LOGITGAZE: PREDICTING HUMAN ATTENTION USING SEMANTIC INFORMATION FROM VISION-LANGUAGE MODELS

Dmitry Lvov
Research Center of
the Artificial Intelligence Institute
Innopolis University,
Innopolis, Russia
d.lvov@innopolis.ru

Ilya Pershin
Research Center of
the Artificial Intelligence Institute
Innopolis University,
Innopolis, Russia
i.pershin@innopolis.ru

ABSTRACT

Modeling human scanpaths remains a challenging task due to the complexity of visual attention dynamics. Traditional approaches rely on low-level visual features, but they often fail to capture the semantic and contextual factors that guide human gaze. To address this, we propose a novel method that integrates LLMs and VLMs to enrich scanpath prediction with semantic priors. By leveraging word-level representations extracted through interpretability tools like the logit lens, our approach aligns spatial-temporal gaze patterns with high-level scene semantics. Our method establishes a new state of the art, improving all key scanpath prediction metrics by approximately 15% on average, demonstrating the effectiveness of integrating linguistic and visual knowledge for enhanced gaze modeling.

1 INTRODUCTION AND RELATED WORKS

Predicting human scanpaths remains crucial for computational vision, with applications spanning healthcare, autonomous systems, and retail. Recent LLM advances in multimodal reasoning create new opportunities to enhance gaze-driven attention modeling, as scanpaths provide implicit cues about task relevance and spatial awareness. Studies, such as GazeGPT(Konrad et al., 2024) and GazeReward(Lopez-Cardona et al., 2024) showcase gaze-enhanced LLM reasoning, while scanpath modeling improvements (Mondal et al., 2025; Yang et al., 2024) highlight persistent challenges in semantic reasoning for multimodal AI.

Traditional models often depend on annotated gaze datasets or object detectors (Cheng et al., 2024), limiting adaptability. GazeFormer (Mondal et al., 2023) addresses this through language-driven target encoding instead of object detection, while DeepGaze III (Kümmerer et al., 2022) effectively models scanpaths via fixation history integration. Interpretability advances like GazeXplain (Chen et al., 2024) further enable natural language explanations alongside predictions.

Vision-Language Models (VLMs) like LLaVA (Liu et al., 2023) have enabled richer multimodal integration for scanpath prediction. Studies (Neo et al., 2024) reveal progressive alignment between VLM visual representations and textual tokens, while techniques like logit-lens (Nostalgebraist, 2020) reveal how transformer representations evolve across layers.

Our framework advances scanpath prediction through LLaVA-derived semantic priors and logit-lens integration. By fusing visual, linguistic, and fixation data, we achieve state-of-the-art performance with 15% average metric improvements, demonstrating effective task-context integration at the patch level.

2 EXPLORATION

2.1 METHODOLOGY

Inspired by advancements in scanpath prediction, we propose a novel approach that leverages LLMs and the logit-lens technique to enhance fixation prediction by integrating semantic priors into the process. Traditional methods like GazeFormer have demonstrated the effectiveness of transformer-based architectures in predicting human attention, but our model extends this foundation by incorporating multimodal information from LLaVA and utilizing logit-lens to extract explicit word-level representations. Instead of relying solely on visual embeddings, we process images at the patch level, allowing each region to acquire a semantically meaningful representation (See Fig. 1 for details).

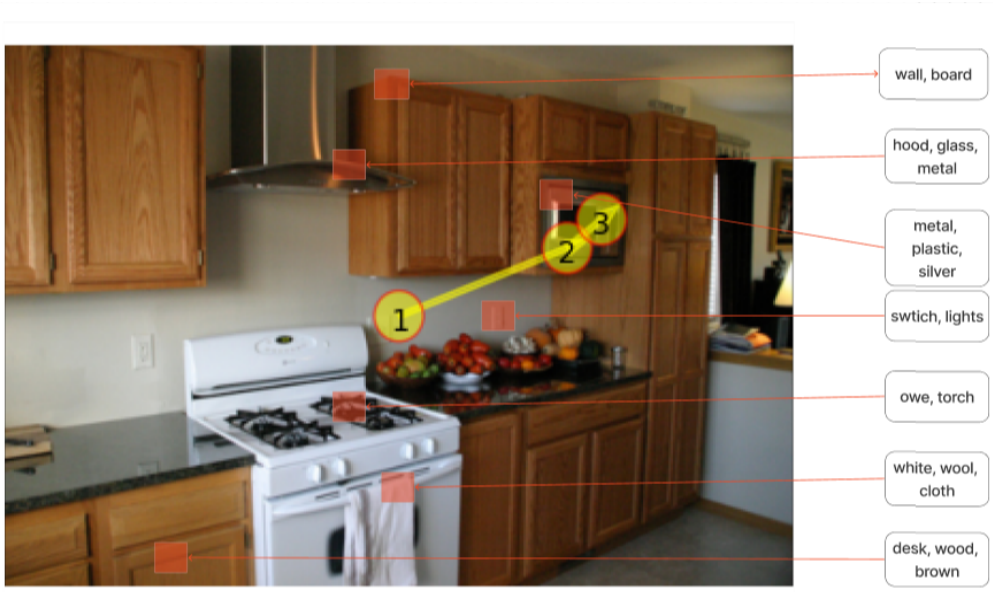


Figure 1: Visualization of a scanpath for the "microwave" search task, where fixations (yellow circles) represent the sequential attention shifts of the observer. Each image patch is decoded into semantic information—such as "wall, board," "hood, glass, metal," and "desk, wood, brown"—using an LLM, aligning gaze prediction with high-level scene understanding.

The input image $\mathcal{I} \in \mathbb{R}^{H\times W\times C}$ is first processed through LLaVA's visual backbone ViT-L/14, generating feature representations at the patch level:

$$\mathbf{F}_{\text{img}} = \mathcal{F}_{\text{LLaVa}}(\mathcal{I}) \in \mathbb{R}^{M\times D}$$

where $M = 576$ (24×24 patches) represents the number of extracted image patches, and $D = 4096$ is the dimensionality of each patch. These features are refined by a 6-layer transformer encoder, using 2D positional embeddings to preserve spatial structure:

$$\mathbf{F}'_{\text{img}} = \mathcal{T}_{\text{enc}}(\mathbf{F}_{\text{img}}) + \mathbf{P}_{\text{2D}}$$

To introduce linguistic context, each patch is mapped to a word distribution via logit-lens. For each feature $\mathbf{f}_m$, a softmax projection assigns token probabilities:

$$P(w_j|\mathbf{f}_m) = \text{softmax}(\mathbf{W}\mathbf{f}_m)$$

where $\mathbf{W}$ is a learnable projection matrix. The most probable word w^*_m is selected as $\arg\max_{w_j} P(w_j|\mathbf{f}_m)$, generating semantic representations $\mathbf{F}_{\text{sem}}$. These are further refined by LLaVA's multimodal encoder:

$$\mathbf{F}'_{\text{sem}} = \mathcal{F}_{\text{LLaVa}}(\mathbf{F}_{\text{sem}})$$

The search query (e.g., "cup") is encoded using a pretrained BERT-base model:

$$\mathbf{F}_{\text{task}} = \mathcal{F}_{\text{LM}}(\text{query})$$

The modalities are fused via concatenation, cross-attention, and projection:

$$\mathbf{F}_{\text{joint}} = \mathcal{F}_{\text{fusion}}(\mathbf{F}'_{\text{img}}, \mathbf{F}'_{\text{sem}}, \mathbf{F}_{\text{task}})$$

A 4-layer transformer decoder, initialized randomly, autoregressively predicts the scanpath:

$$(x_t, y_t, d_t) \sim p(x, y, d|\mathbf{F}_{\text{joint}})$$

where x_t, y_t denote fixation coordinates, d_t represents fixation duration, and $p(\text{valid}|t)$ determines whether the fixation is valid or padded. The predicted scanpath consists of a sequence $\{(x_t, y_t, d_t)\}_{t=0}^{T}$, where T is the number of fixations.

2.2 DATABASE

We will use the COCO-Search18 dataset Chen et al. (2021), which is designed to study goal-directed visual search behavior. Unlike datasets focused on free-viewing or bottom-up attention(Chen et al., 2022), COCO-Search18 captures human eye movements during active search tasks. It consists of over 300,000 fixations collected from 10 participants searching for 18 target objects in 6202 natural images. This dataset provides detailed insights into how humans allocate attention during specific tasks.

To ensure a fair comparison, we use the same metrics as the original framework. Sequence Score (SS)(Yang et al., 2020) evaluates how well predicted fixation sequences match human scanpaths, while Semantic Sequence Score (SemSS)(Yang et al., 2022) extends this by considering fixated objects instead of fixation clusters. Fixation Edit Distance (FED) and Semantic Fixation Edit Distance (SemFED) measure the differences between predicted and actual scanpaths, with SemFED emphasizing semantic alignment. Multimatch (MM)(Anderson et al., 2015) assesses spatial scanpath similarity by comparing shape, direction, length, and position. Higher SS, SemSS, and MM indicate better alignment with human gaze patterns, while lower FED and SemFED suggest more accurate fixation predictions.

2.3 EXPERIMENTS AND RESULT

The results presented in Tables 1 and 2 show that our approach, LogitGaze, outperforms the GazeFormer method across all evaluated metrics for both the target-present and target-absent tasks. Specifically, LogitGaze demonstrates significant improvements in the SS and SemSS metrics. In the target-present task (see Table 1 (a)), LogitGaze achieves an SS of 0.506 (without duration) and a SemSS of 0.525, both surpassing the GazeFormer. Similarly, in the target-absent task (Table 2), LogitGaze achieves a higher SS and SemSS compared to the state-of-the-art.

Table 1: Comparison of model performance trained on the target-present search task. Metrics are provided for both the target-present and target-absent search tasks, with the best performance highlighted in bold.

(a) Tested on TP

Model	**SS ↑**		**SemSS ↑**		**FED ↓**		**SemFED ↓**		**MM ↑**
	w/o dur	w/ dur	w/o dur	w/ dur	w/o dur	w/ dur	w/o dur	w/ dur	
GazeFormer	0.491	0.441	0.495	0.456	2.073	9.887	1.896	7.569	0.816
LogitGaze	**0.506**	**0.453**	**0.525**	**0.474**	**2.035**	**9.613**	**1.726**	**7.295**	**0.862**

(b) Tested on TA

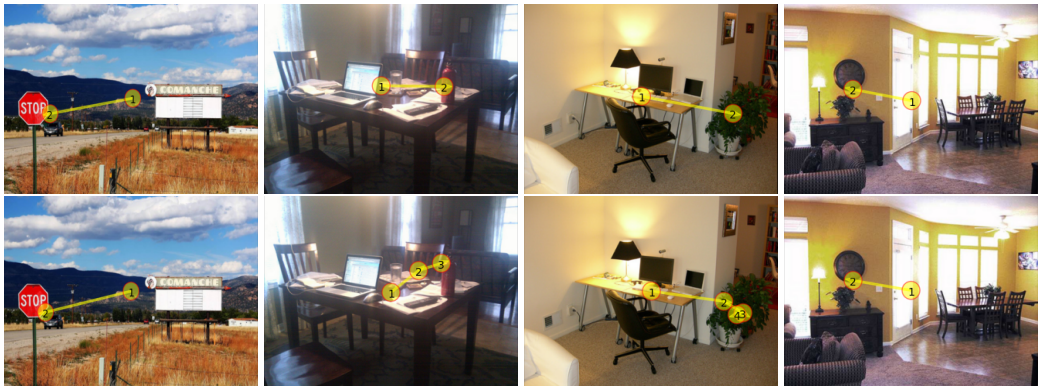
Model	**SS ↑**		**SemSS ↑**		**FED ↓**		**SemFED ↓**		**MM ↑**
	w/o dur	w/ dur	w/o dur	w/ dur	w/o dur	w/ dur	w/o dur	w/ dur	
GazeFormer	0.362	0.351	0.377	0.359	6.002	18.378	4.142	14.642	0.832
LogitGaze	**0.378**	**0.373**	**0.405**	**0.401**	**5.539**	**16.211**	**3.942**	**13.937**	**0.849**

LogitGaze also shows superior performance in the SemFED metric, with a score of 1.726 (without duration) for the target-present task, indicating that it is more effective at predicting semantically relevant fixations. These improvements are attributed to the model's use of semantic information, which enables it to focus on meaningful regions of the scene, aligning more closely with human attention patterns.

Table 2: Comparison of model performance trained and tested on the target-absent search task. Metrics are provided for both cases, with the best performance highlighted in bold.

Model	SS ↑		SemSS ↑		FED ↓		SemFED ↓		MM ↑
	w/o dur	w/ dur	w/o dur	w/ dur	w/o dur	w/ dur	w/o dur	w/ dur	
GazeFormer	0.371	0.371	0.384	0.369	5.078	17.246	3.690	13.550	0.835
LogitGaze	**0.392**	**0.399**	**0.416**	**0.421**	**4.792**	**15.212**	**3.342**	**12.691**	**0.856**

Figure 2 visualizes the alignment between human scan paths and LogitGaze's predictions, demonstrating the model's ability to capture human-like attention patterns. The use of semantic information helps LogitGaze focus on meaningful regions within the scene, resulting in more structured and coherent gaze predictions that closely resemble natural human viewing behavior.



(a) Task: sign **(b)** Task: bottle **(c)** Task: plant **(d)** Task: clock

Figure 2: Comparison of human scan paths (top) and LogitGaze model predictions (bottom) for different tasks. The model demonstrates strong alignment with human gaze patterns, accurately capturing fixation points and transitions.

3 Discussion and Conclusion

We introduced LogitGaze, a novel scanpath prediction model that integrates semantic information from VLMs to enhance fixation modeling. Unlike previous methods that rely primarily on visual embeddings, LogitGaze incorporates explicit semantic priors extracted from logit-lens and multimodal representations from LLaVA.

Our results demonstrate that enriching scanpath modeling with semantic information improves all major metrics by approximately 15%, leading to more human-like gaze patterns. To assess the effectiveness of our approach, we compared LogitGaze with GazeFormer, the current state-of-the-art in scanpath prediction.

However, the logit-lens mechanism also introduces challenges. While it provides transparent word-level representations, it can generate noisy or irrelevant activations, potentially affecting prediction quality. Future work will focus on filtering ambiguous word associations and optimizing the fusion of linguistic and visual features to further refine scanpath modeling.

Acknowledgments

All authors were supported by the Research Center of the Artificial Intelligence Institute of Innopolis University.

REFERENCES

Nicola C Anderson, Fraser Anderson, Alan Kingstone, and Walter F Bischof. A comparison of scanpath comparison methods. *Behavior research methods*, 47:1377–1392, 2015.

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

Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, 11(1):1–11, 2021.

Yupei Chen, Zhibo Yang, Souradeep Chakraborty, Sounak Mondal, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Characterizing target-absent human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5031–5040, 2022.

Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Robert Konrad, Nitish Padmanaban, J Gabriel Buckmaster, Kevin C Boyle, and Gordon Wetzstein. Gazegpt: Augmenting human capabilities using gaze-contingent contextual ai for smart eyewear. *arXiv preprint arXiv:2401.17217*, 2024.

Matthias Kümmerer, Matthias Bethge, and Thomas SA Wallis. Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, 22(5):7–7, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Angela Lopez-Cardona, Carlos Segura, Alexandros Karatzoglou, Sergi Abadal, and Ioannis Arapakis. Seeing eye to ai: Human alignment via gaze-based response rewards for large language models. *arXiv preprint arXiv:2410.01532*, 2024.

Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1441–1450, June 2023.

Sounak Mondal, Seoyoung Ahn, Zhibo Yang, Niranjan Balasubramanian, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Look hear: Gaze prediction for speech-directed human attention. In *European Conference on Computer Vision*, pp. 236–255. Springer, 2025.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models, 2024. URL https://arxiv.org/abs/2410.07149.

Nostalgebraist. Interpreting gpt: the logit lens, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning, 2020. URL https://arxiv.org/abs/2005.14310.

Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Target-absent human attention. In *European Conference on Computer Vision*, pp. 52–68. Springer, 2022.

Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Ruoyu Xue, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Unifying top-down and bottom-up scanpath prediction using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1683–1693, 2024.