand the multimodal information in real-life problem-solving. We envision that this work will spur innovation and encourage developments in problem-solving systems that can positively impact a wide range of applications.

Chapter 5

Every Problem, Every Step, All In Focus: Learning to Solve Vision-Language Problems with Integrated Attention

Integrating information from vision and language modalities has sparked interesting applications in the fields of computer vision and natural language processing. Existing methods, though promising in tasks like image captioning and visual question answering, face challenges in understanding real-life issues and offering step-by-step solutions. In particular, they typically limit their scope to solutions with a sequential structure, thus ignoring complex inter-step dependencies. To bridge this gap, we propose a graph-based approach to vision-language problem solving. It leverages a novel integrated attention mechanism that jointly considers the importance of features within each step as well as across multiple steps. Together with a graph neural network method, this attention mechanism can be progressively learned to predict sequential and non-sequential solution graphs depending on the characterization of the problem-solving process. To tightly couple attention with the problem-solving procedure, we further design new learning objectives with attention metrics that quantify this integrated attention, which better aligns visual and language information within steps, and more accurately captures

Figure 5.1: Problem-solving tasks such as "how to decorate the tables for a vintage-themed wedding" often follow a non-sequential procedure. For example, steps 1, 3, and 4 can be completed in no particular order, as long as step 1 takes place before step 2, step 4 happens before step 5, and all of them take place before step 6. Our method represents such problem-solving procedures in a graph structure. Steps are represented as nodes, and dependent steps are directly connected by edges indicating ordering constraints. In this way, our approach can handle various types of step dependencies in free-formed procedures. Attention is optimized end-to-end over the full graph-based solution structure.

information flow between steps. Experimental results on VisualHow, a comprehensive dataset of varying solution structures, show significant improvements in predicting steps and dependencies, demonstrating the effectiveness of our approach in tackling various vision-language problems.

5.1 Introduction

Recent years have witnessed impressive progress in computer vision and natural language processing, enabling intelligent systems to perform a broad range of joint vision-language tasks, such as image captioning [88, 166, 165, 224, 225, 226], visual storytelling [167, 227], visual question answering [168, 169, 171, 94, 1, 228, 229, 230], visual dialog [159, 160, 231], and natural language generation [232, 233, 234]. However, a major challenge still remains in developing artificial intelligence that can understand vision-language problems and provide procedural solutions with step-by-step instructions. Humans exhibit remarkable ability in visually perceiving problems, comprehending goals, and mapping out plans and procedures to solve them. Developing similar procedural reasoning capabilities in artificial intelligence remains a significant challenge.

Solving vision-language problems requires recognizing important visual details, understanding the multimodal context, and predicting cohesive solutions incorporating visual illustrations and natural language descriptions [52]. Understanding and predicting such multimodal descriptions require an intelligent system to decompose the solution into multiple steps. For example, as shown in Fig. 1, visual illustrations (*e.g.*, flowers, pillows) or natural language descriptions (*e.g.*, "Look for vintage glasses") are used to describe specific steps taken to decorate the tables for a vintage-themed wedding. Existing methods [187, 191, 192, 218, 174, 189, 235, 236, 237, 238, 239] have approached problem-solving with procedure planning, representing each solution as a linear sequence of steps. Such sequential approaches, while convenient, are unable to model complex dependencies across multiple steps. Vision-language problems often involve multiple dependencies between steps, which might not fit neatly into a linear sequence: (1) a step may depend on multiple steps. As shown in Fig. 1, step 6 must depend on the completion of steps 2, 3, and 5, and (2) certain problem-solving steps (*e.g.*, paths 1-2, 3, 4-5 in Fig. 1) can occur simultaneously. A sequential model might oversimplify the relationships and struggle to represent these cases effectively, facing challenges in the following aspects: First, sequential models inherently follow a linear *structure*, processing information in a step-by-step fashion. This linear nature becomes a constraint when dealing with multiple dependencies that don't conform to a straightforward sequence. Second, the *efficiency* of sequential models is compromised when confronted

with interdependencies across multiple steps. Directly converting parallel processes into a fixed-order sequence regardless of variations can lead to suboptimal and inefficient solutions. Third, sequential models often lack the *interpretability* required to understand complex dependencies between different steps of the problem-solving process, diminishing the trust and transparency essential for real-world applications. Therefore, in light of these challenges, our work is motivated by the need for a more flexible and structured approach to vision-language problem-solving. Therefore, in light of these challenges, our work is motivated by the need for a more flexible and nuanced approach to vision-language problem-solving.

In this work, to enable more general and flexible problem solving, we propose a graph neural network approach that represents solutions as graphs. This structured representation allows graph-based models to overcome the limitations of sequential models, providing a more general and effective approach to handling complex problem-solving scenarios. Our method leverages an integrated attention mechanism that jointly models intra-step attention and inter-step attention. This provides a more holistic view compared to isolated step-based attention. To jointly and progressively supervise the integrated attention, we further introduce quantitative metrics that consider attention propagation across the entire graph of solution steps. This graph-based approach with the novel integrated attention mechanism aims to provide an effective framework for modeling complex dependencies across multiple steps and solving real-world problems, such as those in autonomous driving, medical diagnosis, and various other applications.

To summarize, the main contributions of this paper are as follows:

1. We propose a graph neural network approach to represent procedural solutions as graphs, capturing complex step dependencies and enabling an integral understanding of the entire problem-solving procedure.

2. We design an integrated attention mechanism that jointly models the importance of multimodal features within each step as well as across interdependent steps.

3. We introduce quantitative attention metrics to optimize attention propagation across the full solution graph, enabling supervised learning of attention for complex vision-language problem solving.

The remainder of this paper is structured as follows. In Section 5.2, we provide a concise overview of related research pertaining to vision-language problem solving and attention mechanisms in vision-language tasks. Section 5.3 outlines the problem statement, introducing the formulation of the vision-language problem solving task that we aim to address. The details of our proposed method, designed to tackle the aforementioned task, are elaborated in Section 5.4. Extensive experiments are presented in Section 5.5, where we report quantitative and qualitative results, along with comprehensive analyses of our approach's performance. We conclude this paper and discuss its limitations in Section 5.6, while also providing directions for future research and improvements.

5.2 Related Works

Our work is relevant to previous efforts on visual problem-solving, attention mechanisms in vision-and-language tasks, and supervision of attention.

5.2.1 Problem Solving Methods

Procedural problem solving with instructional solutions has gained increasing research attention. Several studies [187, 174, 192, 191, 52, 189, 190, 235, 239] have curated datasets of images or videos demonstrating procedures for daily tasks like cooking, maintenance, sports, and healthcare. These efforts have enabled data-driven approaches to generate solutions for diverse problems. A series of previous methods focus on developing captioning models to summarize instructional text describing procedures [187, 52, 238]. Other works emphasize aligning textual and visual modalities [240, 174, 191, 192, 52]. They retrieve images given instruction text or localize described activities. Alternative approaches factorize solutions into discrete steps and predict structured representations [218, 241, 242, 190, 236, 237, 238, 243]. However, these studies oversimplify real-world solution procedures as sequential activities. Solutions often have complex, free-formed structures with inter-dependencies between steps. Thus, while demonstrating feasibility for varied tasks, existing methods are limited in generalizing across problems regardless of solution structure. They also do not perform joint reasoning over steps and their relationships. Our work addresses these gaps by representing solutions as graphs to capture

step dependencies and provide a comprehensive framework for complex problem solving.

5.2.2 Attention in Vision-Language Tasks

Attention mechanisms have become critical components in vision-language models to effectively couple modalities and identify salient features for various tasks. Prior studies have focused on designing attention for input feature prioritization [88, 1, 104, 244, 245], cross-modal alignment [198, 52], and concept-dependency modeling [246, 247, 248]. Early attention approaches operated on grid-structured inputs like images or text, using convolutional neural networks [249] or Transformers [213, 16], while recent graph-based methods [246, 248] allow modeling attention in structured inputs [250, 251, 252, 253]. However, capturing the complex dependencies across steps in procedural solutions requires structured representations that consider attention shifts across multiple modalities and multiple steps. We advance existing techniques with a novel integrated attention mechanism that enables joint attention modeling for both aspects and leverage this new attention mechanism to progressively construct structured solutions for various problems.

5.2.3 Supervised Learning of Attention

Instead of implicitly learning the attention mechanism with the end objectives of different tasks, prior works have explored explicitly supervising attention mechanisms to improve alignment with regions of interest. Various approaches have been proposed to construct the ground truth attention based on task annotation [1, 245], human attention [224, 225, 244, 254], or adversarial learning [255]. Some supervision methods use single-step supervision based on human annotations of salient image regions [244, 254, 245], while others account for integrating attention across the visual reasoning procedure [1]. However, focusing on local alignments limits modeling relationships between steps in structured problem solving. Without propagating attention, these methods fail to capture complex interdependencies in multi-step procedures. Differently, in this work, we present a new metric that quantitatively measures the contributions of attention for constructing the task solution, and leverage it to progressively supervise both the intra- and inter-step attention. It provides an integral view of problem-solving procedures,

resulting in enhanced performance in formulating a structured representation of the solutions.

5.3 Problem Statement

The vision-language problem solving task involves comprehending general vision-language problems and generating structured instructions to address them, incorporating both visual and textual information [52]. Previous research has explored instructional images [186, 187] or videos [188, 189, 190, 174, 192], but these were limited to predicting sequential instructions for specific task categories. In contrast, our work considers a wide range of problems and their corresponding solution structures. The fundamental goals of our proposed approach are twofold: (1) understanding the input problem description and (2) constructing a solution graph consisting of essential problem-solving steps, each associated with relevant images and captions.

As shown in Fig. 5.1, the input of our proposed approach consists of a problem description $\boldsymbol{g}$, such as "how to decorate the tables for a vintage-themed wedding," and a pool of images $\{I_1, I_2, \cdots, I_N\}$ or captions $\{C_1, C_2, \cdots, C_N\}$. These images and captions serve as candidate steps or actions that could be relevant or irrelevant to solving the given problem. The main challenge in the vision-language problem solving task is to identify the essential steps and their correct order to construct a coherent and effective solution for the problem at hand.

To tackle this challenge, our proposed approach involves creating a solution graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ that encapsulates the problem-solving process. The graph nodes in $\mathcal{V}$ represent essential steps, including the start node (node 0), the end node (node $N+1$), and the nodes corresponding to the candidate steps (nodes $1, \cdots, N$) with their associated image or caption capturing the possible actions that can be taken to solve the problem. The edges in $\mathcal{E}$ represent the dependencies or chronological order between the steps. For instance, a directed edge between the nodes corresponding to "prepare vintage centerpieces for tables" and "add more flowers to the main table centerpiece" indicates that the latter step should happen after the former.

By constructing such a directed graph, our approach can effectively model the logical

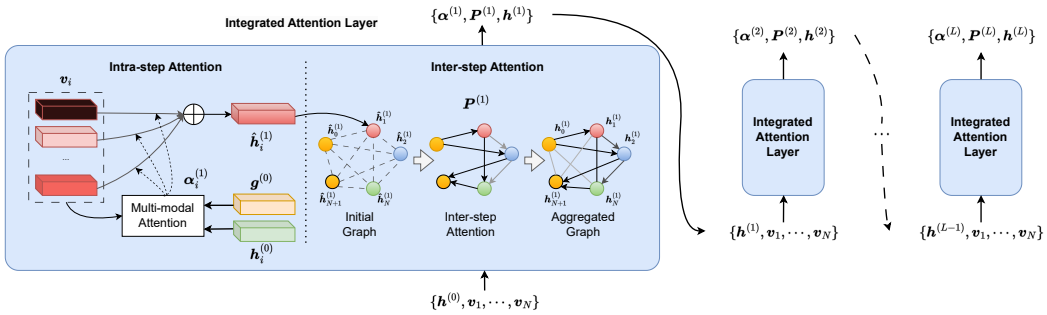

Figure 5.2: Overview of the proposed SGAN architecture. The input consists of node representations $\boldsymbol{h}^{(0)}$, features for images/caption candidates $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N$. The network leverages an integrated attention mechanism that progressively processes the input features and predicts the output intra-step attention $\boldsymbol{\alpha}^{(L)}$ for capturing salient information from the input images or captions, the inter-step attention $\boldsymbol{P}^{(L)}$ characterizing the probabilities of dependencies across different steps, and the final updated node representations $\boldsymbol{h}^{(L)}$.

flow of the problem-solving procedure, enabling a structured and coherent representation of the solution. The directed graph representation also allows for the existence of multiple paths from the start node to the end node, corresponding to different ways of solving the problem. This flexibility in the graph structure is particularly beneficial for handling vision-language problems with multiple viable solutions or alternative sequences of steps.

5.4 Method

Our proposed Solution Graph Attention Network (SGAN) addresses the vision-language problem solving task by leveraging both intra-step and inter-step attention mechanisms to iteratively refine the solution graph. The key technical components of our method are (1) a novel graph neural network approach that progressively predicts solutions with diverse structures, (2) an integrated attention mechanism combining intra-step attention and inter-step attention for a comprehensive understanding of the problem-solving procedure, and (3) new attention metrics and learning objectives to jointly supervise the attention throughout the solution graph by leveraging information propagation.

Together, these components empower SGAN to effectively capture dependencies within individual steps and the relationships between them, providing a powerful ability to handle complex vision-language problems and generate coherent solutions.

5.4.1 Solution Graph Attention Network

In problem-solving scenarios, dependencies between steps can be complex and may not be readily apparent. To address this challenge and predict the solution graph $\mathcal{G}$, SGAN progressively learns integrated attention using a graph attention network, enabling a better understanding of the problem-solving procedure.

As depicted in Fig.5.2, the input features representing the candidates, denoted as $\boldsymbol{v} = \{\boldsymbol{v}_i | i = 1, \cdots, N\}$, are obtained with a pre-trained image encoder (*e.g.*, ResNeXT-101 [6], ViT [256]) for image candidates or a language embedding network (*e.g.*, BERT [213]) for caption candidates. The language embedding $\boldsymbol{g}$ represents the description of the input problem [213, 52]. SGAN is designed with a stack of L graph attention layers, allowing the step-by-step refinement of the solution graph. Specifically, the network iteratively updates the node representations $\boldsymbol{h}^{(\ell)} = \{\boldsymbol{h}_i^{(\ell)} | i = 0, \cdots, N+1\}$, where $\ell = 1, \cdots, L$ indicates the ℓ-th layer. It consists of the updated features of the graph nodes start ($i = 0$), end ($i = N + 1$), and each candidate step ($i = 1, \cdots, N$). The node representations of the previous layer $\boldsymbol{h}^{(\ell-1)}$ are passed to the current layer as the input, while the first layer input is initialized as $\boldsymbol{h}^{(0)} = \{\boldsymbol{g}, \bar{\boldsymbol{v}}_1, \cdots, \bar{\boldsymbol{v}}_N, \boldsymbol{W}_e \boldsymbol{g}\}$, where $\bar{\boldsymbol{v}}_i$ is the average of $\boldsymbol{v}_{i,k}$ across all $k = 1, \cdots, K$ image patches or word tokens, and $\boldsymbol{W}_e$ represents the learnable parameters to transform the language embedding $\boldsymbol{g}$ as the end node representation. Each layer also outputs the corresponding intra-step attention $\boldsymbol{\alpha}^{(\ell)}$ and the inter-step attention $\boldsymbol{P}^{(\ell)}$ (see Section 5.4.2 for details).

To convert the final-layer output $\boldsymbol{P}^{(L)}$ into the solution graph $\mathcal{G}$, we employ the following process. Initially, a heuristic threshold is applied to the dependency matrix $\boldsymbol{P}^{(L)}$ to preserve the most pertinent nodes (see Steps 1-3 in Algorithm 1). Next, these selected nodes are iteratively added into the graph (see Steps 4-5 in Algorithm 1), along with their associated edges featuring the highest values in $\boldsymbol{P}^{(L)}$. This iterative process ensures that the graph remains a directed acyclic graph without loops or isolated nodes. Finally, attention weights $\boldsymbol{\alpha}^{(L)}$ assigned to each step's images and captions offer insights into what demands attention for effectively solving the given problem.

Algorithm 1 Graph post-processing method to obtain the final solution graph

INPUT: Predicted dependency matrix $\boldsymbol{P}^{(L)}$, retrieval threshold λ_r, and dependency threshold λ_d.

1: Filter candidate steps using λ_r over $\boldsymbol{P}^{(L)}_{0,1:N}$ to obtain the node set S, where $\boldsymbol{P}^{(L)}_{0,i} \geq \lambda_r$ for $i \in S$.

2: Remove cycles between nodes i and j in S by updating $\boldsymbol{P}^{(L)}_{i,j} = \max(0, \boldsymbol{P}^{(L)}_{i,j} - \boldsymbol{P}^{(L)}_{j,i})$.

3: Initialize solution graph $\mathcal{G}$ with nodes $\mathcal{V} = \{0, N+1\}$, edges $\mathcal{E} = \{(0, N+1)\}$, and candidate edges $\mathcal{W} = \{(0, N+1)\}$ containing potential edges to add to the graph.

For u in S

4: Find the best node v_b with maximum

$$b = \max_{v \in \mathcal{S}, v \notin \mathcal{V}} \max_{(\bar{v}_1, \bar{v}_2) \in \mathcal{W}} \left(\sum_{\bar{v}_3 \in Pa(\bar{v}_1)} \boldsymbol{P}^{(L)}_{\bar{v}_3, v} + \boldsymbol{P}^{(L)}_{\bar{v}_1, v} + \boldsymbol{P}^{(L)}_{v, \bar{v}_2} + \sum_{\bar{v}_4 \in Ch(\bar{v}_2)} \boldsymbol{P}^{(L)}_{v, \bar{v}_4} \right),$$

where $Pa(\bar{v}_1)$ and $Ch(\bar{v}_2)$ represent the parent set of node $\bar{v}_1$ and child set of node $\bar{v}_2$ in the solution graph $\mathcal{G}$, respectively.

5: If $b > \lambda_d$, Update the edge set $\mathcal{E}$ and node set $\mathcal{V}$ by adding node v_b and candidate edges in $\mathcal{W}$ to ensure the graph remains a directed acyclic graph.

OUTPUT: The final solution graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$

The proposed network is powerful for learning the dependencies between problem-solving steps. By using this iterative approach, the network can generate free-formed solutions with a better understanding of the problem-solving procedure. In the following, we will describe the specific design of our integrated attention mechanism to effectively capture the important contents and dependencies across problem-solving steps.

5.4.2 Integrated Attention Mechanism

Attention is a crucial component that drives advancements in natural language processing and computer vision, which enables models to selectively focus on the most relevant parts of the input data when performing different tasks. In the context of problem-solving, our integrated attention mechanism plays a critical role in identifying the key features and dependencies between the steps involved in a solution. It combines intra-step and inter-step attention to enable the network to capture both the fine-grained details of each step and the broader context in which they exist.

Intra-Step Attention

The intra-step attention focuses on capturing salient information from the input images or captions for understanding and completing each individual step. Specifically, in the ℓ-th layer, for the i-th candidate step, we define the intra-step attention weights as $\boldsymbol{\alpha}_i^{(\ell)}$, which is computed based on the input problem description $\boldsymbol{g}$, the candidate features $\boldsymbol{v}_i$, and the node features $\boldsymbol{h}_i^{(\ell-1)}$:

$$\boldsymbol{a}_{i,k}^{(\ell)} = \boldsymbol{w}_a^T \tanh(\boldsymbol{W}_v \boldsymbol{v}_{i,k} + \boldsymbol{W}_g \boldsymbol{h}_0^{(\ell-1)} + \boldsymbol{W}_h \boldsymbol{h}_i^{(\ell-1)}), \tag{5.1}$$

where $\boldsymbol{w}_a, \boldsymbol{W}_g, \boldsymbol{W}_v, \boldsymbol{W}_h$ are learnable parameters, and $k = 1, \cdots, K$ indicates the k-th element of the input candidate (*i.e.*, image patch or word token).

The attention weights $\boldsymbol{a}_i^{(\ell)}$ are normalized as $\boldsymbol{\alpha}_i^{(\ell)}$ with a masked softmax activation function

$$\boldsymbol{\alpha}_i^{(\ell)} = \text{softmax}(\boldsymbol{a}_i^{(\ell)}, \boldsymbol{m}_i), \tag{5.2}$$

where $\boldsymbol{m}_i$ is a binary vector and $\boldsymbol{m}_{i,k}$ indicates the k-th element (*i.e.*, image patch or word token) of the i-th candidate features is padded or not due to the variable length

of the image or language inputs.

Finally, we apply the attention to the candidate features $\boldsymbol{v}$ to initialize the node representations for the ℓ-th layer:

$$\hat{\boldsymbol{h}}_i^{(\ell)} = \begin{cases} \boldsymbol{h}_i^{(\ell-1)} & \text{if } i = 0 \text{ or } N+1 \\ \sum_k \boldsymbol{\alpha}_{i,k}^{(\ell)} \boldsymbol{v}_{i,k}^{(\ell-1)} & \text{if } i = 1, \ldots, N. \end{cases} \tag{5.3}$$

Inter-Step Attention

The inter-step attention is responsible for capturing the chronological order between different problem-solving steps, providing a coherent and structured representation of the solution. By integrating inter-step attention into our model, we aim to enable more effective joint reasoning across multiple problem-solving steps. Specifically, we compute graph attention weights [246, 248] to estimate the existence of a dependency between each pair of steps based on the initial node features $\hat{\boldsymbol{h}}_i^{(\ell)}$ computed in Equation (5.3):

$$\boldsymbol{P}_{i,j}^{(\ell)} = \sigma\left(\boldsymbol{\gamma}^{(\ell)^T} \text{LeakyReLU}(\boldsymbol{W}_l^{(\ell)} \hat{h}_i^{(\ell)} + \boldsymbol{W}_r^{(\ell)} \hat{\boldsymbol{h}}_j^{(\ell)})\right), \tag{5.4}$$

where $\boldsymbol{\gamma}^{(\ell)}$, $\boldsymbol{W}_l^{(\ell)}$ and $\boldsymbol{W}_r^{(\ell)}$ are learnable parameters and $\sigma(\cdot)$ is the sigmoid function. This computation involves learning parameters that weigh the significance of each step's features in establishing a dependency with another step. The resulting weight matrix $\boldsymbol{P}^{(\ell)}$ explicitly represents the probabilities of dependency between steps in order to construct the final solution graph.

With these inter-step attention weights, we proceed to update the features of each node i by combining information from all graph nodes, which involves measuring how much weight is given to the connection between nodes i and j at the ℓ-th layer and then using these weights to update the features of node i:

$$\boldsymbol{h}_i^{(\ell)} = \text{ELU}\left(\sum_j \frac{\boldsymbol{P}_{i,j}^{(\ell)} \boldsymbol{W}_r^{(\ell)} \hat{\boldsymbol{h}}_j^{(\ell)}}{\sum_{j'} \boldsymbol{P}_{i,j'}^{(\ell)}}\right), \tag{5.5}$$

where ELU is the exponential linear unit function. This feature update allows the model to adaptively refine the representation of each node, incorporating insights from its connections in the solution graph.

By integrating both intra-step attention and inter-step attention mechanisms into SGAN's stack of attention layers, the model achieves a comprehensive understanding of the problem-solving procedure. The iterative refinement of the solution graph across these layers enables SGAN to progressively capture important features within individual steps and the relationships between the steps. This integration introduces a novel and powerful framework for SGAN to generate structured and coherent solutions for a wide range of vision-language problem solving tasks.

5.4.3 Learning Objectives

Our integrated attention mechanism progressively focuses on salient information in visual and textual inputs, capturing step dependencies for effective problem-solving. We propose novel learning objectives, supervising attention to identify important parts of the images and captions, and propagating information across steps for high-quality solution graphs.

Learning Intra-Step Attention

We present the attention learning loss to measure the prediction error of intra-step attention, based on the ground-truth multimodal attention annotations. These annotations are binary masks that indicate important image regions or word tokens in the captions. To measure the prediction error of the intra-step attention $\boldsymbol{\alpha}_i^{(\ell)}$, the intra-step attention loss is defined as

$$L_{att}^{(\ell)} = \sum_{i \in \mathcal{GT}} l_{att}(\boldsymbol{\alpha}_i^{(\ell)}, \boldsymbol{\alpha}_i'), \tag{5.6}$$

where $\mathcal{GT}$ is the set of ground-truth steps and l_{att} is a dissimilarity metric that measures the misalignment between the predicted $\boldsymbol{\alpha}_i^{(\ell)}$ and the softmax-normalized ground-truth attention $\boldsymbol{\alpha}_i'$ [52]. In our implementation, we define l_{att} as a cross-entropy loss:

$$L_{att}(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_i') = -\sum_k \boldsymbol{\alpha}_{i,k}' \log(\boldsymbol{\alpha}_{i,k}). \tag{5.7}$$

Similarly, other attention evaluation metrics like SIM [257], JSD [258, 259], and CC [257]) can also be used to measure the intra-step attention alignment.

Learning Inter-Step Attention

To gain deeper insights into the contributions of attention throughout the entire problem-solving process, we adopt an integrated approach that considers attention allocation across multiple problem-solving steps. Inspired by information retrieval techniques [247, 260], we introduce novel learning objectives that involve propagating the intra-step attention measurements along the edges of the predicted solution graph, quantifying the impact of attention in achieving successful solution prediction.

Specifically, given the ground truth solution graph represented as an adjacency matrix $\boldsymbol{G}$ and the inter-step attention $\boldsymbol{P}^{(\ell)}$ predicted by the ℓ-th layer, we compute $\boldsymbol{F}^{out(\ell)}$ and $\boldsymbol{F}^{in(\ell)}$ that denote the probabilities of information propagation along the ground-truth edges from step i, and those to step j, from out-degree and in-degree perspectives, respectively:

$$\boldsymbol{F}^{out(\ell)}_{i,j} = \frac{\sum_k \boldsymbol{G}_{i,j}\boldsymbol{P}^{(\ell)}_{i,k}}{\sum_k \boldsymbol{P}^{(\ell)}_{i,k}}, \quad j = 0, \cdots, N+1 \tag{5.8}$$

$$\boldsymbol{F}^{in\,(\ell)}_{i,j} = \frac{\sum_k \boldsymbol{G}_{k,j}\boldsymbol{P}^{(\ell)}_{k,j}}{\sum_k \boldsymbol{P}^{(\ell)}_{k,j}}, \quad i = 0, \cdots, N+1 \tag{5.9}$$

Based on these propagation probabilities, we define two inter-step attention scores that quantify the information flow from both in-degree and out-degree perspectives at the ℓ-th layer, respectively:

$$S^{(\ell)}_{out} = \text{mean}[(\boldsymbol{F}^{out(\ell)} \odot \boldsymbol{D}^{(\ell)})^T \boldsymbol{s}^{(\ell)}], \tag{5.10}$$

$$S^{(\ell)}_{in} = \text{mean}[(\boldsymbol{D}^{(\ell)} \odot \boldsymbol{F}^{in(\ell)})^T \boldsymbol{s}^{(\ell)}], \tag{5.11}$$

where $\odot$ represents the Hadamard product, $\boldsymbol{s}^{(\ell)} = [1, \boldsymbol{s}^{(\ell)}_1, \cdots, \boldsymbol{s}^{(\ell)}_N, 0]^T$ denotes an intra-step attention similarity measure, and $\boldsymbol{D}^{(\ell)}$ is a distribution matrix measuring the probability distribution of attention weights from step i to step j:

$$\boldsymbol{D}^{(\ell)}_{i,j} = \frac{\boldsymbol{G}_{i,j}\boldsymbol{P}^{(\ell)}_{i,j}}{\sum_k \boldsymbol{G}_{i,k}}. \tag{5.12}$$

Specifically, the similarity $\boldsymbol{s}^{(\ell)}_i$ is defined as

$$\boldsymbol{s}_i^{(\ell)} = 1 - \frac{\mathrm{JSD}(\boldsymbol{\alpha}_i^{(\ell)}, \boldsymbol{\alpha}_i')}{\ln 2}, \tag{5.13}$$

where JSD is the Jensen–Shannon divergence [258, 259].

The above inter-step attention scores $S_{out}^{(\ell)}$ and $S_{in}^{(\ell)}$ comprehensively quantify the performance of inter-step attention prediction from the out-degree and in-degree perspectives, where higher scores indicate that attention can be more effectively allocated over the important steps and dependencies to build the solution graph, and the maximum score of 1 indicates the perfect alignment with the ground-truth solution graph.

Overall Objectives

Our final objective function is defined as a combination of the binary cross entropy loss L_{BCE} that evaluates the solution graph, the intra-step attention loss ${L_{att}}^{(\ell)}$, and the inter-step attention scores $S_{out}^{(\ell)}$ and $S_{in}^{(\ell)}$ across all graph attention layers:

$$L = L_{\mathrm{BCE}} + \sum_{\ell=1}^{L} L_{att}^{(\ell)} - \sum_{\ell=1}^{L} (S_{out}^{(\ell)} + S_{in}^{(\ell)}), \tag{5.14}$$

where

$$L_{\mathrm{BCE}} = -\sum_{\ell=1}^{L} \sum_{i,j} (\boldsymbol{G}_{i,j} \log \boldsymbol{P}_{i,j}^{(\ell)} + (1 - \boldsymbol{G}_{i,j}) \log(1 - \boldsymbol{P}_{i,j}^{(\ell)})), \tag{5.15}$$

is the binary cross-entropy loss.

With this objective function, our method jointly and progressively supervises both intra-step attention and inter-step attention. It enables an integrated optimization of the solution with respect to multimodal attention alignment within individual problem-solving steps, information propagation for between-step connections, and the final solution graph. With the ability to traverse the graph and selectively aggregate information, our method achieves significant improvement in formulating solutions to various problems.

5.5 Experiments

In this section, we present comprehensive experiments to demonstrate the advantages of our proposed method and assess the contributions of its major components. The experimental results underscore the significance of progressive attention learning and the effectiveness of the proposed objectives, shedding light on the intricacies of complex problem-solving processes. These findings hold promise in substantially advancing the domain of vision-language problem solving and paving the way for more sophisticated intelligent systems.

5.5.1 Experimental Setup

In this subsection, we provide a thorough description of our experiments and implementation details. We introduce the dataset used for our multimodal problem-solving task, the compared state-of-the-art models and baselines, the evaluation methods, and the implementation details of our proposed SGAN method.

Dataset

Our experimental evaluation is conducted on the VisualHow dataset [52], which comprises 20,028 real-life problems categorized hierarchically into 18 main categories and 317 subcategories. The number of problems in each category ranges from 405 to 2,952, providing a diverse set of problem-solving scenarios. Unlike previous datasets [187, 174, 191, 192, 189, 190] that focus solely on sequential procedures, the VisualHow dataset includes a solution graph for each problem, representing the structured dependencies between individual steps. Importantly, a substantial portion of the graphs exhibit non-sequential characteristics, featuring more complex inter-step dependencies. Each solution graph consists of 3 to 10 steps, each described with images and captions. The images encompass a variety of formats, including realistic photos, cartoons, drawings, handwriting, charts, among others. The captions have a vocabulary of 30,000 tokens, ensuring rich and informative descriptions. To facilitate attention learning and evaluation, fine-grained attention annotations are provided for both images and captions.

Models

To evaluate the effectiveness of our method in handling vision-language problem-solving tasks, we compare it with state-of-the-art approaches on the VisualHow dataset [52]. We treat these methods as multi-task models, addressing both the retrieval of the multimodal instructions and the prediction of step dependencies. The compared methods, including SEQ GPO [157], SEQ GAP [52], and SEQ ATT [52], aim to predict individual problem-solving steps and their dependencies using various sequential processes. Specifically, SEQ GPO employs a generalized pooling operator to align visual and language features and jointly aggregates them during feature aggregation. Similarly, SEQ GAP adopts a global average pooling method to process features from different image regions and word tokens independently, without considering their importance. Finally, SEQ ATT utilizes an attention mechanism to highlight important semantics in each modality and then aggregates them based on learned weights, supervised with ground-truth attention annotations from VisualHow [52].

To further investigate the role and significance of the integrated attention mechanism, we conduct a comprehensive ablation study using three variants of our proposed model: SGAN-Base, SGAN-Intra, and SGAN-Inter. SGAN-Base is a basic model that uses the same architecture as SGAN but doesn't rely on any extra attention supervision from outside sources. This helps us understand how well the model performs when it learns attention on its own from the solution graph. For SGAN-Intra and SGAN-Inter, we supervise the model with the intra-step attention loss and inter-step attention loss terms, respectively. By comparing the performance of these three variants with our full SGAN model, which incorporates intra-step and inter-step attention supervision, we can analyze the specific contributions of each attention component.

Evaluation

To ensure a fair comparison with other methods, we adhere to the official training and validation splits provided by the dataset. We construct candidate pools by sampling images and captions from the corresponding subsets. These candidate pools include positive samples corresponding to the given problem and negative samples from other problems randomly sampled from the dataset. Note that the candidate pools contain

only training data during the training phase, and only validation data during the validation phase. Different from the previous study [52] that samples unrelated steps from different problems, in this paper, to obtain negative step dependencies, we sample negative problems first, and include all steps and their dependencies in the negative problems. This approach serves as a suitable test bed for robustly evaluating and justifying the model's performance. Following the VisualHow [52] study and our proposed attention evaluation methods, we evaluate model performances with four categories of metrics:

Retrieval of Steps. To evaluate the performance models in retrieving the correct ground-truth steps, we rank the candidate steps based on their predicted relevance to the input problem (*i.e.*, $\boldsymbol{P}_{0,i}^{(L)},\ i = 1, 2, \cdots, N$). We employ the mean reciprocal rank (MRR) [159, 160, 52], Recall@K [159, 160, 52, 261, 157, 158], and recall sum (RSUM) [261, 157, 158, 52] metrics. The MRR computes the reciprocal rank of a correct step, which is defined as 1 divided by its position in the ranked list. Recall@K measures the presence of the correct step in the top-K ranked steps. The RSUM is defined as the sum of recall metrics at different values of K (*e.g.*, $K = \{1, 5, 10\}$). The combination of these metrics provides a comprehensive summary of the model's overall performance in image and caption retrieval.

Step Dependency Prediction. The prediction of dependencies between steps is evaluated using the area under the ROC curve (AUC) [52, 262], the area under the precision-recall curve (AUPR) [262], and the intersection over union (IoU) [218, 219, 52]. The AUC represents the overall performance of the model in distinguishing positive (correctly predicted edges) from negative (incorrectly predicted edges) dependencies between steps. The AUPR is a useful performance metric for imbalanced data in a setting with a bigger focus on positive examples, which is the case for our experiments. To measure IoU, we apply a threshold (e.g., 0.25, 0.5, 0.75) [52] to the model output $\boldsymbol{P}^{(L)}$ to determine the graph edges and count the edges for the intersection and union between the predicted graph and the ground truth. These metrics enable a comprehensive evaluation of the model's performance in predicting the structure of solutions.

Intra-Step Attention. To evaluate the intra-step attention, the output $\boldsymbol{\alpha}^{(L)}$ is first normalized and converted into an attention map, where each value indicates the attention probability of an image patch or word token. The ground-truth attention

maps are computed similarly as the annotations. Three attention metrics are used to compute the attention maps: the linear Correlation Coefficient (CC) [257, 162] scores are computed as Pearson's linear correlation between the attention maps; the similarity of histogram intersection (SIM) [257] computes the sum of the minimum values at every location; Kullback-Leibler divergence (KL) [257] measures the difference between two distributions based on information theory.

Inter-Step Attention. The inter-step attention is evaluated based on the final-layer outputs $\boldsymbol{\alpha}^{(L)}$ and $\boldsymbol{P}^{(L)}$ simultaneously by three metrics that measure out-degree $S_{out}^{(L)}$ (see Equation (5.10)), in-degree $S_{in}^{(L)}$ (see Equation (5.11)) attention scores, and an overall attention score $S_{all}^{(L)}$ computed as

$$S_{all}^{(L)} = mean[(\boldsymbol{F}^{out(L)} \odot \boldsymbol{D}^{(L)} \odot \boldsymbol{F}^{in(L)})^T \boldsymbol{s}^{(L)}]. \tag{5.16}$$

Implementation Details

To extract discriminative visual-linguistic features, we adopt state-of-the-art pre-trained models. For the visual features, we use ResNeXT-101 [6] (32×8d) trained on Instagram images (WSL) [212], with image size 256×256. Regarding the language features, we use a pre-trained BERT model [213] optimized on a massive corpus of text. We use these models to extract features from the candidate image and caption pools, which are then used as inputs to our SGAN model. We train our model using the Adam [101] optimizer with learning rate 2×10^{-4}, weight decay 10^{-4} and batch size 16. A cosine annealing scheduler schedules the learning rate. We set $L = 3$ as the total number of network layers. To address the imbalance between the positive and negative samples from the solution graph, we train the model with the loss related to the retrieval task for 5 epochs and then train the model with the loss related to the whole solution graph for the remaining 20 epochs. A hard negative mining strategy [263, 264] is also used. The post-processing method to obtain the final solution graph is implemented following Algorithm 1, where we set dependency threshold $\lambda_d = 0.8$ and retrieval threshold $\lambda_r = 0.45$.

Method	Mode	Retrieval ↑					Dependency ↑				
		MRR	R@1	R@5	R@10	RSUM	AUC	AUPR	IoU@0.25	IoU@0.5	IoU@0.75
SEQ GPO [157]	I	0.4529	31.03	61.89	79.99	386.28	0.713	0.768	0.455	0.262	0.102
	C	0.6066	47.25	77.27	88.85		0.816	0.847	0.544	0.443	0.269
SEQ GAP [52]	I	0.5311	39.28	69.55	85.30	404.72	0.685	0.752	0.444	0.250	0.080
	C	0.5935	45.55	76.75	88.29		0.817	0.850	0.557	0.430	0.258
SEQ ATT [52]	I	0.5069	37.10	66.48	82.82	410.77	0.698	0.760	0.422	0.274	0.106
	C	0.6509	51.94	82.02	91.41		0.815	0.849	0.520	0.432	0.263
SGAN-Base	I	0.5524	41.99	71.42	86.95	433.55	0.721	0.777	0.487	0.307	0.152
	C	0.6820	56.66	83.10	93.43		0.799	0.834	0.553	0.481	0.321
SGAN-Intra	I	0.5833	45.03	74.89	**88.93**	451.82	0.700	0.765	0.488	0.316	0.189
	C	0.7247	61.61	86.52	94.84		0.780	0.825	0.578	0.428	0.324
SGAN-Inter	I	0.5703	43.72	73.45	88.26	442.26	**0.801**	0.816	0.577	0.504	0.385
	C	0.6954	57.98	84.76	94.09		0.861	0.863	0.653	0.616	0.538
SGAN	I	**0.5898**	**45.97**	**75.61**	88.89	**455.56**	0.800	**0.817**	**0.580**	**0.508**	**0.394**
	C	**0.7324**	**62.77**	**86.95**	**95.37**		**0.862**	**0.864**	**0.659**	**0.620**	**0.547**

Table 5.1: Solution graph prediction results from retrieval and dependency aspects. In each panel, the first row (I) indicates the image modality and the second row (C) indicates the caption modality. The best results are highlighted in bold.

5.5.2 Quantitative Results

Comparison with the State-of-the-Art

Our approach demonstrates superior performance across all metrics for generalizing solutions to vision-language problems, as shown in Table 5.1. Overall, it outperforms the state-of-the-art SEQ GPO, SEQ GAP, and SEQ ATT methods [52] across all evaluation metrics. In terms of retrieving multimodal instructions for individual problem-solving steps, it achieves an impressive improvement of 11.1% and 12.5% in MRR scores for images and captions, respectively, as well as an improvement of 10.9% in RSUM scores which aggregate the Recall@K scores over both modalities. Further, in terms of predicting the step dependencies, our method exhibits strong capability in capturing the diverse structures of solutions, which has been a challenge for existing methods. It shows 81.0% and 45.3% improvements in the average IoU scores (*i.e.*, 0.25, 0.5, and 0.75) for images and captions, respectively. These observations not only demonstrate the advantages of our approach in solving complex vision-language problems but also highlight the significance of progressively constructing task solutions.

Comparison with Baseline Models

Table 5.1 also compares our proposed SGAN model with different baselines, including the SGAN-Base model that is learned without supervision from attention annotations, the SGAN-Intra model supervised with the intra-step attention loss, and the SGAN-Inter model supervised with the inter-step attention loss. The comparison shows that even without any external supervision, the SGAN-Base can still effectively learn the integrated attention from the ground-truth solution graph, and achieve promising results. Its MRR, RSUM, and IoU scores are all significantly better than those of the SEQ ATT method (*e.g.*, RSUM is improved from 410.77 to 433.55), demonstrating the effectiveness of the proposed network design. Notably, the introduction of either intra-step or inter-step attention supervision leads to substantial improvements. In particular, compared with SGAN-Base, SGAN-Intra achieves an improvement of 5.6% and 6.3% in MRR scores for images and captions, respectively. Its RSUM score is improved from 433.55 to 451.82, outperforming the SGAN-Base by 4.2%. These improvements suggest that the supervision of intra-step attention can benefit the localization of important information in both modalities. Furthermore, SGAN-Inter's performance highlights its practical significance in predicting step dependencies. With inter-step attention supervision, it achieves an impressive average improvement of 37.7% across AUC, AUPR, and IoU scores. This suggests the models' applicability in real-world scenarios where detailed annotations may be limited. Overall, incorporating both types of attention supervision achieves the best results, demonstrating the effectiveness of the integral design of our method in modeling attention for vision-language problem solving.

5.5.3 Qualitative Results

To further understand the proposed integrated attention mechanism and how it contributes to the prediction of problem-solving procedures, we conduct a qualitative comparison of the predicted solution graph and their intra-step attention maps. The qualitative examples are shown in Fig. 5.3, where the proposed SGAN method is compared with the state-of-the-art SEQ ATT [52] method and the ground truth. For a clearer illustration, we present the optimal predicted solution graph obtained from the image or caption candidate pool. The results consist of (1) the final solution graph obtained

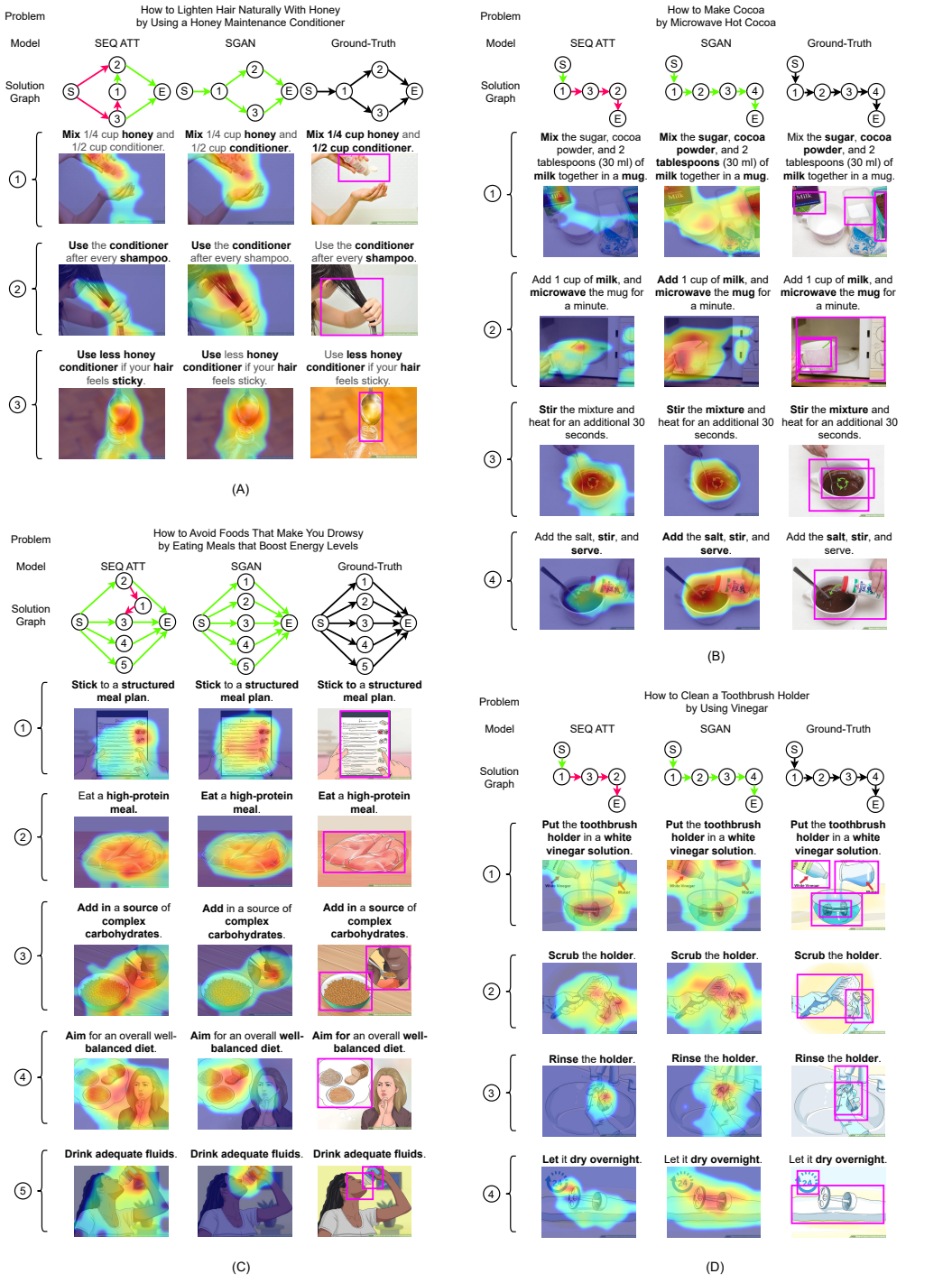

Figure 5.3: Qualitative comparison of the predicted solution graphs and intra-step attention maps. The green edges in the graphs indicate correct predictions, while the red ones indicate wrong predictions.

with Algorithm 1 showing the procedure flows across all steps, and (2) the intra-step attention maps for each problem-solving step overlaid on the images (*i.e.*, hot areas) and the captions (*i.e.*, bold text).

Despite leveraging explicit intra-step attention supervision based on fine-grained annotations, SEQ ATT sometimes fails to adequately attend to crucial objects relevant to problem-solving. As shown in Fig. 5.3, SEQ ATT allocates insufficient attention on the *conditioner* (see Fig. 5.3A, step 1), the *sugar* and *cocoa powder* (see Fig. 5.3B, step 1), the *structured meal plan* (see Fig. 5.3C, step 1), and the *rinse* action (see Fig. 5.3D, step 3). On the contrary, our proposed SGAN exhibits promising performance by attending to essential information within various steps. The comparison of intra-step attention between SEQ ATT and SGAN shows that progressively refining attention is effective in terms of learning accurate attention distribution in images and captions.

Furthermore, the inter-step attention mechanism is also shown to be effective in predicting the solution graph correctly. Because SEQ ATT sequentially predicts the dependencies one step at a time, it results in suboptimal solutions (see Fig. 5.3A-D). Differently, the integration of intra-step attention and inter-step attention in SGAN allows it to better understand the importance of key objects (*e.g.*, *conditioner*, *sugar*, *cocoa powder*, *structured meal plan*, *rinse*, *etc.*) across multiple steps. In addition, the progressive learning of the integrated attention mechanism allows SGAN to improve the solution graph by interactively refining it. Therefore, with a holistic view of the problem-solving procedure and interactive refinement, SGAN manages to predict the dependencies more accurately.

5.5.4 Performance Analyses

We further present extensive analyses to understand the roles and contributions of different components in our proposed approach. Through these in-depth analyses, we aim to gain a deeper understanding of the key factors that contribute to the success of our approach in solving complex vision-language problems.

Sequential and Non-Sequential Solutions

Unlike previous datasets that focus on sequential solutions, VisualHow is a unique dataset that contains a variety of complex problem-solving tasks. To demonstrate

Method	Mode	Sequence		Non-Sequence	
		MRR ↑	IoU@0.5 ↑	MRR ↑	IoU@0.5 ↑
SEQ GPO [157]	I	0.4435	0.190	0.4640	0.337
	C	0.5707	0.326	0.6487	0.564
SEQ GAP [52]	I	0.5338	0.181	0.5280	0.321
	C	0.5584	0.302	0.6346	0.563
SEQ ATT [52]	I	0.4968	0.187	0.5187	0.363
	C	0.6130	0.286	0.6952	0.584
SGAN	I	**0.5823**	**0.404**	**0.5986**	**0.616**
	C	**0.7087**	**0.537**	**0.7602**	**0.706**

Table 5.2: Solution graph prediction results for sequential and non-sequential solutions. The best results are highlighted in bold.

the effectiveness of our proposed method on different types of solution structures, we present the model's performance on both sequential and non-sequential solutions separately. In Table 5.2, we evaluate the performance of our method in both sequential and non-sequential problem-solving scenarios. The results show that our SGAN method outperforms the state-of-the-art methods in both scenarios, achieving the highest MRR and IoU@0.5 scores. This demonstrates that SGAN excels in capturing the structure and dependencies of solution steps, regardless of whether the problem-solving process is sequential or not, making it a versatile approach for a wide range of real-world applications that involve complex structures and diverse multimodal instructions.

Intra-Step Attention

The results presented in Table 5.3 provide insights into the performance of our intra-step attention mechanism. The attention output $\boldsymbol{\alpha}^{(L)}$ is evaluated using three metrics: CC, KLD, and SIM to quantify the quality of intra-step attention learning and help assess the effectiveness of this method in focusing on salient information. The state-of-the-art method SEQ ATT [52], which also learns intra-step attention following a sequential approach, achieves moderate results for both image and caption modalities. However, our proposed SGAN with intra-step attention (SGAN-Intra) outperforms SEQ ATT consistently across almost all the metrics (5/6) for both modalities. This demonstrates that the progressive refinement of the solution graph with intra-step attention enables the

Method	Mode	Intra-Step Attention		
		CC ↑	SIM ↑	KLD ↓
SEQ ATT	I	0.600	0.571	0.705
	C	0.764	0.623	0.616
SGAN-Base	I	0.133	0.401	1.540
	C	0.307	0.337	1.915
SGAN-Intra	I	**0.611**	0.586	0.668
	C	0.768	0.656	0.648
SGAN-Inter	I	0.456	0.523	0.880
	C	0.726	0.625	0.740
SGAN	I	0.608	**0.588**	**0.666**
	C	**0.772**	**0.657**	**0.605**

Table 5.3: Intra-step attention evaluation results. The best results are highlighted in bold.

model to focus on relevant information within each step, leading to improved attention quality. On the other hand, the impact of inter-step attention (SGAN-Inter) alone is not as significant on these evaluation metrics. However, integrating the two attention mechanisms is able to further improve the model's ability in finding important information in the images and captions. This highlights the importance of combining both attention mechanisms to achieve a comprehensive understanding of the problem-solving procedure.

Inter-Step Attention

Understanding how attention is aligned across multiple steps in complex problem-solving is crucial for developing effective learning models. Here, we provide a detailed analysis of our method by examining the attention alignment between steps. Table 5.4 presents the results of the inter-step attention evaluation, which sheds light on the model's ability to capture dependencies between problem-solving steps. The metrics used to evaluate inter-step attention include $S_{in}^{(L)}$, $S_{out}^{(L)}$, and $S_{all}^{(L)}$, which quantify the quality of attention propagation within the solution graph. The state-of-the-art method SEQ ATT [52] exhibits limited performance in capturing inter-step dependencies, as evidenced by the

Method	Mode	Inter-Step Attention		
		$S_{in}^{(L)} \uparrow$	$S_{out}^{(L)} \uparrow$	$S_{all}^{(L)} \uparrow$
SEQ ATT	I	0.0696	0.0551	0.0102
	C	0.1313	0.0998	0.0244
SGAN-Base	I	0.0766	0.0627	0.0209
	C	0.1102	0.0934	0.0309
SGAN-Intra	I	0.0878	0.0654	0.0125
	C	0.1354	0.0971	0.0217
SGAN-Inter	I	0.2017	0.2217	0.1366
	C	0.3081	0.3158	0.2288
SGAN	I	**0.2220**	**0.2360**	**0.1493**
	C	**0.3214**	**0.3274**	**0.2390**

Table 5.4: Inter-step attention evaluation results. The best results are highlighted in bold.

relatively low values of all metrics for both image and caption modalities. This is because the sequential design of SEQ ATT cannot effectively propagate attention to other steps across multiple steps, resulting in suboptimal predictions. However, the most significant improvement is observed with the addition of inter-step attention in the SGAN-Inter model. The values of $S_{in}^{(L)}$, $S_{out}^{(L)}$, and $S_{all}^{(L)}$ for SGAN-Inter are notably higher than those of SEQ ATT, SGAN-Base, and SGAN-Intra. The full SGAN model, which combines both intra-step and inter-step attention mechanisms, achieves the best results among all methods and modalities across all metrics. These observations indicate that the inter-step attention mechanism effectively captures the dependencies between problem-solving steps, allowing the attended information to effectively propagate across multiple steps, leading to improved reasoning about the chronological order of various solution steps.

Correlation Between Attention Performance and Task Performance

To further investigate how the intra-step and inter-step attention contribute to the model performance in tackling vision-language problems, we compute the Pearson's r between the attention evaluation scores CC, $S_{all}^{(L)}$ and task evaluation scores MRR and

Attention Type	Mode	Sequence		Non-Sequence	
		MRR	IoU@0.5	MRR	IoU@0.5
Intra-step	I	-0.068	**0.212**	-0.001	0.001
	C	-0.091	**0.236**	-0.048	-0.058
Inter-step	I	**0.677**	**0.417**	**0.625**	**0.738**
	C	**0.629**	**0.435**	**0.607**	**0.732**

Table 5.5: Pearson's r between attention evaluation score and our proposed SGAN model's performance. Bold numbers indicate significant positive correlations ($p < 0.05$).

Number	Mode	Retrieval ↑					Dependency ↑				
		MRR	R@1	R@5	R@10	RSUM	AUC	AUPR	IoU@0.25	IoU@0.5	IoU@0.75
1	I	0.5649	42.90	73.11	87.52	432.90	0.781	0.795	0.501	0.427	0.308
	C	0.6624	53.62	82.62	93.13		0.845	0.847	0.582	0.515	0.413
2	I	0.5846	45.00	75.35	88.46	451.77	0.796	0.809	0.553	0.482	0.369
	C	0.7224	61.26	86.83	94.86		0.855	0.854	0.629	0.583	0.504
3	I	**0.5898**	**45.97**	**75.61**	**88.89**	**455.56**	**0.800**	**0.817**	**0.580**	**0.508**	**0.394**
	C	**0.7324**	**62.77**	**86.95**	**95.37**		**0.862**	**0.864**	**0.659**	**0.620**	**0.547**
4	I	0.5815	44.98	74.38	88.08	448.05	0.799	0.811	**0.580**	0.491	0.363
	C	0.7169	60.89	85.36	94.36		0.861	0.858	0.655	0.618	0.532

Table 5.6: Ablation study of the number of integrated attention layers. The best results are highlighted in bold.

IoU@0.5. Table 5.5 shows the Pearson's correlation coefficient (r) between the attention evaluation scores and the performance of our proposed SGAN model on predicting sequential solutions and non-sequential ones. For the intra-step attention evaluation, we observe a significant positive correlation between attention performance and model's ability to predict the dependencies in sequential solutions. The correlation coefficients for IoU@0.5 are 0.212 and 0.238 for the image and caption modalities, respectively. On non-sequential problems, the correlation coefficients are close to zero, indicating a weak correlation between intra-step attention performance and model performance. The weak correlations suggest that in the final SGAN model, the quality of attention within individual steps has limited impacts on the model's performance. In contrast, the inter-step attention evaluation shows strong positive correlations between attention performance and model performance on both sequential and non-sequential solutions. In particular, for non-sequential ones, attention performance is highly correlation with the IoU@0.5,

Layer	Mode	Intra-Step Attention			Inter-Step Attention			Retrieval	Dependency
		CC ↑	SIM ↑	KLD ↓	$S_{in}^{(L)}$ ↑	$S_{out}^{(L)}$ ↑	$S_{all}^{(L)}$ ↑	MRR ↑	IoU@0.5 ↑
1	I	0.607	0.587	0.669	0.1994	0.2039	0.1366	0.5867	0.450
	C	0.764	0.656	0.608	0.2981	0.2993	0.2275	0.7324	0.569
2	I	0.607	0.588	0.667	0.2201	0.2297	**0.1512**	0.5862	0.488
	C	0.765	0.656	0.606	0.3191	0.3210	**0.2421**	**0.7327**	0.601
3	I	**0.608**	**0.588**	**0.666**	**0.2220**	**0.2360**	0.1493	**0.5898**	**0.508**
	C	**0.772**	**0.657**	**0.605**	**0.3214**	**0.3274**	0.2390	0.7324	**0.620**

Table 5.7: Evaluations on Intra-step attention, Inter-step attention, and solution graph prediction results across layers. The best results are highlighted in bold.

Proportion	Mode	Retrieval ↑					Dependency ↑				
		MRR	R@1	R@5	R@10	RSUM	AUC	AUPR	IoU@0.25	IoU@0.5	IoU@0.75
0%	I	0.5524	41.99	71.42	86.95	433.55	0.721	0.777	0.487	0.307	0.152
	C	0.6820	56.66	83.10	93.43		0.799	0.834	0.553	0.481	0.321
20%	I	0.5551	42.12	71.84	87.00	434.13	0.774	0.803	0.546	0.405	0.286
	C	0.6829	56.65	83.50	93.03		0.834	0.849	0.612	0.453	0.383
40%	I	0.5875	45.48	75.35	**89.35**	453.59	0.786	0.810	0.548	0.423	0.299
	C	0.7272	61.95	86.58	94.87		0.846	0.857	0.635	0.565	0.443
60%	I	0.5890	45.94	75.14	89.12	454.29	0.793	0.814	0.560	0.457	0.335
	C	0.7286	62.26	86.92	94.92		0.853	0.859	0.643	0.593	0.488
80%	I	0.5863	45.52	74.78	89.02	454.21	0.796	0.815	0.570	0.484	0.367
	C	**0.7334**	**63.16**	86.61	95.12		0.859	0.863	0.658	0.613	0.527
100%	I	**0.5898**	**45.97**	**75.61**	88.89	**455.56**	**0.800**	**0.817**	**0.580**	**0.508**	**0.394**
	C	0.7324	62.77	**86.95**	**95.37**		**0.862**	**0.864**	**0.659**	**0.620**	**0.547**

Table 5.8: Ablation study of the proportion of fine-grained data annotations. The best results are highlighted in bold.

with values of 0.738 and 0.732 for the image and caption modalities, respectively. The strong positive correlations suggest that the quality of inter-step attention is closely related to the model's ability to capture dependencies between problem-solving steps and predict coherent and structured solutions. These results indicate that the inter-step attention mechanism plays a crucial role in improving the model's performance on both sequential and non-sequential problems.

Number of Attention Layers

Progressively refining attention is a fundamental component of our proposed SGAN architecture, which enables the network to iteratively focus on key information within from visual and textual inputs and discover the dependencies between the steps. To verify the effect of the number of integrated attention layers, we conduct experiments with four variants of our models. As shown in Table 5.6, for the retrieval of the most relevant images and captions, increasing the number of attention layers consistently improves the model's performance. We observe that with three attention layers, the SGAN model achieves the highest MRR, Recall@K, and RSUM scores for both the image and caption modalities. However, adding more layers does not lead to further improvements in the retrieval performance. Similar trends are observed for the evaluation of step dependencies, with AUC, AUPR, and IoU scores. Overall, this ablation study demonstrates that a three-layer SGAN model results in the right balance between capturing relevant information within individual steps and modeling the dependencies between steps. This configuration achieves the best performance for both retrieval and dependency aspects, indicating its effectiveness in tackling complex multimodal problem-solving tasks.

Progressive Attention Refinement Across Layers

Gradually refining attention constitutes an important element within our proposed SGAN method, empowering the model to progressively concentrate on key information across visual and textual inputs, unraveling inter-step dependencies. To illustrate the effectiveness of progressively refining attention in our proposed SGAN, we compare the outputs of different layers, including the intra-step attention $\boldsymbol{\alpha}^{(\ell)}$ and the inter-step attention $\boldsymbol{P}^{(\ell)}$ ($\ell = 1, 2, 3$). As shown in Table 5.7, we find that the attention alignments (Intra-Step Attention and Inter-Step Attention) exhibit a progressive enhancement as the layers delve deeper. This suggests that, with each subsequent layer, the model refines its ability to focus on relevant information, capturing more detailed relationships. This refinement in attention aligns with an observed improvement in prediction performance metrics, including MRR and IoU@0.5, suggesting the significance of this progressive attention mechanism in the success of problem solving.

Proportion of Attention Annotations

In Table 5.1, we have demonstrated that SGAN-Base can self-learn attention from the solution graph, which has performed better than the SEQ ATT [52] model that requires additional attention annotations, while learning from annotations with the proposed objectives can further improve the model's performance. To study the impact of the annotations on model performance, we use different proportions of annotations in training, ranging from 0% to 100%, and evaluate the model's performance using various metrics. Table 5.8 presents the results of our ablation study on the proportion of fine-grained data annotations used in training the SGAN model. For both the retrieval and the dependency evaluations, we observe that all evaluation scores increase steadily with a higher proportion of fine-grained annotations. This indicates that providing more detailed annotations enhances the model's ability to accurately retrieve multimodal instructions for individual problem-solving steps, as well as to better predict the structured dependencies between steps.

Using Pre-Trained Grounding As Attention Annotations

Although providing more attention annotations can improve model performance, the practicality of obtaining such annotations may raise scalability concerns. To address this, instead of leveraging human annotations, we generate ground-truth attention annotations using a pre-trained GLIP [29] model, which exhibits strong zero-shot and few-shot transferability to diverse object-level recognition tasks. As shown in Table 5.9, the GLIP-generated annotations demonstrate comparable performance as the human annotations from the VisualHow dataset. This consistency suggests that large pre-trained vision-language models can provide sufficient attention annotations for modeling intra-step attention across various problems, offering a viable approach to scalability.

Multimodal Procedure Planning Models

Table 5.10 compares the performance of our method with state-of-the-art multimodal procedure planning models, including Text-Image Prompting (TIP) [265] and Skip-Plan [243]. TIP generates a sequence of step captions using the text-davinci-003 model [28],

Source	Mode	Retrieval ↑					Dependency ↑				
		MRR	R@1	R@5	R@10	RSUM	AUC	AUPR	IoU@0.25	IoU@0.5	IoU@0.75
GLIP [29]	I	**0.5907**	**46.06**	75.39	**89.28**	453.21	0.797	0.812	0.572	0.507	**0.395**
	C	0.7238	61.52	86.32	94.65		0.852	0.856	0.648	0.495	0.451
VisualHow [52]	I	0.5898	45.97	**75.61**	88.89	**455.56**	**0.800**	**0.817**	**0.580**	**0.508**	0.394
	C	**0.7324**	**62.77**	**86.95**	**95.37**		**0.862**	**0.864**	**0.659**	**0.620**	**0.547**

Table 5.9: Solution graph prediction results with different sources of attention annotations. In each panel, the first row (I) indicates the image modality and the second row (C) indicates the caption modality. The best results are highlighted in bold.

and subsequently converting these captions into images using Stable Diffusion [25]. Skip-Plan learns to predict solutions by breaking down a long chain of steps into several reliable sub-chains, addressing error accumulation in long sequence predictions. Since these sequential methods cannot handle complex graph structures, we only compare them with our method by evaluating them through image and caption retrieval. As shown in Table 5.10, there is a notable discrepancy in the retrieval capabilities of the TIP model between image and caption retrieval tasks, indicating a greater proficiency in processing and extracting information from textual data compared to visual inputs. On the other hand, the Skip-Plan model exhibits an improved retrieval performance, a result of its end-to-end training on the VisualHow [52] dataset. However, these state-of-the-art procedure planning methods still underperform our SGAN model, because of their sequential nature. The graph-based model architecture and the novel attention mechanisms allow SGAN to capitalize on the extensive in-domain problem-solving knowledge embedded in the VisualHow dataset [52], achieving a significant performance improvement. This improvement solidifies SGAN's status as a promising solution for effectively addressing multimodal complexities in problem solving.

Similarity Functions used in Attention Learning

In this ablation study, we investigate the impact of adopting different attention evaluation metrics on attention learning. We consider three widely used similarity functions: SIM [257], JSD [258, 259], and CC [257], which are applied to supervise the inter-step attention mechanism in our proposed SGAN model. The results in Table 5.11 demonstrate that our attention supervision method is robust against the choice of similarity

Model	Mode	Retrieval ↑				
		MRR	R@1	R@5	R@10	RSUM
TIP [265]	I	0.4046	29.09	50.72	64.06	377.67
	C	0.7124	62.08	82.40	89.32	
Skip-Plan [243]	I	0.4819	33.72	64.21	82.25	384.08
	C	0.5571	40.92	74.24	88.76	
SGAN	I	**0.5898**	**45.97**	**75.61**	**88.89**	**455.56**
	C	**0.7324**	**62.77**	**86.95**	**95.37**	

Table 5.10: Comparison of multimodal procedure planning models. The best results are highlighted in bold.

Similarity	Mode	Retrieval ↑					Dependency ↑				
		MRR	R@1	R@5	R@10	RSUM	AUC	AUPR	IoU@0.25	IoU@0.5	IoU@0.75
SIM [257]	I	0.5896	45.81	75.74	89.03	455.44	0.796	0.815	0.576	0.485	0.377
	C	0.7310	62.48	87.22	95.17		0.861	0.865	0.657	0.615	0.531
JSD [258, 259]	I	0.5898	45.97	75.61	88.89	455.56	0.800	0.817	0.580	0.508	0.394
	C	0.7324	62.77	86.95	95.37		0.862	0.864	0.659	0.620	0.547
CC [257]	I	0.5921	46.24	75.52	88.89	456.62	0.798	0.816	0.579	0.507	0.389
	C	0.7370	63.47	87.14	95.37		0.862	0.863	0.659	0.621	0.553

Table 5.11: Ablation study of similarity functions used in the proposed evaluation metrics.

function, as all three metrics produce similar performance. This consistency in performance indicates that our method effectively captures the attention alignment from different perspectives, leading to comparable results regardless of the selected similarity function. Based on these findings, we adopt JSD similarity in the measurement of inter-step attention. Overall, these results affirm the effectiveness of our approach in measuring attention alignment from multiple angles. This versatility is crucial for the success of our SGAN model in solving complex multimodal problem-solving tasks, as it allows the model to capture fine-grained dependencies between individual solution steps, leading to more accurate and coherent predictions.

λ_d	λ_r	Precision	Recall	$F1$
0.2		0.477	0.605	0.499
0.5		0.477	0.605	0.499
0.8	0.45	0.482	0.601	**0.500**
1.1		0.490	0.586	0.498
1.4		0.495	0.529	0.473
	0.05	0.444	0.650	0.493
	0.25	0.462	0.629	0.499
0.8	0.45	0.482	0.601	**0.500**
	0.65	0.505	0.557	0.494
	0.85	0.525	0.473	0.463

Table 5.12: Ablation study on different combinations of dependency threshold λ_d and retrieval threshold λ_r. The best results are highlighted in bold.

Graph Post-Processing Thresholds

Finally, we investigate the impacts of the thresholds (*i.e.*, dependency threshold λ_d and retrieval threshold λ_r) on the predicted solution graph. It is noteworthy that following the VisualHow [52] benchmark, quantitative results presented in this paper, including the evaluation of retrieval, dependency, intra-step attention, and inter-step attention, are based on the probabilistic output $\boldsymbol{P}^{(L)}$. The dependency threshold λ_d and retrieval threshold λ_r are only used to binarize the soft probabilities into the final deterministic solution graph. In Table 5.12, we show various threshold combinations and their corresponding precision, recall, and $F1$ scores computed with the final solution graph. These scores are derived from comparing the ground-truth solution graph with binarized solution graphs after post-processing. The analysis reveals that the final solution graphs are not significantly affected by the choice of the dependency threshold λ_d ($0.2 \leq \lambda_d \leq 1.1$). The retrieval threshold λ_r acts as a balancing factor between precision and recall, and the final solution graphs are not sensitive to the choice of it ($0.05 \leq \lambda_r \leq 0.65$). Based on this observation, we empirically choose $\lambda_d = 0.8$ and $\lambda_r = 0.45$ for our experiment.

5.6 Conclusion

In this paper, we focus on addressing existing gaps in understanding and providing effective step-by-step instructions for problem-solving in vision-and-language applications. Our contribution is a novel Solution Graph Attention Network (SGAN) approach that takes into account both intra-step and inter-step attention mechanisms, enabling a progressive construction of solutions by refining the dependencies between relevant problem-solving steps. The flexibility of our method allows for the formulation of solutions with various structures, accommodating both sequential and non-sequential patterns. In order to enhance the accuracy of attention in the problem-solving process, we have introduced quantitative metrics to study the role of attention in task accomplishment. These metrics serve as valuable tools for attention supervision, providing insights into how attention mechanisms can be leveraged effectively.

Our experimental results showcase the advantages of our proposed method in tackling a wide range of vision-language problems. By employing our model, we achieved significant improvements in formulating solutions with complex graph structures. Moreover, our findings shed light on the crucial components that contribute to successful problem-solving, thus offering valuable insights for future research and applications. We believe that the insights gained from our work will have a profound impact on solving intricate visual problems and providing effective guidance for various daily-life activities. Our method not only advances the state-of-the-art in vision-language problem solving, but also lays the groundwork for the development of more powerful and flexible attention mechanisms. With the hope that our work will inspire further advancements in this field, we envision that our proposed GNN-based model and attention supervision techniques will continue to drive progress in solving problems more effectively and efficiently.

While our proposed method shows promising results in tackling vision-language problem-solving tasks, it also has several limitations and opens up interesting avenues for future research. One limitation is that our method relies on annotated data for training and supervision. We have explored GLIP-generated annotations to reduce the data dependency and improve the generalization capabilities of our model, which has shown promising results. Another challenge we face in this work is that the dependencies

between steps may not always be clear-cut. There can be cases where multiple possible dependencies exist, leading to ambiguity in constructing the solution graph. Developing methods to handle such ambiguity and effectively capture uncertain dependencies is an important direction for future research.

Chapter 6

GazeXplain: Learning to Predict Natural Language Explanations of Visual Scanpaths

While exploring visual scenes, humans' scanpaths are driven by their underlying attention processes. Understanding visual scanpaths is essential for various applications. Traditional scanpath models predict the where and when of gaze shifts without providing explanations, creating a gap in understanding the rationale behind fixations. To bridge this gap, we introduce GazeXplain, a novel study of visual scanpath prediction and explanation. This involves annotating natural-language explanations for fixations across eye-tracking datasets and proposing a general model with an attention-language decoder that jointly predicts scanpaths and generates explanations. It integrates a unique semantic alignment mechanism to enhance the consistency between fixations and explanations, alongside a cross-dataset co-training approach for generalization. These novelties present a comprehensive and adaptable solution for explainable human visual scanpath prediction. Extensive experiments on diverse eye-tracking datasets demonstrate the effectiveness of GazeXplain in both scanpath prediction and explanation, offering valuable insights into human visual attention and cognitive processes.

Figure 6.1: This example reveals how observers strategically investigate a scene to find out if the person is walking on the sidewalk. Fixations (circles) start centrally, locating a driving car, then shifting to the sidewalk to find the person, and finally looking down to confirm if the person is walking. By annotating observers' scanpaths with detailed explanations, we enable a deeper understanding of the "what" and "why" behind fixations, providing insights into human decision-making and task performance.

6.1 Introduction

Picture yourself driving through a bustling city at dusk, with your eyes scanning the surroundings for critical details like pedestrian crossings, brake lights, and turn signals. These seemingly random glances are guided by an internal dialogue questioning your environment. As depicted in figure 6.1, when determining if a person on the sidewalk is standing or walking, our gaze naturally shifts from the car ahead to the sidewalk. We may fixate on their upper body to start with, and then move downward to assess their movement. Understanding this implicit language of gaze and translating it into explicit explanations, such as whether we correctly deduced the person's movement or overlooked subtle details, holds significant potential for enhancing human-machine interaction.

Research on human attention modeling builds upon decades of study in psychology and cognitive science, aiming to understand how humans allocate their attention to visual stimuli [138, 133]. Recent studies have shifted from static fixation distribution modeling to dynamic gaze patterns, known as scanpaths [70, 65, 50, 144]. Current scanpath models excel at tracking visual exploration trajectories, predicting "when" and "where" people shift their attention. However, scanpath prediction models fall short of explicitly explaining the "what" and "why" – the insights behind each fixation.

This lack of explainability hampers the understanding and potential applications of these models in real-world scenarios.

To bridge this explainability gap, we introduce GazeXplain, a novel study that goes beyond predicting where people look; it demands models to explain them in natural language, weaving a narrative thread that connects fixations to their underlying meaning. Particularly, GazeXplain features several key distinctions from existing scanpath prediction methods: (1) We annotate ground-truth explanations for scanpaths over diverse eye-tracking datasets. These annotations build a strong foundation for modeling scanpath explanation, unlocking explainable methods that understand user attention in applications. (2) We introduce a general model architecture with an attention-language decoder jointly predicting scanpaths and natural language explanations. (3) We present a novel semantic alignment mechanism that promotes consistency between the vision and language modalities, guiding the model toward generating explanations that faithfully reflect the fixated visual information. (4) While existing models target single task-specific datasets, such as free-viewing, object search, or visual question answering (VQA), we generalize scanpath prediction and explanation with a cross-dataset co-training technique, overcoming data and task-specific biases.

In summary, the contributions of this paper are outlined as follows:

1. We introduce a novel task aiming to jointly predict and explain scanpaths, offering a deeper semantic understanding of what people look.

2. We annotate ground-truth explanations on three public eye-tracking datasets, providing detailed fixation-level explanations.

3. We propose a general model architecture with an attention-language decoder that jointly predicts scanpath and explanation. It incorporates a novel semantic alignment mechanism for consistent fixation-explanation alignment, along with cross-dataset co-training for enhanced generalizability.

4. Comprehensive experiments across various datasets demonstrate GazeXplain's effectiveness in generating accurate scanpaths and explanations, highlighting the importance of explanation prediction, semantic alignment, and cross-dataset co-training on model performance.

6.2 Related Work

Visual Scanpath Prediction. Understanding human visual attention requires insight into the dynamic sequence of eye fixations. While static saliency prediction has been extensively studied [136, 137, 138, 139, 140, 141, 142, 143], dynamic scanpath prediction remains relatively underexplored due to its complexity influenced by various factors. Early studies employed heuristics or statistical priors to generate scanpaths [62, 63, 64, 65, 66], while recent models leverage machine learning techniques, including supervised learning [136, 148, 144, 266, 59, 58, 155] and reinforcement learning [50, 60, 146], achieving promising results [50, 60, 146, 148, 144, 51, 267]. However, these methods lack interpretability and struggle to explain the predicted fixations. Our method, GazeXplain, stands out in two aspects: Firstly, it generates natural language explanations for predicted fixations, going beyond mere scanpath prediction. Secondly, it ensures generalizability across visual tasks by training on a combination of datasets. This improved explainability and generalizability represent significant advancements in understanding human visual attention processes.

Explanations. Automatic reasoning and explanation [268] initially rely on rules or predefined templates to explain medical diagnosis [269], simulator actions [270, 271, 272, 273] and robot movements [274] *etc*. Recent explanation models explored deep learning-based natural language generation, with successful applications in producing natural language justifications for object classification [275, 276, 277], visual reasoning [278, 279, 280, 281, 282, 283, 284], recommendation systems [285], and sentiment analysis [286], *etc*. Different from these studies, we for the first time explore natural language explanations of eye-tracking data to facilitate a deeper understanding of human visual behaviors. Our proposed GazeXplain model simultaneously predicts scanpaths and explanations, establishing a direct semantic connection to jointly improve the scanpath prediction and explanation accuracy.

Vision and Language Models. GazeXplain is inspired by the success of deep vision-language models [287, 249, 288, 77, 289, 225, 52, 53, 290]. These models, trained on multimodal image and language datasets [175, 178, 52], are able to generate fluent and accurate descriptions of visual information. The recent advent of transformer architecture [16, 291, 292] marked a significant breakthrough, providing a robust framework

for handling intricate relationships and long-range dependencies. This advancement facilitated the development of large-scale vision-language models that excel in translating visual information into natural language descriptions [293, 294, 23, 22]. While these models have achieved impressive results in characterizing vision-language features, scanpath models haven't fully leveraged this capability to enhance human attention prediction. Unlike existing scanpath models, GazeXplain builds upon the strengths of vision-language models, incorporating explainability in scanpath prediction. By leveraging the capabilities of vision-language models for both ground-truth annotations and language modeling, GazeXplain deciphers the attention and reasoning behind fixations, bridging the gap between visual attention and language understanding.

6.3 GazeXplain

Human visual attention is a complex interplay across multiple visual features and cognitive factors [64, 295, 296] (*e.g.*, low-level contrasts, objects, semantics, goals, and prior knowledge, *etc.*) However, existing deep learning-based scanpath models lack transparency in explaining how different factors influence their predictions. Our work tackles this challenge through novel dataset construction and modeling approaches: (1) We annotate new scanpath explanations based on existing eye-tracking datasets, offering ground-truth explanations for fixations across diverse tasks like VQA, free-viewing, and search. (2) We propose the first scanpath prediction and explanation model generating natural language explanations alongside predicted scanpaths, featuring novel techniques including attention-language decoder, semantic alignment, and cross-dataset co-training.

6.3.1 Data

We propose data annotation to offer ground-truth explanations for fixations across various eye-tracking datasets. Compared to previous image-to-language datasets, it has two key distinctions: (1) We present the first natural-language annotations on scanpaths, offering explanations for each specific fixation within the scanpath, rather than image-wise descriptions such as image captioning [175, 178] and visual storytelling [167]. This

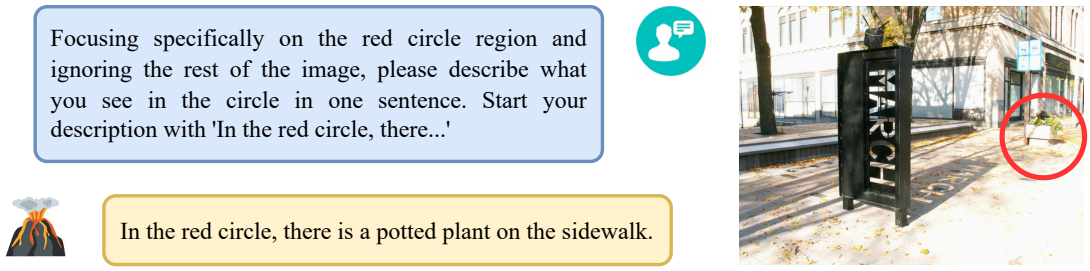

Figure 6.2: LLaVA generates the ground-truth explanation for each fixation using an input image with a red circle marking the fixation. The model's response provides information within the marked area, serving as a basis for further refinement.

granular level of detail offers deeper insights into the cognitive process behind each fixation. (2) While most image-to-language datasets focus on specific tasks, ours comprise a wider range of visual tasks, including free-viewing [105], object search [60, 146], and VQA [1]. This ensures the diversity of explanations, allowing models to be co-trained across multiple datasets to enhance their generalizability.

While the manual annotation of fixation-level explanations is subjective and time-consuming, we employ a novel semi-automated approach, leveraging the power of large vision-language models to efficiently generate accurate explanations for every eye fixation. figure 6.2 illustrates our paradigm for annotating explanations. We utilize LLaVA-1.6 [22] (with a Mistral-7B [297] base language model) for its renowned ability to understand and describe visual information. We combine visual and language prompts to guide the model's description: Firstly, we generate a visual prompt by enclosing each fixation within a red circle [298], mirroring the size of the human fovea [299] (*i.e.*, a diameter of 5 degrees), thereby directing LLaVA's attention to the fixated region. Complementing this, we crafted a language prompt that instructs LLaVA to describe the image information within the circled area in one sentence (see figure 6.2). These prompts guide LLaVA to generate concise and contextually relevant descriptions centered solely on the fixation. Preliminary evaluations have demonstrated the effectiveness of this prompting technique compared to alternative methods, such as describing multiple fixations simultaneously or generating foveated images as prompts, where the proposed one avoids issues related to information overload of multiple explanations or the complexity

Dataset	Task	Images	Scanpaths	Length of Scanpath	Words per Fixation	Words per Scanpath
AiR-D	VQA	987	13,903	10.17 ± 2.23	10.79 ± 3.46	109.81 ± 31.27
OSIE	Free Viewing	700	10,500	9.36 ± 1.88	11.43 ± 3.99	107.07 ± 31.26
COCO-Search18 TP	Object Search	3,101	30,998	3.48 ± 1.82	9.84 ± 3.14	34.28 ± 20.55
COCO-Search18 TA	Object Search	3,101	31,006	5.86 ± 4.07	10.61 ± 3.45	62.21 ± 45.85

Table 6.1: Statistics of the eye-tracking datasets with annotated explanations.

of computing foveated images. Finally, these generated descriptions are combined in the order of fixations to describe the full scanpath, enabling the extraction of meaningful insights into the dynamic shift of attention.

While LLaVA's capabilities are impressive, it may exhibit limited robustness in handling noisy or ambiguous visual inputs, such as small objects, text, or complex scenes with cluttered backgrounds. Therefore, manual quality control remains crucial for ensuring accuracy and objectivity. To improve the data quality, we review and revise generated explanations based on the following criteria: Firstly, any reference to the red circle is eliminated to ensure that descriptions accurately reflect the information of the fixated regions. Secondly, for consistency and readability across datasets, the generated descriptions are maintained within a specific length (*e.g.*, 5-20 words), facilitating subsequent analysis and interpretation. Thirdly, in images containing English text, the text recognition is manually verified and corrected. Finally, to ensure the consistency of explanations of fixations on the same object or region, we apply MeanShift [300] clustering to fixation positions and manually correct semantically different explanations in each cluster without sacrificing linguistic diversity. This quality control process enhances the overall accuracy, objectivity, and reliability of the annotations, mitigating potential errors introduced by automated processes.

By leveraging the combined strengths of LLaVA and human expertise, we annotat ground-truth explanations for three different eye-tracking datasets: AiR-D [1], OSIE [105], as well as COCO-Search18 [60] including target-present (TP) and target-absent (TA) subsets. As shown in table 6.1, this results in a rich collection of natural-language explanations annotated on 7,004 images and 86,407 fixations across diverse visual tasks. The explanations are concise, with lengths falling within 10.66 ± 3.54 words each. The AiR-D dataset, involving question-answering scenarios, exhibited a

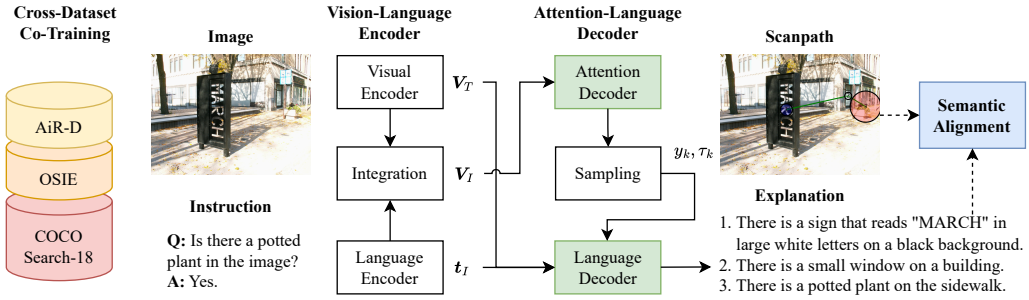

Figure 6.3: GazeXplain's architecture consists of a general vision-language encoder and a novel attention-language decoder. The decoder outputs an explanation for each fixation in the predicted scanpath, with a semantic alignment mechanism facilitating the semantic consistency between fixations and explanations. The model is developed on three public datasets using a cross-dataset co-training technique.

range of explanation lengths (*i.e.*, 10.79 per fixation), likely reflecting the varied complexity of the questions and corresponding fixations. Explanations for free-viewing tasks in OSIE tended to be slightly longer (*i.e.*, 11.43 words per fixation) compared to search-oriented tasks in COCO-Search18 (*i.e.*, 10.33 words per fixation). This aligns with the inherent differences in information processing during free exploration versus focused object search. Overall, the annotated explanations offer a valuable resource for researchers studying visual attention and its connection to language.

6.3.2 Model

The core challenge in scanpath explanation is the mapping ambiguity: translating brief fixations with limited context into clear natural language descriptions. This difficulty stems from inherent subjectivity in visual perception and the lack of explicit semantic meaning behind each fixation. To address this, GazeXplain presents a three-fold solution (see figure 6.3):

GazeXplain is built on top of a general vision-language encoder [144, 161]: Given an image (*i.e.*, the visual stimuli) and a language instruction (*i.e.*, the task context) as inputs, the encoder computes the image features $\boldsymbol{V}_T \in \mathbb{R}^{d \times hw}$, the semantic embedding $\boldsymbol{t}_I \in \mathbb{R}^{d_{\text{text}}}$, and the joint visual-semantic embedding $\boldsymbol{V}_I \in \mathbb{R}^{d \times hw}$, where $h \times w$ is the

size of the image feature maps, d_{text} is the semantic embedding dimensionality, and d is the joint embedding dimensionality.

Our **attention-language decoder** employs these features in predicting explanations alongside fixations, leveraging a novel **semantic alignment** mechanism to ensure that explanations accurately reflect fixated information. GazeXplain's language input generalizes it to a wider range of eye-tracking tasks, allowing it to be trained on various eye-tracking datasets with different task designs. This **cross-dataset co-training** equips GazeXplain with a broader range of knowledge across different tasks and prevents overfitting to specific datasets, improving model robustness and generalizability.

Attention-Language Decoder.

GazeXplain goes beyond conventional scanpath models by introducing a novel attention-language decoder to bridge the gap between visual attention patterns and natural language explanations.

The **attention decoder** utilizes a transformer model to generate feature vectors $\{\boldsymbol{s}_k|\boldsymbol{s}_k \in \mathbb{R}^d\}_{k=1}^K$, indicating salient features at each temporal step, where K is the maximum number of fixations. With a cross-attention mechanism, it computes the cosine similarity between $\boldsymbol{s}_k$ and the joint vision-language embedding $\boldsymbol{V}_I$ to predict the spatiotemporal distribution of fixations, denoted as $\{\boldsymbol{m}_k|\boldsymbol{m}_k \in \mathbb{R}^{h\times w}\}_{k=1}^K$. Additionally, it predicts parameters $\{\mu_k, \sigma_k^2\}_{k=1}^K$ characterizing the log-normal distribution of fixation durations, along with a binary indicator $\{e_k\}_{k=1}^K$ denoting the end of the scanpath. Following [50], we perform Monte Carlo sampling to obtain fixation positions $\{y_k\}_{k=1}^{K'}$ and durations $\{\tau_k\}_{k=1}^{K'}$, where K' is the length of the sampled scanpath.

The **language decoder** in GazeXplain is a novel and distinguishing component designed to provide comprehensive semantic explanations for fixated regions, accomplished through three key steps:

1. From the visual encoder's output $\boldsymbol{V}_T$, it extracts the local features according to the fixation position y_k, which results in the fixated features $\boldsymbol{g}_k \in \mathbb{R}^d$ that captures the salient information within the fixated region, emphasizing localization over the entire image.

2. To integrate visual features $\boldsymbol{g}_k$ and semantic embedding $\boldsymbol{t}_I$ effectively, we transform

them into the same dimensionality $\boldsymbol{g}_k^d \in \mathbb{R}^{d_{\text{text}}} = \boldsymbol{W}_d \boldsymbol{g}_k + \boldsymbol{v}_I$ and $\boldsymbol{t}_I^d \in \mathbb{R}^{d_{\text{text}}} = \boldsymbol{t}_I + \boldsymbol{v}_T$, through learnable parameters $\boldsymbol{W}_d \in \mathbb{R}^{d_{\text{text}} \times d}$ and positional encodings $\boldsymbol{v}_I, \boldsymbol{v}_T \in \mathbb{R}^{d_{\text{text}}}$, allowing for the integration of both visual and textual information. This integration facilitates the description of local visual information in the context of task instruction.

3. To generate the description, the features $\boldsymbol{g}_k^d$ and $\boldsymbol{t}_I^d$ are stacked and fed into a pre-trained language model (*e.g.*, BLIP [293]), leveraging its contextual understanding and linguistic capabilities. This enables the generation of detailed and informative explanations $\{\boldsymbol{w}_\ell^k\}_{\ell=1}^L$ for each fixation, where L represents the length of the generated explanation.

By integrating visual and semantic features and incorporating language models, our language decoder enables the explanations of the scanpath predictions.

Semantic Alignment.

We propose a semantic alignment mechanism to ensure the semantic consistency between predicted fixations, explanations, and visual features. It operates by computing the cosine similarity $S_{\cos}(\cdot, \cdot)$ of different categories of features between the i-th and the j-th fixations of a scanpath:

1. The **visual similarity** serves as pseudo labels for supervising the semantic alignment. It is computed as $s_{i,j}^r = S_{\cos}(\boldsymbol{r}_i, \boldsymbol{r}_j)$, where $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ represent the local image features at the fixation points, obtained from a pre-trained and frozen ResNet [6] model.

2. The **explanation similarity**, computed as $s_{i,j}^e = S_{\cos}(\boldsymbol{e}_i^p, \boldsymbol{e}_j^p)$, measures how closely the explanations of different fixations resemble each other semantically, where $\boldsymbol{e}_i^p$ and $\boldsymbol{e}_j^p$ represents the language features of the corresponding explanations, obtained from the language decoder.

3. The **fixation similarity**, computed as $s_{i,j}^f = S_{\cos}(\boldsymbol{g}_i, \boldsymbol{g}_j)$, compares the fixated features acquiring global image information from the visual encoder. It measures whether the two fixations focus on similar visual information.

4. The **multimodal similarity**, computed as $s^m_{i,j} = S_{\cos}(\boldsymbol{e}^p_i, \boldsymbol{g}_j)$, measures the gap between the language features $\boldsymbol{e}^p_i$ and the visual features $\boldsymbol{g}_j$, evaluating how well the explanations align with the visual information fixated upon.

Based on the similarity measures, the semantic alignment loss is denoted as

$$\mathcal{L}_{\text{aln}} = \frac{1}{K'^2} \sum_{i=1}^{K'} \sum_{j=1}^{K'} \Big((s^e_{i,j} - s^r_{i,j})^2 + (s^f_{i,j} - s^r_{i,j})^2 + (s^m_{i,j} - s^r_{i,j})^2 \Big), \tag{6.1}$$

which compares similarities $s^e_{i,j}$, $s^f_{i,j}$, $s^m_{i,j}$ against their corresponding pseudo labels $s^r_{i,j}$. Minimizing this loss during the optimization process encourages alignment of semantic representations across fixations, explanations, and visual features, ensuring consistency in the understanding of the scanpath, fostering explanations of the visual scene throughout the scanpath. Our final training objective combines this loss with a traditional scanpath prediction loss [50] and a language generation loss [88, 301], jointly optimizing scanpath prediction and explanation.

Cross-Dataset Co-Training.

Prior studies commonly focus on single dataset training [59, 50, 144]. For example, ChenLSTM relies on external VQA models to predict scanpaths on the AiR-D dataset [1], while Gazeformer targets search-related tasks offered by COCO Search-18 [60]. Such model and task dependencies limit their broader applicability. To address this, we propose cross-dataset co-training, enabling models to learn from multiple datasets simultaneously. We standardize inputs across datasets, ensuring compatibility and meaningful interaction. On the one hand, images and scanpaths are scaled to 384×512 resolution. On the other hand, task-specific instructions are structured into a standard VQA format. For example, for free-viewing, a general question "What do you see in the image?" is asked, while object search instructions are converted to a question "Is there a `[search target]` in the image?" with a binary "yes/no" answer. Optionally, on datasets with behavioral responses (*e.g.*, AiR-D, COCO-Search18), the observer's answer is also added to the instruction, which helps the model to understand inter-observer variations. Different from general co-training techniques relying on structured input formats [267], GazeXplain's free-formed input captures rich semantics for

scanpath explanation, allowing the model to understand the specific contexts and goals. In this way, models can be trained with a combination of multiple datasets, unlocking their full potential in generalization across various tasks.

6.4 Experiments

We evaluate GazeXplain through comprehensive experiments: (1) performance comparison against state-of-the-art methods, (2) ablation studies to understand component contributions, (3) evaluation of generated explanations, and (4) qualitative analysis of predicted scanpaths and explanations.

6.4.1 Experimental Setup

Datasets. Our experiments utilize a combination of eye-tracking datasets. AiR-D [1] provides insights into human gaze behavior in VQA [302, 94], capturing gaze patterns aligned with complex visual reasoning processes. OSIE [105] enriches our evaluation with eye-tracking data from free-viewing scenarios, ensuring a comprehensive assessment of our model's predictive capabilities amidst multiple salient objects. COCO-Search18 [60] expands our evaluation to include both target-present and target-absent scenarios. The target-present subset focuses on gaze behavior when the search target is present, while the target-absent subset assesses our model's ability to predict gaze patterns without the target.

Compared Models. We compare GazeXplain against human ground truths and a diverse range of scanpath prediction models, including SaltiNet [57], PathGAN [56], IOR-ROI [59], ChenLSTM [50], Gazeformer [144].

Evaluation Metrics. We comprehensively evaluate GazeXplain using a diverse set of metrics evaluating three aspects of models: First, with established metrics, including ScanMatch (SM) [55], MultiMatch (MM) [97], SED [99, 98, 100], SS [60, 146, 144] and SemSS [146, 144], we assess scanpath models' ability to accurately predict the temporal dynamics of gaze patterns. In addition, we aggregate the sampled fixations into a smoothed saliency map [149], and incorporate saliency metrics, including CC [138, 133], NSS [138, 133], AUC [257], and sAUC [257], to assess the spatial accuracy of the prediction. Finally, to measure the linguistic quality of the generated textual explanations,

we adopt BLEU [220], METEOR [221], ROUGE [222] and CIDEr-R [223, 303]. This comprehensive suite of metrics allows us to assess how well the model captures the temporal, spatial, and semantic accuracies in the fixations and explanations.

6.4.2 Scanpath Prediction Results

GazeXplain demonstrates remarkable spatiotemporal accuracy in scanpath prediction, consistently surpassing state-of-the-art methods across various datasets. As shown in table 6.2, GazeXplain's promising performance in **scanpath metrics** suggests its excellence in capturing spatial, temporal, and semantic aspects of human gaze behavior. In addition, its dominance in **saliency metrics** also indicates its ability to highlight visually important image regions. These comprehensive results suggest that GazeXplain effectively captures the underlying patterns of human visual attention across diverse tasks and datasets, demonstrating its robustness and generalizability. The performance improvements suggest the significant role of integrated attention-language decoder, semantic alignment mechanism, and cross-dataset co-training strategy in characterizing human attention dynamics, particularly in tasks requiring semantic-level cognitive processing.

6.4.3 Ablation Study for Scanpath Prediction and Explanation

Our GazeXplain features three key components: the language decoder for scanpath explanations (EXP), the semantic alignment mechanism (ALN), and the cross-dataset co-training (CT). The ablation study conducted on the AiR-D dataset, as shown in table 6.3, reveals the role of each component and their joint impacts on the accuracy of scanpath prediction and explanation. To evaluate the linguistic quality of a baseline, we directly crop fixated image regions and describe them with a pre-trained BLIP captioner [293].

Language Decoder. table 6.3 presents notable improvements achieved by integrating the language decoder into the model architecture. Even in the absence of semantic alignment, GazeXplain exhibits considerable improvements in scanpath prediction accuracy by explaining the scanpath. For instance, the inclusion of fixation-based explanations elevates the SM score from 0.356 to 0.378, which emphasizes the role of semantic

Dataset	Method	Scanpath					Saliency			
		SM ↑	MM ↑	SED ↓	SS ↑	SemSS ↑	CC ↑	NSS ↑	AUC ↑	sAUC ↑
AiR-D [1]	Human	0.403	0.803	8.110	0.336	-	0.830	2.328	0.879	0.797
	SaltiNet	0.106	0.750	10.749	0.117	-	-0.014	-0.021	0.506	0.502
	PathGAN	0.151	0.733	9.407	0.079	-	0.134	0.280	0.558	0.503
	IOR-ROI	0.209	0.795	8.883	0.176	-	0.342	0.743	0.708	0.571
	ChenLSTM	0.350	0.808	7.881	0.283	-	0.629	1.727	0.806	0.702
	Gazeformer	0.357	0.811	7.962	0.287	-	0.550	1.512	0.760	0.670
	GazeXplain	**0.386**	**0.817**	**7.489**	**0.308**	-	**0.662**	**1.851**	**0.808**	**0.719**
OSIE [105]	Human	0.386	0.808	7.481	0.332	-	0.903	2.976	0.912	0.867
	SaltiNet	0.149	0.745	8.768	0.166	-	0.230	0.556	0.659	0.596
	PathGAN	0.056	0.744	9.392	0.135	-	-0.091	-0.199	0.448	0.494
	IOR-ROI	0.290	0.790	7.826	0.232	-	0.499	1.426	0.776	0.673
	ChenLSTM	0.377	0.805	7.244	0.316	-	0.722	2.488	0.813	0.770
	Gazeformer	0.372	0.805	7.298	0.315	-	0.685	2.308	0.793	0.739
	GazeXplain	**0.380**	**0.806**	**7.228**	**0.317**	-	**0.748**	**2.530**	**0.839**	**0.786**
COCO-Search18 Target-Present [60]	Human	0.427	0.810	1.957	0.510	0.401	0.861	3.675	0.944	0.836
	SaltiNet	0.127	0.715	3.827	0.269	0.205	0.425	1.923	0.680	0.578
	PathGAN	0.213	0.716	2.461	0.318	0.268	0.377	1.465	0.720	0.591
	IOR-ROI	0.137	0.770	6.990	0.198	0.162	0.301	0.836	0.748	0.569
	ChenLSTM	0.448	0.803	**1.932**	0.475	0.406	0.802	3.376	0.903	0.814
	Gazeformer	0.433	0.800	2.224	0.470	0.394	0.712	2.990	0.872	0.785
	GazeXplain	**0.480**	**0.807**	1.981	**0.541**	**0.443**	**0.809**	**3.529**	**0.915**	**0.836**
COCO-Search18 Target-Absent [60]	Human	0.330	0.802	5.539	0.353	0.341	0.800	2.351	0.872	0.765
	ChenLSTM	0.366	0.810	4.345	0.371	0.359	0.701	2.036	0.796	0.703
	Gazeformer	0.354	0.812	4.492	0.366	0.353	0.632	1.837	0.774	0.681
	GazeXplain	**0.373**	**0.813**	**4.307**	**0.382**	**0.365**	**0.716**	**2.089**	**0.811**	**0.721**

Table 6.2: Scanpath prediction results. The best results are highlighted in bold.

Method			Scanpath				Saliency				CIDEr-R ↑
EXP	ALN	CT	SM ↑	MM ↑	SED ↓	SS ↑	CC ↑	NSS ↑	AUC ↑	sAUC ↑	
			0.337	0.805	8.197	0.274	0.582	1.582	0.794	0.693	61.9
✓			0.339	0.805	8.216	0.280	0.614	1.674	0.806	0.706	91.9
✓	✓		0.346	0.806	8.250	0.284	0.631	1.733	0.807	0.713	115.1
		✓	0.356	0.812	7.834	0.292	0.582	1.597	0.781	0.688	66.7
✓		✓	0.378	**0.819**	7.693	0.299	0.647	1.797	0.806	0.713	97.3
✓	✓	✓	**0.386**	0.817	**7.489**	**0.308**	**0.662**	**1.851**	**0.808**	**0.719**	**123.1**

Table 6.3: Ablation study on AiR-D [1] for the proposed technical components: language decoder (EXP), semantic alignment (ALN), and cross-dataset co-training (CT). The best results are highlighted in bold.

comprehension in fostering precise and interpretable scanpath predictions. Compared

to the off-the-shelf BLIP captioner used in the baseline, the CIDEr-R score is improved from 66.7 to 97.3, demonstrating the effects of our model design and training on individual datasets. These results suggest that by providing explanations for individual fixations, the model gains deeper insights into the underlying visual semantics, thereby refining its predictive capabilities.

Semantic Alignment. The semantic alignment mechanism further improves the model's accuracy in identifying fixated visual semantics and generating coherent descriptions. Aligning the semantics of fixations with their corresponding explanations not only improves the precision of explanations, as observed in the improved CIDEr-R scores from 97.3 to 123.1, but also guides the model to produce more accurate fixations, reflected in the scanpath and saliency metrics (*e.g.*, SM from 0.378 to 0.386, CC from 0.647 to 0.662). This indicates the importance of semantic coherence in guiding attention prediction models.

Cross-Dataset Co-Training. Scanpath prediction research typically tackles individual tasks in isolation, each relying on its own dataset. However, our approach diverges by training a unified model across multiple datasets, harnessing shared knowledge and contemporary features to enhance performance. By leveraging diverse data sources, our model achieves notable improvements in performance across various datasets. For instance, we observe a substantial enhancement in the SM score (from 0.346 to 0.386) as well as CIDEr-R (from 115.1 to 123.1) This demonstrates the effectiveness of integrating diverse data sources for robust scanpath prediction and explanation.

6.4.4 Scanpath Explanation Results

We evaluate GazeXplain's explanatory capabilities through three main analyses: (1) assessing agreement with ground-truth annotations using language evaluation metrics, (2) analyzing the diversity and informativeness of explanations, and (3) examining its ability to accurately describe fixated objects.

Language Evaluation. table 6.4 comprehensively evaluates the agreement between generated explanations and ground-truth annotations with language metrics. GazeXplain consistently outperforms its variants (without alignment, without co-training, or both) across all datasets. The semantic alignment mechanism results in consistent performance gains across datasets (*e.g.*, BLEU-4 from 27.7 to 30.7 and CIDEr-R from 97.3

Dataset	Method	B-4	M	R	C-R	Voc.	Len.	UnP%
AiR-D [1]	w/o CT & ALN	27.6	20.5	50.1	91.9	557	100.8	30.92
	w/o CT	30.4	21.7	51.6	115.1	668	100.4	39.51
	w/o ALN	27.7	20.6	50.3	97.3	541	91.8	35.74
	GazeXplain	**30.7**	**21.9**	**51.7**	**123.1**	579	88.3	40.34
OSIE [105]	w/o CT & ALN	12.4	16.5	40.2	23.6	633	103.4	42.08
	w/o CT	16.1	17.4	41.7	37.4	760	105.9	44.20
	w/o ALN	15.7	20.4	41.7	37.2	569	94.4	42.17
	GazeXplain	**16.7**	**21.1**	**42.0**	**48.6**	614	90.9	44.76
COCO-Search18 TP [60]	w/o CT & ALN	23.3	15.4	52.4	111.2	304	27.3	64.67
	w/o CT	26.0	16.2	54.2	133.2	401	26.0	70.41
	w/o ALN	26.8	18.1	54.5	130.9	505	28.0	68.83
	GazeXplain	**28.2**	**19.5**	**55.3**	**139.6**	560	28.4	71.30
COCO-Search18 TA [60]	w/o CT & ALN	15.6	20.9	43.2	77.0	514	35.8	58.35
	w/o CT	17.2	22.5	43.8	91.9	583	35.9	67.03
	w/o ALN	16.3	26.4	43.2	92.9	566	33.3	66.04
	GazeXplain	**18.5**	**27.5**	**44.5**	**106.5**	685	35.5	71.35

Table 6.4: Explanation prediction results and diversity analysis. The best results are highlighted in bold.

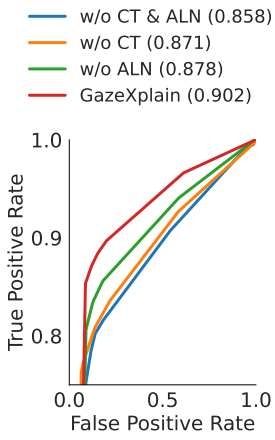

Figure 6.4: ROC analysis of fixations and explanations.

to 123.1 on AiR-D), suggesting its significance in generating natural and fluent explanations. The co-training is more effective on OSIE (free-viewing) and COCO-Search18 (target-absent) datasets involving less structured exploration compared to the other datasets where specific objects need to be identified. It allows the model to exploit the combined information from all available data sources to learn diverse visual and linguistic relationships under these more challenging scenarios (*e.g.*, CIDEr-R is 48.6 on OSIE, compared to the 139.6 on COCO-Search18 target-present dataset).

Diversity. To assess explanation diversity with three metrics: vocabulary size (Voc.), total explanation length per scanpath (Len.), and the percentage of unique sentences per scanpath (UnP%). Table 6.4 reveals that incorporating semantic alignment significantly improved both vocabulary size and UnP%. For example, on the COCO-Search18 (target-absent) dataset, vocabulary size increased from 566 to 685 words, and UnP% increased from 66.04% to 71.35%. Notably, this improvement in diversity occurred while maintaining consistent explanation lengths. The COCO-Search18 dataset, known for its shorter scanpaths, naturally yielded a smaller vocabulary size, shorter explanations, and a higher percentage of unique sentences. Our co-training method, while consistently boosting UnP%, also helped balance vocabulary sizes and explanation lengths across datasets. These findings highlight the importance of semantic

alignment and co-training in promoting both diverse and specific explanations.

Faithfulness. We evaluate the faithfulness of explanations in describing the search targets of the COCO-Search18 dataset. Specifically, we examine whether the explanation describes the search target when it is fixated on, and refrain from falsely describing it when fixations are elsewhere. To achieve this, we employ two key metrics: fixation proximity to the search target, quantified as the distance between fixations and the bounding box of the target, and semantic similarity between the generated explanation and the target, computed as the cosine distance between their embeddings using state-of-the-art techniques such as SBERT [304]. By varying spatial and semantic distance thresholds, we construct ROC curves and calculate the area under the curve (AUC) as a performance metric. Our findings, shown in figure 6.4, indicate that both semantic alignment and co-training lead to improved agreements between explanations and fixations, with AUC values increasing from 0.878 to 0.902 and 0.871 to 0.902, respectively. It suggests the significance of these techniques in aligning explanations with fixated objects.

6.4.5 Qualitative Analysis

figure 6.5 presents qualitative examples of GazeXplain's scanpaths and explanations. For the Gazeformer model, we directly crop fixated image regions and describe them with a pre-trained BLIP captioner [293]. For illustration, we select two explanations describing task-relevant fixations. We observe GazeXplain's enhanced capability in predicting fixations on key objects crucial for answering questions, mirroring human gaze behavior during high-level cognitive processing. For instance, in figure 6.5a, GazeXplain accurately identifies the cake on the left side. Similarly, in figure 6.5b, the model focuses on the dog, while in figure 6.5c, it prioritizes the trash can. This alignment with human scanpaths demonstrates GazeXplain's capability of characterizing complex gaze patterns associated with cognitive tasks. Regarding explanations, while Gazeformer wrongly describes its fixations (*e.g.*, figure 6.5a: "a plate of food with a fork and knife" while there is no fork or knife present, figure 6.5b: "a man riding a horse" while the man is walking, and figure 6.5c: "a red wall with a black chair and a black chair" while the chair is not black), GazeXplain provides more accurate and specific fixation descriptions. Particularly in scenes with multiple relevant objects (*e.g.*, different types

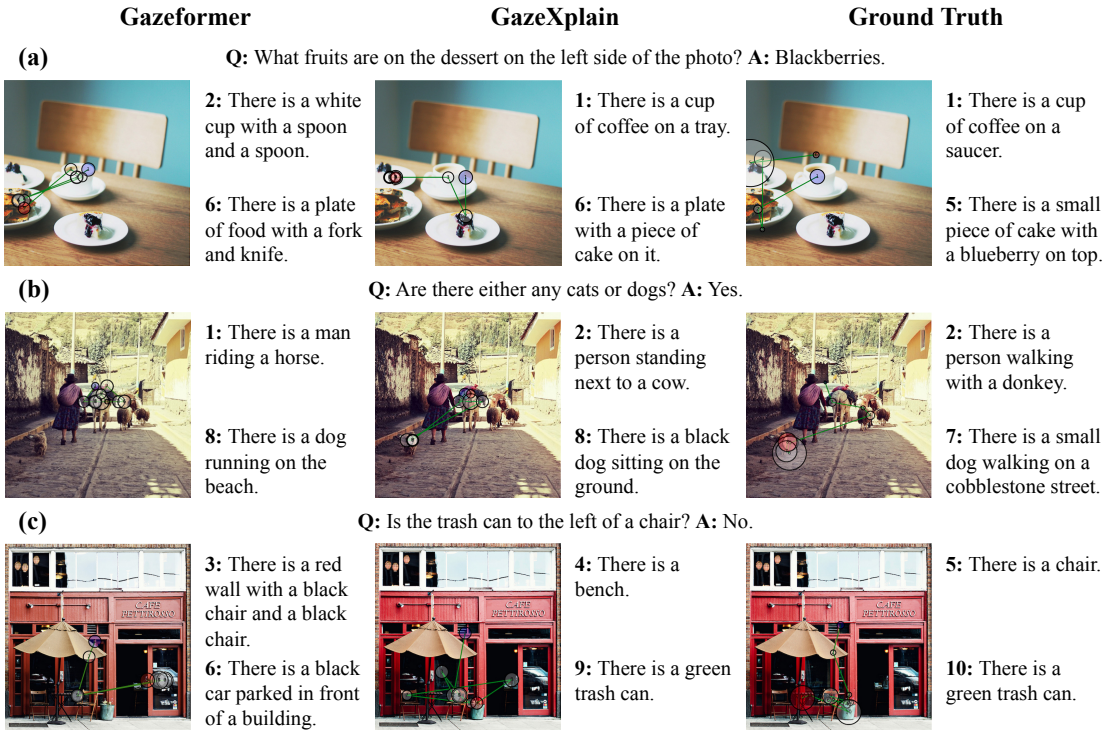

Figure 6.5: Quantitative examples from GazeXplain compared to Gazeformer and the ground truth. Each row shows scanpaths and explanations of two key fixations.

of desserts and animals in figure 6.5a-b), GazeXplain successfully distinguishes them, demonstrating robust semantic understanding. These examples illustrate GazeXplain's effectiveness in integrating visual exploration with semantic understanding, yielding more explainable and robust scanpath predictions.

6.5 Conclusion

We introduce GazeXplain, a novel scanpath explanation task to understand human visual attention. We provide ground-truth explanations on various eye-tracking datasets and develop a model architecture for predicting scanpaths and generating natural language explanations. The model features an attention-language decoder with a unique semantic alignment mechanism ensuring fixation-explanation consistency. Additionally,

our proposed cross-dataset co-training approach enhances generalizability by leveraging diverse training datasets. Extensive experiments demonstrate GazeXplain's superior performance in both scanpath prediction and explanation, suggesting not only scanpath modeling benefits from language explanations but also GazeXplain's explanations can be integrated with other language-driven user environments. We anticipate that GazeXplain will catalyze the development of interpretable attention models, fostering advancements in human visual behavior understanding.

Chapter 7

Conclusion and Discussion

In this dissertation, we summarize our prior and ongoing efforts in modeling human visual attention, aligning machine attention with human attention, and developing an explainable visual intelligence system. By treating visual attention as a key interface connecting perception, reasoning, and performance, we explored methods to address complex real-life tasks and uncover the rationale behind human attention deployment.

Our findings demonstrate that human attention plays an important role in guiding machines to pay attention to relevant information, which enhances their ability to perform general tasks and quantify reasoning capabilities. With a profound understanding of visual attention, we curated an open-ended dataset to study real-life problems and their evolution by incorporating key components across multiple modalities, focusing on attention deployment and complex step dependency relationships. Results show that current machine attention significantly differs from human attention, since machine attention relies heavily on top-down statistical correlations rather than reasoning based on justifiable evidence. This limitation emphasizes the challenges faced by existing vision and language models in generalizing to open-ended tasks and explaining the rationale behind their decisions.

These observations provide a principled framework for understanding how humans dynamically integrate visual stimuli and reasoning processes to make decisions while also justifying the rationale behind their visual fixations. Bringing perception and explanation together gives us a practical starting point for designing the next generation of visual intelligence systems. Such systems should be able to interpret human gaze

patterns and shed light on the reasoning that drives them. Such systems have the possibility to address the complexity of real-world environments and produce significant societal benefits in areas such as healthcare, education, and autonomous systems.

While our studies present significant progress in bridging the gap between human and machine intelligence, there is still room for further improvement. Most notably, our reliance on fine-grained annotated data highlights the need for models that can generalize beyond specific datasets. Recent developments in foundational models, such as GPT [32, 33] and DINO [305, 306, 307], suggest promising avenues for distilling world knowledge and reasoning across modalities, while also offering practical tools for generating fine-grained annotations. The next generation of artificial general intelligence lies in developing interpretable and explainable systems that evolve from black-box statistical models to white-box frameworks capable of genuine reasoning with multimodal inputs. In the future, I plan to advance human-like machine intelligence by curating high-quality multimodal data and leveraging foundational models for impactful applications beneficial for human beings, with a particular emphasis on healthcare. By overcoming the challenges of AI deployment and placing transparency at the forefront, we can enable transformative applications that fully realize the societal benefits of AI innovations.

Chapter 8

List of All Publications

Below, I have included a complete list of publications where I served as the lead or co-first author during my doctoral studies.

Journal

1. **Xianyu Chen**, Jinhui Yang, Shi Chen, Louis Wang, Ming Jiang, and Qi Zhao, Every Problem, Every Step, All In Focus: Learning to Solve Real-World Problems with Integrated Attention. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024.

Conference

1. **Xianyu Chen**, Ming Jiang and Qi Zhao. GazeXplain: Learning to Predict Natural Language Explanations of Visual Scanpaths. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. (**Oral Paper, 2.3% acceptance rate**)

2. **Xianyu Chen**, Ming Jiang and Qi Zhao. Beyond Average: Individualized Visual Scanpath Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

3. Jinhui Yang*, **Xianyu Chen***, Ming Jiang, Shi Chen, Louis Wang and Qi Zhao.

VisualHow: Multimodal Problem Solving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. (*Co-first authors/Equal contribution)

4. **Xianyu Chen**, Ming Jiang and Qi Zhao. Leveraging Human Attention in Novel Object Captioning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

5. **Xianyu Chen**, Ming Jiang and Qi Zhao. Predicting Human Scanpaths in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

6. **Xianyu Chen**, Ming Jiang and Qi Zhao. Self-Distillation for Few-Shot Image Captioning. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

Preprint

1. **Xianyu Chen**, Ming Jiang and Qi Zhao. Leveraging Bottom-Up and Top-Down Attention for Few-Shot Object Detection. In *arXiv preprint arXiv:2007.12104*, 2020.

References

[1] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. AiR: Attention with reasoning capability. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[2] Shuo Wang, Ming Jiang, Xavier Morin, Duchesne, Elizabeth A. Laugeson, Daniel P. Kennedy, Ralph Adolphs, and Qi Zhao. Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron*, 2015.

[3] Alex Krizhevsky, Ilya Sutskever, , and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2012.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2018.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015.

[8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2017.

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

[11] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[12] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 2016.

[13] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai andArthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 2018.

[14] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 2019.

[15] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[17] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[18] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[19] Llama Team. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[20] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[21] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue,

Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[24] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[26] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

[27] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[28] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack

Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[29] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. rounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[30] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. GLIPv2: Unifying localization and vision-language understanding. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[31] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. Detecting everything in the open world: Towards universal object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[32] Tianyang Zhong, Zhengliang Liu, Yi Pan, Yutong Zhang, Yifan Zhou, Shizhe Liang, Zihao Wu, Yanjun Lyu, Peng Shu, Xiaowei Yu, Chao Cao, Hanqi Jiang, Hanxu Chen, Yiwei Li, Junhao Chen, Huawen Hu, Yihen Liu, Huaqin Zhao, Shaochen Xu, Haixing Dai, Lin Zhao, Ruidong Zhang, Wei Zhao, Zhenyuan Yang, Jingyuan Chen, Peilong Wang, Wei Ruan, Hui Wang, Huan Zhao, Jing Zhang, Yiming Ren, Shihuan Qin, Tong Chen, Jiaxi Li, Arif Hassan Zidan, Afrar Jahin, Minheng Chen, Sichen Xia, Jason Holmes, Yan Zhuang, Jiaqi Wang, Bochen Xu, Weiran Xia, Jichao Yu, Kaibo Tang, Yaxuan Yang, Bolun Sun, Tao Yang, Guoyu Lu, Xianqiao Wang, Lilong Chai, He Li, Jin Lu, Lichao Sun, Xin Zhang, Bao Ge, Xintao Hu, Lian Zhang, Hua Zhou, Lu Zhang, Shu Zhang, Ninghao Liu, Bei Jiang, Linglong Kong, Zhen Xiang, Yudan Ren, Jun Liu, Xi Jiang, Yu Bao, Wei Zhang, Xiang Li, Gang Li, Wei Liu, Dinggang Shen, Andrea Sikora, Xiaoming Zhai, Dajiang Zhu, and Tianming Liu. Evaluation of openai o1: Opportunities and challenges of agi. *arXiv preprint arXiv:2409.18486*, 2024.

[33] OpenAI. Introducing GPT-5. `https://openai.com/index/introducing-gpt-5/`.

[34] Zhenjie Yang, Xiaosong Jiaand Hongyang Li, and Junchi Yan. Llm4drive: A survey of large language models for autonomous driving. *arXiv preprint arXiv:2409.18486*, 2024.

[35] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2024.

[36] Sirvan Khalighi, Kartik Reddy, Abhishek Midya, Krunal Balvantbhai Pandav, Anant Madabhushi, and Malak Abedalthagafi. Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment. *npj Precision Oncology*, 2024.

[37] Mohammad Amin Kuhaila, Nazik Alturkib, Justin Thomasc, Amal K. Alkhalifad, and Amal Alshardan. Human-human vs human-ai therapy: An empirical study. *International Journal of Human–Computer Interaction*, 2024.

[38] Anne M Treisman and Nancy G Kanwisherf. Perceiving visually presented objects: recognition, awareness, and modularity. *Current Opinion in Neurobiology*, 1998.

[39] J. E. (Hans). Korteling, G. C. van de Boer-Visschedijk, R. A. M. Blankendaal, R. C. Boonekamp, and A. R. Eikelboom. Human- versus artificial intelligence. *Frontiers in Artificial Intelligence*, 2021.

[40] Nils J. Nilsson. Human-level artificial intelligence? be serious! *AI Magazine*, 2005.

[41] Hyoung Seok Shin. Reasoning processes in clinical reasoning: from the perspective of cognitive psychology. *Korean Journal of Medical Education*, 2019.

[42] Hans Rudi Fischer. Abductive reasoning as a way of worldmaking. *Foundations of Science*, 2001.

[43] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On Pearl's hierarchy and the foundations of causal inference. In *Technique Report*, 2020.

[44] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

[45] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *arXiv preprint arXiv:2102.11107v1*, 2021.

[46] Anthony M. Zador. A critique of pure learning: What artificial neural networks can learn from animal brains. *Nature Communications*, 2019.

[47] Robert Geirhos, Carlos R. Medina Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in humans and deep neural networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[48] Nicholas T. Franklin and Michael J. Frank. Generalizing to generalize: Humans flexibly switch between compositional and conjunctive structures during reinforcement learning. *PLoS Computational Biology*, 2020.

[49] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[50] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[51] Xianyu Chen, Ming Jiang, and Qi Zhao. Beyond average: Individualized visual scanpath prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[52] Jinhui Yang, Xianyu Chen, Ming Jiang, Shi Chen, Louis Wang, and Qi Zhao. VisualHow: Multimodal problem solving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[53] Xianyu Chen, Jinhui Yang, Shi Chen, Louis Wang, Ming Jiang, and Qi Zhao. Every problem, every step, all in focus: Learning to solve real-world problems with integrated attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2024.

[54] Xianyu Chen, Ming Jiang, and Qi Zhao. GazeXplain: Learning to predict natural language explanations of visual scanpaths. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[55] Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D Gilchrist. ScanMatch: A novel method for comparing fixation sequences. *Behavior Research Methods (BRM)*, 2010.

[56] Marc Assens, Xavier Giro-i-Nieto, Kevin McGuinness, and Noel E. O'Connor. PathGAN: Visual scanpath prediction with generative adversarial networks. In *Proceedings of the European Conference on Computer Vision Workshop (ECCVW)*, 2018.

[57] Marc Assens, Kevin McGuinness, Xavier Giro-i-Nieto, and Noel E. O'Connor. SaltiNet: Scan-path prediction on 360 degree images using saliency volumes. In *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2017.

[58] Zhenzhong Chen and Wanjie Sun. Scanpath prediction for visual attention using IOR-ROI LSTM. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[59] Wanjie Sun, Zhenzhong Chen, and Feng Wu. Visual scanpath prediction using IOR-ROI recurrent mixture density network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2019.

[60] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[61] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2013.

[62] Dirk Brockmann and Theo Geisel. The ecology of gaze shifts. *Neurocomputing*, 2000.

[63] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research (VR)*, 2000.

[64] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience (NRN)*, 2001.

[65] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 1998.

[66] Dirk Walther and Christof Koch. Modeling attention to salient proto-objects. *Neural Networks (NN)*, 2006.

[67] Giuseppe Boccignonea and Mario Ferraro. Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications (PHYSICA A)*, 2004.

[68] Xiaodi Hou and Liqing Zhang. Dynamic visual attention: searching for coding length increments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2008.

[69] Huiying Liu, Dong Xu, Qingming Huang, Wen Li, Min Xu, and Stephen Lin. Semantically-based human scanpath estimation with HMMs. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

[70] Olivier Le Meur and Zhi Liu. Saccadic model of eye movements for free-viewing condition. *Vision Research (VR)*, 2015.

[71] Xiaoshuai Sun, Hongxun Yao, Rongrong Ji, and Xian-Ming Liu. Toward statistical modeling of saccadic eye-movement and visual saliency. *IEEE Transactions on Image Processing (IEEE TIP)*, 2014.

[72] Hamed Rezazadegan Tavakoli, Esa Rahtu, and Janne Heikkilä. Stochastic bottom-up fixation prediction and saccade generation. *Image Vision Computing (IVC)*, 2013.

[73] Wei Wang, Cheng Chen, Yizhou Wang, Tingting Jiang, Fang Fang, and Yuan Yao. Simulating human saccadic scanpaths on natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[74] Yixiu Wang, Bin Wang, Xiaofeng Wu, and Liming Zhang. Scanpath estimation based on foveated image saliency. *Cognitive Processing (CP)*, 2017.

[75] Calden Wloka, Iuliia Kotseruba, and John K. Tsotsos. Active fixation control to predict saccade sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[76] Ming Jiang, Shi Chen, Jinhui Yang, and Qi Zhao. Fantastic answers and where to find them: Immersive question-directed visual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[77] Hamed R. Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. Paying attention to descriptions generated by image captioning models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[78] Ming Jiang, Xavier Boix, Gemma Roig, Juan Xu, Luc Van Gool, and Qi Zhao. Learning to predict sequences of human visual fixations. *IEEE Transactions on Neural Networks and Learning Systems (IEEE TNNLS)*, 2016.

[79] Silviu Minut and Sridhar Mahadevan. A reinforcement learning model of selective visual attention. In *Proceedings of the International Conference on Autonomous Agents (AGENTS)*, 2001.

[80] Dimitri Ognibene, Christian Balkenius, and Gianluca Baldassarre. A reinforcement-learning model of top-down attention based on a potential-action map. *The Challenge of Anticipation Anticipatory Approach (CAAA)*, 2008.

[81] Dimitri P. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific, Belmont, Massachusetts, USA, 1st edition, 2019.

[82] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, Massachusetts, USA, 2nd edition, 2018.

[83] Stefan Mathe and Cristian Sminchisescu. Action from still image dataset and inverse optimal control to learn task specific visual scanpaths. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[84] Mai Xu, Yuhang Song, Jianyi Wang, Minglang Qiao, Liangyu Huo, and Zulin Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2019.

[85] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2004.

[86] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *arXiv preprint arXiv:1806.06877*, 2018.

[87] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[88] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[89] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[90] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. Exploring human-like attention supervision in visual question answering. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[91] Gary Feng. Eye movements as time-series random variables: A stochastic model of eye movement control in reading. *Cognitive Systems Research (CSR)*, 2006.

[92] Arthur J. Lugtigheid. Distributions of fixation durations and visual acquisition rates. *Ph.D. dissertation*, 2007.

[93] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning (ML)*, 1992.

[94] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[95] Hiroyuki Sogo. Gazeparser: an open-source and multiplatform library for low-cost eye tracking and analysis. *Behavior Reserch Methods (BRM)*, 2013.

[96] Saul B.Needleman and Christian D.Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology (JMB)*, 1970.

[97] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior Research Methods (BRM)*, 2012.

[98] Stephan A. Brandt and Lawrence W. Stark. Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience (JCN)*, 1997.

[99] Tom Foulsham and Geoffrey Underwood. What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision (JoV)*, 2008.

[100] Lapo Faggi, Alessandro Betti, Dario Zanca, Stefano Melacci, and Marco Gori. Wave propagation of visual stimuli in focus of attention. *arXiv preprint arXiv:2006.11035*, 2020.

[101] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[102] Zhi Zhang, Tong He, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of freebies for training object detection neural networks. *arXiv preprint arXiv:1902.04103*, 2019.

[103] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[104] Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[105] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision (JoV)*, 2014.

[106] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[107] Ke Gu, Shiqi Wang, Huan Yang, Weisi Lin, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Saliency-guided quality assessment of screen content images. *IEEE Transactions on Multimedia (TMM)*, 2016.

[108] Patrick Le Callet and Ernst Niebur. Visual attention and applications in multimedia technologies. *Proceedings of the Institution of Electrical Engineers*, 2013.

[109] Leida Li, Yu Zhou, Weisi Lin, Jinjian Wu, Xinfeng Zhang, and Beijing Chen. No-reference quality assessment of deblocked images. *Neurocomputing*, 2016.

[110] Tommy Strandvall. Eye tracking in human-computer interaction and usability research. In *IFIP Conference on Human-Computer Interaction*, 2009.

[111] Thomas E. Hutchinson, K. Preston White, Worthy N. Martin, Kelly C. Reichert, and Lisa A. Frey. Human-computer interaction using eye-gaze input. *IEEE Transactions on Systems, Man, and Cybernetics (TSMC)*, 1989.

[112] Uchenna Chinyere Onyemauche, Samuel Makuochi Nkwo, and Charity Elochukwu Mbanusi. Towards the use of eye gaze tracking technology: Human computer interaction (hci) research. In *African Human-Computer Interaction Conference: Inclusiveness and Empowerment*, 2021.

[113] Anjana Sharma and Pawanesh Abrol. Eye gaze techniques for human computer interaction: A research survey. *International Journal of Computer Applications*, 2013.

[114] Yue Jiang, Luis A. Leiva, Hamed R. Tavakoli, Paul R. B. Houssel, Julia Kylmälä, and Antti Oulasvirta. UEyes: Understanding visual saliency across user interface types. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2023.

[115] Thammathip Piumsomboon, Gun Lee, Robert W. Lindeman, and Mark Billinghurst. Exploring natural eye-gaze-based interaction for immersive virtual reality. In *IEEE Symposium on 3D User Interfaces (3DUI)*, 2017.

[116] Isayas Berhe Adhanom, Paul MacNeilage, and Eelke Folmer. Eye gaze techniques for human computer interaction: A research survey. *Virtual Reality*, 2023.

[117] Kun Qian, Tomoki Arichi, Anthony Price, Sofia Dall'Orso, Jonathan Eden, Yohan Noh, Kawal Rhode, Etienne Burdet, Mark Neil, A. David Edwards, and Joseph V. Hajnal. An eye tracking based virtual reality system for use inside magnetic resonance imaging systems. *Scientific Reports*, 2021.

[118] Xinyue Gui, Koki Toda, Stela Hanbyeol Seo, Chia-Ming Chang, and Takeo Igarashi. "I am going this way": Gazing eyes on self-driving car show multiple driving directions. In *International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2022.

[119] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. Predicting driver attention in critical situations. In *Asian Conference on Computer Vision (ACCV)*, 2018.

[120] Ye Xia, Jinkyu Kim, John Canny, Karl Zipser, Teresa Canas-Bajo, and David Whitney. Periphery-fovea multi-resolution driving model guided by human attention. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[121] Ming Jiang, Sunday M Francis, Angela Tseng, Diksha Srishyla, Megan DuBois, Katie Beard, Christine Conelea, Qi Zhao, and Suma Jacob. Predicting core characteristics of asd through facial emotion recognition and eye tracking in youth. In *International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020.

[122] Shi Chen and Qi Zhao. Attention-based autism spectrum disorder screening with privileged modality. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[123] Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Zhaohui Che, Yi Fang, Xiaokang Yang, Jesús Gutiérrez, and Patrick Le Callet. A dataset of eye movements for the children with autism spectrum disorder. In *ACM Multimedia Systems Conference (MMSys)*, 2019.

[124] Evan F. Risko, Nicola C. Anderson, Sophie Lanthier, and Alan Kingstone. Curious eyes: Individual differences in personality predict eye movement behavior in scene-viewing. *Cognition*, 2012.

[125] Matthew F. Peterson and Miguel P. Eckstein. Individual differences in eye movements during face identification reflect observer-specific optimal points of fixation. *Psychological Science*, 2013.

[126] Bahman Abdi Sargezeh, Niloofar Tavakoli, and ohammad Reza Daliri. Gender-based eye movement differences in passive indoor picture viewing: An eye-tracking study. *Physiology & Behavior*, 2019.

[127] Felix Joseph Mercer Moss, Roland Baddeley, and Nishan Canagarajah. Eye movements to natural images as a function of sex and personality. *PLoS One*, 2012.

[128] Negar Sammaknejad, Hamidreza Pouretemad, Changiz Eslahchi, Alireza Salahirad, and Ashkan Alinejad. Gender classification based on eye movements: A processing effect during passive face viewing. *Advances in Cognitive Psychology*, 2017.

[129] Young Hoon Oh and Da Young Ju. Age-related differences in fixation pattern on a companion robot. *Sensors*, 2020.

[130] Ze-Yu Wang and Ji Young Cho. Older adults' response to color visibility in indoor residential environment using eye-tracking technology. *Sensors*, 2022.

[131] Mikle South, Sally Ozonoff, and William M. McMahon. Repetitive behavior profiles in asperger syndrome and high-functioning autism. *Journal of Autism and Developmental Disorders*, 2005.

[132] Mark H. Lewis and James W. Bodfish. Repetitive behavior disorders in autism. *Developmental Disabilities Research Reviews*, 1998.

[133] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. SALICON: Saliency in context. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[134] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.

[135] Ali Borji and Laurent Itti. CAT2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581v1*, 2015.

[136] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016.

[137] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing (IEEE TIP)*, 2018.

[138] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[139] Camilo Fosco, Vincent Casser, Amish Kumar Bedi, Peter O'Donovan, Aaron Hertzmann, and Zoya Bylinskii. Predicting visual importance across graphic design types. In *ACM Symposium on User Interface Software and Technology*, 2020.

[140] Souradeep Chakraborty, Zijun Wei, Conor Kelton, Seoyoung Ahn, Aruna Balasubramanian, Gregory J. Zelinsky, and Dimitris Samaras. Predicting visual attention in graphic design documents. *IEEE Transactions on Multimedia (TMM)*, 2022.

[141] Sen Jia and Neil D. B. Bruce. EML-NET:an expandable multi-layer network for saliency prediction. *Image and Vision Computing*, 2020.

[142] Shi Chen, Nachiappan Valliappan, Shaolei Shen, Xinyu Ye, Kai Kohlhoff, and Junfeng He. Learning from unique perspectives: User-aware saliency modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[143] Bahar Aydemir, Ludo Hoffstetter, Tong Zhang, Mathieu Salzmann, and Sabine Susstrunk. TempSAL - uncovering temporal information for deep saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[144] Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[145] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Predicting human attention using computational attention. *arXiv preprint arXiv:2303.09383*, 2023.

[146] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Target-absent human attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[147] Ryan Anthony Jalova de Belen, Tomasz Bednarz, and Arcot Sowmya. Scanpathnet: A recurrent mixture density network for scanpath prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2022.

[148] Mengyu Qiu, Yi Guo, Mingguang Zhang, Jingwei Zhang, Tian Lan, and Zhilin Liu. Simulating human visual system based on vision transformer. In *Proceedings of the 2023 ACM Symposium on Spatial User Interaction*, 2023.

[149] Xiangjie Sui, Yuming Fang, Hanwei Zhu, Shiqi Wang, and Zhou Wang. ScanDMM: A deep markov model of scanpath prediction for 360° images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[150] Yanyu Xu, Shenghua Gao, Junru Wu, Nianyi Li, and Jingyi Yu. Personalized saliency and its prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2018.

[151] Yanyu Xu, Nianyi Li, Junru Wu, Jingyi Yu, and Shenghua Gao. Beyond universal saliency: Personalized saliency prediction with multi-task cnn. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[152] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Few-shot personalized saliency prediction based on adaptive image selection considering object and visual attention. *IEEE International Conference on Consumer Electronics*, 2020.

[153] Aoqi Li and Zhenzhong Chen. Individual trait oriented scanpath prediction for visual attention analysis. In *IEEE International Conference on Image Processing (ICIP)*, 2017.

[154] Xinhui Luo, Zhi Liu, Weijie Wei, Linwei Ye, Tianhong Zhang, Lihua Xu, and Jijun Wang. Few-shot personalized saliency prediction using meta-learning. *Image and Vision Computing*, 2022.

[155] Matthias Kümmerer, Matthias Bethge, and Thomas S. A. Wallis. Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision (JoV)*, 2022.

[156] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[157] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[158] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[159] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[160] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Stefan Lee, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2019.

[161] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[162] Alexander Toet. Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2011.

[163] Torbjörn Falkmer, Katie Anderson, Marita Falkmer, and Chiara Horlin. Diagnostic procedures in autism spectrum disorders: a systematic literature review. *European Child & Adolescent Psychiatry*, 2013.

[164] Ming Jiang and Qi Zhao. Learning visual attention to identify people with autism spectrum disorder. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[165] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. M^2: Meshed-memory transformer for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[166] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[167] Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, 2016.

[168] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[169] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[170] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[171] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[172] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[173] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K. Wong, and Qi Wu. Cops-Ref: A new dataset and task on compositional referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[174] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[175] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325v2*, 2015.

[176] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[177] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research (JAIR)*, 2013.

[178] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2014.

[179] Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. Recursive visual attention in visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[180] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[181] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[182] Yue Wang, Shafiq Joty, Michael R. Lyu, Irwin King, Caiming Xiong, and Steven C.H. Hoi. VD-BERT: A unified vision and dialog transformer with bert. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[183] Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. Multi-modal graph neural network for joint reasoning on vision and scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[184] Michael Cogswell, Jiasen Lu, Rishabh Jain, Stefan Lee, Devi Parikh, and Dhruv Batra. Dialog without dialog data: Learning visual dialog agents from VQA data. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[185] Tao Tu, Qing Ping, Govindarajan Thattai, Gokhan Tur, and Prem Natarajan. Learning better visual dialog agents with pretrained visual-linguistic representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[186] Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[187] Khyathi Raghavi Chandu, Ruo-Ping Dong, and Alan Black. Reading between the lines: Exploring infilling in visual narratives. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[188] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. MERLOT: Multimodal neural script knowledge models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[189] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[190] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[191] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. COIN: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[192] Yansong Tang, Jiwen Lu, and Jie Zhou. Comprehensive instructional video analysis: The COIN dataset and performance evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2020.

[193] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference (BMVC)*, 2018.

[194] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[195] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[196] Alec Radford, Jong Wook Kim, Aditya Ramesh Chris Hallacy, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[197] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[198] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[199] Aisha Urooj Khan, Amir Mazaheri, Niels da Vitoria Lobo, and Mubarak Shah. MMFT-BERT: Multimodal fusion transformer with bert encodings for visual question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[200] Zhe Gan, Yu Cheng, Ahmed El Kholy, Linjie Li1, Jingjing Liu, and Jianfeng Gao. Multi-step reasoning via recurrent dual attention for visual dialog. In *Annual Conference of the Association for Computational Linguistics (ACL)*, 2019.

[201] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[202] Raeid Saqur and Karthik Narasimhan. Multimodal graph networks for compositional generalization in visual question answering. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[203] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2019.

[204] Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. DeViSE: A deep visual-semantic embedding model. 2013.

[205] Ryan Kiros and Richard S. Zemel Ruslan Salakhutdinov. Unifying visual-semantic embeddings with multimodal neural language models. *Conference on Neural Information Processing Systems Workshops (NeurIPSW)*, 2014.

[206] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[207] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[208] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[209] Irshad Bhat, Talita Anthonio, and Michael Roth. Towards modeling revision requirements in wikiHow instructions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[210] Mahnaz Koupaee and William Yang Wang. WikiHow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*, 2018.

[211] Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. Visual goal-step inference using wikiHow. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[212] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[213] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[214] Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Annual Conference of the Association for Computational Linguistics (ACL)*, 2018.

[215] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[216] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[217] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2018.

[218] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[219] Darwin Saire Pilco and Adín Ramírez Rivera. Graph learning network: A structure learning algorithm. In *Proceedings of the International Conference on Machine Learning Workshop (ICMLW)*, 2019.

[220] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Annual Conference of the Association for Computational Linguistics (ACL)*, 2002.

[221] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Annual Conference of the Association for Computational Linguistics Workshop (ACLW)*, 2005.

[222] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Annual Conference of the Association for Computational Linguistics Workshop (ACLW)*, 2004.

[223] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[224] Shi Chen and Qi Zhao. Boosted attention: Leveraging human attention for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[225] Xianyu Chen, Ming Jiang, and Qi Zhao. Leveraging human attention in novel object captioning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[226] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2023.

[227] Tengpeng Li, Hanli Wang, Bin He, and Chang Wen Chen. Knowledge-enriched attention network with group-wise semantic for visual storytelling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2023.

[228] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. Attention in reasoning: Dataset, analysis, and modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2021.

[229] Long Chen, Yuhang Zheng, Yulei Niu, Hanwang Zhang, and Jun Xiao. Counterfactual samples synthesizing and training for robust visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2023.

[230] Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Deyu Meng, Yue Gao, and Chunhua Shen. Plenty is plague: Fine-grained learning for visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2022.

[231] Dan Guo, Hui Wang, and Meng Wang. Context-aware graph inference with knowledge distillation for visual dialog. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2022.

[232] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv preprint arXiv:2202.06935*, 2022.

[233] Shereen Oraby, Vrindavan Harrison, Abteen Ebrahimi, and Marilyn Walker. Curate and generate: A corpus and method for joint control of semantics and style in neural NLG. In *Annual Conference of the Association for Computational Linguistics (ACL)*, 2019.

[234] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[235] Antoine Miech, Dimitri Zhukov, and Jean-Baptiste Alayrac. The language of actions: Recovering the syntax and semantics of goal-directed human activities.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[236] Honglu Zhou, Roberto Martín-Martín, Mubbasir Kapadia, Silvio Savarese, and Juan Carlos Niebles. Procedure-aware pretraining for instructional video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[237] Hanlin Wang, Yilu Wu, Sheng Guo, and Limin Wang. PDPP: Projected diffusion for procedure planning in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[238] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oğuz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[239] Fadime Sener, Rishabh Saraf, and Angela Yao. Transferring knowledge from text to video: Zero-shot anticipation for procedural actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2023.

[240] Nikita Dvornik, Isma Hadji, Hai Pham, Dhaivat Bhatt, Brais Martinez, Afsaneh Fazly, and Allan D. Jepson. Graph2Vid: Flow graph to video grounding for weakly-supervised multi-step localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[241] Jiankai Sun, De-An Huang, Bo Lu, Yun-Hui Liu, Bolei Zhou, and Animesh Garg. PlaTe: Visually-grounded planning with transformers in procedural tasks. *IEEE Robotics and Automation Letters*, 2021.

[242] He Zhao, Isma Hadji, Nikita Dvornik, Konstantinos G. Derpanis, Richard P. Wildes, and Allan D. Jepson. P^3IV: Probabilistic procedure planning from instructional videos with weak supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[243] Zhiheng Li, Wenjia Geng, Muheng Li, Lei Chen, Yansong Tang, Jiwen Lu, and Jie Zhou. Skip-Plan: Procedure planning in instructional videos via condensed

action space learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

[244] Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a HINT: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[245] Jialin Wu and Raymond Mooney. Self-critical reasoning for robust visual question answering. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[246] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[247] Paul A. Gagniuc. *Markov Chains: From Theory to Implementation and Experimentation*. John Wiley & Sons, 2017.

[248] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[249] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *International Conference on Machine Learning (ICML)*, 2015.

[250] Anurag Arnab, Chen Sun, and Cordelia Schmid. Unified graph structured models for video understanding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[251] Feilong Chen, Xiuyi Chen, Fandong Meng, Peng Li, and Jie Zhou. GoG: Relation-aware graph-over-graph network for visual dialog. In *Findings of Annual Conference of the Association for Computational Linguistics (Findings of ACL)*, 2021.

[252] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[253] Liang Peng, Shuangji Yang, Yi Bin, and Guoqing Wang. Progressive graph attention network for video question answering. In *Proceedings of the International Conference on Multimedia (MM)*, 2021.

[254] Hamed R. Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. Paying attention to descriptions generated by image captioning models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[255] Badri N. Patro, Anupriy, and Vinay P. Namboodiri. Explanation vs attention: A two-player game to obtain attention for VQA. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020.

[256] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[257] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Fredo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2019.

[258] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

[259] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory (IEEE TIT)*, 1991.

[260] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. In *The Web Conference*, 1999.

[261] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[262] Yang Yang, Ryan N. Lichtenwalter, and Nitesh V. Chawla. Evaluating link prediction methods. *Knowledge and Information Systems*, 2015.

[263] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[264] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[265] Yujie Lu, Pan Lu, Zhiyu Chen, Wanrong Zhu, Xin Eric Wang, and William Yang Wang. Multimodal procedural planning via dual text-image prompting. *arXiv preprint arXiv:2305.01795*, 2023.

[266] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Predicting human attention using computational attention. *arXiv preprint arXiv:2303.09383v2*, 2023.

[267] Peizhao Li, Junfeng He, Gang Li, Rachit Bhargava, Shaolei Shen, Nachiappan Valliappan, Youwei Liang, Hongxiang Gu, Venky Ramachandran, Golnaz Farhadi, Yang Li, Kai J Kohlhoff, and Vidhya Navalpakkam. UniAR: Unifying human attention and response prediction on visual content. *arXiv preprint arXiv:2312.10175*, 2023.

[268] Lisa Anne Hendricks, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Zeynep Akata. Generating visual explanations with natural language. *Applied AI Letters*, 2021.

[269] Edward H. Shortliffe and Bruce G. Buchanan. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 1975.

[270] H. Chad Lane, Mark Core, Michael van Lent, Steve Solomon, and Dave Gomboc. Explainable artificial intelligence for training and tutoring. In *International Conference on Artificial Intelligence in Education*, 2005.

[271] Mark G. Core, H. Chad Lane, Michael van Lent, Dave Gomboc, Steve Solomon, and Milton Rosenberg. Building explainable artificial intelligence systems. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2006.

[272] Michael van Lent, William Fisher, and Michael Mancuso. An explainable artificial intelligence system for small-unit tactical behavior . In *National Conference on Artificial Intelligence*, 1999.

[273] W. Lewis Johnson. Agents that learn to explain themselves. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 1994.

[274] Meghann Lomas, Robert Chevalier, Ernest Vincent Cross II, Robert Christopher Garrett, John Hoare, and Michael Kopack. Explaining robot actions. In *ACM/IEEE International Conference on Human-Robot Interaction*, 2012.

[275] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Bernt Schiele Jeff Donahue, and Trevor Darrell. Generating visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[276] Kosuke Nishida, Kyosuke Nishida, and Shuichi Nishioka. Improving few-shot image classification using machine- and user-generated natural language descriptions. In *Findings of the Association for Computational Linguistics (NAACL)*, 2022.

[277] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Generating counterfactual explanations with natural language. In *Proceedings of the International Conference on Machine Learning Workshop (ICMLW)*, 2018.

[278] Shi Chen and Qi Zhao. REX: Reasoning-aware and grounded explanation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[279] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[280] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. VQA-E: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[281] Jialin Wu and Raymond Mooney. Faithful multimodal explanation for visual question answering. In *Annual Conference of the Association for Computational Linguistics (ACL)*, 2019.

[282] Ana Marasović, Chandra Bhagavatula, Jae Sung Park, Ronan Le Bras, Noah A. Smith, and Yejin Choi. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[283] Radhika Dua, Sai Srinivas Kancheti, and Vineeth N Balasubramanian. Beyond VQA: Generating multi-word answers and rationales to visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2021.

[284] Jialin Wu and Raymond J. Mooney. Self-critical reasoning for robust visual question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[285] Hanxiong Chen, Xu Chen, Shaoyun Shi, and Yongfeng Zhang. Generate natural language explanations for recommendation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval Workshops (SIGIRW)*, 2019.

[286] Zunwang Ke, Jiabao Sheng, Zhe Li, Wushour Silamu, and Qinglang Guo. Knowledge-guided sentiment analysis via learning from natural language explanations. *IEEE ACCESS*, 2021.

[287] Iro Laina, Christian Rupprecht, and Nassir Navab. Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[288] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[289] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[290] Xianyu Chen, Ming Jiang, and Qi Zhao. Self-distillation for few-shot image captioning. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[291] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[292] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. M^2: Meshed-memory transformer for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[293] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

[294] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[295] Jeremy M Wolfe and Todd S Horowitz. Five factors that guide attention in visual search. *Nature Human Behaviour*, 2017.

[296] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 2006.

[297] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

[298] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does CLIP know about a red circle? visual prompt engineering for VLMs. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

[299] Jeffrey S. Perry and Wilson S. Geisler. Gaze-contingent real-time simulation of arbitrary visual fields. In *International Society for Optics and Photonics*, 2002.

[300] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2002.

[301] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[302] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[303] Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. CIDEr-R: Robust consensus-based image description evaluation. In *Conference on Empirical Methods in Natural Language Processing Workshop (EMNLPW)*, 2021.

[304] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[305] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[306] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[307] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.