GazeXplain: Learning to Predict Natural Language Explanations of Visual Scanpaths

European Conference on Computer Vision ECCV 2024

Xianyu Chen

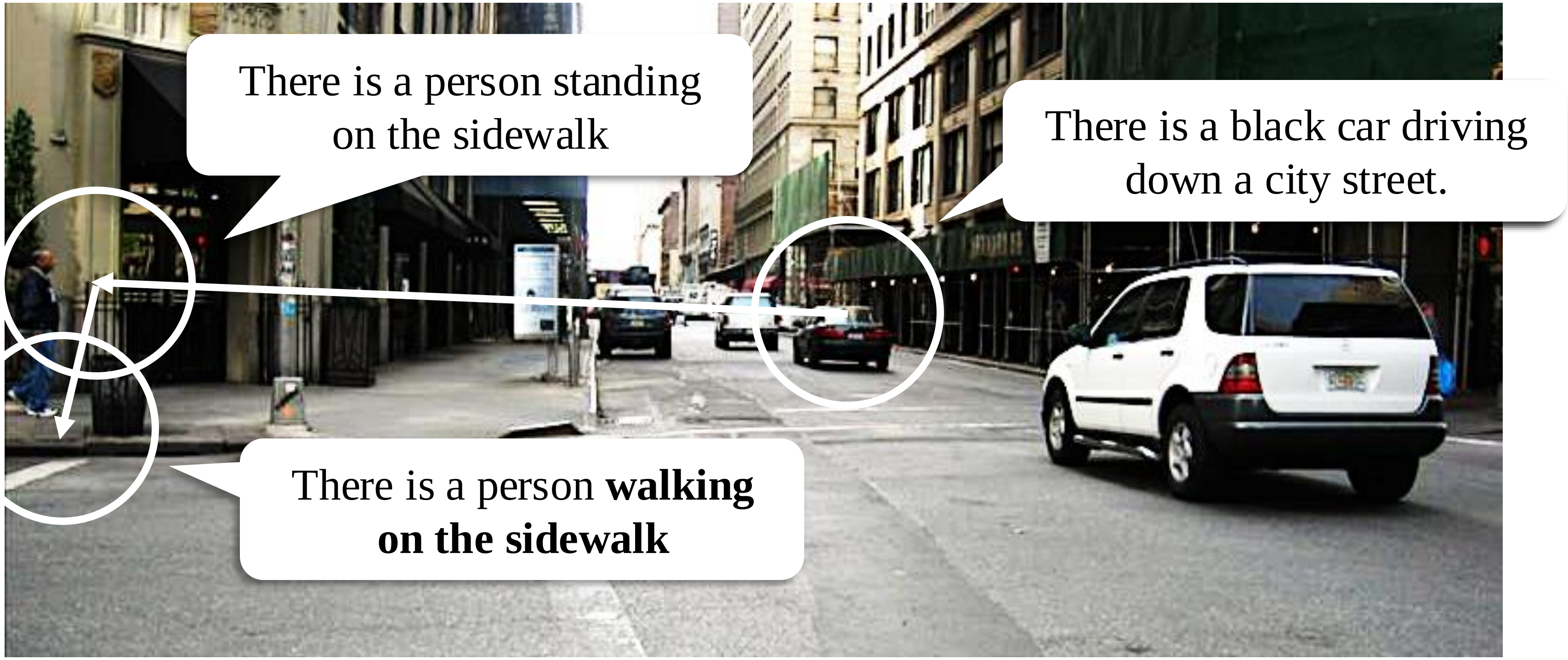
Ming Jiang

Catherine Qi Zhao

Motivation: Interpretable Visual Scanpaths

Q: Does the person on the sidewalk appear to be walking?

A: Yes



Scanpath prediction models are black boxes

Free-viewing

SaltiNet [M. Assens et al., ICCV Workshops, 2017]

PathGAN [M. Assens et al., ECCV Workshops, 2018]

IOR-ROI [W. Sun et al., TPAMI, 2019]

DeepGaze III [M. Kummerer et al., JoV, 2022]

Visual Search

IRL [Z. Yang et al., CVPR 2020]

ChenLSTM [X. Chen et al., CVPR 2021]

FFMs [Z. Yang et al., ECCV 2022]

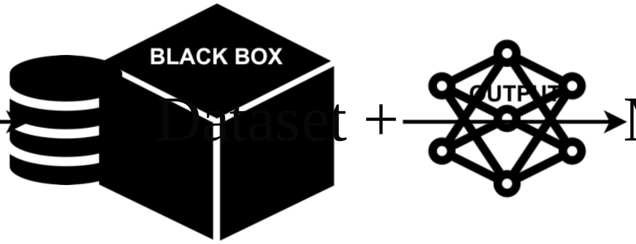
Gazeformer [S. Mondal et al., CVPR 2023]

VQA

ChenLSTM [X. Chen et al., CVPR 2021]

ISP [X. Chen et al., CVPR 2024]

Is the trash can to the
left of a chair?



GazeXplain – Dataset

A rich collection of natural-language explanations annotated on **7,004** images and **86,407** fixations across diverse visual tasks. The explanations are concise, with lengths falling within **10.66 ± 3.54** words each.

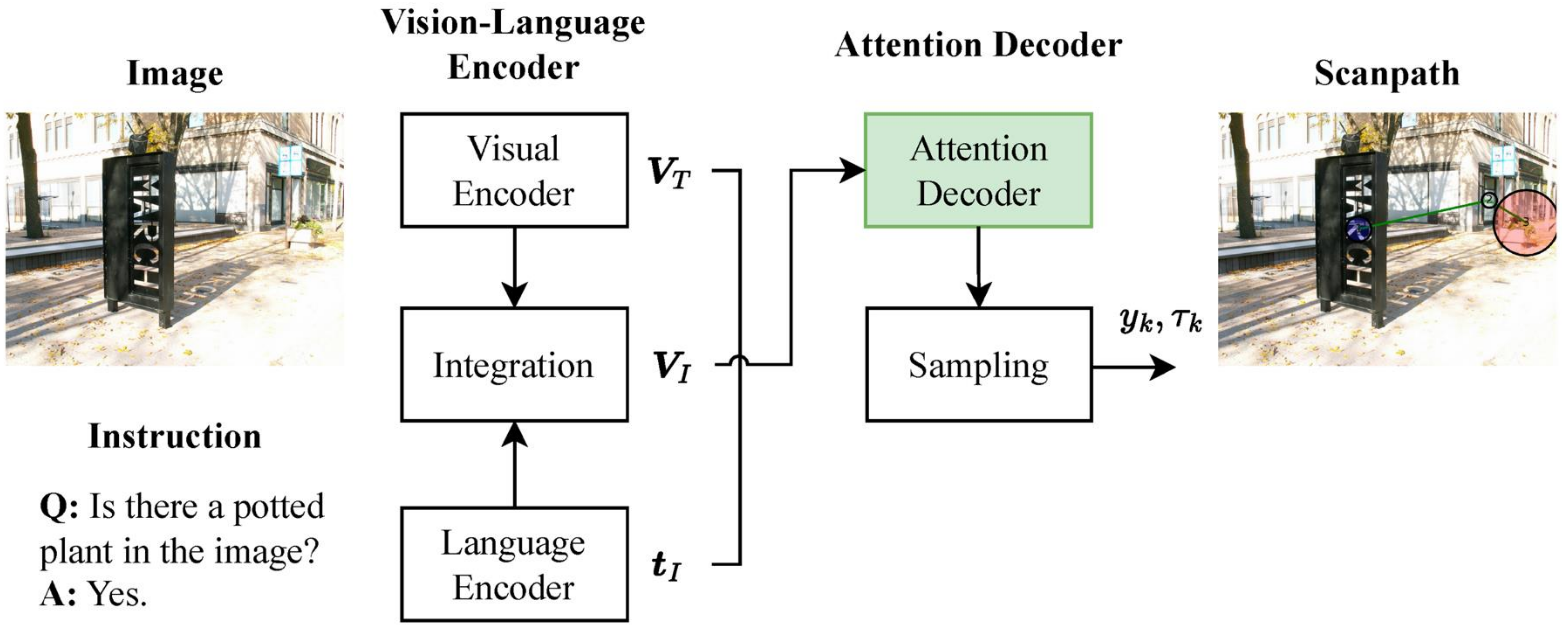
Scanpath

Explanation

1. There is a sign that reads "MARCH" in large white letters on a black background.
2. There is a small window on a building.
3. There is a potted plant on the sidewalk.

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

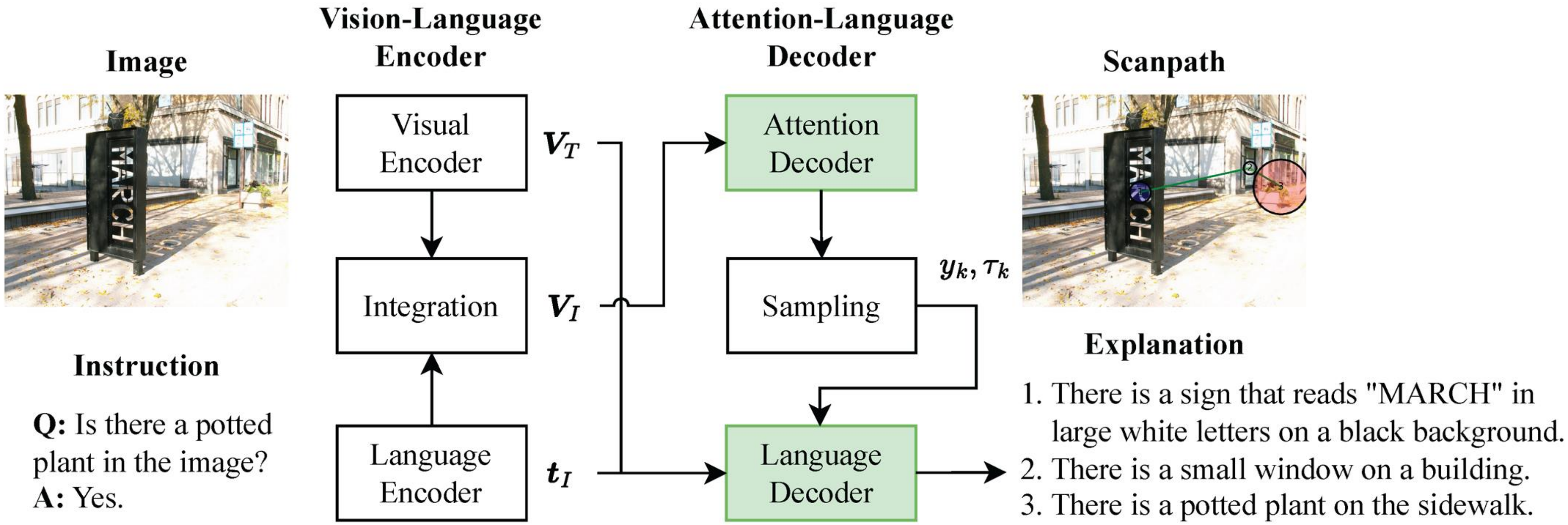
GazeXplain – Method



Baseline: Gazeformer [S. Mondal et al., CVPR 2023]

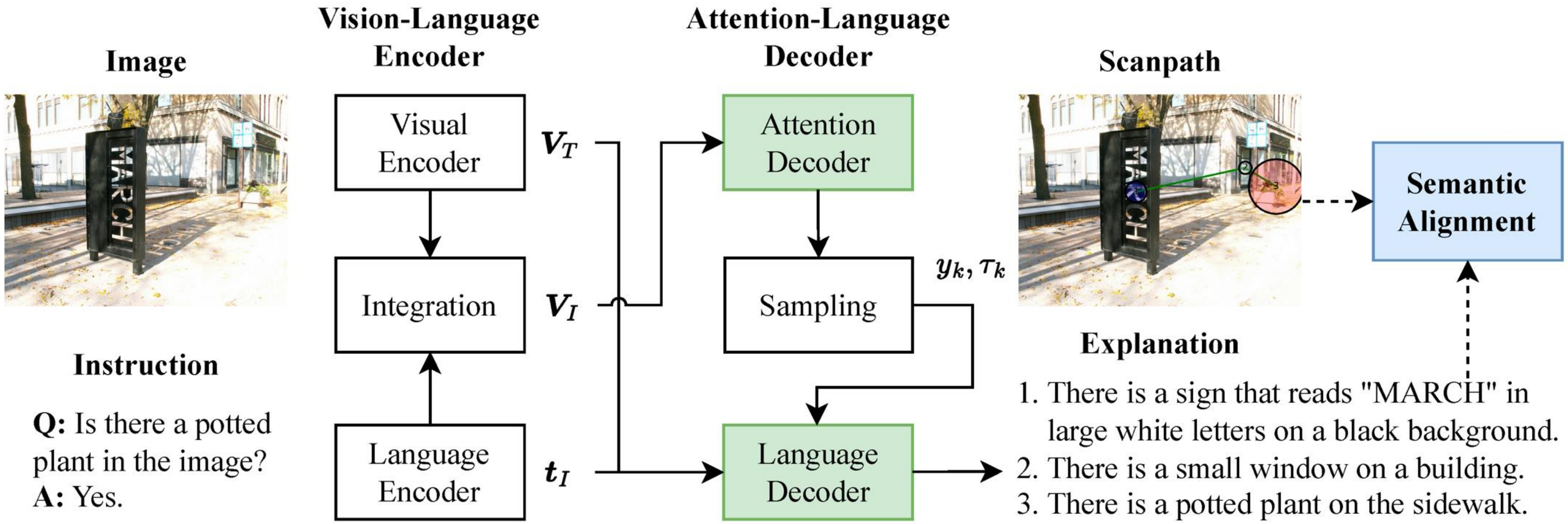
- Visual Encoder: ResNet-50 (Frozen) and Transformer
- Language Encoder: RoBERTa
- Attention Decoder: Transformer with Grid Classification

GazeXplain – Method



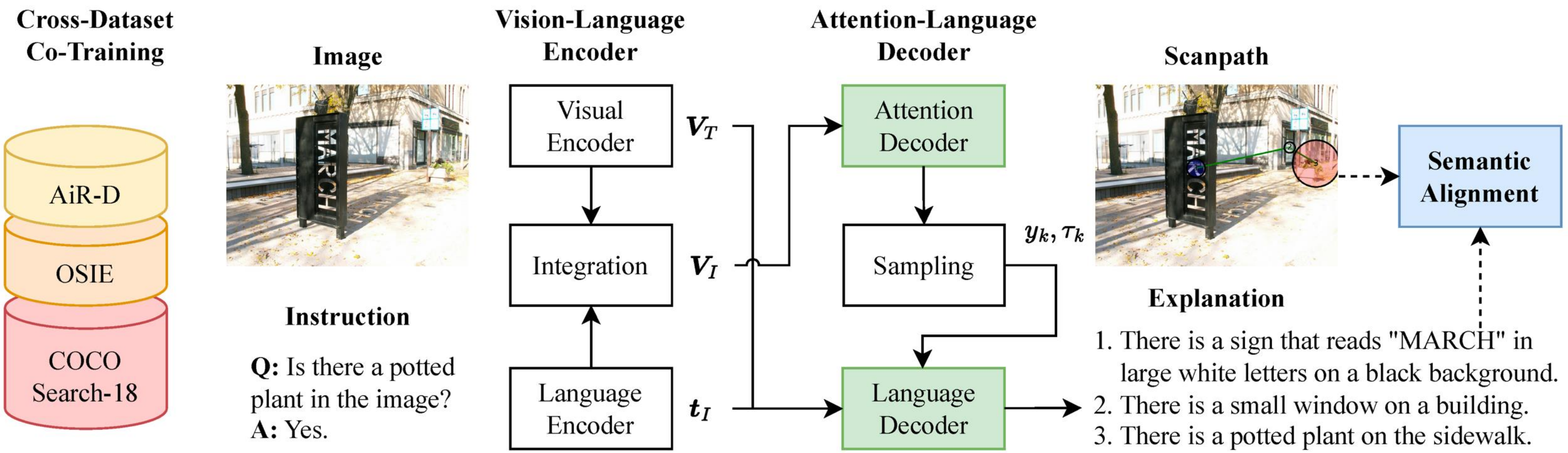
1. **Attention-Language Decoder (EXP):** Explanation generation with a language decoder (BLIP) to provide comprehensive semantic explanations for fixated regions.

GazeXplain – Method



2. **Semantic Alignment (ALN):** Compute and optimize the cosine similarity to ensure the consistency between predicted fixations, explanations and visual features.

GazeXplain – Method



3. **Cross-Dataset Co-Training (CT):** Enable models to learn from multiple datasets simultaneously with scaling the image and scanpath into same resolution and structure the task-specific instructions into the standard VQA format.

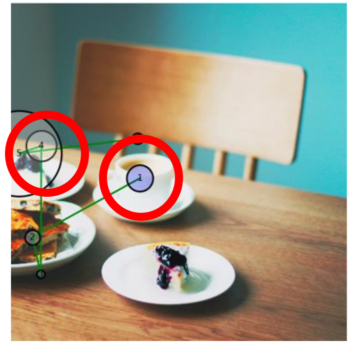
Scanpath Prediction Results

Method	Scanpath				Saliency			
	SM ↑	**MM ↑**	**SED ↓**	**SS ↑**	**CC ↑**	**NSS ↑**	**AUC ↑**	**sAUC ↑**
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
ChemLSTM	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**

Qualitative Examples

Question: What fruits are on the dessert on the left side of the photo? **Answer:** Blackberries

Ground Truth



Gazeformer

GazeXplain

1: There is a cup of coffee on a saucer.

5: There is a small piece of cake with a blueberry on top.

2: There is a white cup with a spoon and a spoon.

6: There is a plate of food with a fork and knife.

1: There is a cup of coffee on a tray.

6: There is a plate with a piece of cake on it.

Ablation Study

Method			Scanpath				Saliency				CIDEr-R
1. EXP	2. ALN	3. 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**

1. **Attention-Language Decoder (EXP):** 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.

Ablation Study

Method			Scanpath				Saliency				CIDEr-R
1. EXP	2. ALN	3. 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**

2. **Semantic Alignment (ALN):** This indicates the importance of semantic coherence in guiding attention prediction models.

Ablation Study

Method			Scanpath				Saliency				CIDEr-R
1. EXP	**2. ALN**	**3. 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**

3. **Cross-Dataset Co-Training (CT):** This demonstrates the effectiveness of integrating diverse data sources for robust scanpath prediction and explanation.

Summary of Contributions

1. **Joint Prediction & Explanation of Scanpaths**
   A novel task that jointly predicts and explains scanpaths, offering a natural interface for understanding the underlying rationales for visual behaviors and reasoning processes to perform tasks.

2. **Ground-Truth Explanations**
   Provides detailed fixation-level explanations across three public eye-tracking datasets, offering detailed insights into the cognitive processes driving gaze behavior.

3. **General Model Architecture**
   Features an attention-based language decoder that predicts both scanpaths and their explanations, bridging the gap between visual patterns and their semantic interpretations.

COLLEGE OF Science & Engineering
UNIVERSITY OF MINNESOTA

ECCV

GazeXplain: Learning to Predict Natural Language Explanations of Visual Scanpaths

European Conference on Computer Vision ECCV 2024

https://github.com/chenxy99/GazeXplain

This work is supported by NSF Grant 2143197.