GazeXplain: Learning to Predict Natural Language Explanations of Visual Scanpaths

Xianyu Chen, Ming Jiang, and Qi Zhao

University of Minnesota, Minneapolis MN 55455, USA
{chen6582,mjiang}@umn.edu, qzhao@cs.umn.edu

Abstract. 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.

Keywords: Scanpath · Explanation · Interpretability · Eye-Tracking

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 Fig. 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

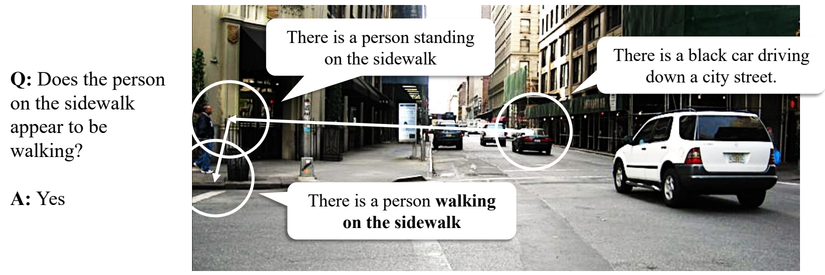

Fig. 1: This example reveals how observers strategically investigate a scene to find out if the person is walking on the sidewalk. Fixations (circles) start centrally, locating a driving car, then shifting to the sidewalk to find the person, and finally looking down to confirm if the person is walking. 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.

attention to visual stimuli [39, 46]. Recent studies have shifted from static fixation distribution modeling to dynamic gaze patterns, known as scanpaths [15,43, 65,66]. 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.

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 [4, 9, 12, 23, 30, 39, 44, 51], dynamic scanpath prediction remains relatively underexplored due to its complexity influenced by various factors. Early studies employed heuristics or statistical priors to generate scanpaths [7, 41–43, 83], while recent models leverage machine learning techniques, including supervised learning [20, 50, 51, 66, 71, 77, 92] and reinforcement learning [15, 90, 91], achieving promising results [15, 17, 57, 66, 71, 90, 91]. 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 [36] initially rely on rules or predefined templates to explain medical diagnosis [74], simulator actions [22, 47, 53, 54] and robot movements [63] *etc*. Recent explanation models explored deep learning-based natural language generation, with successful applications in producing natural language justifications for object classification [34, 35, 67], visual reasoning [13, 27, 58, 64, 69, 85, 86], recommendation systems [10], and sentiment analysis [48], *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 [14,16,18,29,49,52,78,88,89]. These models, trained on multimodal image and language datasets [19, 89, 93], are able to generate fluent and accurate descriptions of visual information. The recent advent of transformer architecture [24, 37, 80] 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 [55,56,60,61]. 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.

3 GazeXplain

Human visual attention is a complex interplay across multiple visual features and cognitive factors [42, 79, 84] (*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.

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 [19, 93] and visual storytelling [38]. This 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 [87], object search [90, 91], and VQA [11]. 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. Fig. 2 illustrates our paradigm for annotating explanations. We utilize LLaVA-1.6 [61] (with a Mistral-7B [45] 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

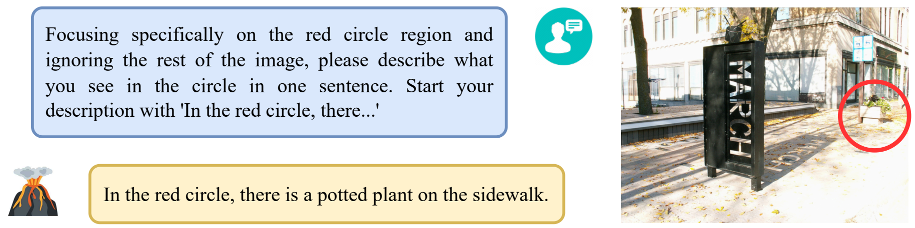

Fig. 2: LLaVA generates the ground-truth explanation for each fixation using an input image with a red circle marking the fixation. The model's response provides information within the marked area, serving as a basis for further refinement.

generate a visual prompt by enclosing each fixation within a red circle [75], mirroring the size of the human fovea [70] (*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 Fig. 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 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 [21] 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 [11], OSIE [87], as well as COCO-Search18 [90] including target-present (TP) and target-absent (TA) subsets. As shown in Tab. 1, this results in a rich col-

Table 1: Statistics of the eye-tracking datasets with annotated explanations.

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

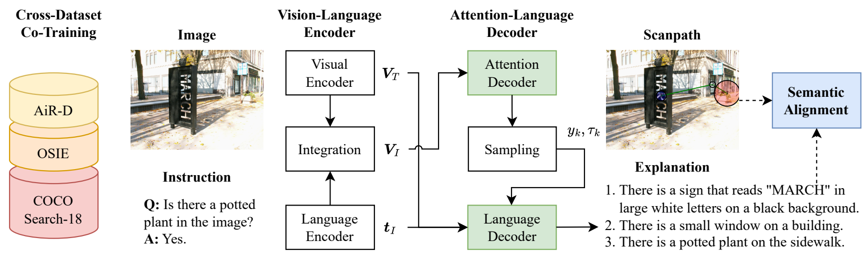

Fig. 3: GazeXplain's architecture consists of a general vision-language encoder and a novel attention-language decoder. The decoder outputs an explanation for each fixation in the predicted scanpath, with a semantic alignment mechanism facilitating the semantic consistency between fixations and explanations. The model is developed on three public datasets using a cross-dataset co-training technique.

lection 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 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.

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 Fig. 3):

GazeXplain is built on top of a general vision-language encoder [62,66]: 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 [15], 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 [56]), 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 $s^r_{i,j} = S_{\cos}(\boldsymbol{r}_i, \boldsymbol{r}_j)$, 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 [33] model.
2. The **explanation similarity**, computed as $s^e_{i,j} = S_{\cos}(\boldsymbol{e}^p_i, \boldsymbol{e}^p_j)$, measures how closely the explanations of different fixations resemble each other semantically, where $\boldsymbol{e}^p_i$ and $\boldsymbol{e}^p_j$ represents the language features of the corresponding explanations, obtained from the language decoder.
3. The **fixation similarity**, computed as $s^f_{i,j} = S_{\cos}(\boldsymbol{g}_i, \boldsymbol{g}_j)$, 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

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

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 [15] and a language generation loss [1, 82], jointly optimizing scanpath prediction and explanation. Please refer to the Supplementary Materials for the implementation details.

Cross-Dataset Co-Training. Prior studies commonly focus on single dataset training [15, 66, 77]. For example, ChenLSTM relies on external VQA models to predict scanpaths on the AiR-D dataset [11], while Gazeformer targets search-related tasks offered by COCO Search-18 [90]. 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 [57], 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.

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. Further results, analyses, and implementation details are reported in the Supplementary Materials.

4.1 Experimental Setup

Datasets. Our experiments utilize a combination of eye-tracking datasets. AiR-D [11] provides insights into human gaze behavior in VQA [32,40], capturing gaze patterns aligned with complex visual reasoning processes. OSIE [87] 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 [90] 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 [3], PathGAN [2], IOR-ROI [77], ChenLSTM [15], Gazeformer [66].

Evaluation Metrics. We comprehensively evaluate GazeXplain using a diverse set of metrics evaluating three aspects of models: First, with established metrics, including ScanMatch (SM) [25], MultiMatch (MM) [26], SED [6,28,31], SS [66,90,91] and SemSS [66,91], 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 [76], and incorporate saliency metrics, including CC [39, 46], NSS [39, 46], AUC [8], and sAUC [8], to assess the spatial accuracy of the prediction. Finally, to measure the linguistic quality of the generated textual explanations, we adopt BLEU [68], METEOR [5], ROUGE [59] and CIDEr-R [73, 81]. 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.

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 Tab. 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.

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 Tab. 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 [56]. Please refer to the Supplementary Materials for ablation studies on the other datasets.

Language Decoder. Tab. 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 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.

Table 2: Scanpath prediction results. The best results are highlighted in bold.

Dataset	Method	Scanpath					Saliency			
		SM ↑	MM ↑	SED ↓	SS ↑	SemSS ↑	CC ↑	NSS ↑	AUC ↑	sAUC ↑
AiR-D [11]	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 [87]	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 [90]	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 [90]	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 3: Ablation study on AiR-D [11] for the proposed technical components: language decoder (EXP), semantic alignment (ALN), and cross-dataset co-training (CT). 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**

Cross-Dataset Co-Training. Scanpath prediction research typically tackles individual tasks in isolation, each relying on its own dataset. However, our

Table 4: Explanation prediction results and diversity analysis. The best results are highlighted in bold.

Dataset	Method	B-4	M	R	C-R	Voc.	Len.	UnP%
AiR-D [11]	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 [87]	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 [90]	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 [90]	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

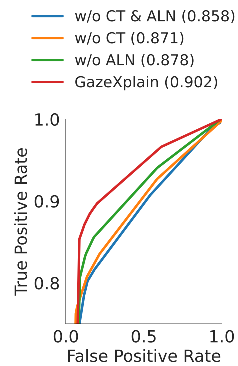

Fig. 4: ROC analysis of fixations and explanations.

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.

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. Tab. 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 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 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 [72]. 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 Fig. 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.

4.5 Qualitative Analysis

Fig. 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 [56]. 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 Fig. 5a, GazeXplain accurately identifies the cake on the left side. Similarly, in Fig. 5b, the model focuses on the dog, while in Fig. 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.*, Fig. 5a: "a plate of food with a fork and knife" while there is no fork or knife present, Fig. 5b: "a man riding a horse" while the man is walking, and Fig. 5c: "a red wall with a black chair and a black chair" while the

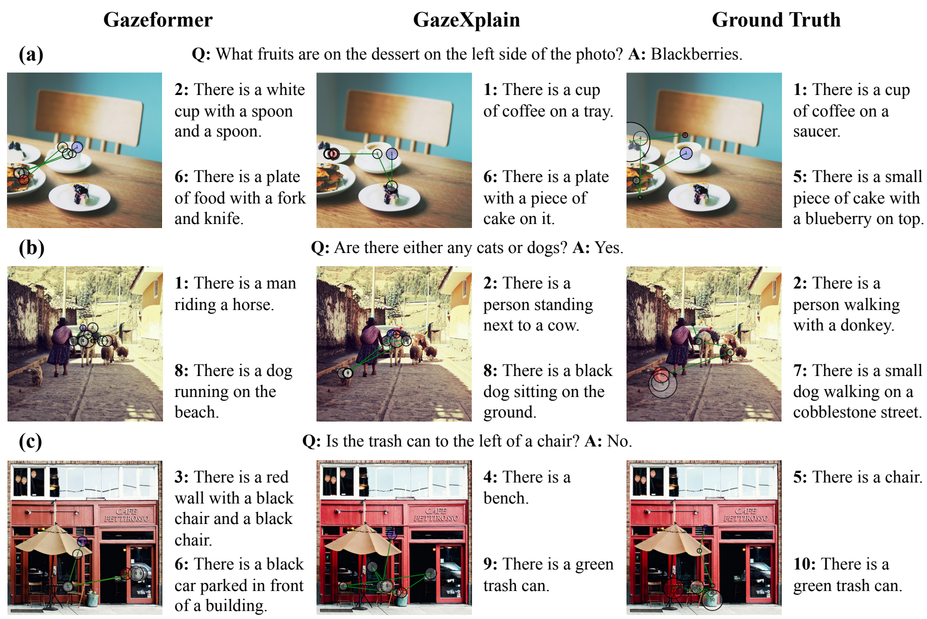

Fig. 5: Quantitative examples from GazeXplain compared to Gazeformer and the ground truth. Each row shows scanpaths and explanations of two key fixations.

chair is not black), GazeXplain provides more accurate and specific fixation descriptions. Particularly in scenes with multiple relevant objects (*e.g.*, different types of desserts and animals in Fig. 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.

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.

Acknowledgements

This work is supported by NSF Grant 2143197.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: 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)
2. Assens, M., Giro-i-Nieto, X., McGuinness, K., O'Connor, N.E.: PathGAN: Visual scanpath prediction with generative adversarial networks. In: Proceedings of the European Conference on Computer Vision Workshop (ECCVW) (2018)
3. Assens, M., McGuinness, K., Giro-i-Nieto, X., O'Connor, N.E.: SaltiNet: Scan-path prediction on 360 degree images using saliency volumes. In: Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW) (2017)
4. Aydemir, B., Hoffstetter, L., Zhang, T., Salzmann, M., Susstrunk, S.: TempSAL - uncovering temporal information for deep saliency prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
5. Banerjee, S., Lavie, A.: 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)
6. Brandt, S.A., Stark, L.W.: Spontaneous eye movements during visual imagery reflect the content of the visual scene. Journal of Cognitive Neuroscience (JCN) (1997)
7. Brockmann, D., Geisel, T.: The ecology of gaze shifts. Neurocomputing (2000)
8. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI) (2019)
9. Chakraborty, S., Wei, Z., Kelton, C., Ahn, S., Balasubramanian, A., Zelinsky, G.J., Samaras, D.: Predicting visual attention in graphic design documents. IEEE Transactions on Multimedia (TMM) (2022)
10. Chen, H., Chen, X., Shi, S., Zhang, Y.: Generate natural language explanations for recommendation. In: International ACM SIGIR Conference on Research and Development in Information Retrieval Workshops (SIGIRW) (2019)
11. Chen, S., Jiang, M., Yang, J., Zhao, Q.: AiR: Attention with reasoning capability. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
12. Chen, S., Valliappan, N., Shen, S., Ye, X., Kohlhoff, K., He, J.: Learning from unique perspectives: User-aware saliency modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
13. Chen, S., Zhao, Q.: REX: Reasoning-aware and grounded explanation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
14. Chen, X., Jiang, M., Zhao, Q.: Leveraging human attention in novel object captioning. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) (2021)
15. Chen, X., Jiang, M., Zhao, Q.: Predicting human scanpaths in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

16. Chen, X., Jiang, M., Zhao, Q.: Self-distillation for few-shot image captioning. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV) (2021)
17. Chen, X., Jiang, M., Zhao, Q.: Beyond average: Individualized visual scanpath prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
18. Chen, X., Yang, J., Chen, S., Wang, L., Jiang, M., Zhao, Q.: 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)
19. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollar, P., Zitnick, C.L.: Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325v2 (2015)
20. Chen, Z., Sun, W.: Scanpath prediction for visual attention using IOR-ROI LSTM. In: International Joint Conference on Artificial Intelligence (IJCAI) (2018)
21. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI) (2002)
22. Core, M.G., Lane, H.C., van Lent, M., Gomboc, D., Solomon, S., Rosenberg, M.: Building explainable artificial intelligence systems. In: Association for the Advancement of Artificial Intelligence (AAAI) (2006)
23. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Predicting human eye fixations via an lstm-based saliency attentive model. IEEE Transactions on Image Processing (IEEE TIP) (2018)
24. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: M^2: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
25. Cristino, F., Mathôt, S., Theeuwes, J., Gilchrist, I.D.: ScanMatch: A novel method for comparing fixation sequences. Behavior Research Methods (BRM) (2010)
26. Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., Holmqvist, K.: It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. Behavior Research Methods (BRM) (2012)
27. Dua, R., Kancheti, S.S., Balasubramanian, V.N.: 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)
28. Faggi, L., Betti, A., Zanca, D., Melacci, S., Gori, M.: Wave propagation of visual stimuli in focus of attention. arXiv preprint arXiv:2006.11035 (2020)
29. Feng, Y., Ma, L., Liu, W., Luo, J.: Unsupervised image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
30. Fosco, C., Casser, V., Bedi, A.K., O'Donovan, P., Hertzmann, A., Bylinskii, Z.: Predicting visual importance across graphic design types. In: ACM Symposium on User Interface Software and Technology (2020)
31. Foulsham, T., Underwood, G.: What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. Journal of Vision (JoV) (2008)
32. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: 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)

33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
34. Hendricks, L.A., Akata, Z., Rohrbach, M., Jeff Donahue, B.S., Darrell, T.: Generating visual explanations. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016)
35. Hendricks, L.A., Hu, R., Darrell, T., Akata, Z.: Generating counterfactual explanations with natural language. In: Proceedings of the International Conference on Machine Learning Workshop (ICMLW) (2018)
36. Hendricks, L.A., Rohrbach, A., Schiele, B., Darrell, T., Akata, Z.: Generating visual explanations with natural language. Applied AI Letters (2021)
37. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
38. Huang, T.H.K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C.L., Parikh, D., Vanderwende, L., Galley, M., Mitchell, M.: Visual storytelling. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT) (2016)
39. Huang, X., Shen, C., Boix, X., Zhao, Q.: 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)
40. Hudson, D.A., Manning, C.D.: 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)
41. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research (VR) (2000)
42. Itti, L., Koch, C.: Computational modelling of visual attention. Nature Reviews Neuroscience (NRN) (2001)
43. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI) (1998)
44. Jia, S., Bruce, N.D.B.: EML-NET:an expandable multi-layer network for saliency prediction. Image and Vision Computing (2020)
45. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7B. arXiv preprint arXiv:2310.06825 (2023)
46. Jiang, M., Huang, S., Duan, J., Zhao, Q.: SALICON: Saliency in context. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
47. Johnson, W.L.: Agents that learn to explain themselves. In: Association for the Advancement of Artificial Intelligence (AAAI) (1994)
48. Ke, Z., Sheng, J., Li, Z., Silamu, W., Guo, Q.: Knowledge-guided sentiment analysis via learning from natural language explanations. IEEE ACCESS (2021)
49. Kim, D.J., Choi, J., Oh, T.H., Kweon, I.S.: Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2019)
50. Kümmerer, M., Bethge, M., Wallis, T.S.A.: Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. Journal of Vision (JoV) (2022)

51. Kümmerer, M., Wallis, T.S.A., Bethge, M.: DeepGaze II: Reading fixations from deep features trained on object recognition. arXiv preprint arXiv:1610.01563 (2016)
52. Laina, I., Rupprecht, C., Navab, N.: Towards unsupervised image captioning with shared multimodal embeddings. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
53. Lane, H.C., Core, M., van Lent, M., Solomon, S., Gomboc, D.: Explainable artificial intelligence for training and tutoring. In: International Conference on Artificial Intelligence in Education (2005)
54. van Lent, M., Fisher, W., Mancuso, M.: An explainable artificial intelligence system for small-unit tactical behavior . In: National Conference on Artificial Intelligence (1999)
55. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
56. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of the International Conference on Machine Learning (ICML) (2022)
57. Li, P., He, J., Li, G., Bhargava, R., Shen, S., Valliappan, N., Liang, Y., Gu, H., Ramachandran, V., Farhadi, G., Li, Y., Kohlhoff, K.J., Navalpakkam, V.: UniAR: Unifying human attention and response prediction on visual content. arXiv preprint arXiv:2312.10175 (2023)
58. Li, Q., Tao, Q., Joty, S., Cai, J., Luo, J.: VQA-E: Explaining, elaborating, and enhancing your answers for visual questions. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
59. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Annual Conference of the Association for Computational Linguistics Workshop (ACLW) (2004)
60. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
61. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
62. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
63. Lomas, M., Chevalier, R., II, E.V.C., Garrett, R.C., Hoare, J., Kopack, M.: Explaining robot actions. In: ACM/IEEE International Conference on Human-Robot Interaction (2012)
64. Marasović, A., Bhagavatula, C., Park, J.S., Bras, R.L., Smith, N.A., Choi, Y.: 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)
65. Meur, O.L., Liu, Z.: Saccadic model of eye movements for free-viewing condition. Vision Research (VR) (2015)
66. Mondal, S., Yang, Z., Ahn, S., Zelinsky, G., Samaras, D., Hoai, M.: 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)
67. Nishida, K., Nishida, K., Nishioka, S.: Improving few-shot image classification using machine- and user-generated natural language descriptions. In: Findings of the Association for Computational Linguistics (NAACL) (2022)

68. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Annual Conference of the Association for Computational Linguistics (ACL) (2002)
69. Park, D.H., Hendricks, L.A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., Rohrbach, M.: Multimodal explanations: Justifying decisions and pointing to the evidence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
70. Perry, J.S., Geisler, W.S.: Gaze-contingent real-time simulation of arbitrary visual fields. In: International Society for Optics and Photonics (2002)
71. Qiu, M., Guo, Y., Zhang, M., Zhang, J., Lan, T., Liu, Z.: Simulating human visual system based on vision transformer. In: Proceedings of the 2023 ACM Symposium on Spatial User Interaction (2023)
72. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese bert-networks. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2019)
73. dos Santos, G.O., Colombini, E.L., Avila, S.: CIDEr-R: Robust consensus-based image description evaluation. In: Conference on Empirical Methods in Natural Language Processing Workshop (EMNLPW) (2021)
74. Shortliffe, E.H., Buchanan, B.G.: A model of inexact reasoning in medicine. Mathematical Biosciences (1975)
75. Shtedritski, A., Rupprecht, C., Vedaldi, A.: 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)
76. Sui, X., Fang, Y., Zhu, H., Wang, S., Wang, Z.: 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)
77. Sun, W., Chen, Z., Wu, F.: Visual scanpath prediction using IOR-ROI recurrent mixture density network. IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI) (2019)
78. Tavakoli, H.R., Shetty, R., Borji, A., Laaksonen, J.: Paying attention to descriptions generated by image captioning models. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
79. Torralba, A., Oliva, A., Castelhano, M.S., Henderson, J.M.: Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological Review (2006)
80. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Conference on Neural Information Processing Systems (NeurIPS) (2017)
81. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
82. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
83. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks (NN) (2006)
84. Wolfe, J.M., Horowitz, T.S.: Five factors that guide attention in visual search. Nature Human Behaviour (2017)
85. Wu, J., Mooney, R.: Faithful multimodal explanation for visual question answering. In: Annual Conference of the Association for Computational Linguistics (ACL) (2019)

86. Wu, J., Mooney, R.J.: Self-critical reasoning for robust visual question answering. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
87. Xu, J., Jiang, M., Wang, S., Kankanhalli, M.S., Zhao, Q.: Predicting human gaze beyond pixels. Journal of Vision (JoV) (2014)
88. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the International Conference on Machine Learning (ICML) (2015)
89. Yang, J., Chen, X., Jiang, M., Chen, S., Wang, L., Zhao, Q.: VisualHow: Multimodal problem solving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
90. Yang, Z., Huang, L., Chen, Y., Wei, Z., Ahn, S., Zelinsky, G., Samaras, D., Hoai, M.: Predicting goal-directed human attention using inverse reinforcement learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
91. Yang, Z., Mondal, S., Ahn, S., Zelinsky, G., Hoai, M., Samaras, D.: Target-absent human attention. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
92. Yang, Z., Mondal, S., Ahn, S., Zelinsky, G., Hoai, M., Samaras, D.: Predicting human attention using computational attention. arXiv preprint arXiv:2303.09383v2 (2023)
93. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics (TACL) (2014)