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

ECCV 2024

什么是扫视路径预测？

1. 全景图像中，自由观看时的视线移动情况（图1）
2. 目标类别导向的扫视路径预测（图2）
3. 增量式指代目标导向的视线预测（图3）

图1

找寻：car

图2

找寻：[BOT] kid on the far right [EOT]

图3

背景：
传统的扫描路径模型预测注视点的“何时”和“何处”，但不提供解释，这在理解注视点背后的原因方面存在空白。
为了填补这一空白，文章提出了GazeXplain任务，旨在同时预测扫描路径并生成自然语言解释，以揭示每个注视点背后的“是什么”和“为什么”

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.

Contributions

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

Data

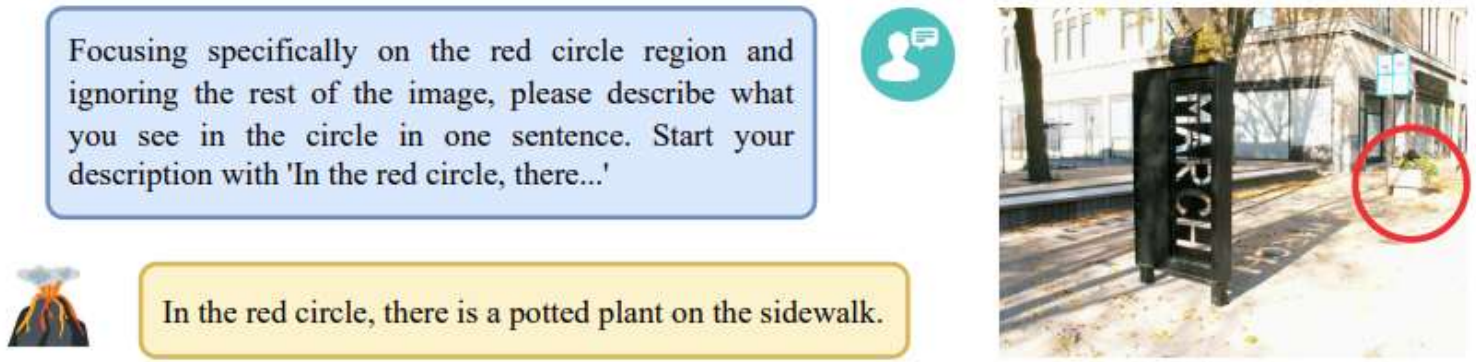


Fig. 2: LLaVA generates the ground-truth explanation for each fixation using an input image with a red circle marking the fixation. The model's response provides information within the marked area, serving as a basis for further refinement.

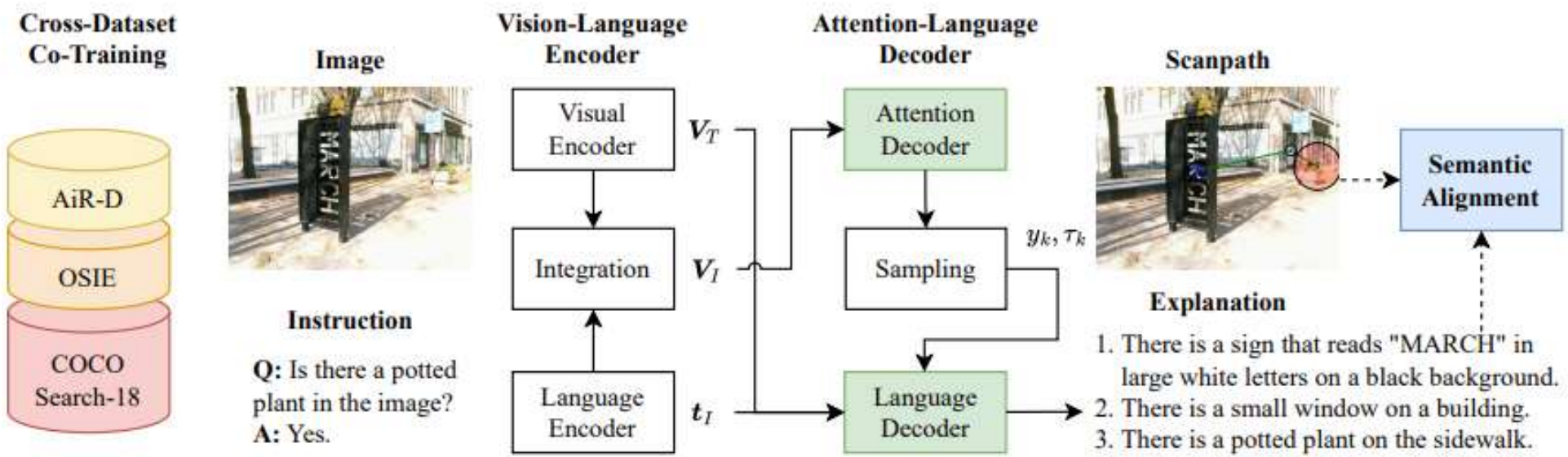
1. We present the first natural-language annotations on scanpaths, offering explanations for each specific fixation within the scanpath, rather than image-wise descriptions such as image captioning.
2. While most image-to-language datasets focus on specific tasks, ours comprise a wider range of visual tasks, including free-viewing, object search, and VQA

we annotate ground-truth explanations for three different eye-tracking datasets: AiR, OSIE, as well as COCO-Search18 including target-present (TP) and target-absent (TA) subsets.

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

Architecture



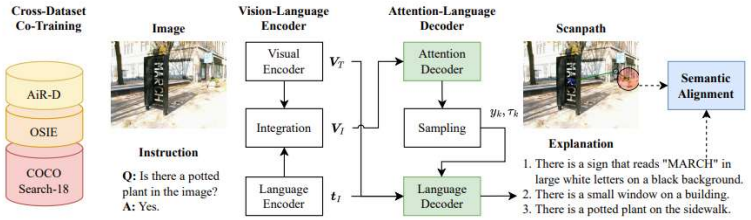
Semantic Alignment

We propose a semantic alignment mechanism to ensure the semantic consistency between predicted fixations, explanations, and visual features.

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 [36] 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

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



Results

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 [12]	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 [95]	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 [98]	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 [98]	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**

Ablation Study

Language Decoder（EXP）: Tab. 3 presents notable improvements achieved by integrating the language decoder into the model architecture.

Semantic Alignment(ALN): The semantic alignment mechanism further improves the model's accuracy in identifying fixated visual semantics and generating coherent descriptions.

Co-Train(CT): our approach diverges by training a unified model across multiple datasets, harnessing shared knowledge and contemporary features to enhance performance

Table 3: Ablation study on AiR-D [12] 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**