Gaze-Language Alignment for Zero-Shot Prediction of Visual Search Targets from Human Gaze Scanpaths

Supplementary Material

Sounak Mondal[1,2*] Naveen Sendhilnathan[2] Ting Zhang[2] Yue Liu[2] Michael Proulx[2]
Michael Louis Iuzzolino[2] Chuan Qin[2] Tanya R. Jonker[2]

[1]Stony Brook University [2]Meta Reality Labs Research

In this document, we provide additional experiments, visualizations, analysis, and details of our work on zero-shot gaze target prediction using GLAM and GLAD. The specific sections of this document are listed below.

- We present empirical evidence of CLIP [12]-based contrastive learning being ineffective when used to train GLAM instead of GLAD (Section 1).
- We detail our procedure of prompting a large language model (LLM) to generate visual search process descriptions for object categories (Section 2).
- We examine how our method predicts target categories for scanpaths using only a certain percentage of initial fixations, despite being trained on only complete scanpaths (Section 3).
- We provide additional metrics for the quantitative experiments discussed in the main text (Section 4).
- We provide additional implementation and design details for GLAM and GLAD (Section 5).
- We discuss the insights guiding our design of GLAD's pre-training stage (Section 6).
- We analyze the effects of parafoveal information on the gaze target prediction capability of GLAM (Section 7).
- We investigate the category-distinguishability of scanpaths in COCO-Search18. (Section 8)

1. Inefficacy of CLIP's contrastive learning strategy for Gaze-Language Alignment

CLIP [12]'s contrastive learning strategy treats similarity as binary, i.e. "positive pairs" are perfectly similar to each other, while "negative pairs" are completely dissimilar. When naively applied to our method, CLIP's contrastive learning strategy assumes that $\mathbf{F}_{gaze}$ and $\mathbf{F}_{lang}$ for the same batch sample are completely similar, while being entirely dissimilar to representations from other batch samples. We empirically show how this formulation of CLIP is incompatible with the Gaze Target Prediction problem. We substitute our novel strategy, GLAD, with CLIP-style training in the "GLAM + GLAD + LLM-Generated Prompts" variant. Specifically, the pre-training step is discarded and the training loss $\mathcal{L}_{align}$ defined in Equations 4-6 in the main text is modified as follows:

$$\mathcal{L}_{align} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{\texttt{SIM}(\mathbf{F}_{gaze}^{(i)}, \mathbf{F}_{lang}^{(i)})}{\sum_{l=1}^{N} \texttt{SIM}(\mathbf{F}_{gaze}^{(i)}, \mathbf{F}_{lang}^{(l)})} \quad (1)$$

$$\text{where } \texttt{SIM}(a, b) = exp(\langle a, b \rangle / \tau') \quad (2)$$

Here, τ' is a learnable temperature parameter and N is the batch size. The results for this ablation are in Table 1. As observed, GLAM's target prediction performance declines in both Zero-Shot and Fully-Supervised setups when trained with CLIP's contrastive learning strategy, performing worse than CWCL [13], and far worse than our GLAD training strategy. We attribute this to the fact that COCO-Search18 [3] contains 18 target categories, which is smaller than typical batch sizes. This leads to multiple batch elements sharing the same Language Encoder prompt due to their association with the same ground truth object category, contradicting CLIP's assumption that each language prompt (or text input) in a batch is unique.

Training Strategy	Target Prediction Acc. (%)		Presence Prediction Acc. (%)
	Zero-Shot	Fully Supervised	
CLIP [12]	8.39	13.72	83.12
CWCL [13]	23.82	52.95	**84.46**
GLAD	**30.17**	**58.22**	83.48

Table 1. A comparison of training strategies for GLAM along with LLM-generated prompts in terms of target prediction and target presence prediction accuracies. While our proposed strategy, GLAD, significantly outperforms the other two strategies in terms of target prediction, CLIP [12] performs the worst by a considerable margin.

*Work done during an internship at Meta Reality Labs Research

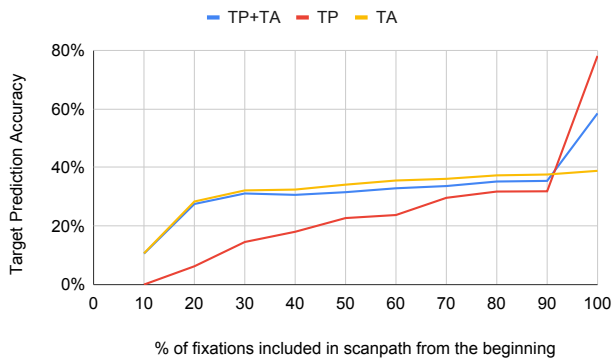

Figure 1. Evaluation of "GLAM + GLAD + LLM-Generated Prompts" model variant in terms of target prediction accuracy, across different evaluation conditions: combination of both target-present and target-absent scanpaths (TP+TA), only target-present scanpaths (TP), and only target-absent scanpaths (TA). This evaluation was performed using varying percentages of initial fixations included in the scanpath input to the model.

2. Details of LLM Prompting

We prompted the LLM using a text prompt to generate JSON objects for each of the 2,456 object categories used in GLAD pre-training, including the 18 COCO-Search18 categories used in GLAD training. An example prompt to the LLM for category "bottle":

```
Describe how I would search for
"bottle" in a scene in terms of
appearance, and contextual cues.
Respond in JSON format without
mentioning the name of the object.
```

These JSON objects were processed to form *search process descriptions* describing the visual search process using the following template: `Object Name:[category name]. Appearance:[appearance info]. Contextual Cues:[contextual info].` These text strings form the visual search descriptions that are used as input to GLAM's Language Encoder.

3. GLAM's predictions for incomplete scanpaths

We investigated how GLAM (specifically, the "GLAM + GLAD + LLM-Generated Prompts" variant) predicted target categories for scanpaths, but instead of including all fixations as input to GLAM, we provided only a certain percentage of fixations from the beginning. This percentage ranged from 10% to 100%. This ablation probed the model's ability to predict the target category for incomplete scanpaths despite being trained on only complete scanpaths, and was performed in the Fully Supervised setting. Plots showing the trends in change of target prediction accuracy with varying percentages of fixations included in scanpath from the beginning are in Fig. 1. It was observed that for target absent scanpaths, the model's performance increased substantially to approximately 30% after the first 20% fixations, indicating that the model picks up on contextual cues early in the scanpath. Since target-present scanpaths are shorter than target-absent scanpaths (average length of target-present test scanpaths is 2.89 whereas average length of target-absent test scanpaths is 5.85), target-present scanpaths required at least 90% fixations (i.e. 2-3 fixations on average) for the model to predict the correct category.

4. Additional metrics for Experimental Results

In the Zero-Shot setting, where we evaluate zero-shot target prediction on each COCO-Search18 category in an N-fold cross-validation manner, we report average recall across all categories, referred to as "Recall", along with the accuracy metric. In the fully-supervised setting, we evaluate target prediction and target presence prediction by reporting precision (P), recall (R), and F_1 score (F_1). These metrics are averaged across 18 target categories for target prediction and across the two possibilities (present/absent) for target presence prediction, and reported alongside the previously reported accuracy metrics. We report the aforementioned additional metrics for the results in Table 1 of the main text in Table 2. We observe that the trends in Table 1 are repeated for the additional metrics. Specifically, "GLAM+GLAD+LLM-Generated Prompts" variant outperforms other variants in almost every metric for target prediction, whereas "GLAM+GLAD+Category-Label based prompts" variant achieves the best metrics for target presence prediction. Note that the aforementioned additional metrics for COCO-Search18 have not been reported previously for baselines GST [11], GazeGNN [14] and BoVW [1]. Since implementations of GST [11], GazeGNN [14] (specifically, Nishiyasu *et al.*'s implementation[1] adapted for COCO-Seach18) and BoVW [1] are not publicly available, we are unable to report the aforementioned additional metrics for these baselines.

5. Implementation Details

In this section, we provide additional implementation details of GLAM and GLAD.

GLAM. A dropout of 0.3 is applied to the Gaze Encoder layers while a dropout of 0.2 is applied to $\mathbf{F}_{image}$ and $\mathbf{F}_{fix}$ prior to input to the gaze encoder. Patch size p for $\mathbf{F}_{image}$

[1]Our implementation of GazeGNN for COCO-Search18 yielded poor performance, i.e. 30.28% (compared to 38.77% as reported by Nishiyasu *et al.*) for target prediction accuracy, and 63.13% (compared to 73.28% as reported by Nishiyasu *et al.*) for target presence prediction accuracy. Hence, to avoid confusion, we refrain from reporting the results of our implementation in the tables of this document.

			Target Prediction						Presence Prediction			
			Zero-Shot		Fully-Supervised				Fully-Supervised			
Architecture	Training Strategy	Prompts	Acc. (%)	Recall	Acc. (%)	P	R	F_1	Acc. (%)	P	R	F_1
Random	N/A	N/A	5.16	0.05	5.16	0.051	0.051	0.050	50.23	0.502	0.502	0.502
GLAM Gaze Encoder	Classification	N/A	N/A	N/A	58.19	***0.623***	0.551	0.566	84.25	0.848	0.843	0.842
GLAM	CWCL [13]	Category Label	22.34	0.226	57.95	0.575	0.585	0.576	***85.07***	***0.851***	***0.851***	***0.851***
GLAM	CWCL [13]	LLM-Generated	23.82	0.241	52.95	0.508	0.503	0.507	84.46	0.845	0.846	0.845
GLAM	**GLAD**	LLM-Generated	***30.17***	***0.316***	***58.22***	0.573	***0.586***	***0.577***	83.48	0.838	0.835	0.836

Table 2. Additional metrics – Precision (P), Recall (R), F_1 score (F_1) – are added to the accuracy metrics (denoted here as "Acc. (%)") reported in Table 1 of the main text for comparing different methods for target prediction and target presence prediction (excluding methods without any public implementation or published metrics) on a combination of both target-present and target-absent scanpaths of the COCO-Search18 test set. Both Zero-Shot and Fully Supervised setups are used for evaluation. Bold indicates the methods proposed in this research, and bold italics highlight the best prediction metric in each evaluation setup.

Model variant	No. of parameters	Training latency (sec/batch)	Inference latency (sec/sample)
GLAM Gaze Encoder + Classification	37M	1.1	0.091
GLAM + CWCL + Category Label	147M	1.6	0.115
GLAM+CWCL+LLM-Generated Prompts	147M	2.0	0.147
GLAM+GLAD+LLM-Generated Prompts	147M	1.9	0.132

Table 3. Model complexity and latency metrics for all GLAM variants. All model variants except the classification variant contains an MP-Net language encoder which contributes to 110M parameters. All variants are trained with a batch size of 256. During inference, we process one sample (image-scanpath pair) at a time."GLAM+GLAD+LLM-Generated Prompts" variant is the main proposed model of our paper.

is 16, and is used in the fixation encoder to obtain patch embeddings of fixation patches. Output dimension of the Language Encoder d_{lang} is 768. The output embedding from the attentional pooling layer $\mathbf{f}_{attpool}$ having dimensionality $d_{lang} = 768$ gets projected by $\mathbf{f}^{(1)}_{lang}$ to $\mathcal{F}_{search} \in \mathbb{R}^d$ where $d = 256$. $\mathbf{f}^{(2)}_{lang}$ subsequently projects $\mathcal{F}_{search}$ to another d-dimensional embedding, which is then added with $\mathcal{F}_{search}$ and normalized to yield $\mathbf{F}_{gaze} \in \mathbb{R}^d$. The two FFN layers in gaze encoder sub-modules each consist of two fully connected layers: the first layer projecting d-dimensional vectors to d_{ffn}-dimensional vectors to which ReLU [9] activation is applied, followed by the second layer projecting the resultant d_{ffn}-dimensional vectors to d-dimensional vectors. The hyperparameter d_{ffn} was set to 512.

GLAD Pre-training. For optimization during the pre-training stage of GLAD, we used an AdamW [6] optimizer with a learning rate of 1e-4 and weight decay of 0.01. The pre-training process was performed for 500 epochs with a batch size of 128. The learnable temperature parameters τ and τ_C were initialized to 0.07. Note that the CLIP text encodings used in this stage have a dimensionality of 512, and these are projected by $\mathbf{f}_{cat}$ and subsequently normalized to yield $\mathcal{F}_{cat} \in \mathbb{R}^d$ where $d = 256$. In this stage, $\mathcal{F}_{search}$ generated by $\mathbf{f}^{(1)}_{lang}$ is also normalized.

GLAD Training. For optimization during the training stage of GLAD, we used an AdamW [6] optimizer with a learning rate of 1e-4 and weight decay of 0.01. All variants of GLAM were trained for a maximum of 30 epochs with a batch size of 256. The learnable temperature parameters τ' and τ'_C were initialized to 0.07. Our fixation-based masking strategy creates higher number of unique training samples than images, thereby allowing us to train both visual and language encoders with a limited number of images without overfitting.

Model Complexity. The model complexity metrics for all GLAM variants are in Table 3. The "GLAM Gaze Encoder + Classification" variant contains 37M parameters. Other GLAM variants share the same 147M-parameter architecture with 110M parameters coming from MP-Net.

Latency Analysis. The model latency metrics for all GLAM variants are in Table 3. We train all variants of GLAM with a batch size of 256, and for a realistic inference setting, process one sample (image-scanpath pair) at a time. Time metrics are similar for the "GLAM+CWCL+LLM-Generated Prompts" (training latency:2s/batch, inference latency:147ms/sample) and "GLAD+GLAD+LLM-Generated Prompts" (training latency: 1.9s/batch, inference latency: 132ms/sample) variants. The "GLAM + CWCL + Category Label"

variant (training latency: 1.6s/batch, inference latency: 115ms/sample) and "GLAM Gaze Encoder + Classification" variant (train latency: 1.1s/batch, inference latency: 91ms/sample) variants are faster due to shorter label-based prompts and the absence of MP-Net, respectively.

6. Insights guiding the design of GLAD Pre-training Phase

In this section, we discuss the insights guiding our design of GLAD pre-training stage. We have shown empirically in the main text (Sec. 4.1) that category label embeddings are inadequate, potentially because existing pre-trained language encoders encode generic object properties, not explicit human search behavior. In this regard, visual search descriptions containing explicit cues used by humans for search are more appropriate. However, it is essential for a gaze target prediction model to know which cues are necessary for category disambiguation (*e.g*. "bowl" vs. "bottle"), especially for zero-shot inference. This necessitates the pre-training stage of GLAD where GLAM learns to focus on target-discriminative features in search descriptions for search target categories that might even lack gaze annotations. When encountering a scanpath originating from search for a novel category lacking prior gaze annotations (*i.e*. during zero-shot inference), the GLAD pre-training stage would therefore allow the model to leverage its pre-trained knowledge of the target-discriminative features in the novel category's search description embedding to correlate with the gaze embedding of the scanpath and consequently disambiguate the correct target category. CWCL [13] used for learning during the pre-training stage encourages the model to be aware of inter-category similarities (*e.g*. that "chair" and "couch" are very similar).

7. Effects of Parafoveal Information on Gaze Target Prediction

COCO-Search18 was collected with horizontal and vertical visual angles of 54° and 35°, respectively. Our Gaze encoder views images as 32 horizontal patches and 20 vertical patches, so the optimal peripheral window of size=1 (as shown in Sec. 4.3 of the main text), spanning 3 patches horizontally and vertically, covers visual angles $(3/32)\times 54^\circ = 5.06^\circ$ horizontally and $(3/20)\times 35^\circ = 5.25^\circ$ vertically – lower than parafovea's anatomical span of $\sim 10^\circ$ but greater than foveal span of $\sim 2^\circ$. This suggests that parafoveal cues aid search, but critical information comes from within $\sim 5^\circ$ visual angle – likely due to eccentricity effects.

8. Category-distinguishability of Search Scanpaths in COCO-Search18

A COCO-Search18 data sample consists of a scanpath $\mathbf{S}$ of eye fixations made by a human instructed to search for a target category $\mathcal{C}$ in an image $\mathbf{I}$ containing either a single (Target-Present) or zero (Target-Absent) object belonging to $\mathcal{C}$. Hence, COCO-Search18 provides precise correspondence of the real search target with the pair $(\mathbf{S}, \mathbf{I})$. However, we analytically investigated whether scanpaths of a category $\mathcal{C}$ in COCO-Search18 are indeed distinguishable from scanpaths of other categories (*i.e*. *not* $\mathcal{C}$).

Now, to validate our hypothesis that scanpaths of category $\mathcal{C}$ are distinguishable from scanpaths of other search categories within COCO-Search18, we show that for the same visual stimulus (*i.e*. image $\mathbf{I}$) and scanpaths from multiple participants searching for categories $\mathcal{C}_1$, $\mathcal{C}_2$, ..., $\mathcal{C}_n$, intra-category scanpath similarity scores are statistically significantly higher than inter-category scanpath similarity scores. In other words, we show that scanpaths from multiple participants searching for the *same* target category in the same image are more similar to each other than with scanpaths from those participants searching for a *different* target category in the same image. For every image-target category pair, COCO-Search18 contains one scanpath from each participant in the participant pool of 10 individuals, enabling us to perform this analysis.

We selected images from both target-present and target-absent scenarios for each of which scanpaths (from all 10 participants) for multiple categories were available in COCO-Search18. We used Semantic Sequence Score (SemSS) [15] metric, a non-parametric metric widely used in the evaluation of state-of-the-art scanpath prediction models [2, 8, 15]. SemSS computes similarities between two scanpaths by converting those scanpaths into strings of fixated objects in the scene, and consequently using a string matching algorithm [10] to measure similarity between that pair of strings. SemSS can be used both with and without fixation duration component, similar to ScanMatch [4]. Higher SemSS score indicates greater similarity.

We conducted an analysis of SemSS similarity scores for both Target-Present and Target-Absent scenarios (separately, not combined), comparing intra-category SemSS similarity scores (scanpaths from multiple participants searching for the *same* target category in the same image) and inter-category SemSS similarity scores (for scanpaths from multiple participants searching for two *different* target categories in the same image) using the non-parametric Mann-Whitney U test [7]. Results revealed that for both Target-Present and Target-Absent scenarios (both with and without the fixation duration component), intra-category SemSS similarity scores are *statistically significantly larger* ($p < 0.05$) than inter-category SemSS sim-

ilarity scores. This finding suggests that scanpaths corresponding to a category are distinct from those of other categories and an ideal model will be able to discern an unambiguous target category for a given scanpath and an image. We also conducted the same analysis using SemFED [8], another non-parametric scanpath similarity metric based on edit distances [5], instead of SemSS. With SemFED similarity metric, we again found that a scanpath is statistically significantly more similar to other scanpaths originating from search for the *same* category, than it is to scanpaths originating from search for a *different* category.

However, please note that one cannot simply predict the target category for a test scanpath using a scanpath similarity metric like SemSS and SemFED, as this would also require scanpaths from other participants for the same search category and image, which we usually do not have access to in real-world application scenarios.

References

[1] Michael Barz, Sven Stauden, and Daniel Sonntag. Visual search target inference in natural interaction settings with machine learning. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–8, 2020. 2

[2] Xianyu Chen, Ming Jiang, and Qi Zhao. Gazexplain: Learning to predict natural language explanations of visual scanpaths. *arXiv preprint arXiv:2408.02788*, 2024. 4

[3] 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):8776, 2021. 1

[4] Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D Gilchrist. Scanmatch: A novel method for comparing fixation sequences. *Behavior research methods*, 42(3): 692–700, 2010. 4

[5] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710, 1965. 5

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 3

[7] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947. 4

[8] 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*, 2023. 4, 5

[9] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 3

[10] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970. 4

[11] Takumi Nishiyasu and Yoichi Sato. Gaze scanpath transformer: Predicting visual search target by spatiotemporal semantic modeling of gaze scanpath. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 625–635, 2024. 2

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021. 1

[13] Rakshith Sharma Srinivasa, Jaejin Cho, Chouchang Yang, Yashas Malur Saidutta, Ching-Hua Lee, Yilin Shen, and Hongxia Jin. Cwcl: Cross-modal transfer with continuously weighted contrastive loss. *Advances in Neural Information Processing Systems*, 36, 2023. 1, 3, 4

[14] Bin Wang, Hongyi Pan, Armstrong Aboah, Zheyuan Zhang, Elif Keles, Drew Torigian, Baris Turkbey, Elizabeth Krupinski, Jayaram Udupa, and Ulas Bagci. Gazegnn: A gaze-guided graph neural network for chest x-ray classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2194–2203, 2024. 2

[15] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Target-absent human attention. In *European Conference on Computer Vision*, 2022. 4