Gaze-to-text Generation: Beyond Categorical Decoding of Human Attention

Sounak Mondal[1], Dimitris Samaras[1], Gregory Zelinsky[1], and Minh Hoai[2]

[1] Stony Brook University, Stony Brook, New York, USA
[2] Australian Institute for Machine Learning, Adelaide University, Australia

Abstract. We introduce a novel learning problem: decoding gaze into natural language descriptions of human goals across diverse visual tasks. Unlike prior work, which frames gaze decoding as a discriminative task over predefined categories, we formulate it as a generative learning problem: training a model to produce free-form descriptions that capture the rich nuances and open-ended nature of human intentions beyond fixed labels. To this end, we introduce *Gazette*, the first gaze-to-text decoding framework. Based on multimodal large language models (MLLMs), Gazette learns to decode gaze scanpaths into natural language for goals that may extend beyond categorical labels and require articulation in natural language. To help Gazette filter out individual differences in gaze behavior and learn the goal-specific spatiotemporal dynamics crucial for generating accurate natural language goal descriptions, we propose a novel strategy that leverages the encyclopedic knowledge and reasoning abilities of a large language model to synthesize natural language explanations of goal-directed attentional behavior called *think-aloud transcripts*. Instruction tuning on these synthetic narratives allows Gazette to achieve state-of-the-art performance in gaze decoding across multiple tasks, demonstrating its generalizability and versatility, thereby enabling gaze to serve as a powerful, non-intrusive cue for inferring human goals and intentions in diverse scenarios. Code is available at `https://github.com/cvlab-stonybrook/Gazette`.

Keywords: Multimodal LLM · Instruction Tuning · Human Attention

1 Introduction

The ability to infer and analyze human attention and intention underpins a wide range of applications, including user engagement analysis [9, 30], driver monitoring [65], human–robot interaction [34], hands-free assistive technologies [50], and psychological diagnosis [36]. One practical approach is to track a user's gaze and decode it to estimate their underlying intent and goals. This process, known as gaze decoding, offers a non-intrusive and scalable alternative to neural decoding methods based on EEG [16, 25], fMRI [64, 74], MEG [6, 17], and ECoG [15, 32], which typically require bulky, costly, and sometimes invasive equipment.

Gaze decoding is an important research problem, and existing approaches [4, 45, 48, 54–56, 68] formulate gaze decoding as a classification task over predefined

categories, addressing questions such as "Does the gaze behavior reflect free viewing or goal-directed behavior?" (i.e., binary classification) or "Which object is the user searching for?" (i.e., multi-class classification over a fixed object set), with the model outputting a single label from a predefined category space.

However, such formulations are inherently constrained by a fixed label set. When the number of categories is small, outputs tend to be overly coarse and lack the fine-grained detail required for many downstream applications; when the label space is large, the model demands extensive annotated data to achieve adequate coverage. More fundamentally, any predefined category set is limited in expressive power and cannot capture the richness of natural language. In this work, we move beyond category-based gaze decoding and study the task of inferring a person's goal from attention sequences, where the output may range from simple open-vocabulary object labels (e.g., "bowl", "car") to free-form, context-rich descriptions such as "red bowl to the left of the cup in the middle."

This natural language-based formulation of gaze decoding enables a broad range of real-world applications. In e-commerce, descriptive language-based representations of gaze patterns allow systems to capture nuanced user intent when interacting with novel products that cannot be described by a fixed taxonomy. Instead of merely identifying attention to a "chair," decoding gaze into descriptions such as "mid-century walnut armchair with woven rattan back" provides richer semantic signals about style, material, and aesthetic preference. In educational settings, gaze descriptions such as "the equation's denominator term" or "the shaded region under the curve" may reveal conceptual focus or misunderstanding more precisely than coarse object tags. In assistive technologies using augmented reality glasses, generating descriptions such as "the user is approaching and looking at the top of the escalator" can help infer navigation intent. By decoding gaze behavior into expressive natural language, the model unlocks fine-grained, open-vocabulary semantic signals that enable reasoning over new, rare, or previously unseen concepts and improve adaptability across domains.

To operationalize this formulation, we propose ***Gaze-to-text*** (*Gazette*), a multimodal large language model (MLLM)-based generative framework that decodes goal-directed human attention across diverse gaze behaviors. Gazette takes as input an image I and a language instruction $T_{GazeDec}$ textually representing the gaze scanpath S, and the output is text D describing the *cognitive context* of the human. We define cognitive context at two levels: a coarse *behavior type* (visual search, object referral, or visual question answering) and a fine-level *stimulus*, that triggers attentional movements. The latter ranges from a search target's category label to free-form text such as a referring expression or question.

Gazette is essentially an MLLM for the task of gaze decoding, yet the standard training paradigm for such models does not directly transfer. The conventional approach—instruction tuning [35, 40]—starts from an MLLM, constructs labeled instruction data, and fine-tunes the model using these supervision signals. In the context of gaze decoding, the instruction-tuning data would consist of N triplets of an image, a gaze scanpath, and the corresponding description of the goal, $\{(I_i, S_i, D_i)\}_{i=1}^{N}$, collected by first defining a goal D_i (e.g., "red car on right") for a

given image I_i and then recording the gaze behavior S_i of a participant pursuing that goal. However, adapting an MLLM to perform gaze decoding—i.e., learning a mapping from an input image and scanpath (I_i, S_i) to the goal description D_i—is fundamentally challenging because the supervision is inherently weak. Although the goal description D_i reflects the task-driven objective, the observed gaze signal S_i is not purely goal-driven; it is also substantially influenced by individual differences among participants. In other words, for the same goal D_i, different participants may exhibit different scanpaths S_i's. The model must therefore learn a mapping from partially confounded inputs to clean task-level outputs. This disentanglement problem is further exacerbated by the lack of domain-specific priors in MLLMs [19,22,43], particularly regarding gaze behavior, making it difficult to separate goal-relevant signals from participant-specific variation. As a result, models trained solely through standard instruction tuning exhibit suboptimal performance on the gaze decoding task.

Building on this insight, we argue that effective gaze decoding requires factoring out the goal-driven signal from participant-specific confounding factors. At first glance, this appears intractable—akin to solving an underdetermined system with a single equation and multiple unknowns. However, when multiple gaze behaviors are observed for the same visual stimulus under a shared goal, the goal-driven component remains invariant while participant-specific variation changes across instances. This shared invariance enables the common attentional objective to be isolated despite individual differences.

Concretely, given scanpaths from multiple participants sharing a goal, we generate *think-aloud transcripts* [20,58]—narratives that surface goal-relevant attentional patterns while filtering out individual variability. We obtain these by prompting GPT-4 [1] to interpret the shared structure across scanpaths and extract the common *top-down attention allocation strategy*, then use the transcripts as auxiliary instruction-tuning data for Gazette. Crucially, multiple scanpaths and transcript generation are needed only during training; at inference, the model decodes the goal from a single scanpath. This auxiliary think-aloud objective yields substantial gains in gaze decoding performance.

In summary, the contributions of this paper are threefold. First, we introduce the task of unconstrained decoding of top-down attention, where goals are expressed in natural language. Second, we propose *Gazette*, a text-generative MLLM-based framework designed for this task. Third, we introduce an auxiliary *think-aloud transcript* prediction objective that encourages Gazette to capture goal-specific attentional patterns, thereby improving decoding performance across a broad range of top-down attentional tasks.

2 Related Work

Goal-directed Human Attention. Goal-directed attention, in contrast to bottom-up attention [29,41], is the top-down control exerted by frontal-parietal brain areas that modulates processing of sensory input based on current task demand, prior knowledge and expectations [24,31]. Several datasets have been col-

lected to study various facets of goal-directed gaze behavior. COCO-Search18 [13] is a dataset of visual search gaze fixations from 10 participants performing Target-Present/Absent tasks on 6,202 natural images spanning 18 target categories. RefCOCO-Gaze [44] contains gaze scanpaths from 220 participants viewing 2,094 COCO images while simultaneously hearing referring expressions grounding objects in the images. AiR-D [11] is a dataset with scanpaths of 20 participants performing the visual question answering task for 195 image–question pairs. In this study, we use COCO-Search18, RefCOCO-Gaze and AiR-D.

Gaze Decoding. Gaze measured by eye-trackers can be a noninvasive means of understanding human intention. Yarbus [77] showed that eye movements depend strongly on the viewer's task, with fixations varying across instructions for the same visual stimulus. Building on this, prior work has decoded goal-directed tasks from eye movements [2,7,8,79,80], reconstructed images from fixations [61–63,71], and inferred user tasks or activities from gaze [5,10,14,27,60]. For search-target inference specifically, Sattar et al. [54–56] addressed both images and categories, and later methods used pre-trained CNN features [4,59]. More recent models add structure: GST [48] fuses category semantics with scanpaths to predict COCO-Search18 search targets, and GazeGNN [68], a radiological diagnosis model, can be adapted for gaze target decoding. All frame decoding as classification – selecting one goal from a fixed set – limiting real-world applicability. Relatedly, Chen *et al*. [12] jointly predicted scanpaths and per-fixation natural-language explanations, but took scanpaths as output rather than input. Our work instead accepts human scanpaths as input and decodes the cognitive context behind top-down attention. Mondal *et al*. [45] enabled zero-shot search-target decoding via gaze–language alignment, but were limited to a fixed vocabulary of target descriptions at inference. More recently, Xue et al. [75] absorbed training-time attention into a subject embedding to personalize image descriptions, conditioning on *who* is describing rather than decoding a goal, and thus requiring no gaze at inference. Gazette instead takes the scanpath as input and decodes the underlying goal, treating individual variation as a confound to filter out. To our knowledge, Gazette, is the first generative, gaze-to-text decoding framework.

Instruction Tuning of MLLMs. Instruction tuning or supervised fine-tuning (SFT) for multimodal large language models (MLLMs) adapts pre-trained vision-language architectures for downstream tasks by training on instruction-response pairs, enhancing model performance on specific tasks and general instruction-following [39,40,52]. Methods like LLaVA [40] and VIGC [69] generate large-scale visual instruction datasets, often with proprietary LLMs like GPT-4 endowed with image context expressed via object bounding box annotations in the scene. Visual instruction tuning has been applied in diverse domains. Instruction-tuned MLLMs have excelled in robotics [18, 35], healthcare [57], and e-commerce [39], demonstrating how instruction tuning transforms general-purpose MLLMs into specialized, domain-adapted agents. Inspired by these works, we curate the first instruction tuning dataset for gaze decoding, enabling the adaptation of general-purpose MLLMs for the decoding of gaze behavior.

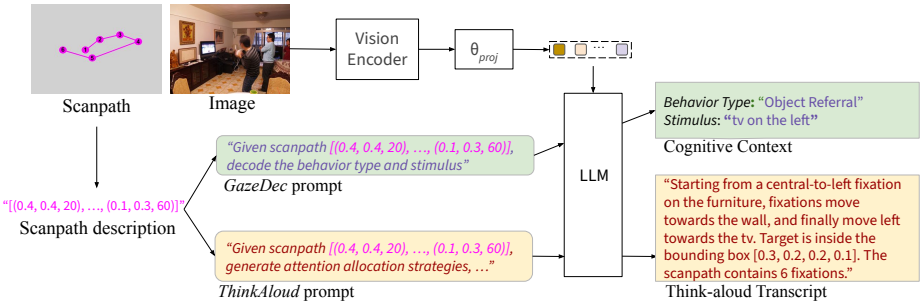


Fig. 1: ***Gazette***: A text-generative decoding framework for top-down attention. For an input of an image and a language instruction conveying scanpath information, a textual response is generated by the Multimodal LLM comprising a Vision Encoder, an LLM, and a linear projection θ_{proj} interfacing the Vision Encoder and the LLM. Language instruction can correspond to either the primary gaze decoding task *GazeDec*, or the auxiliary think-aloud transcript generation task *ThinkAloud*.

3 *Gazette*: Gaze-to-Text Decoding of Human Attention

Gazette is an MLLM-based generative framework that decodes diverse gaze behaviors into natural language. Specifically, Gazette extends the LLaVA-1.5-7B foundation model [39] and is fine-tuned on our novel visual instruction-tuning dataset, which consists of tasks closely aligned with gaze understanding and is described in detail in this section. The input to Gazette consists of an image and a gaze scanpath, and the output is a hierarchical representation capturing two facets of top-down attentional control: (i) a coarse-level *behavior type*, representing the category of goal-directed behavior (e.g., Target-Present Visual Search, Target-Absent Visual Search, object referral, or VQA), and (ii) a fine-level *stimulus*, specifying the concrete top-down goal. For visual search, the stimulus corresponds to the target category (e.g., "car"); for object referral, it corresponds to a referring expression (e.g., "red car on the right"); and for VQA, it corresponds to the question itself (e.g., "Is there a red car on the road?").

The remainder of this section is organized as follows. We first describe the architecture of Gazette and its processing pipeline. We then present *think-aloud transcript generation*, our proposed method for extracting narratives that highlight goal-relevant attentional patterns while filtering out individual variability. Finally, we discuss the training and inference procedures.

3.1 Architecture and Decoding Pipeline

Built upon the LLaVA foundation model [39], Gazette adopts its multimodal architecture, consisting of a *Vision Encoder* [51] coupled with a large language model (LLM). The input image $I \in \mathbb{R}^{3\times H\times W}$ is processed by the Vision Encoder, and the resulting visual features are projected via a lightweight MLP θ_{proj} into the LLM's token embedding space. A gaze scanpath $S = \{f_i \mid i = 0, \ldots, N-1\}$

is a sequence of N fixations $f_i = (x_i, y_i, t_i)$, where (x_i, y_i) denotes the normalized spatial coordinates (scaled to $[0, 1]$) and t_i the fixation duration. Following prior visual instruction-tuning work [35, 39, 52], we encode S textually using a prompt template to produce a language instruction $T_{GazeDec}$ (e.g., "Given scanpath $[(0.4, 0.4, 20), \ldots, (0.1, 0.3, 60)]$, decode the behavior type and stimulus"). The projected visual tokens and textual prompt are concatenated and fed into the LLM, which autoregressively generates natural language output D describing the cognitive context. We refer to this primary decoding task as *GazeDec*. The generated response D can be parsed into two components: D_{type}, representing the behavior type, and D_{goal}, specifying the stimulus or goal. The components of Gazette and the overall processing pipeline are illustrated in Fig. 1.

In addition to the primary decoding task *GazeDec*, Gazette also supports an auxiliary think-aloud task that decodes the attention allocation strategy (please find more detailed description in the next subsection). For this task, the inputs remain the image I and the scanpath S, but the prompt template is modified to produce $T_{ThinkAloud}$, which textualizes the scanpath and instructs the model to generate an attention allocation strategy (e.g., "Given scanpath $[(0.4, 0.4, 20), \ldots, (0.1, 0.3, 60)]$, generate the attention allocation strategy ..."). The model then autoregressively generates the corresponding think-aloud transcript D_{TaT}. This auxiliary task is also illustrated in Fig. 1.

3.2 ThinkAloud Transcript Generation

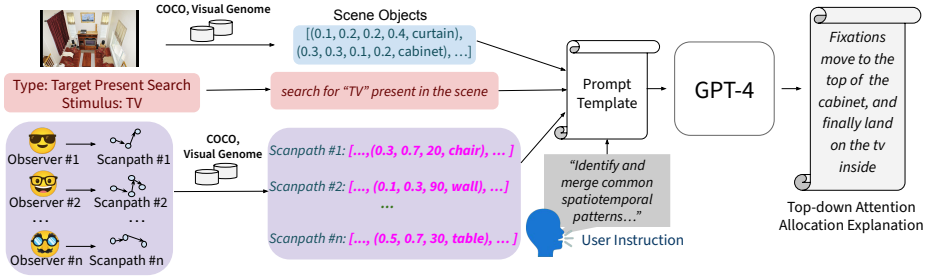
Assume we are given training data consisting of N triplets of an image, a gaze scanpath, and the corresponding goal description, $(I_i, S_i, D_i)_{i=1}^{N}$. The primary gaze decoding task is to learn a mapping from (I_i, S_i) to D_i. However, defining this mapping as an MLLM and fine-tuning it solely on the primary objective yields suboptimal performance, because the observed scanpath is not purely goal-driven, and general-purpose MLLMs lack prior knowledge of human attentional control, making it difficult to disentangle goal-specific signals from individual variability in scanpaths. This difficulty can be viewed through the lens of an underdetermined inverse problem. For a single participant, the observed scanpath S_i can be conceptualized as a function of both the latent goal D_i and participant-specific factors U_i, i.e., $S_i = f(I_i, D_i) + U_i$. Although supervision is provided for D_i, the nuisance variable U_i remains unobserved and entangled with the goal signal, making it difficult to recover a clean mapping from (I_i, S_i) to D_i. In this setting, attempting to infer $f(I_i, D_i)$ from a single equation involving both $f(I_i, D_i)$ and U_i is fundamentally ill-posed. However, when multiple scanpaths S_i's are observed for the same visual stimulus under a shared goal, the participant-specific factors vary while the underlying goal remains invariant. These multiple observations introduce additional structure, allowing the goal-related signal to be isolated from individual variation. Intuitively, aggregating diverse yet goal-consistent behaviors makes the disentanglement of the shared objective feasible.

Suppose the training data can be partitioned into groups such that if indices i and j belong to the same group G, then $I_i = I_j$ and $D_i = D_j$, while $S_i \neq S_j$ because the scanpaths are collected from different participants. With slight abuse

of notation, we denote such a group as $(I, D, \{S_i \mid i \in G\})$, where I and D represent the shared stimulus and goal. Although individual scanpaths vary, we posit that they share a common component attributable to the goal-driven attention process, which we interpret as the underlying *attention allocation strategy*. This hypothesis is grounded in classic findings that eye movement patterns vary systematically with task demands [77], yet are simultaneously influenced by multiple factors, including bottom-up salience and idiosyncratic viewer characteristics. Beyond this descriptive observation, the convergence we exploit has a principled theoretical basis. Najemnik & Geisler's Bayesian ideal-observer model [47] shows that the optimal fixation policy depends on the target template, the scene prior, and the observer's visibility map; the first two are shared across observers performing the same task, while only the third varies mildly between individuals. Cross-observer convergence on a goal-carrying fixation pattern is therefore a theoretical prediction rather than a mere empirical regularity, with individual idiosyncrasies acting as approximately orthogonal noise. Hayhoe & Ballard's Cognitive Relevance Theory [23] extends this account beyond visual search through a shared, goal-determined relevance signal, predicting the same convergence for object referral and VQA. Thus, while isolating goal-specific information from a single scanpath is difficult, the shared structure across multiple participants engaged in the same task provides a principled basis for extracting the common goal-driven signal.

To operationalize this idea, we leverage GPT-4 [1] to derive *think-aloud transcripts* that capture this shared attention allocation strategy. Each transcript is structured into three idea units—(i) a top-down attention allocation explanation, (ii) the inferred target location, and (iii) the scanpath length—following principles from think-aloud protocol analysis [21]. These transcripts serve as supervision for the auxiliary *ThinkAloud* task.

We employ GPT-4 [1] to analyze scanpaths from multiple participants who share the same cognitive context for a given image and to summarize their common spatiotemporal patterns (see Fig. 2). We select GPT-4 for its strong capability in identifying latent patterns in structured data [42, 72] and its extensive world knowledge. As noted in existing literature [69], currently available multimodal large language models (MLLMs) remain less capable and more costly than their text-only LLM counterparts. Therefore, following LLaVA [40], we adopt text-only GPT-4 for its accessibility, cost-effectiveness, and computational efficiency. We provide scene context in the form of bounding-box annotations of objects in the scene derived from COCO [38] for COCO-Search18, RefCOCO-Gaze, and COCO-originated images in the AiR-D dataset, and from Visual Genome [33] scene graphs for the Flickr-originated images in the AiR-D dataset sourced from the GQA dataset [28]. Again, to counter scale variation, we normalize bounding-box parameters. Additionally, we provide scanpath information for all observers in the form of normalized fixation co-ordinates, raw fixation durations, and category labels of the fixated objects (derived from COCO and Visual Genome bounding box annotations). Finally, cognitive context is also included in the prompt. Prompt details are given in the supplement.



(a)

For stimulus: *Object Referral* for *"car on the right"*.
GPT-4 generated: *Most humans initially fixate on the large vertical parking meter spanning a significant portion of the left-central part of the image, then shift focus toward the large car located on the right side, spending varying durations there.*

(b)

Fig. 2: A novel prompting strategy for GPT-4 to extract top-down attention allocation explanation for think-aloud transcript generation task (*ThinkAloud*). (a) This strategy instructs GPT-4 to summarize common spatiotemporal patterns in scanpaths of n participants, given the task type and goal, along with scene information via scene object bounding boxes from COCO [38] and Visual Genome [33]. The response is used to construct the think-aloud transcript. (b) Sample scanpaths from several users for an image-stimulus pair and the corresponding GPT-4 generated response.

GPT-4's response reflects common patterns across scanpaths of several participants, which we hypothesize indicates the top-down attentional allocation strategies utilized by the human visual system, therefore serving as pseudo-annotation for the "top-down attention allocation explanation" idea unit of the think-aloud transcript. Because this response captures the *common* patterns, it remains consistent across *all* scanpaths from multiple participants, and can be used as pseudo-annotation for scanpaths from *each* individual participant. Additionally, following previous visual instruction tuning research [35, 52] adapting MLLMs for specialized tasks, we propose two fundamental scanpath comprehension tasks as "idea units" in the think-aloud transcript: target localization, and counting fixations in a scanpath. Since MLLMs struggle in object localization [52], a crucial task underlying gaze behavior for visual search and object referral [46, 76], we posit that Gazette will benefit from learning target localization, particularly for Visual Search and Object Referral tasks. Following previous MLLM literature [35, 39, 52], we normalize bounding box parameters $[x, y, w, h]$ (where x, y locate the upper-left corner, and w and h are the width and height of the bounding box, respectively) to deal with scale variation. This task also

accounts for target-absent cases (null scenario), where Gazette must predict the absence of the target instead of a bounding box. The second task is inspired by previous research [53, 66, 73] suggesting that scanpath length indicates many properties of gaze behavior, *e.g.* the degree of exploratory behavior compared to focused behavior. Prompts for both *GazeDec* and *ThinkAloud* along with human evaluation of the GPT-4-generated top-down attention allocation explanation pseudo-annotations are in the supplement.

3.3 Model Training and Inference

We detail Gazette's training and inference procedures below, with more details in the supplement.

Training. We initialize our model with pre-trained LLaVA-1.5-7B model [39] weights, and fine-tune it via Low-Rank Adaptation (LoRA) [26], reducing the number of trainable parameters, mitigating overfitting on limited gaze training data. We avoid newer MLLMs to ensure fair comparison and isolate the impact of think-aloud transcript–based tuning from backbone improvements. We use an auto-regressive language modeling objective [40] for instruction tuning on both *GazeDec* and *ThinkAloud* tasks.

Inference. We prompt Gazette using $T_{GazeDec}$ and decode the response text using a greedy decoding strategy. The response text is then parsed to obtain the gaze behavior type D_{type} and the stimulus D_{goal}. D_{type} is encoded using a language encoder [70] and matched to label vocabulary by computing cosine similarities between each text embedding and each label's language embedding, where label vocabulary is {``Target-Present Search'', ``Target-Absent Search'', ``Object Referral'', or ``Visual Question Answering''}. Similarly, for COCO-Search18, D_{goal} is matched with the 18 target categories of the dataset.

4 Experiments

Gazette can decode gaze scanpaths for multiple gaze behaviors, including categorical visual search, object referral, and visual question answering (VQA). In this section, we evaluate Gazette's gaze decoding capabilities using a wide range of metrics across multiple gaze behaviors. Gaze behavior type prediction is trivial for Gazette (see supplement), perhaps due to artifacts in scanpaths from distinct data collection setups. Here, we focus on stimulus decoding. We train our model on two NVIDIA RTX A6000 GPUs with a total batch size of 32 and a learning rate of 2e-5. Training is done on the training splits of COCO-Search18 (both Target-Present and Target-Absent trials), RefCOCO-Gaze and AiR-D, and evaluation is done on the corresponding test splits. To evaluate and analyze the efficacy of our proposed auxiliary task, we compare performances of our full model Gazette trained on both sets of instructions $\mathbf{T}_{GazeDec}$ and $\mathbf{T}_{ThinkAloud}$ (corresponding to *GazeDec* and *ThinkAloud*, respectively), and a variant where the MLLM is trained only on $\mathbf{T}_{GazeDec}$ but not $\mathbf{T}_{ThinkAloud}$. The latter variant is simply denoted as "*w/o ThinkAloud*".

While discriminative gaze decoding methods exist [45, 48, 68], Gazette is the first generative gaze decoding model to our knowledge, so we constructed baselines based on the LLaVA [39] model. For Object Referral and Target-Present Visual Search, a majority of last fixations land on the target, so we fine-tune LLaVA-1.5 to describe the object where the final fixation lands, to yield the gaze stimulus. We call this baseline *LLaVA-last*. Another baseline is a frozen *LLaVA-1.5* [39] prompted with scanpath information and instructions to describe the goal. We also compare Gazette with Mondal *et al*. [45], GST [48] and GazeGNN [68] (adapted by us for visual search gaze decoding) in visual search gaze decoding with COCO-Search18 [13]. More details of Gazette and the aforementioned baselines are in the supplement.

4.1 Evaluation Metrics

We curated metrics suited to each gaze behavior. Visual search is decoded as a target category from a predefined set, whereas object referral and VQA are decoded into free-form expressions and questions; the metrics differ accordingly.

Text Generation evaluation metrics for Object Referral and VQA. We evaluate model text-generative capabilities on Object Referral gaze decoding and VQA gaze decoding using two paradigms: (1) using standard lexical overlap metrics, and (2) using GPT-4 to assess generated texts using a set of abstract rubrics, also known as LLM-as-a-Judge paradigm.

Lexical Overlap Metrics: To assess the text generation-based decoding capabilities of Gazette, we use a broad set of standard metrics used to evaluate the lexical overlap of generated text from text generation models with the ground truth text. **BLEU-1 to BLEU-4** [49] evaluate machine-generated text by comparing its n-gram precision (n=1–4) against reference translations and apply a brevity penalty to discourage overly short output. **METEOR** [3] aligns candidate and reference translations via exact, stem, synonym, and paraphrase matches, then computes a recall-weighted harmonic mean of unigram precision and recall with a fragmentation penalty to preserve word order. **ROUGE-L** [37] measures the longest common subsequence between candidate and reference texts, combining recall and precision into an F-score to capture sentence-level structure and word-order overlap. **CIDEr** [67] captures human consensus by computing the TF–IDF–weighted cosine similarity of n-gram vectors between a candidate and multiple references, highlighting terms frequent in the candidate and distinctive in the references. While there is only one reference question for VQA samples in AiR-D, there are multiple reference referring expressions in RefCOCO [78] annotated by multiple annotators for the same object in an image. This allows us to compute lexical overlap metrics for each referring expression as the candidate, while reserving the others as ground truth. This is repeated for all referring expressions; the average defines the RefCOCO Inter-Annotator Consistency (**RefCOCO-IAC**), a noise ceiling approximating human annotator variability (similar to Inter-Observer Consistency in behavioral literature).

LLM-as-a-Judge: In this evaluation paradigm popularized by recent text-generative research [40, 81], a very large language model (such as GPT-4) is used as an external evaluator to assess the quality of the generated outputs, providing a scalable and cost-effective alternative to human evaluation. Additionally, this allows more interpretable evaluation rubrics to be used (*e.g.* fluency, helpfulness, relevance), allowing for a more nuanced, and context-aware, approach to assessing model-generated text that might not be possible with standard lexical overlap metrics. Note that in this evaluation setting, GPT-4 is used only to evaluate predicted referring expressions/questions against ground truth from RefCOCO-Gaze/AiR-D, not the think-aloud transcripts, avoiding any circular dependency via GPT-4 that may bias evaluation.

For each scanpath, GPT-4 is prompted to evaluate texts using a scale from 1 to 10 for each rubric, where 1 is poor and 10 is excellent. The prompt also contains these key information: **(1)** referring expressions/questions generated by all models to be evaluated, **(2)** ground truth referring expressions/questions, **(3)** scene context in the form of bounding boxes of every object in the scene (similar to the method described in Sec. 3.2), **(4)** rubrics to evaluate the generated texts on. The rubrics are distinct for object referral and VQA and are detailed below, and the detailed prompt to GPT-4 is provided in the supplement.

For object referral, the rubrics are: (1) **Expression Overlap** - How much does the generated referring expression match the ground truth referring expressions, especially in terms of coverage of the entities, their attributes and spatial relationships? (2) **Referential Equivalence** - Does either of the ground truth expressions refer to the same object in the image as the generated expression? (3) **Category Correctness** - Is the type or category of the referred object mentioned correctly?

On the other hand, for VQA, the rubrics are: (1) **Question Overlap** - How much does the generated question match the ground truth question, especially in terms of coverage of the entities, their attributes and spatial relationships? (2) **Answer Equivalence** - Does the generated question and the ground truth question have the same correct answer on the image?

Target category prediction metrics for Visual Search Tasks. For COCO-Search18's 18 target categories, we assess decoding via **precision**, **recall**, $\mathbf{F}_1$ **score**, and **accuracy**. As the test set is imbalanced across categories, we macro-average precision, recall, and F_1 over the 18 categories, weighting each equally. Accuracy ignores this imbalance, but we report it to compare with state-of-the-art methods Mondal *et al.* [45] and GST [48], neither of which has a public implementation. Mondal *et al.* report precision, recall, and F_1 only on the combined target-present/target-absent test set, not individually.

4.2 Results

We evaluate the efficacy of Gazette in decoding gaze scanpaths from four top-down attention tasks: Task A – Object Referral, Task B – Visual Question Answering

(VQA), Task C – Target-Present Visual Search, and Task D – Target-Absent Visual Search. We evaluate generative gaze decoding on Object Referral (using RefCOCO-Gaze [44] dataset) and VQA (using AiR-D [11] dataset) tasks under two paradigms: (1) Lexical Overlap Metric-based evaluation (Table 1) and (2) LLM-as-a-Judge-based evaluation (Table 2).

Table 1: Performance of Gazette and baselines on Task A – Object Referral Gaze Decoding evaluated using RefCOCO-Gaze [44], and Task B – VQA Gaze Decoding evaluated using AiR-D [11] on the basis of lexical overlap metrics (B-1 through B-4 denote BLEU-1 through BLEU-4, M denotes METEOR, R-L denotes ROUGE-L, and C denotes CIDEr). Best results are highlighted in bold. Results exceeding RefCOCO Inter-Annotator Consistency (RefCOCO-IAC) values are underlined. Percentage improvements of Gazette over variant w/o *ThinkAloud* are provided in parentheses.

Task	Method	B-1	B-2	B-3	B-4	M	R-L	C
Object Referral (Task A)	RefCOCO-IAC	0.500	0.282	0.148	0.057	0.244	0.452	1.115
	LLaVA-1.5 [39]	0.066	0.018	0.006	0.0	0.077	0.105	0.062
	LLaVA-last	0.170	0.080	0.039	0.014	0.113	0.221	0.113
	w/o ThinkAloud	0.479	0.263	0.145	0.070	0.232	0.443	0.872
	Gazette	**0.519**	**0.305**	**0.175**	**0.098**	**0.248**	**0.480**	**0.974**
		(+8.36%)	(+15.97%)	(+20.69%)	(+40.00%)	(+6.90%)	(+8.36%)	(+11.70%)
VQA (Task B)	LLaVA-1.5 [39]	0.124	0.030	0.012	0.006	0.039	0.103	0.047
	w/o ThinkAloud	0.329	0.222	0.144	0.095	0.147	0.286	0.263
	Gazette	**0.364**	**0.268**	**0.202**	**0.159**	**0.160**	**0.324**	**0.367**
		(+10.64%)	(+20.72%)	(+40.28%)	(+67.37%)	(+8.84%)	(+13.29%)	(+39.54%)

As shown in Table 1, Gazette trained with the auxiliary *ThinkAloud* task significantly outperforms the variant not trained on *ThinkAloud* ("*w/o ThinkAloud*"), and baselines LLaVA-1.5 [39] and *LLaVA-last* for both Task A and Task B on every lexical overlap metric. Under the LLM-as-a-Judge paradigm (Table 2), the reported scores follow the same trend seen in lexical overlap metrics, with Gazette outperforming the variant not trained on *ThinkAloud* ("*w/o ThinkAloud*") by large margins on every human-interpretable rubric. This shows the efficacy of deeper reasoning about the top-down attentional processes via the *ThinkAloud* task. *LLaVA-last* baseline performs poorly in Tables 1 and 2, suggesting that complete scanpaths provide crucial context for referring expression generation. In the supplement, we explore the individual effects of idea units within think-aloud transcripts, and show that for Object Referral, learning both attention allocation explanation generation and target localization enhances performance. For VQA (where target localization is not relevant), only attention allocation explanation generation is crucial, as it helps Gazette learn complex reasoning processes underlying the visual question answering task.

Table 2: Performance of Gazette and baselines on Task A – Object Referral Gaze Decoding evaluated using RefCOCO-Gaze dataset [44], and Task B – VQA Gaze Decoding evaluated using AiR-D dataset [11] by GPT-4 [1] on a scale of 1-10 under the LLM-as-a-Judge setting. Scores are averaged with best scores in bold.

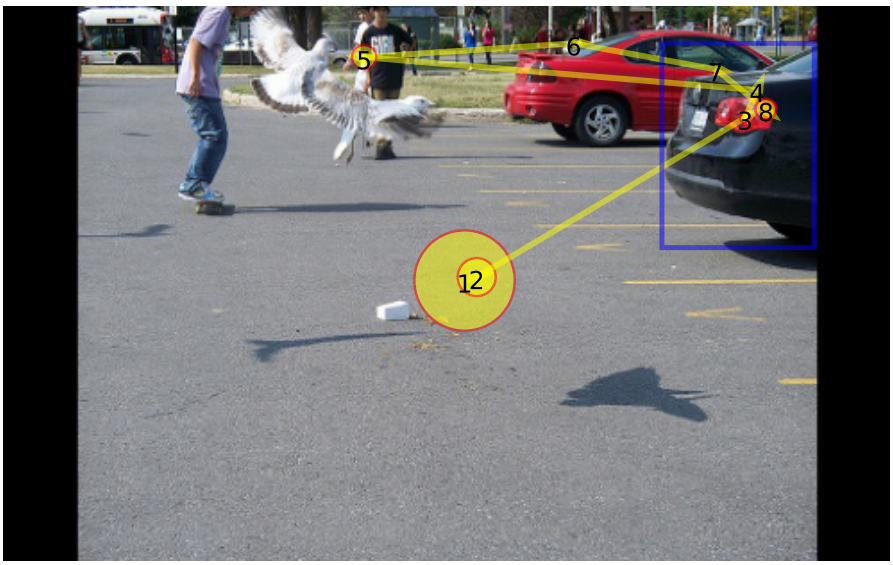
		Method		
Task	Rubric	*LLaVA-last*	*w/o ThinkAloud*	*Gazette*
Object Referral (Task A)	Expression Overlap	3.515	5.019	**5.980**
	Referential Equivalence	4.100	5.771	**6.638**
	Category Correctness	7.577	8.376	**8.743**
VQA (Task B)	Question Overlap	-	2.103	**2.792**
	Answer Equivalence	-	1.642	**2.135**

Table 3: Performance of Gazette and baselines on Task C (Target-Present) and Task D (Target-Absent) Visual Search Gaze Decoding, both evaluated using COCO-Search18 dataset [13]. Best results are in bold, and second best are underlined.

Task	**Method**	**Precision**	**Recall**	**F_1**	**Accuracy**
Target-Present Visual Search (Task C)	LLaVA-1.5 [39]	0.448	0.436	0.406	0.402
	LLaVA-last	0.688	0.649	0.628	0.618
	GazeGNN [68]	0.338	0.345	0.319	0.335
	GST [48]	-	-	-	0.544
	Mondal *et al.* [45]	-	-	-	**0.776**
	w/o ThinkAloud	<u>0.768</u>	<u>0.746</u>	<u>0.742</u>	0.742
	Gazette	**0.786**	**0.775**	**0.775**	<u>0.773</u>
Target-Absent Visual Search (Task D)	LLaVA-1.5 [39]	0.128	0.102	0.077	0.088
	GazeGNN [68]	0.241	0.251	0.218	0.270
	GST [48]	-	-	-	0.385
	Mondal *et al.* [45]	-	-	-	0.387
	w/o ThinkAloud	<u>0.437</u>	<u>0.424</u>	<u>0.420</u>	**0.445**
	Gazette	**0.438**	**0.426**	**0.424**	<u>0.443</u>

Next, we focus on evaluating the target prediction capabilities of Gazette (trained with and without *ThinkAloud*), and compare with baselines GazeGNN [68] (which we adapt for target prediction), GST [48], Mondal *et al.* [45], LLaVA-1.5 [39], and *LLaVA-last*, on the COCO-Search18 benchmark [13]. The results are shown in Table 3. The implementations for GST and Mondal *et al.* are not publicly available, therefore we are unable to furnish metrics not reported by the authors (*i.e.* precision, recall, F_1). Gazette significantly outperforms prior methods across all metrics on the Target-Absent Visual Search task, and achieves accuracy in the Target-Present setting that is nearly on par with the state-of-the-art visual search gaze decoding method from Mondal *et al.* (0.773 vs. 0.776). *LLaVA-last* fares well in recognizing Target-Present gaze targets, but still lags behind Gazette, suggesting the importance of scanpath context. In the supplement,

Gazette Predicts: Object Referral for **"black car top right"**.
Gazette w/o ThinkAloud Predicts: Object Referral for **"red car"**.



Gazette-generated explanation: Most humans initially fixate on the road before shifting their attention to the car located in the upper right corner of the image, often after scanning other cars and occasionally focusing on people in the vicinity.

Fig. 3: Qualitative Results. Comparison of Gazette variants on decoding a gaze scanpath for **Object Referral** target **"black car on right"** (blue bounding box). We show predictions from Gazette and its variant not trained on *ThinkAloud*, along with the attention allocation explanation from the Gazette-generated think-aloud transcript.

we show that similar to Object Referral, both target localization and attention allocation explanation generation idea units are important for Target-Present search gaze decoding. For Target-Absent trials, training with *ThinkAloud* results in modest and mixed performance improvements. We attribute this to poor agreement between observers in Target-Absent search [76], where gaze behavior increasingly resembles free-viewing as search progresses [14], resulting in low commonality across participants and weak shared attentional patterns — potentially degrading GPT-4's responses when prompted by our strategy. We show in the supplement that predicting the absence of the target and measuring the length of the scanpath improves Target-Absent search gaze decoding.

Fig. 3 qualitatively analyzes generated texts from Gazette trained either with or without *ThinkAloud* instructions to decode a scanpath of a human grounding the referring expression "black car on right". Without a deeper understanding of the verification and scanning strategies exhibited by the scanpath through *ThinkAloud* instruction tuning, the model misinterprets the "red car" to be the target, whereas our full Gazette model trained on *ThinkAloud* instructions

identifies the correct car by analyzing the gaze patterns, as evidenced by the attention allocation explanation text generated by Gazette. In the supplement, we qualitatively analyze additional samples of model-decoded cognitive contexts and think-aloud transcripts generated by Gazette.

5 Conclusion

In this work, we explored the capabilities of Multimodal LLMs in understanding human attention behavior, specifically top-down attention control. We proposed *Gazette*, a novel text-generative gaze decoding framework to decode a wide array of top-down attention behaviors. Rather than naively instruction-tuning MLLMs on the primary gaze decoding task, we built an auxiliary dataset for generating top-down attentional allocation explanations – termed *think-aloud transcripts* – to encourage Gazette to explicitly separate goal-specific attentional dynamics from individual idiosyncrasies. To generate pseudo-annotations for these think-aloud transcripts, we prompted GPT-4 using a novel prompting strategy which exploits commonalities in gaze patterns across participants engaged in the same attentional task for the same image. We showed that when trained with a combination of our proposed primary and auxiliary tasks, Gazette achieved significant performance boost in both generative and predictive gaze decoding tasks over naive instruction tuning strategies and state-of-the-art methods, as evaluated by our rigorous evaluation scheme. Gazette opens the door to applications that hinge on decoding precisely what a person is looking for, down to fine-grained details, bringing such capabilities within the reach of real-world systems. While we demonstrated generalization across a diverse set of top-down attentional behaviors, Gazette can potentially be extended to psychological analysis, driving, and assistive healthcare—domains that require precise decoding of human mental states. We expect that Gazette, together with our novel GPT-4 prompting strategy for generating think-aloud transcript pseudo-annotations, will inspire future work exploring new avenues of gaze understanding with LLMs and MLLMs.

Limitations. Our method is not without limitations. Its reliance on inter-observer consistency to extract top-down attentional control patterns means efficacy degrades on tasks with low observer agreement, *e.g.* Target-Absent Visual Search. In addition, because the think-aloud transcripts are GPT-4-generated, they inherit its training priors rather than participants' true cognition, raising a concern of *bias* toward a majority-weighted notion of attention allocation. We leave addressing these limitations to future work.

Acknowledgements. This project was supported by US National Science Foundation Award IIS-1763981, IIS-2123920, DUE-2055406, and the SUNY2020 Infrastructure Transportation Security Center, and a gift from Adobe. We are grateful to Abe Leite for his invaluable help with the human evaluation of the GPT-4-generated top-down attention allocation explanation pseudo-annotations, to Niranjan Balasubramanian for his generous guidance and discussion, and to Xiang Li for his crucial help with platform issues during MLLM training.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bahle, B., Mills, M., Dodd, M.D.: Human classifier: Observers can deduce task solely from eye movements. Attention, Perception, & Psychophysics **79**, 1415–1425 (2017)
3. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Association for Computational Linguistics, Ann Arbor, Michigan (2005)
4. Barz, M., Stauden, S., Sonntag, D.: Visual search target inference in natural interaction settings with machine learning. In: ACM Symposium on Eye Tracking Research and Applications (2020)
5. Bektas, K., Strecker, J., Mayer, S., Garcia, K.: Gaze-enabled activity recognition for augmented reality feedback. Computers & Graphics **119**, 103909 (2024)
6. Benchetrit, Y., Banville, H., King, J.R.: Brain decoding: toward real-time reconstruction of visual perception. arXiv preprint arXiv:2310.19812 (2023)
7. Borji, A., Itti, L.: Defending Yarbus: Eye movements reveal observers' task. Journal of Vision **14**(3), 29–29 (2014)
8. Borji, A., Lennartz, A., Pomplun, M.: What do eyes reveal about the mind?: Algorithmic inference of search targets from fixations. Neurocomputing **149**, 788–799 (2015)
9. Bühler, B., Bozkir, E., Deininger, H., Gerjets, P., Trautwein, U., Kasneci, E.: On task and in sync: Examining the relationship between gaze synchrony and self-reported attention during video lecture learning. Proceedings of the ACM on Human-Computer Interaction **8**(ETRA) (2024)
10. Bulling, A., Ward, J.A., Gellersen, H., Tröster, G.: Eye movement analysis for activity recognition using electrooculography. IEEE transactions on pattern analysis and machine intelligence **33**(4), 741–753 (2010)
11. Chen, S., Jiang, M., Yang, J., Zhao, Q.: AiR: Attention with reasoning capability. In: European Conference on Computer Vision (2020)
12. Chen, X., Jiang, M., Zhao, Q.: GazeXplain: Learning to predict natural language explanations of visual scanpaths. European Conference on Computer Vision (2024)
13. Chen, Y., Yang, Z., Ahn, S., Samaras, D., Hoai, M., Zelinsky, G.: COCO-Search18 fixation dataset for predicting goal-directed attention control. Scientific reports **11**(1), 8776 (2021)
14. Chen, Y., Yang, Z., Chakraborty, S., Mondal, S., Ahn, S., Samaras, D., Hoai, M., Zelinsky, G.: Characterizing target-absent human attention. In: Proceedings of CVPR International Workshop on Gaze Estimation and Prediction in the Wild (2022)
15. Chestek, C.A., Gilja, V., Blabe, C.H., Foster, B.L., Shenoy, K.V., Parvizi, J., Henderson, J.M.: Hand posture classification using electrocorticography signals in the gamma band over human sensorimotor brain areas. Journal of neural engineering **10**(2), 026002 (2013)
16. Daly, I.: Neural decoding of music from the EEG. Scientific Reports **13**(1), 624 (2023)

17. Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., King, J.R.: Decoding speech perception from non-invasive brain recordings. Nature Machine Intelligence **5**(10), 1097–1107 (2023)
18. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P.: PaLM-E: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
19. Duan, Z., Cheng, H., Xu, D., Wu, X., Zhang, X., Ye, X., Xie, Z.: CityLLaVA: Efficient fine-tuning for VLMs in city scenario. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
20. Ericsson, K.A., Simon, H.A.: Protocol Analysis: Verbal Reports as Data. MIT Press, Cambridge, MA (1984)
21. Fonteyn, M.E., Kuipers, B., Grobe, S.J.: A description of think aloud method and protocol analysis. Qualitative Health Research **3**(4), 430–441 (1993)
22. Hamza, A., Ahn, Y.H., Lee, S., Kim, S.T., et al.: LLaVA needs more knowledge: Retrieval augmented natural language generation with knowledge graph for explaining thoracic pathologies. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025)
23. Hayhoe, M., Ballard, D.: Eye movements in natural behavior. Trends in cognitive sciences **9**(4), 188–194 (2005)
24. Henderson, J.M., Brockmole, J.R., Castelhano, M.S., Mack, M.: Visual saliency does not account for eye movements during visual search in real-world scenes. In: Eye movements, pp. 537–III. Elsevier (2007)
25. Hollenstein, N., Renggli, C., Glaus, B., Barrett, M., Troendle, M., Langer, N., Zhang, C.: Decoding EEG brain activity for multi-modal natural language processing. Frontiers in Human Neuroscience **15**, 659410 (2021)
26. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
27. Hu, Z., Bulling, A., Li, S., Wang, G.: EHTask: Recognizing user tasks from eye and head movements in immersive virtual reality. IEEE Transactions on Visualization and Computer Graphics **29**(4), 1992–2004 (2021)
28. Hudson, D.A., Manning, C.D.: GQA: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2019)
29. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11), 1254–1259 (1998)
30. Khokhar, A., Yoshimura, A., Borst, C.: Eye-gaze-triggered visual cues to restore attention in educational vr. In: 2019 IEEE conference on virtual reality and 3D user interfaces (VR), poster (2019)
31. Koehler, K., Guo, F., Zhang, S., Eckstein, M.P.: What do saliency models predict? Journal of Vision **14**(3), 14–14 (2014)
32. Komeiji, S., Mitsuhashi, T., Iimura, Y., Suzuki, H., Sugano, H., Shinoda, K., Tanaka, T.: Feasibility of decoding covert speech in ECoG with a transformer trained on overt speech. Scientific Reports **14**(1), 11491 (2024)
33. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual Genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017)

34. Li, S., Zhang, X.: Implicit intention communication in human–robot interaction through visual behavior studies. IEEE Transactions on Human-Machine Systems **47**(4), 437–448 (2017)
35. Li, X., Mata, C., Park, J., Kahatapitiya, K., Jang, Y.S., Shang, J., Ranasinghe, K., Burgert, R., Cai, M., Lee, Y.J., et al.: LLaRA: Supercharging robot learning data for vision-language policy. International Conference on Learning Representations (2025)
36. Liaqat, S., Wu, C., Duggirala, P.R., Cheung, S.c.S., Chuah, C.N., Ozonoff, S., Young, G.: Predicting ASD diagnosis in children with synthetic and image-based eye gaze data. Signal Processing: Image Communication **94**, 116198 (2021)
37. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. Association for Computational Linguistics, Barcelona, Spain (2004)
38. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision. pp. 740–755. Springer (2014)
39. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
40. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
41. Masciocchi, C.M., Mihalas, S., Parkhurst, D., Niebur, E.: Everyone knows what is interesting: Salient locations which should be fixated. Journal of Vision **9**(11), 25–25 (2009)
42. Mirchandani, S., Xia, F., Florence, P., Ichter, B., Driess, D., Arenas, M.G., Rao, K., Sadigh, D., Zeng, A.: Large language models as general pattern machines. arXiv preprint arXiv:2307.04721 (2023)
43. Mohbat, F., Zaki, M.J.: LLaVA-Chef: A multi-modal generative model for food recipes. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (2024)
44. Mondal, S., Ahn, S., Yang, Z., Balasubramanian, N., Samaras, D., Zelinsky, G., Hoai, M.: Look hear: Gaze prediction for speech-directed human attention. In: European Conference on Computer Vision (2024)
45. Mondal, S., Sendhilnathan, N., Zhang, T., Liu, Y., Proulx, M., Iuzzolino, M.L., Qin, C., Jonker, T.R.: Gaze-language alignment for zero-shot prediction of visual search targets from human gaze scanpaths. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (October 2025)
46. Mondal, S., Yang, Z., Ahn, S., Samaras, D., Zelinsky, G., Hoai, M.: 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)
47. Najemnik, J., Geisler, W.S.: Optimal eye movement strategies in visual search. Nature **434**(7031), 387–391 (2005)
48. Nishiyasu, T., Sato, Y.: 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 (2024)
49. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (2002)

50. Perfect, E., Hoskin, E., Noyek, S., Davies, C.T.: Outcome measures and uptake barriers when children and youth with complex disabilities use eye gaze assistive technology. Developmental Neurorehabilitation **23**(3), 145–159 (2020)
51. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021)
52. Ranasinghe, K., Shukla, S.N., Poursaeed, O., Ryoo, M.S., Lin, T.Y.: Learning to localize objects improves spatial reasoning in visual-LLMs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
53. Sabab, S.A., Kabir, M.R., Hussain, S.R., Mahmud, H., Rubaiyeat, H.A., Hasan, M.K.: Vis-iTrack: Visual intention through gaze tracking using low-cost webcam. IEEE Access **10**, 70779–70792 (2022)
54. Sattar, H., Bulling, A., Fritz, M.: Predicting the category and attributes of visual search targets using deep gaze pooling. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (2017)
55. Sattar, H., Fritz, M., Bulling, A.: Deep gaze pooling: Inferring and visually decoding search intents from human gaze fixations. Neurocomputing **387**, 369–382 (2020)
56. Sattar, H., Muller, S., Fritz, M., Bulling, A.: Prediction of search targets from fixations in open-world settings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
57. Singhal, K., Azizi, S., Tu, T., et al.: Large language models encode clinical knowledge. Nature **620**, 297–302 (2023)
58. van Someren, M.W., Barnard, Y.F., Sandberg, J.A.C.: The Think Aloud Method: A Practical Guide to Modelling Cognitive Processes. Academic Press, London (1994)
59. Stauden, S., Barz, M., Sonntag, D.: Visual search target inference using bag of deep visual words. In: KI 2018: Advances in Artificial Intelligence: 41st German Conference on AI, Berlin, Germany, September 24–28, 2018, Proceedings 41. Springer (2018)
60. Steil, J., Bulling, A.: Discovery of everyday human activities from long-term visual behaviour using topic models. In: Proceedings of the 2015 acm international joint conference on pervasive and ubiquitous computing (2015)
61. Strohm, F., Bâce, M., Bulling, A.: Usable and fast interactive mental face reconstruction. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (2023)
62. Strohm, F., Sood, E., Mayer, S., Müller, P., Bâce, M., Bulling, A.: Neural photofit: gaze-based mental image reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
63. Strohm, F., Sood, E., Thomas, D., Bâce, M., Bulling, A.: Facial composite generation with iterative human feedback. In: Annual Conference on Neural Information Processing Systems. PMLR (2023)
64. Takagi, Y., Nishimoto, S.: High-resolution image reconstruction with latent diffusion models from human brain activity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
65. Tawari, A., Chen, K.H., Trivedi, M.M.: Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation. In: 17th International IEEE conference on intelligent transportation systems (ITSC). IEEE (2014)
66. Tsang, H.Y., Tory, M., Swindells, C.: eseetrack—visualizing sequential fixation patterns. IEEE Transactions on Visualization and Computer Graphics **16**(6), 953–962 (2010)

67. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
68. Wang, B., Pan, H., Aboah, A., Zhang, Z., Keles, E., Torigian, D., Turkbey, B., Krupinski, E., Udupa, J., Bagci, U.: 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 (2024)
69. Wang, B., Wu, F., Han, X., Peng, J., Zhong, H., Zhang, P., Dong, X., Li, W., Li, W., Wang, J., et al.: VIGC: Visual instruction generation and correction. In: Proceedings of the AAAI Conference on Artificial Intelligence (2024)
70. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: Advances in Neural Information Processing Systems (2020)
71. Wang, X., Ley, A., Koch, S., Lindlbauer, D., Hays, J., Holmqvist, K., Alexa, M.: The mental image revealed by gaze tracking. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (2019)
72. Weber, I.: Large language models are pattern matchers: Editing semi-structured and structured documents with ChatGPT. In: AKWI Jahrestagung 2024 (2024)
73. Williams, C.C., Castelhano, M.S.: The changing landscape: High-level influences on eye movement guidance in scenes. Vision **3**(3), 33 (2019)
74. Xia, W., de Charette, R., Oztireli, C., Xue, J.H.: DREAM: Visual decoding from reversing human visual system. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2024)
75. Xue, R., Le, H., Xu, J., Mondal, S., Leite, A., Zelinsky, G., Hoai, M., Samaras, D.: Personalized image descriptions from attention sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2026)
76. Yang, Z., Mondal, S., Ahn, S., Zelinsky, G., Hoai, M., Samaras, D.: Target-absent human attention. In: European Conference on Computer Vision (2022)
77. Yarbus, A.L.: Eye Movements and Vision. Plenum Press, New York (1967)
78. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European Conference on Computer Vision (2016)
79. Zelinsky, G., Zhang, W., Samaras, D.: Eye can read your mind: Decoding eye movements to reveal the targets of categorical search tasks. Journal of Vision **8**(6), 380–380 (2008)
80. Zelinsky, G.J., Peng, Y., Samaras, D.: Eye can read your mind: Decoding gaze fixations to reveal categorical search targets. Journal of Vision **13**(14), 10–10 (2013)
81. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Gonzalez, J.E., Stoica, I.: Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685 (2023)