GazeXplain: Learning to Predict Natural Language Explanations of Visual Scanpaths (Supplementary Materials)

Xianyu Chen, Ming Jiang, and Qi Zhao

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

1 Introduction

In the main paper, we have introduced GazeXplain, a novel study of visual scanpath and prediction. It involves an annotation of ground-truth explanations for diverse eye-tracking datasets related to scanpath, a general model architecture with an attention-language decoder simultaneously predicting scanpaths and the corresponding natural language explanations, a novel semantic alignment mechanism for consistent fixation-explanation alignment, and a cross-dataset co-training to generalize the scanpath prediction and explanation as well as overcome data and task-specific biases. Our experimental results demonstrate that the proposed method achieves competitive performance and strong generalizability. The supplementary materials provide further details and additional results to support these findings:

1) Sec. 2 elaborates on the specific details of the proposed GazeXplain model, including the vision-language encoding module and the objective functions.
2) Sec. 3 presents the implementation details regarding the setting of hyperparameters and the training method of the proposed GazeXplain.
3) Sec. 4 presents supplementary ablation studies conducted on all three eye-tracking datasets (AiR-D [2], OSIE [19], and COCO-Search18 [20]). These studies evaluate the effectiveness of the three key technical components of our approach:
   - Language Decoder for Scanpath Explanations (EXP)
   - Semantic Alignment Mechanism (ALN)
   - Cross-Dataset Co-training (CT)
4) Sec. 5 presents additional quantitative results by analyzing the generated explanations from various large vision-language models, including our GazeXplain. We provide comprehensive experiments on different prompt settings, with or without observer answers to the prompts, varied training strategies of competitors, and a more diverse range of eye-tracking datasets. These results highlight the robustness and effectiveness of our model across various scenarios.
5) Sec. 6 presents additional qualitative results comparing GazeXplain's scanpaths and explanations with those generated by state-of-the-art scanpath

prediction methods. These results further emphasize the superior performance of GazeXplain on the OSIE (free-viewing) and COCO-Search18 (visual search) datasets, highlighting its adaptability to various real-world visual tasks.

2 Supplementary Method

We have introduced the novel components of our GazeXplain model architecture to address the scanpath explanation problem, including an attention-language decoder, a semantic alignment mechanism, and cross-dataset co-training. In this section, we elaborate on further details of GazeXplain's architecture, specifically focusing on the vision-language encoding process and the objective function used for training the model (as briefly mentioned in Section 3.2 of the main paper).

2.1 Vision-Language Encoding

GazeXplain adopts a vision encoder and a language encoder to effectively capture both the inherent visual cues within an image (bottom-up processing) and the higher-level cognitive influences stemming from the task instructions (top-down processing).

Vision Encoding. To characterize the bottom-up stimulus-driven attention, the vision encoding involves the extraction of local image features and refining the features considering the global context:

To extract local image features, the input image is processed with a pre-trained convolutional neural network (CNN), such as the well-established ResNet-50 architecture [7]. The final convolutional-layer outputs of the network are extracted, denoted as $\boldsymbol{V}_R \in \mathbb{R}^{C \times hw}$, where C is the number of channels and h and w indicate the height and width of the feature map, respectively. The extracted features represent localized details scattered across the image, providing a foundational understanding of the visual content.

While $\boldsymbol{V}_R$ captures localized details, it lacks a holistic understanding of the scene. To address this, GazeXplain employs a Transformer encoder [6,12,17] that excels at capturing the relationships between these local features, resulting in the refined visual features denoted as $\boldsymbol{V}_T \in \mathbb{R}^{d \times hw}$, representing the visual content independent of the specific task at hand, where d is the feature dimensionality.

Language Encoding. Human visual attention is not solely driven by the raw visual stimuli. GazeXplain incorporates the influence of task instructions by accepting a general task description as input. It is formatted as a question, such as "What do you see in the image?" or "Is there a `[search target]` in the image?"

The task instruction is fed through a tokenizer [18], which breaks it down into a sequence of meaningful units. The tokens are then processed by a transformer-based language model, such as the powerful RoBERTa architecture [11]. This

stage generates instructional features, denoted as $\boldsymbol{t}_I \in \mathbb{R}^{d_{\text{text}}}$, where d_{text} is the feature dimensionality. Thus, the features $\boldsymbol{t}_I$ encapsulate the semantic meaning and intent conveyed by the task instruction.

Multimodal Integration. Following these independent encoding stages, GazeXplain merges the bottom-up visual features ($\boldsymbol{V}_T$) and the top-down instructional features ($\boldsymbol{t}_I$) through a concatenation operation. This combined representation, denoted as $\boldsymbol{V}_I \in \mathbb{R}^{d \times hw}$, serves as the foundation for GazeXplain's subsequent processing steps, enabling the model to leverage both visual information and task-specific guidance for accurate scanpath prediction and explanation generation.

2.2 Objectives

GazeXplain tackles the dual challenge of predicting scanpaths and generating explanations concurrently. To achieve this, it employs a combined loss function that guides the training process and optimizes model performance for both tasks. Given the ground-truth scanpath $\{y_k, \tau_k\}_{k=1}^{K'}$ and the language explanation $\{\boldsymbol{w}^k\}_{k=1}^{K'}$ with the length of scanpath K', where y_k indicates the fixation position, τ_k indicates its duration, and $\boldsymbol{w}^k$ is its corresponding explanation, the final training objective is a combined loss function to optimize for both scanpath prediction and explanation

$$\mathcal{L} = \mathcal{L}_{\text{fix}} + \mathcal{L}_{\text{exp}} + \mathcal{L}_{\text{aln}}, \tag{1}$$

where $\mathcal{L}_{\text{fix}}$ is the standard scanpath prediction loss, $\mathcal{L}_{\text{exp}}$ is the standard language prediction loss, and $\mathcal{L}_{\text{aln}}$ is the semantic alignment loss as detailed in Section 3.2 of the main paper, which encourages the model to ensure that the generated explanations exhibit a strong semantic connection with the visual features associated with each fixation. By carefully balancing these loss terms during training, GazeXplain not only predicts scanpaths accurately but also generates explanations that illuminate the rationale behind those fixations.

Scanpath Prediction Loss. Given the ground truth scanpath $\{y_k, \tau_k\}_{k=1}^{K'}$, and the corresponding duration parameters $\{\mu_k, \sigma_k^2\}_{k=1}^{K'}$ of log-normal distribution from the output of GazeXplain, the scanpath prediction loss is defined as

$$\mathcal{L}_{\text{fix}} = -\sum_{k=1}^{K'+1} \log p_k^y(y_k|y_1, \cdots, y_{k-1}; \boldsymbol{\theta}) - \sum_{k=1}^{K'} \log p_k^\tau(\tau_k|\mu_k, \sigma_k^2; \boldsymbol{\theta}), \tag{2}$$

where $\boldsymbol{\theta}$ represents the learnable parameters of GazeXplain, $\log p_k^y$ is the parametric conditioned probability of fixation position y_k, and $\log p_k^\tau$ is the parametric log-normal function [3]. This standard scanpath prediction loss term acts as a guiding force, encouraging the model to predict fixations that closely resemble the actual sequence of fixations observed in the ground truth data.

Language Prediction Loss. This standard language prediction loss term ensures that the generated explanations are not only grammatically correct but also semantically consistent with the predicted scanpath and the provided task instruction.

$$\mathcal{L}_{\text{exp}} = \frac{1}{LK'} \sum_{k=1}^{K'} \sum_{\ell=1}^{L} -\log p(\boldsymbol{w}_{\ell}^{k} | \boldsymbol{g}_{k}^{d}, \boldsymbol{t}_{I}^{d}, \boldsymbol{w}_{0:\ell-1}^{k}; \boldsymbol{\theta}), \tag{3}$$

where $\boldsymbol{\theta}$ represents the learnable parameters of GazeXplain, $\boldsymbol{g}_{k}^{d}$ and $\boldsymbol{t}_{I}^{d}$ represents the encoded integration of visual and textual information mentioned in Section 3.2 of the main paper, $\boldsymbol{w}^{k}$ is the ground truth language explanation of the k-th fixation with length L and $\boldsymbol{w}_{\ell}^{k}$ represent the ℓ-th token of the explanation $\boldsymbol{w}^{k}$. This loss term promotes the generation of explanations that accurately reflect what the model sees at each fixation point.

3 Implementation Details

We adhere to the original dataset splits [3, 12, 21], maintaining consistency with prior research. During training, we conduct supervised learning for 8 epochs using the Adam [8] optimizer with specific hyperparameters: a learning rate of 4×10^{-4}, weight decay of 5×10^{-5}, and batch size of 16. Subsequently, we integrate self-critical sequence training (SCST) [3,15] for the remaining 2 epochs to enhance the model's ability to sample scanpaths and generate explanations. In SCST, the learning rate linearly decays from 10^{-5}, with a batch size of 8, facilitating further refinement of the model's performance. The minimum and maximum lengths of the fixations for the generated scanpath are set to 1 and 16, respectively. All compared models are adapted following the same settings for fairness [3].

4 Supplementary Ablation Study

In Tab. 3 of the main paper, we have conducted a comprehensive ablation study on the AiR-D [2] dataset to demonstrate the effectiveness of three key components of our proposed GazeXplain: language decoder (EXP), semantic alignment (ALN), and cross-dataset co-training (CT). In this section, to demonstrate the generalizability of our GazeXplain model and provide further insights into the contributions of these components, we conduct comprehensive ablation studies on all datasets: AiR-D [2], OSIE [19] and COCO-Search18 [20] (see Tab. 1). Similar to the findings reported in Section 4.3 of the main paper, these results show that EXP, ALN, and CT play complementary roles in significantly enhancing overall performance on our GazeXplain:

Language Decoder. Across all datasets, incorporating the language decoder yields significant improvements in scanpath prediction, spatial saliency, and explanation quality. This highlights the importance of explaining fixations for the

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

Dataset	Modules			Scanpath					Saliency				Explanation ↑			
	EXP	ALN	CT	SM ↑	MM ↑	SED ↓	SS ↑	SemSS ↑	CC ↑	NSS ↑	AUC ↑	sAUC ↑	B-4	M	R	C-R
AiR-D [2]				0.337	0.805	8.197	0.274	-	0.582	1.582	0.794	0.693	19.5	18.5	45.0	61.9
	✓			0.339	0.805	8.216	0.280	-	0.614	1.674	0.806	0.706	27.6	20.5	50.1	91.9
	✓	✓		0.346	0.806	8.250	0.284	-	0.631	1.733	0.807	0.713	30.4	21.7	51.6	115.1
			✓	0.356	0.812	7.834	0.292	-	0.582	1.597	0.781	0.688	18.6	18.1	44.4	66.7
	✓		✓	0.378	**0.819**	7.693	0.299	-	0.647	1.797	0.806	0.713	27.7	20.6	50.3	97.3
	✓	✓	✓	**0.386**	0.817	**7.489**	**0.308**	-	**0.662**	**1.851**	**0.808**	**0.719**	**30.7**	**21.9**	**51.7**	**123.1**
OSIE [19]				0.364	0.804	7.588	0.301	-	0.674	2.272	0.805	0.754	13.9	14.2	38.6	24.0
	✓			0.366	0.803	7.561	0.312	-	0.701	2.380	0.824	0.768	12.4	16.5	40.2	23.6
	✓	✓		0.369	0.804	7.633	0.315	-	0.728	2.414	0.826	0.769	16.1	17.4	41.7	37.4
			✓	0.358	0.804	7.431	0.305	-	0.682	2.304	0.807	0.755	13.7	14.2	39.0	26.2
	✓		✓	0.372	0.805	7.392	0.314	-	0.730	2.471	0.829	0.776	15.7	20.4	41.7	37.2
	✓	✓	✓	**0.380**	**0.806**	**7.228**	**0.317**	-	**0.748**	**2.530**	**0.839**	**0.786**	**16.7**	**21.1**	**42.0**	**48.6**
COCO-Search18 Target-Present [20]				0.415	0.791	2.043	0.477	0.387	0.662	2.859	0.864	0.772	22.0	19.4	48.6	69.9
	✓			0.433	0.795	2.122	0.499	0.407	0.718	3.074	0.891	0.808	23.3	15.4	52.4	111.2
	✓	✓		0.449	0.798	1.983	0.513	0.424	0.772	3.298	0.908	0.827	26.0	16.2	54.2	133.2
			✓	0.419	0.800	2.216	0.487	0.385	0.675	2.887	0.874	0.777	22.4	19.0	48.1	67.6
	✓		✓	0.476	**0.809**	**1.966**	0.535	0.440	0.804	3.503	0.913	0.831	26.8	18.1	54.5	130.9
	✓	✓	✓	**0.480**	0.807	1.981	**0.541**	**0.443**	**0.809**	**3.529**	**0.915**	**0.836**	**28.2**	**19.5**	**55.3**	**139.6**
COCO-Search18 Target-Absent [20]				0.328	0.801	4.430	0.342	0.338	0.628	1.737	0.779	0.680	10.2	12.8	39.7	61.8
	✓			0.342	0.806	4.489	0.352	0.345	0.682	1.891	0.804	0.706	15.6	20.9	43.2	77.0
	✓	✓		0.349	0.810	4.409	0.362	0.354	0.692	1.948	0.805	0.711	17.2	22.5	43.8	91.9
			✓	0.345	0.805	4.414	0.359	0.340	0.609	1.739	0.772	0.680	10.2	12.7	39.6	62.2
	✓		✓	0.368	0.811	**4.282**	0.378	0.362	0.704	2.055	0.802	0.712	16.3	26.4	43.2	92.9
	✓	✓	✓	**0.373**	**0.813**	4.307	**0.382**	**0.365**	**0.716**	**2.089**	**0.811**	**0.721**	**18.5**	**27.5**	**44.5**	**106.5**

model to gain a deeper understanding of the underlying visual semantics, leading to more refined predictions. In particular, when co-training is applied, there is a consistent improvement in the SM scores (0.01+ on OSIE and 0.02+ on all datasets) and CIDEr-R scores (11.0 on OSIE and 30.0+ on the other datasets). Similarly, SS, SemSS, CC, NSS and *etc.* scores all see a substantial increase on all the datasets, indicating that explanations enhance the model's ability to not only predict fixations accurately but also describe them in a way that is consistent with human understanding.

Semantic Alignment. Including semantic alignment further enhances performance. We observe improvements in most metrics on all the datasets, indicating that aligning the semantics of fixations with their explanations improves both the precision of explanations and the accuracy of fixations. Across all datasets, semantic alignment yields a boost in CIDEr-R scores (about 10.0+ on all the datasets) and an improvement on almost the scanpath and saliency metric across all the datasets (0.018 increase of CC on OSIE dataset). This suggests that ensuring semantic coherence between fixations and their corresponding descriptions not only improves the quality of the explanations themselves but also guides the model to generate more accurate fixations.

Cross-Dataset Co-Training. Co-training the model across diverse datasets consistently improves performance. This is evident from the overall increase in scores across all metrics on most datasets. Co-training allows the model to leverage complementary information from various data sources, leading to more robust scanpath prediction and explanation generation. For instance, on the COCO-Search18 Target-Present dataset, co-training results in significant improvements in both scanpath prediction (SM increases from 0.449 to 0.480) and explanation quality (CIDEr-R increases from 133.2 to 139.6). This highlights the effectiveness of co-training in enhancing the model's generalizability.

Overall, the ablation study highlights the effectiveness of each core component in GazeXplain. Language decoding empowers explanation, semantic alignment fosters coherence, and cross-dataset co-training promotes generalizability. By incorporating all three components, GazeXplain achieves superior performance in scanpath prediction, saliency prediction, and explanation generation across diverse datasets.

5 Supplementary Quantitative Results

We have presented comprehensive quantitative results in the main paper, including scanpath prediction results, an ablation study of our proposed GazeXplain, and scanpath explanation results. In this section, we elaborate on further analyses and quantitative results of generated explanations from large vision-language models, explore the inclusion of observer answers during the training and inference stages, and investigate cross-dataset training strategies for competitors as well as the generalizability of GazeXplain across datasets. These analyses serve as complementary quantitative results to the main paper.

Analyses on the Generated Explanations from Large Vision-Language Models. In the main paper, we intend to summarize the natural advantages of model-generated descriptions from large vision-language models (LVLM) over those labeled by humans, where the former is automatic, cost-effective, scalable, and possibly more consistent.To further demonstrate the quality and accuracy of the LLaVA [10] generated descriptions in the main paper, we conduct a systematic evaluation by comparing LLaVA [10] and GPT-4V [13] descriptions of 201 red-circled COCO-Search18 objects with human annotations from Visual Genome [9], using CIDEr-R (C-R) [16] and Sentence Similarity (SenS) [14] scores. The experimental result shows that LLaVA generates reasonably accurate descriptions (C-R=110.4, SenS=0.606), better than GPT-4V (C-R=99.1, SenS=0.592), while GazeXplain generates similarly accurate descriptions (C-R=106.3, SenS=0.590). This demonstrates that LLaVA generates more reasonable descriptions aligned with human annotations, and our GazeXplain has a similar ability to describe fixation positions by learning from the curated dataset.

This work establishes the foundation for modeling scanpath explanations by utilizing LLaVA-generated explanations. However, there are some limitations to the LLaVA-generated explanations. For example, rephrased LLaVA outputs exist

due to the variability of fixations in the same region, and our manual corrections addressed outliers (less than 0.58%).

Exploration of Observer Answer. The AiR-D (VQA) dataset collects observers' answer during eye-tracking [2–4], which can be different from the ground-truth. This creates a new scenario for training scanpath models to be aware of task performance. As shown in Tab. 2, GazeXplain can flexibly handle different scenarios w/ or w/o observer answers: 1. When a particular observer's answer is present, it predicts the observer's scanpaths. 2. When the answer is absent, it predicts general scanpaths. The main paper presents the first scenario, where SM=0.386 and NSS=1.851. Removing the answer from the test set results in a similar performance (SM=0.385, NSS=1.845). Removing the answers from both the training and test sets leads to a slight decrease (SM=0.380, NSS=1.810), but it still outperforms the compared models. This demonstrates GazeXplain's ability to capture inter-observer attention variations and provide a flexible interface for predicting either observer-specific scanpath patterns or general scanpath patterns.

Table 2: Ablation study on AiR-D [2] for the absence of observer answers in the training set and/or the test set. The best results are highlighted in bold.

Answer Absent		Scanpath				Saliency				CIDEr-R ↑
Training	Test	SM ↑	MM ↑	SED ↓	SS ↑	CC ↑	NSS ↑	AUC ↑	sAUC ↑	
		0.386	0.817	**7.489**	0.308	**0.662**	**1.851**	**0.808**	**0.719**	**123.1**
	✓	0.385	0.816	7.539	**0.310**	0.659	1.845	0.805	0.717	119.6
✓	✓	0.380	**0.817**	7.684	0.307	0.653	1.810	0.801	0.711	114.4

Cross-Dataset Training for Competitors. To investigate whether retraining other models (ChenLSTM [3] and Gazeformer [12]) on more datasets can improve their performance, we adjusted the settings of these models to be trained on various scanpath datasets. As shown in Tab. 3, directly combining all training datasets results in lower performance compared to single-dataset training. This suggests the challenge of leveraging data from distinct tasks and settings in training. However, GazeXplain can address this challenge due to its unique model design and co-training strategy.

Generalizability across Datasets. To demonstrate the generalizability across different datasets, we also consider the COCO-FreeView [5] and WebSaliency [1] datasets. COCO-FreeView [5] enlarges the scale of free-viewing eye fixations, offering a more appropriate testbed for free-viewing scenarios. WebSaliency [1] extends the scope of natural image analysis to include webpage images and

Table 3: Ablation study on the cross-dataset training strategy for all the datasets (AiR-D [2], OSIE [19], and COCO-Search18 [20]). The best results are highlighted in bold. † indicates the model trained with the cross-dataset training strategy.

Method	SM ↑				NSS ↑			
(†cross-dataset training)	AiR-D	OSIE	TP	TA	AiR-D	OSIE	TP	TA
ChenLSTM †	0.325	0.344	0.358	0.333	1.790	2.406	2.694	1.819
Gazeformer †	0.356	0.358	0.419	0.345	1.597	2.304	2.887	1.739
ChenLSTM	0.350	0.377	0.448	0.366	1.727	2.488	3.376	2.036
Gazeformer	0.357	0.372	0.433	0.354	1.512	2.308	2.990	1.837
GazeXplain †	**0.386**	**0.380**	**0.480**	**0.373**	**1.851**	**2.530**	**3.529**	**2.089**

Table 4: Scanpath prediction results on two additional datasets (COCO-FreeView [5] and WebSaliency [1]). The best results are highlighted in bold.

Dataset	Method	Scanpath				Saliency			
		SM ↑	MM ↑	SED ↓	SS ↑	CC ↑	NSS ↑	AUC ↑	sAUC ↑
COCO-FreeView [5]	Human	0.340	0.814	12.782	0.325	0.830	1.998	0.869	0.719
	ChenLSTM	0.360	0.827	12.243	0.351	0.790	1.879	0.820	0.692
	Gazeformer	0.364	0.826	12.207	0.349	0.790	1.850	0.822	0.692
	GazeXplain	**0.375**	**0.828**	**12.125**	**0.353**	**0.804**	**1.909**	**0.832**	**0.701**
WebSaliency [1]	Human	0.331	0.838	18.858	0.213	0.819	1.720	0.842	0.768
	ChenLSTM	0.302	0.819	16.927	0.199	0.746	1.348	0.775	0.679
	Gazeformer	0.284	**0.831**	17.106	**0.218**	0.714	1.328	0.777	0.702
	GazeXplain	**0.329**	0.828	**16.820**	0.217	**0.754**	**1.516**	**0.789**	**0.715**

graphic designs, ensuring a thorough evaluation of our model's generalizability to non-natural images, which often contain a mix of text, images, logos, and banners. As shown in Tab. 4, GazeXplain consistently outperforms the competitors across all datasets, demonstrating promising performance in both scanpath metrics and saliency metrics.

6 Supplementary Qualitative Results

In addition to the qualitative examples presented in Fig. 5 of the main paper, we present more qualitative results, involving a thorough comparison of the Gazeformer model, GazeXplain, and human ground truth, covering a range of visual tasks based on the OSIE [19], COCO-Search18 Target-Present [20] and Target-Absent [20] datasets. GazeXplain consistently enhances the capability to predict fixations on key objects in these diverse tasks. These qualitative examples demon-

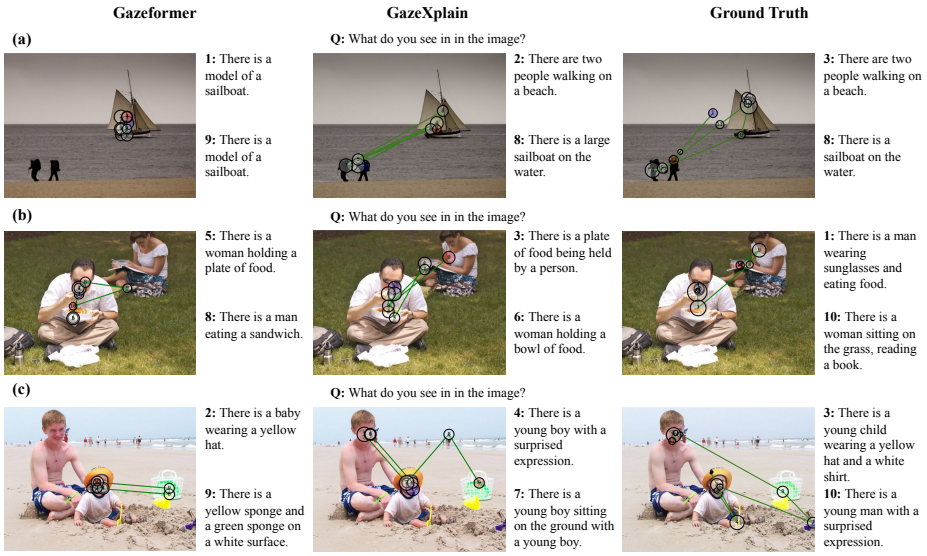

Fig. 1: Quantitative examples from GazeXplain compared to Gazeformer and the ground truth on the OSIE dataset. Each row shows scanpaths and explanations of two key fixations.

strate the potential of our GazeXplain model as a promising and interpretable tool for unraveling the mechanisms of visual perception and attention.

Results on OSIE Dataset. Fig. 1 presents qualitative examples on the OSIE (free-viewing) dataset [19]. Free-viewing tasks involve natural scene exploration, where observers freely gaze at a stimulus without explicit instructions. Understanding these gaze patterns is crucial for tasks like scene understanding and image retrieval. Our qualitative observations from Fig. 1 demonstrate GazeXplain's effectiveness in free-viewing scenarios.

We observe GazeXplain's improved ability to predict and explain fixations on salient objects. In Fig. 1a, GazeXplain accurately identifies the two people in the bottom-left corner, mimicking human focus on social elements within a scene. Similarly, Fig. 1b and Fig. 1c demonstrate the model's ability to detect people (a woman and a young boy) that naturally attract human attention during free-viewing. This alignment with human gaze patterns highlights GazeXplain's capability of capturing the semantic-level saliency.

Beyond fixation prediction, GazeXplain also generates accurate explanations for these fixations. Compared to Gazeformer, GazeXplain offers more precise and semantically relevant narratives. For instance, Gazeformer makes errors in all three examples: In Fig. 1a, it mistakenly describes a real sailboat as a "model of a sailboat." Similarly, it assigns incorrect genders and objects in Fig. 1b and Fig. 1c. In contrast, GazeXplain provides accurate descriptions, demonstrating a deeper semantic understanding of the scene. This is particularly evident in com-

Fig. 2: Quantitative examples from GazeXplain compared to Gazeformer and the ground truth on the COCO-Search18 dataset. Each row shows scanpaths and explanations of two key fixations.

plex scenes with multiple people (*e.g.*, Fig. 1b and Fig. 1c), where GazeXplain successfully distinguishes between individuals. These instances highlight GazeXplain's success in melding visual exploration with semantic insight to predict more accurate scanpaths and explanations.

Results on COCO-Search18 Datasets. Fig. 2 presents a qualitative comparison on the COCO-Search18 [20] Target-Present and Target-Absent datasets, which feature an object search task – finding a specific target object within an image. Our qualitative observations from Fig. 2 demonstrate GazeXplain's effectiveness in modeling these gaze patterns.

We observe that GazeXplain accurately predicts fixations on image regions likely to contain the target object, mimicking human search strategies. For instance, when searching for a potted plant (see Fig. 2a and Fig. 2b), GazeXplain focuses on areas where a plant might typically be placed, such as the desk, floor, table, and nightstand. Similarly, in the search for a fork (see Fig. 2c and Fig. 2d), the model actively explores the table, a common location for forks. This alignment with human search behavior highlights GazeXplain's ability to capture the cognitive process behind object search.

Beyond fixation prediction, GazeXplain's explanations are semantically aligned with the fixated objects, providing insight into the model's reasoning process. This is in contrast to Gazeformer, which often generates inaccurate descriptions

(all four examples in Fig. 2). For example, GazeXplain effectively explains its fixations while searching for the plant (*e.g.*, "desk" in Fig. 2a, or "nightstand" in Fig. 2b), whereas Gazeformer makes irrelevant suggestions (*e.g.* "cat" and "piano keyboard" in Fig. 2a or "hair" in Fig. 2b). Similarly, GazeXplain offers clear explanations during the fork search (*e.g.*, "table" in both Fig. 2c and Fig. 2d), while Gazeformer struggles (referring to non-existent objects, *e.g.*, Fig. 2c: "a painting of a man with a hat on" and Fig. 2d: "a man sitting at a desk with a laptop,"). These results highlight GazeXplain's capability to not only predict search fixations accurately but also to explain the rationale behind them.

References

1. 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) (2023)
2. Chen, S., Jiang, M., Yang, J., Zhao, Q.: AiR: Attention with reasoning capability. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
3. 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)
4. 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)
5. Chen, Y., Yang, Z., Chakraborty, S., Mondal, S., Ahn, S., Samaras, D., Hoai, M., Zelinsky, G.: Characterizing target-absent human attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW) (2022)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021)
7. 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)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015)
9. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Li, F.F.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision (IJCV) (2017)
10. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
11. 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)
12. 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)

13. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
14. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese bert-networks. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2019)
15. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
16. 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)
17. 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)
18. Webster, J.J., Kit, C.: Tokenization as the initial phase in NLP. In: International Conference on Computational Linguistics (COLING) (1992)
19. Xu, J., Jiang, M., Wang, S., Kankanhalli, M.S., Zhao, Q.: Predicting human gaze beyond pixels. Journal of Vision (JoV) (2014)
20. 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)
21. 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)