BEYOND MODEL-INTERNAL SIGNALS: STEERING VLLM ATTENTION WITH PREDICTED HUMAN SALIENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

To reduce hallucinations in Vision Large Language Models (VLLMs), training-free methods steer attention toward relevant image regions, identified from model-internal statistics alone. In this work, we ask whether an external reference, human visual attention, can serve this role. To answer this question, we introduce and release VLFEEDBACK-ETC, an eye-tracking corpus of 21 participants annotating preferences on VQA samples from VLFeedback, a widely used VLLM alignment dataset. Since gaze is unavailable at inference, we also introduce QASALFORMER, a question-aware saliency model for natural images, and benchmark it against existing predictors on our corpus. Using the corpus to compare human gaze with LLaVA-1.5 attention, we find that they are weakly correlated, with large differences between layers. We then propose GAZESTEER, which adds predicted saliency to the pre-softmax attention scores over image tokens. On CHAIR, predicted saliency reduces object hallucination beyond a spatially uniform map that adds the same attention mass, and the reduction is not explained by shorter captions. Interestingly, the utility of a map is not determined by its global alignment with human gaze. Instead, useful maps combine image-specific structure with enough coverage. Moreover, predicted saliency does not replace cross-head consensus, but combining both is at least as effective as consensus alone, suggesting that external saliency is a complementary signal to model-internal ones.

1 INTRODUCTION

Large Language Models (LLMs) have advanced significantly in recent years, demonstrating strong performance across a wide range of Natural Language Processing (NLP) tasks (OpenAI, 2023). Leveraging this progress, Multimodal Large Language Models (MLLMs) have extended language modelling beyond text to incorporate heterogeneous modalities such as images, audio, and video, enabling cross-modal understanding and generation (Yin et al., 2024). When the input modality includes visual information, these models are commonly referred to as VLLMs, which constitute the main focus of this work. Although progress in VLLMs is rapid and grounded in advances in LLMs (Touvron et al., 2023; Bai et al., 2022; Dubey et al., 2024), the core challenges that affect LLMs, like alignment with human values (Casper et al., 2023), persist and are further compounded by the additional visual modality. Despite strong benchmark performance, current VLLMs are often poorly aligned with users' attentional focus and intentions, leading to imprecise or unhelpful responses in everyday scenarios (Yan et al., 2024), and hallucinations are especially acute (Yin et al., 2024).

While the internal reasoning mechanisms that enable visual–textual competence remain largely unknown (Shen, 2025), recent works suggest that many visual hallucinations are closely connected to model *attention sinks* in the self-attention over image tokens (Zhang et al., 2025; Wang et al., 2025; Yang et al., 2025; Kang et al., 2025). In these cases, a small subset of hidden-state dimensions becomes massively activated, pulling disproportionate attention to a few tokens irrespective of visual content, an effect analogous to attention sinks previously observed in text-only LLMs. These findings have motivated a new line of methods that explicitly analyse and manipulate attention (Zhang et al., 2025; Wang et al., 2025; Yang et al., 2025; Kang et al., 2025; Chuang et al., 2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

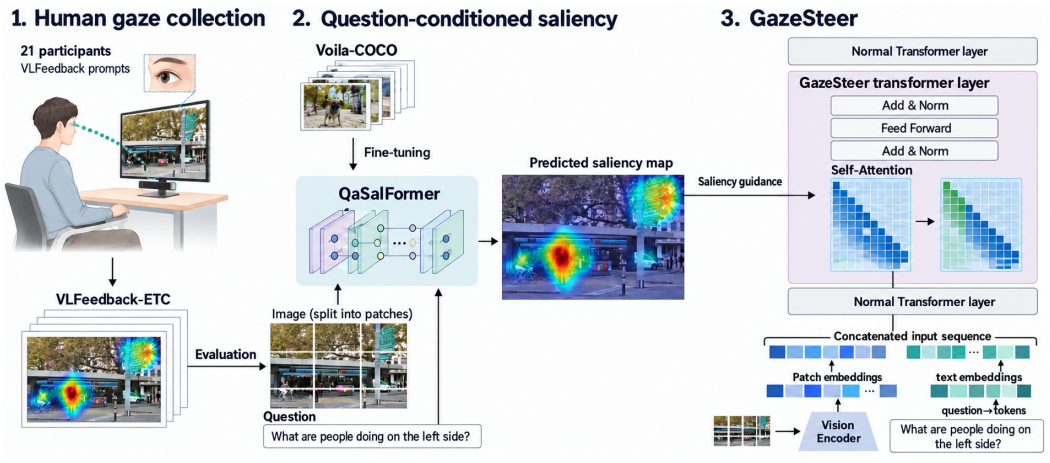


Figure 1: Overview of our contributions. **(1)** We record ET data from 21 participants ranking VLLM responses to VQA prompts, released as VLFEEDBACK-ETC. **(2)** We fine-tune VisSalFormer on VOILA-COCO to obtain QASALFORMER, a question-aware saliency model, evaluated against our organic data. **(3)** GAZESTEER adds the predicted saliency map to the pre-softmax attention scores over image tokens in the middle layers, steering attention toward regions a human would attend to.

On the other hand, cognitive signals have shown promising results in improving LLMs' performance. By augmenting datasets with user-derived behavioural and neurophysiological responses, they can potentially improve alignment while mitigating hallucinations (Lopez-Cardona et al., 2025a). For instance, Eye-tracking (ET) data have been successfully leveraged in a range of NLP tasks (Huang et al., 2023; Khurana et al., 2023; Yang & Hollenstein, 2023; Deng et al., 2023; McGuire & Tomuro, 2021), and recent work has begun to explore its potential in human alignment tasks (Kiegeland et al., 2024; Lopez-Cardona et al., 2025c;b) for LLMs. Specifically, ET offers a non-invasive, accurate way for inferring visual attention patterns from oculomotor behaviour and leveraging it as an explicit feedback signal (Li et al., 2024b). This innovation has driven advancements in applications like Visual Question Answering (VQA) (Sood et al., 2023) with recent integration into VLLMs, as shown by Yan et al. (2024).

However, cognitive data acquisition in a laboratory setting introduces challenges related to cost, need for specialised equipment, and potential privacy risks (Khurana et al., 2023). To overcome these limitations, generative models for ET have been developed, capable of producing synthetic data and broadening the range of potential applications (Cartella et al., 2024). Motivated by recent works linking ET signals to human alignment, and by the growing success of attention interventions in reducing hallucinations in VLLMs, this work makes the following contributions (Figure 1):

- We curate and publicly release the VLFEEDBACK-ETC dataset, which captures ET signals from a sample of 21 participants who answered the VQA preference annotations of VLFeedback (Li et al., 2024a), a widely used VLLM alignment dataset (Section 3).
- We introduce and release QASALFORMER, a question-conditioned saliency model for VQA (Section 4), fine-tuned on VOILA-COCO (Yan et al., 2024) and benchmarked with our dataset against existing saliency models.
- We compare human gaze and LLaVA attention maps, demonstrating that the two are only weakly correlated: empirical evidence of the "grounding gap" in VLLMs (Section 5).
- We introduce GAZESTEER (Section 6), a method inspired by Jiang et al. (2025) to steer attention with predicted human saliency, instead of model-internal statistics, to reduce hallucinations in VLLMs.

2 RELATED WORK

2.1 HALLUCINATIONS AND VISUAL ATTENTION IN VLLMS

VLLMs are prone to hallucinate, i.e. generate content that is inconsistent with or unsupported by the visual input (Zhou et al., 2024; Li et al., 2024c). Recent research has linked hallucinations to attention patterns: *attention sinks*, tokens that attract disproportionate attention despite low semantic content (Kang et al., 2025), and to attention degradation over generation (Zhang et al., 2026a; Yang et al., 2025). Other works have detected hallucinations directly from attention maps (Chuang et al., 2024; Binkowski et al., 2025). A growing body of training-free methods mitigates hallucinations by intervening on the attention computation (Jiang et al., 2025; Kang et al., 2025; Zhu et al., 2025; Tang et al., 2025b;a; Zhang et al., 2025; Cheng et al., 2026), on the output distribution during decoding (Leng et al., 2024; Yang et al., 2025; Xu et al., 2025; Li et al., 2025; Huang et al., 2024), or on both (Zhang et al., 2026b; Wang et al., 2026). For instance, Jiang et al. (2025) demonstrate that real object tokens receive higher attention than hallucinated ones in the LLM's middle layers, a finding we build on (Section 6), and amplify visual attention using the consensus across attention heads as a guide. However, every method identifies relevant visual tokens from model-internal statistics alone (sink positions, head consensus, or the vision encoder's own attention) with no external reference of visual relevance, a gap that we address in this work (Section 6).

2.2 HUMAN GAZE FOR VISION-LANGUAGE MODELS

Datasets. Most ET corpora in NLP have been curated for reading tasks (Shubi et al., 2025). The closest dataset to ours is OASST-ETC (Lopez-Cardona et al., 2025b), which provides gaze-based data during preference judgements of text-only LLM responses; VLFEEDBACK-ETC (Section 3) extends this paradigm to multimodal prompts. For VQA, several datasets predate VLLMs, such as VQA-HAT (Das et al., 2016), AiR (Chen et al., 2020) and VQA-MHUG (Sood et al., 2021), while VOILA-COCO (Yan et al., 2024)—currently the largest dataset—consists of mouse cursor trajectories of Localized Narratives (Pont-Tuset et al., 2020) instead of eye movements, paired with GPT-4-generated questions (OpenAI, 2023).

Saliency models. Gaze data are largely unavailable at inference time. Therefore, recent works (Lopez-Cardona et al., 2025a) have used generative models for synthesizing signals that can act as proxies of attention, such as scanpaths (Martin et al., 2022; Kümmerer et al., 2022; Wang et al., 2024b; Cartella et al., 2025) or saliency maps. For instance, MD-SEM (Fosco et al., 2020) predicts free-viewing saliency on natural images. In addition, Gazeformer (Mondal et al., 2023) predicts goal-directed scanpaths for object search, while VisSalFormer (Wang et al., 2024c) predicts question-driven saliency, but only on information visualisations. Currently, no existing model combines question conditioning with natural-image saliency for VQA: the gap that the proposed QASALFORMER model addresses (Section 4).

Gaze in (V)LLMs. Gaze signals have been used to align text-only LLMs (Kiegeland et al., 2024; Lopez-Cardona et al., 2025c; Galliamov et al., 2025) and to improve Question Answering (QA) (Malmaud et al., 2020; Zhang & Hollenstein, 2024). In VQA, human-like attention has been integrated into models (Sood et al., 2023; Inadumi et al., 2024), trace data has served as a gaze proxy for training VLLMs (Yan et al., 2024), and gaze has regularised egocentric understanding (Pani & Yang, 2025). When comparing attention directly, Sood et al. (2021) found that alignment with human text attention predicts VQA accuracy, while Yamamoto et al. (2025) reported human-like attention heads in DINO (Caron et al., 2021), and Salamatian et al. (2025) (ChartGaze) studied attention patterns in chart understanding. To our knowledge, no prior work has compared human with VLLM attention for general-purpose VQA (Section 5), nor used it to guide attention interventions at inference time (Section 6).

3 VLFEEDBACK-ETC DATASET CONSTRUCTION

We recorded ET from 21 participants (*age* = 30.4 ± 5.0, 19.0% F) while they read and evaluated pairs of responses to VQA prompts generated by VLLMs. Our stimuli are drawn from VLFEEDBACK (Li

et al., 2024a), a widely used VLLM preference dataset (Wang et al., 2024a) built under a Reinforcement Learning from AI Feedback (RLAIF) protocol: responses to 80k multi-modal instructions were generated by a pool of 12 VLLMs and rated by GPT-4 along helpfulness, *visual faithfulness* (consistency with the image, targeting ungrounded hallucinations), and ethical considerations, yielding ranked preference pairs. After filtering responses by display length (10–700 characters), we randomly sampled 30 triplets (image + question, preferred response, rejected response) as experimental stimuli. See Figure 6 for an example stimulus and Appendix B for exclusion criteria and statistics.

Participants followed a streamlined version of the guidelines used for GPT-4 annotation of VLFEEDBACK (Figure 5). As shown in Figure 2, for each prompt, participants rated the candidate responses and selected the best response, while we recorded their eye movements using a Tobii Pro Fusion eye tracker (details of the apparatus and processing pipeline are provided in Appendix A). Inter-annotator agreement was moderate (Fleiss' $\kappa = 0.453$), and GPT-4 matched the user majority in 76.7% of cases, versus the 83.1% reported in the original paper. We restrict our analyses to fixations within the image region and convert these fixations into saliency maps following standard practice. For completeness, the dataset release also includes the raw gaze data.

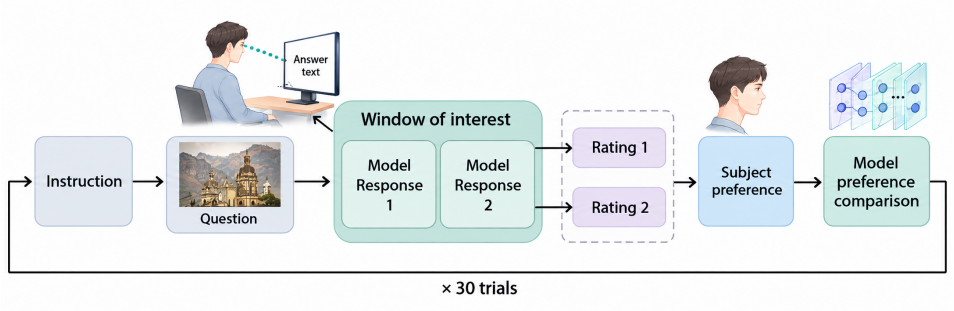


Figure 2: Task procedure. Each trial presents an image and a question, then the two responses with their ratings, and ends with the preference choice. Eye movements are recorded throughout.

4 QASALFORMER: QUESTION-AWARE VISUAL SALIENCY MODEL

We argue that leveraging human attention in VLLMs will likely require models that can accurately predict human saliency conditioned on the question, which does not currently exist on natural images (Subsection 2.2). To address this gap, we introduce QASALFORMER, a variant obtained by fine-tuning VisSalFormer on the VOILA-COCO dataset (Yan et al., 2024).

4.1 METHODOLOGY

To develop QASALFORMER (Figure 1), we maintain the original training configuration but perform fine-tuning on the VOILA-COCO (Yan et al., 2024) dataset to better align the model with natural image VQA tasks. The original VisSalFormer implementation utilises a Swin-T backbone (Liu et al., 2021), initialised with ImageNet-1K (Deng et al., 2009) pre-trained weights, alongside the *bert-base-uncased* weights for text encoding (Devlin et al., 2019). Given an image and its associated question, the model predicts a continuous saliency map.

Datasets. We fine-tune the model on VOILA-COCO, whose synthetic construction enables large-scale supervision for VQA-related tasks. We then assess whether this fine-tuning improves alignment with the human attention data in VLFEEDBACK-ETC (Section 3).

Baselines. We evaluate the multi-duration saliency model (MD-SEM) (Fosco et al., 2020) as a question-agnostic baseline that takes only the visual stimulus as input. To assess goal-directed attention, we include Gazeformer (Mondal et al., 2023), a visual search model, and construct saliency maps from its predicted fixation sequences. We also evaluate VisSalFormer (Wang et al., 2024c), which was originally trained on information visualisations from SalChartQA. Further model details are provided in Appendix C.1. To assess generalisation to question-conditioned attention on natural images, we evaluate all three models on VLFEEDBACK-ETC, using human saliency maps

pooled across all 21 participants as the reference. We also estimate an *inter-observer ceiling* using a leave-one-subject-out procedure: each participant's map is evaluated against the pooled map of the remaining 20 participants, and scores are averaged across participants. This provides an empirical benchmark for inter-observer agreement rather than a strict upper bound on model performance.

Evaluation metrics and training objective. We report standard saliency metrics (Li et al., 2024b; Wang et al., 2024c): Pearson's Correlation Coefficient (CC), Similarity or histogram intersection (SIM) and Kullback–Leibler (KL), evaluated against continuous human saliency maps, and Normalized Scanpath Saliency (NSS) and Area Under the Curve (AUC), evaluated against human fixation maps (see Appendix C.1 for definitions). We additionally report two measures. *Image specificity* (Spec) is the difference between the CC of a predicted map with the human map for the same image and its mean CC with the human maps for all other images. This measure quantifies image-specific correspondence while accounting for shared spatial structure, such as centre bias. *Coverage* (SC) is the percentage of pixels in the predicted saliency map whose values exceed the Otsu threshold. For a predicted saliency map S and a ground-truth human map G, we use a training objective adapted from Wang et al. (2024c), as defined in Equation (1) (see Appendix C.1.1 for further details).

$$\mathcal{L} = \lambda_{\mathrm{KL}}\,\mathrm{KL}(G \,\|\, S) - \lambda_{\mathrm{CC}}\,\mathrm{CC}(S, G) - \lambda_{\mathrm{SIM}}\,\mathrm{SIM}(S, G), \tag{1}$$

4.2 Results

Table 1 reports agreement between model-generated and human saliency maps, alongside image specificity and Saliency Coverage (SC). The best and second-best results are indicated in **bold** and underlined, respectively. An illustrative comparison is shown in Figure 3 (a)–(e). Among question-conditioned models, QaSalFormer consistently performs best across all saliency evaluation metrics, approximately doubling the CC and image specificity of VisSalFormer. However, the free-viewing model MD-SEM performs best overall, except in KL, and its scores improve with viewing duration. This suggests that gaze during preference evaluation shares characteristics with extended free viewing. However, with only one question per image, our corpus cannot isolate the contribution of question conditioning. QaSalFormer produces broader saliency distributions, covering approximately 43% of the image, compared with 21% for human maps and 9% for MD-SEM. We examine how these differences affect attention guidance in Section 6. Finally, the pre-trained baselines obtain lower scores than those reported on their original evaluation datasets (Wang et al., 2024c; Fosco et al., 2020), consistent with a domain shift between those datasets and our corpus.

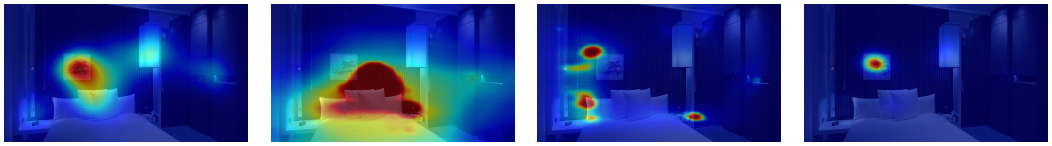
Table 1: Saliency models vs VLFeedback-ETC humans (pooled GT). Spec: image specificity; SC: Otsu coverage (%). Human: leave-one-subject-out ceiling. **Bold**/underline: best/second model.

Model	SIM↑	CC↑	KLD↓	NSS↑	AUC↑	Spec↑	SC
Human (inter-obs.)	*0.661*	*0.727*	*1.122*	*1.159*	*0.779*	*0.518*	*21.1*
Gazeformer	0.324	0.300	5.469	0.436	0.624	0.070	8.1
VisSalFormer	0.538	0.233	0.674	0.345	0.618	0.088	17.9
QaSalFormer	0.599	0.506	**0.513**	0.682	0.692	0.197	42.9
MD-SEM-3000	0.610	0.616	0.626	1.064	0.769	0.308	8.5
MD-SEM-5000	**0.630**	**0.625**	0.559	**1.084**	**0.772**	**0.322**	9.1

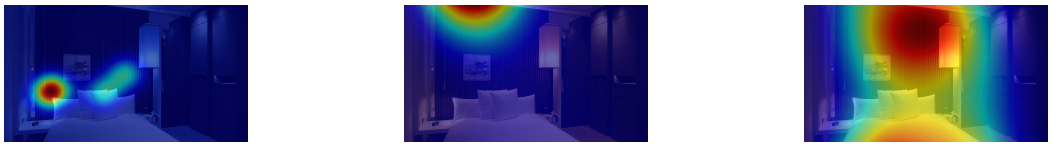
5 Human- vs. model-based attention comparison

As reviewed in Subsection 2.1, there is growing interest in understanding and manipulating the visual attention of VLLMs. However, existing comparisons between model attention and human gaze (Subsection 2.2) focus on specific domains, such as information visualisations (Salamatian et al., 2025), or on models that predate VLLMs (Sood et al., 2021). Therefore, using VLFeedback-ETC we provide, to our knowledge, the first comparison of VLLM attention with recorded human gaze during preference annotation in VQA.

Several techniques have been adapted to study multimodal Transformers, including attention-based analyses, gradient-based attribution and input-importance methods, with recent examples such as



(a) Human (b) QaSalFormer (c) VisSalFormer (d) MD-SEM



(e) Gazeformer (f) LLaVA-1.5-7B (g) LLaVA-1.5-13B

Figure 3: Human saliency, saliency model predictions (a–e) and VLLM attention (f–g, rollout) for prompt 9: "Describe the overall appearance and atmosphere of the bedroom in the image".

GLIMPSE (Shen, 2025). Alongside layer-wise comparisons, we apply *attention rollout* (Abnar & Zuidema, 2020) to VLLMs during decoding. This method has been shown to correlate more closely with input-token importance than raw attention, while remaining computationally lightweight. Attention rollout approximates information flow across layers by incorporating residual connections into head-averaged attention and normalising the resulting matrices to be row-stochastic. These matrices are then composed across layers to trace attention back to the input. In our decoder setting, we extract the rollout values associated with the first generated token.

Visual input. We obtain patch-level attention by selecting, from the rollout row associated with the first generated token, the columns corresponding to image-patch tokens. We then normalise these values over image patches to obtain a visual attention map. We compare the resulting maps with the human saliency maps described in Section 3, using the evaluation metrics from Section 4 to ensure consistency across analyses. Further details are provided in Appendix D.

Models. We evaluate the 7B and 13B variants of LLaVA-1.5 (Liu et al., 2024a) for the following reasons. First, LLaVA-1.5 is widely used in the literature on attention and hallucination reviewed in Subsection 2.1, facilitating comparison with prior work and linking this analysis to our attention-guidance intervention in Section 6. Second, it is one of the models used to generate the original dataset (Section 3). Third, its architecture inserts image-patch tokens directly into the text-token sequence, making it straightforward to trace attention from generated tokens back to individual image patches.

Results. Table 2 and Figure 7 report agreement between model attention and human saliency for attention rollout and individual layers, respectively. Representative examples are shown in Figure 3 (f)–(g). Correlations vary substantially across layers, reaching approximately 0.3 in layers 20–25 but remaining near zero or negative in early layers (3–6). Both models exhibit this pattern, although their correlations remain below those of the best-performing saliency models. Prior work has reported unintuitive visual attention patterns in LLMs, with attention disproportionately concentrated on a small subset of visual tokens (Kang et al., 2025). Likewise, our results indicate limited agreement with human saliency, particularly when attention is aggregated across layers using rollout. However, the metrics in Table 2 do not by themselves establish whether this mismatch arises from excessive concentration on a small subset of tokens.

Table 2: Comparison metrics between VLLM rollout attention and human saliency maps on VLFeedback-ETC (mean ± std over the 30 stimuli). Same saliency metrics as Table 1.

Model	SIM↑	CC↑	KLD↓	NSS↑	AUC↑
LLaVA-1.5-7B	0.466 ± 0.110	0.090 ± 0.274	0.918 ± 0.545	0.143 ± 0.455	0.553 ± 0.141
LLaVA-1.5-13B	0.445 ± 0.127	0.092 ± 0.293	1.042 ± 0.797	0.136 ± 0.482	0.560 ± 0.151

6 GAZESTEER: GAZE-GUIDED ATTENTION INTERVENTION

As discussed in Subsection 2.1, existing attention interventions rely on identifying visual tokens that carry relevant evidence, typically using model-internal signals such as fixed sink positions and large activations (Kang et al., 2025), or insensitivity to instruction changes (Li et al., 2025). We, instead, investigate predicted saliency as an external reference. Human attention has been shown to benefit LLMs and VLLMs (Subsection 2.2), while our analysis in Section 5 shows limited agreement between VLLM attention and human saliency for the same images. Together, these findings motivate testing whether an external estimate of human attention can improve visual grounding.

6.1 METHODOLOGY

We build on the attention-editing method of Jiang et al. (2025). VLLMs process visual tokens $\{v_i\}_{i=1}^{n} \subset \mathbb{R}^d$ alongside text prompt tokens $\{t_j\}_{j=1}^{m}$. Using an autoregressive Transformer decoder with L layers and H heads, the model generates token y_k at step k conditioned on all prior input:

$$\mathcal{X}_k = \{v_1, \ldots, v_n, t_1, \ldots, t_m, y_1, \ldots, y_{k-1}\}. \tag{2}$$

For layer $\ell \in [L]$ and head $h \in [H]$, let $\mathbf{S}_k^{(\ell,h)} \in \mathbb{R}^{a_k \times a_k}$ denote the pre-softmax attention scores at step k, where $a_k = n + m + k - 1$, with attention matrix $A_k^{(\ell,h)} = \mathrm{softmax}\big(\mathbf{S}_k^{(\ell,h)}\big)$. Jiang et al. (2025) amplify visual attention in the layers responsible for visual information enrichment (layers 5–18 in LLaVA-1.5) by adding, to every head's pre-softmax attention score toward image token i, a guidance term computed as the cross-head consensus of the current attention magnitudes (Equation (3)). This term is entirely model-internal: a patch is reinforced only if the model's own heads already agree it matters. Let $\widehat{\mathrm{Sal}} : \{1, \ldots, n\} \to [0, 1]$ denote the saliency map predicted for the current image, normalised by its maximum and resampled to the visual token grid, and let $\bar{\mathbf{S}}_k^{(\ell)}$ be the mean absolute pre-softmax score over image tokens at step k, used to match the scale of $\widehat{\mathrm{Sal}}(i)$ to the model's native score range. GAZESTEER (Figure 1) modifies the attention scores as:

$$\mathbf{S}_k^{(\ell,h)}(a_k, i) \;\leftarrow\; \mathbf{S}_k^{(\ell,h)}(a_k, i) \;+\; \alpha \underbrace{\frac{1}{H} \sum_{h'=1}^{H} \left| \mathbf{S}_k^{(\ell,h')}(a_k, i) \right|}_{\text{consensus (Jiang et al., 2025)}} \;+\; \beta \underbrace{\widehat{\mathrm{Sal}}(i) \cdot \bar{\mathbf{S}}_k^{(\ell)}}_{\text{saliency}} \tag{3}$$

Equation (3) generalises two variants of GAZESTEER, controlled by (α, β): **saliency-only** ($\alpha = 0$), where attention is guided only by the saliency map, independent of the model's own consensus; and **additive** ($\alpha, \beta > 0$), which combines the model-internal consensus term with a saliency term.

Implementation details. We use the 7B and 13B variants of LLaVA (Liu et al., 2024a) again and compare against the original consensus method (Jiang et al., 2025), which has been reported to outperform OPERA (Huang et al., 2024), VCD (Leng et al., 2024), and PAI (Liu et al., 2024b). Following Jiang et al. (2025) and Huang et al. (2024), our evaluation uses a random subset of 500 images from the COCO 2014 validation benchmark, with greedy decoding and up to 512 new tokens. We quantify object hallucination via CHAIR (Rohrbach et al., 2018), which computes the fraction of cited objects missing from the ground-truth annotations, evaluated at both the instance level (C_I) and the sentence level (C_S):

$$C_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{total objects mentioned}\}|}, \qquad C_S = \frac{|\{\text{captions with hallucinations}\}|}{|\{\text{total captions}\}|}, \tag{4}$$

where smaller values correspond to higher fidelity. We additionally report F_1 of the mentioned objects against the ground-truth annotations to assess caption richness and coverage, together with the mean caption length in words. Since CHAIR relies on generic image captioning rather than question-answering, we condition QASALFORMER on a fixed, domain-neutral prompt (*"What is in the image?"*) rather than a task-specific question (see Appendix E.1 for details).

Controls. We compare QaSalFormer saliency with that of the models evaluated in Section 4. Liu et al. (2024b) show that shifting attention from text to image tokens reduces hallucination. However, recent work (Cheng et al., 2026) shows that visual attention mass may not align with model decisions. To distinguish the effects of spatial layout and added mass, we also evaluate a constant map that assigns the image's mean predicted saliency to each image token, preserving mass while removing spatial structure. Since CHAIR favours short captions, we compare each method with greedy decoding truncated to the same caption length (Appendix E.3). Because different maps add different amounts of attention for a given β, all comparisons use matched F1 or recall, never matched β.

6.2 RESULTS

In Figure 4, we can see, for LLaVA-1.5-7B, the comparison of C_S and C_I for the different ways of amplifying attention (consensus, saliency and combined) in the first row of figures, and the same for the saliency controls in the second row. Complementary figures are provided in Appendix E.3 (Figures 8 to 15), where we compare C_S and C_I against caption length, recall and F1 for both models and both guidance regimes. For each configuration, we ran different values of β, aiming to cover the regions of F1 close to the greedy baseline and to consensus (Jiang et al., 2025). The complete sweeps for every family and both models are reported in Appendix E.4. In Appendix E.2 (Tables 7 to 10), for each method we report the configuration with the lowest hallucination whose F1 drops at most half a point, one, two or three points below greedy decoding, respectively, a range in which generation remains comparable to greedy. We discuss the results mainly from the figures and use the tables as support: a single selected configuration does not represent the behaviour of a family, and with 500 images the C_S error is around two points, so the selection can change results.

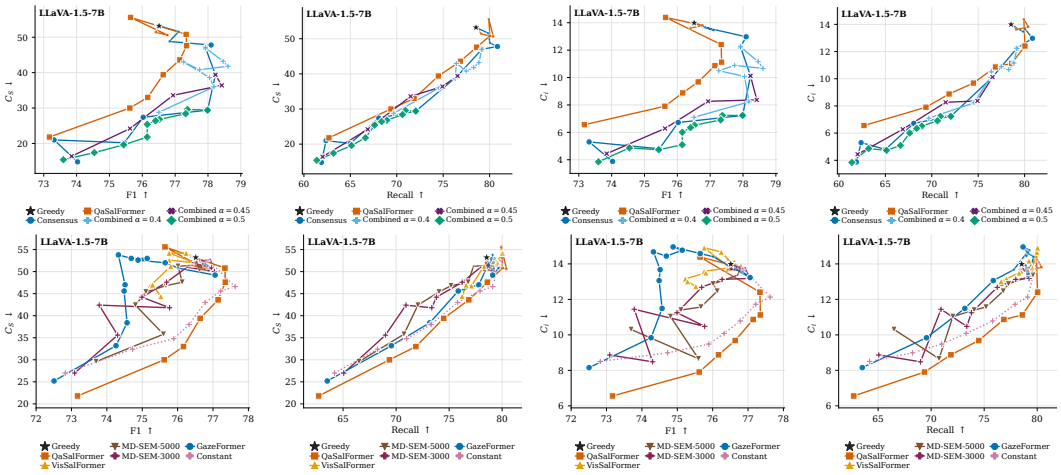


Figure 4: Hallucination–richness trade-off on LLaVA-1.5-7B (CHAIR, 500 COCO images, layers 5–18). Each marker is one configuration; lines connect configurations of increasing strength, starting from greedy decoding ($\star$). Lower C_S/C_i and higher F1/recall are better. First row: Model-internal consensus (Jiang et al., 2025), QASALFORMER saliency, and their combination (α=0.4, 0.45, 0.5, sweep β). Second row: Saliency sources (α=0); the constant map gives more weight to the image without any spatial structure. Configurations with F1 $<$ 72 are omitted.

Attention mass and caption length. A constant map already reduces hallucination at higher guidance strengths, indicating that increasing attention to the image can itself be beneficial. However, comparisons at matched F1, recall and caption length show that QASALFORMER achieves lower C_S and C_I than the constant map. Among the selected configurations, it is the only saliency source that outperforms the constant map at every F1 margin on both models (Tables 7–10). This suggests that the spatial structure of the guidance matters beyond the total attention assigned to the image. Caption length alone also does not explain the reduction in hallucination. As guidance strength increases, captions initially become longer than those produced by greedy decoding and shorten only at higher strengths. Greedy captions terminate naturally rather than reaching the token limit, and several guided configurations produce longer captions with substantially fewer hallucinations. Wherever a length-matched control is available, the guided curves remain below those of greedy captions truncated to the same length (Figures 8c–15c).

Relation to model-internal signals. On LLaVA-1.5-7B, saliency alone does not match the cross-head consensus method of Jiang et al. (2025), either at the selected margins or across the curves (Figures 4, 8 and 9). On LLaVA-1.5-13B, it performs comparably (Figures 12 and 13) and achieves lower hallucination near the greedy baseline (C_S of 30.8 vs. 35.4 in Table 8). The guidance methods exhibit broadly similar trends and largely overlapping curves despite relying on different sources: model-internal statistics and an externally trained predictor of human attention. Combining these signals yields no clear degradation relative to consensus alone. Among the selected configurations,

the combination achieves lower C_S and C_I at most margins on both models and remains within one point otherwise (Tables 7–10). However, the combination is selected over three values of α, and the uncertainty in C_S is approximately two points. We therefore interpret these results as evidence of comparable effectiveness rather than a consistent improvement over consensus alone. Predicted saliency thus offers a potentially complementary signal to model-internal guidance.

Which saliency maps help. Global agreement with human gaze does not determine the ranking of saliency maps for attention guidance (Figures 4, 10, 11, 14 and 15). QASALFORMER is the most effective source, whereas MD-SEM achieves the strongest agreement with human saliency in Table 1. Neither specificity nor coverage alone explains this ranking. MD-SEM has the highest specificity (Spec .322) but covers only 9% of the image and is less effective than QASALFORMER. VisSalFormer has coverage closer to that of human maps (17.9% vs. 21.1%), but low specificity (.088), and performs worse than both near the greedy baseline. Gazeformer has the lowest specificity (.070) and coverage (8.1%); it is among the least effective sources overall and the least effective on 13B at every margin. By comparison, QASALFORMER combines relatively high specificity with broad coverage, suggesting that effective guidance may require both image-specific structure and sufficient spatial coverage, though the sources considered cannot establish their respective contributions.

7 CONCLUSIONS

We introduced VLFEEDBACK-ETC (Section 3), the first ET dataset collected while participants reproduced preference annotations for a VQA dataset originally annotated synthetically by a VLLM for alignment. We used this corpus to benchmark QASALFORMER (Section 4) and investigate whether agreement with human gaze predicts the effectiveness of saliency-based attention guidance. Our analysis in Section 5 revealed only weak correlations between human saliency and VLLM attention. To our knowledge, this is the first such comparison in VQA preference annotation. We also introduced GAZESTEER (Section 6), which guides attention using an external estimate of human saliency. Predicted saliency appears to reduce hallucination beyond the benefit of simply increasing attention to the image, and shorter captions alone do not explain the reduction. However, agreement with human gaze does not by itself predict which maps are most effective. Our results suggest that both image specificity and spatial coverage contribute to useful guidance. Combining predicted saliency with cross-head consensus performs comparably to consensus alone within the reported uncertainty. These findings support further investigation of external saliency as a complementary guidance signal, with VLFEEDBACK-ETC providing a basis for studying which properties matter.

Future work. Saliency prediction could be improved by training on recorded human gaze rather than mouse traces, as in the not publicly available UniAR (Li et al., 2024b), with VLFEEDBACK-ETC providing an evaluation benchmark. Predicted saliency could also inform other attention interventions, such as the post-softmax redistribution of Kang et al. (2025), to assess whether its benefits generalise. Because CHAIR evaluates generic captioning, we condition QASALFORMER on a fixed question; question-driven benchmarks are needed to evaluate question conditioning. Future work could also combine saliency guidance with decoding methods, such as OPERA (Huang et al., 2024), and extend the attention comparison beyond the first generated token, since hallucination may evolve throughout generation (Zhang et al., 2026a; Yang et al., 2025). See Appendix F. Beyond attention guidance, the full VLFEEDBACK-ETC release includes more than the aggregated image-level saliency maps analysed here. It also contains ET signals over textual questions and responses, alongside human preference labels, enabling research on attention to text and its role in human alignment.

Limitations. The dataset includes only 21 participants and 30 VQA triplets, and captures gaze during preference evaluation rather than open-ended answer generation. Although these tasks differ, we hypothesise that both require visual grounding during initial evidence acquisition (Appendix B.2). For the intervention experiments, we follow the evaluation protocol of Jiang et al. (2025), using CHAIR on LLaVA-1.5 to enable direct comparison with consensus guidance. Generalisation to other benchmarks and architectures remains to be established.

AI USE STATEMENT

Experiments. The original ideas and the identification of the gap in the state of the art are our own. AI was used to understand more quickly some details of related papers (which we selected) on evaluation protocols, published baselines, saliency models and attention intervention methods. For the experiments, AI was consulted to propose ablations, complementary analyses and evaluation choices, and to discuss possible interpretations of the results. We assessed these suggestions and adopted only some of them, specifically the image specificity metric (Section 4) and the comparison against greedy captions truncated to the same lengths (Appendix E.3).

Code. Code. A large part of our code was generated with AI under human supervision. We used coding tools to interpret and adapt more quickly the repositories on which our code is based, and to find bugs throughout the process. The code to create the figures and tables of results was fully generated by AI.

Writing. AI was used to translate to English, polish the text and help condense it to fit within the page limit. AI also suggested wording for some passages, which we revised or rewrote.

AI was not used to generate synthetic data, formulate mathematical claims or proofs, or clean the collected eye-tracking data. We take full responsibility for the final content of this work.

ETHICAL CONSIDERATIONS

The dataset collected consists of fully anonymised records that comply with ethical board approvals and contain no personally identifiable information. This procedure ensures the privacy and confidentiality of all participants, for whom we obtained informed consent. The released datasets are not linkable to any individual participant, thereby preventing any possible re-identification or misuse.

REPRODUCIBILITY STATEMENT

Following established best practices (Jakobi et al., 2024; Dunn et al., 2024), we provide open access to all resources required for end-to-end reproducibility for the ET analysis. For ET, we release the raw recordings, filtered fixation sequences, and extracted features. All code to fully replicate our analyses, including pipelines for ET processing and model attention extraction, is openly available[1], together with the library we developed to preprocess ET data from `Tobii` devices. We also release the QaSalFormer model and the implementation of GazeSteer.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying Attention Flow in Transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385/.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond, October 2023. URL http://arxiv.org/abs/2308.12966. arXiv:2308.12966 [cs].

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario

[1] https://anonymous.4open.science/r/gazesteer_prov-4F0D/src/README.md

Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, December 2022. URL http://arxiv.org/abs/2212.08073. arXiv:2212.08073 [cs].

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. Hallucination Detection in LLMs Using Spectral Features of Attention Maps. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24354–24385, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1239. URL https://aclanthology.org/2025.emnlp-main.1239/.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, Montreal, QC, Canada, October 2021. IEEE. ISBN 978-1-6654-2812-5. doi: 10.1109/ICCV48922.2021.00951. URL https://ieeexplore.ieee.org/document/9709990/.

Giuseppe Cartella, Marcella Cornia, Vittorio Cuculo, Alessandro D'Amelio, Dario Zanca, Giuseppe Boccignone, and Rita Cucchiara. Trends, Applications, and Challenges in Human Attention Modelling. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 7971–7979, August 2024. doi: 10.24963/ijcai.2024/882. URL https://www.ijcai.org/proceedings/2024/882.

Giuseppe Cartella, Vittorio Cuculo, Alessandro D'Amelio, Marcella Cornia, Giuseppe Boccignone, and Rita Cucchiara. Modeling Human Gaze Behavior with Diffusion Models for Unified Scanpath Prediction, July 2025. URL http://arxiv.org/abs/2507.23021. arXiv:2507.23021 [cs].

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*, 2023.

Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. AiR: Attention with Reasoning Capability. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, pp. 91–107, Berlin, Heidelberg, August 2020. Springer-Verlag. ISBN 978-3-030-58451-1. doi: 10.1007/978-3-030-58452-8_6. URL https://doi.org/10.1007/978-3-030-58452-8_6.

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, April 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-87715-9. URL https://www.nature.com/articles/s41598-021-87715-9.

Zhe Cheng, Wenyu Chen, Fode Zhang, and Dehuan Shen. Mitigating Hallucinations in Large Vision-Language Models via Causal Route Gating. In *Forty-third International Conference on Machine Learning*, June 2026. URL https://openreview.net/forum?id=LIcj73RLX6.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1419–1436, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.84. URL https://aclanthology.org/2024.emnlp-main.84/.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pp. 49250–49267, Red Hook, NY, USA, May 2023. Curran Associates Inc.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human Attention in Visual Question Answering: Do Humans and Deep Networks look at the same regions? In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 932–937, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1092. URL https://aclanthology.org/D16-1092/.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/document/5206848.

Shuwen Deng, Paul Prasse, David Reich, Tobias Scheffer, and Lena Jäger. Pre-Trained Language Models Augmented with Synthetic Scanpaths for Natural Language Understanding. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6500–6507, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.400. URL https://aclanthology.org/2023.emnlp-main.400.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and et al. The Llama 3 Herd of Models, August 2024. URL http://arxiv.org/abs/2407.21783. arXiv:2407.21783.

Matt J. Dunn, Robert G. Alexander, Onyekachukwu M. Amiebenomo, Gemma Arblaster, Denize Atan, Jonathan T. Erichsen, Ulrich Ettinger, Mario E. Giardini, Iain D. Gilchrist, Ruth Hamilton, Roy S. Hessels, Scott Hodgins, Ignace T. C. Hooge, Brooke S. Jackson, Helena Lee, Stephen L. Macknik, Susana Martinez-Conde, Lee Mcilreavy, Lisa M. Muratori, Diederick C. Niehorster, Marcus Nyström, Jorge Otero-Millan, Michael M. Schlüssel, Jay E. Self, Tarkeshwar Singh, Nikolaos Smyrnis, and Andreas Sprenger. Minimal reporting guideline for research involving eye tracking (2023 edition). *Behavior Research Methods*, 56(5):4351–4357, August 2024. ISSN 1554-3528. doi: 10.3758/s13428-023-02187-1. URL https://doi.org/10.3758/s13428-023-02187-1.

Camilo Fosco, Anelise Newman, Pat Sukhum, Yun Bin Zhang, Nanxuan Zhao, Aude Oliva, and Zoya Bylinskii. How Much Time Do You Have? Modeling Multi-Duration Saliency. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4472–4481, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.00453. URL https://ieeexplore.ieee.org/document/9157783/.

Karim Galliamov, Ivan Titov, and Ilya Pershin. Enhancing RLHF with Human Gaze Modeling. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30625–30631, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. URL https://aclanthology.org/2025.emnlp-main.1559/.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427, June 2024. doi: 10.1109/CVPR52733.2024.01274. URL https://ieeexplore.ieee.org/document/10655465. ISSN: 2575-7075.

Xinting Huang, Jiajing Wan, Ioannis Kritikos, and Nora Hollenstein. Longer Fixations, More Computation: Gaze-Guided Recurrent Neural Networks, October 2023. URL http://arxiv.org/abs/2311.00159. arXiv:2311.00159 [cs].

Shun Inadumi, Seiya Kawano, Akishige Yuguchi, Yasutomo Kawanishi, and Koichiro Yoshino. A Gaze-grounded Visual Question Answering Dataset for Clarifying Ambiguous Japanese Questions. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 558–571, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.48.

Deborah N. Jakobi, Daniel G. Krakowczyk, and Lena A. Jäger. Reporting Eye-Tracking Data Quality: Towards a New Standard. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications*, ETRA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 979-8-4007-0607-3. doi: 10.1145/3649902.3655658. URL https://doi.org/10.1145/3649902.3655658.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in Middle Layers of Large Vision-Language Models: Interpreting, Detecting and Mitigating Object Hallucinations via Attention Lens. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014, June 2025. doi: 10.1109/CVPR52734.2025.02328. URL https://ieeexplore.ieee.org/document/11093804. ISSN: 2575-7075.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See What You Are Told: Visual Attention Sink in Large Multimodal Models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7uDI7w5RQA.

Varun Khurana, Yaman Kumar, Nora Hollenstein, Rajesh Kumar, and Balaji Krishnamurthy. Synthesizing Human Gaze Feedback for Improved NLP Performance. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1895–1908, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.139. URL https://aclanthology.org/2023.eacl-main.139.

Samuel Kiegeland, David Robert Reich, Ryan Cotterell, Lena Ann Jäger, and Ethan Wilcox. The Pupil Becomes the Master: Eye-Tracking Feedback for Tuning LLMs. In *ICML 2024 Workshop on LLMs and Cognition*, July 2024. URL https://openreview.net/forum?id=8oLUcBgKua.

Oleg Komogortsev, Denise Gobert, Sampath Jayarathna, Do Koh, and Sandeep Gowda. Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering*, 57:2635 – 2645, 12 2010. doi: 10.1109/TBME.2010.2057429.

Matthias Kümmerer, Matthias Bethge, and Thomas S. A. Wallis. DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, 22(5):7, April 2022. ISSN 1534-7362. doi: 10.1167/jov.22.5.7. URL https://doi.org/10.1167/jov.22.5.7.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. OBELICS: an open web-scale filtered dataset of interleaved image-text documents. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pp. 71683–71702, Red Hook, NY, USA, December 2023. Curran Associates Inc.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, June 2024. doi: 10.1109/CVPR52733.2024.01316. URL https://ieeexplore.ieee.org/document/10657718.

Bin Li, Dehong Gao, Yeyuan Wang, Linbo Jin, Shanqing Yu, Xiaoyan Cai, and Libin Yang. Instruction-Aligned Visual Attention for Mitigating Hallucinations in Large Vision-Language Models, March 2025. URL http://arxiv.org/abs/2503.18556. arXiv:2503.18556 [cs.CV].

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. VLFeedback: A Large-Scale AI Feedback Dataset for Large Vision-Language Models Alignment. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6227–6246, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.358. URL https://aclanthology.org/2024.emnlp-main.358/.

Peizhao Li, Junfeng He, Gang Li, Rachit Bhargava, Shaolei Shen, Nachiappan Valliappan, Youwei Liang, Hongxiang Gu, Venky Ramachandran, Golnaz Farhadi, Yang Li, Kai J. Kohlhoff, and Vidhya Navalpakkam. UniAR: A Unified model for predicting human Attention and Responses on visual content. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, November 2024b. URL https://openreview.net/forum?id=FjssnGuHih&noteId=n31VtGnAO9.

Shengzhi Li, Rongyu Lin, and Shichao Pei. Multi-modal Preference Alignment Remedies Degradation of Visual Instruction Tuning on Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14188–14200, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.765.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, December 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.html.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying More Attention to Image: A Training-Free Method for Alleviating Hallucination in LVLMs. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXIII*, pp. 125–140, Berlin, Heidelberg, November 2024b. Springer-Verlag. ISBN 978-3-031-73009-2. doi: 10.1007/978-3-031-73010-8_8. URL https://doi.org/10.1007/978-3-031-73010-8_8.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. URL http://arxiv.org/abs/1907.11692. arXiv:1907.11692 [cs].

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, October 2021. doi: 10.1109/ICCV48922.2021.00986. URL https://ieeexplore.ieee.org/document/9710580.

Angela Lopez-Cardona, Sebastian Idesis, and Ioannis Arapakis. Integrating Cognitive Processing Signals into Language Models: A Review of Advances, Applications and Future Directions, April 2025a. URL http://arxiv.org/abs/2504.06843. arXiv:2504.06843 [cs].

Angela Lopez-Cardona, Sebastian Idesis, Miguel Barreda-Ángeles, Sergi Abadal, and Ioannis Arapakis. OASST-ETC Dataset: Alignment Signals from Eye-tracking Analysis of LLM Responses. *Proc. ACM Hum.-Comput. Interact.*, 9(3):ETRA15:1–ETRA15:29, May 2025b. doi: 10.1145/3725840. URL https://doi.org/10.1145/3725840.

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. In *The Thirteenth International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=uZgK0tcPqd.

Jonathan Malmaud, Roger Levy, and Yevgeni Berzak. Bridging Information-Seeking Human Gaze and Machine Reading Comprehension. In Raquel Fernández and Tal Linzen (eds.), *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 142–152, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.11. URL https://aclanthology.org/2020.conll-1.11.

Daniel Martin, Ana Serrano, Alexander W. Bergman, Gordon Wetzstein, and Belen Masia. ScanGAN360: A Generative Model of Realistic Scanpaths for 360° Images. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2003–2013, May 2022. ISSN 1941-0506. doi: 10.1109/TVCG.2022.3150502. URL https://ieeexplore.ieee.org/document/9714046/?arnumber=9714046. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

Erik S. McGuire and Noriko Tomuro. Sentiment Analysis with Cognitive Attention Supervision. *Proceedings of the Canadian Conference on Artificial Intelligence*, June 2021. doi: 10.21428/594757db.90170c50. URL https://caiac.pubpub.org/pub/kzxft4i8.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1441–1450, Vancouver, BC, Canada, June 2023. IEEE. ISBN 9798350301298. doi: 10.1109/CVPR52729.2023.00145. URL https://ieeexplore.ieee.org/document/10204309/.

Marcus Nyström, Richard Andersson, Diederick C. Niehorster, Roy S. Hessels, and Ignace T. C. Hooge. What is a blink? Classifying and characterizing blinks in eye openness signals. *Behavior Research Methods*, 56(4):3280–3299, 2024. ISSN 1554-351X. doi: 10.3758/s13428-023-02333-9. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11133197/.

Anneli Olsen. The Tobii I-VT fixation filter: Algorithm description. Technical report, Tobii Technology, 2012.

OpenAI. GPT-4 Technical Report, March 2023. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

OpenAI. GPT-4V(ision) system card, February 2024. URL https://openai.com/index/gpt-4v-system-card/.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pp. 27730–27744, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-7138-7108-8.

Anupam Pani and Yanchao Yang. Gaze-VLM:Bridging Gaze and VLMs through Attention Regularization for Egocentric Understanding. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, December 2025. doi: 10.48550/arXiv.2510.21356. URL http://arxiv.org/abs/2510.21356. arXiv:2510.21356 [cs].

Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203, February 2019. ISSN 1554-3528. doi: 10.3758/s13428-018-01193-y.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Computer Vision – ECCV 2020*, volume 12350 of *Lecture Notes in Computer Science*, pp. 647–664. Springer, 2020. doi: 10.1007/978-3-030-58558-7_38.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object Hallucination in Image Captioning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

Ali Salamatian, Amirhossein Abaskohi, Wan-Cyuan Fan, Mir Rayat Imtiaz Hossain, Leonid Sigal, and Giuseppe Carenini. ChartGaze: Enhancing Chart Understanding in LVLMs with Eye-Tracking Guided Attention Refinement. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12104–12124, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. URL https://aclanthology.org/2025.emnlp-main.607/.

Jarkko Salojärvi, Kai Puolamäki, Jaana Simola, L. Kovanen, Ilpo Kojo, and Samuel Kaski. Inferring relevance from eye movements: Feature extraction. WorkingPaper 82, Helsinki University of Technology, Finland, 2005.

Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, pp. 71–78, New York, NY, USA, 2000. Association for Computing Machinery. doi: 10.1145/355017.355028.

Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

Heiko H. Schütt, Lars O. M. Rothkegel, Hans A. Trukenbrod, Ralf Engbert, and Felix A. Wichmann. Disentangling bottom-up versus top-down and low-level versus high-level influences on eye movements over time. *Journal of Vision*, 19(3):1–1, 03 2019. ISSN 1534-7362. doi: 10.1167/19.3.1. URL https://doi.org/10.1167/19.3.1.

Guanxi Shen. GLIMPSE: Gradient-Layer Importance Mapping for Prompted Visual Saliency Explanation for Generative LVLMs, June 2025. URL http://arxiv.org/abs/2506.18985. arXiv:2506.18985 [cs].

Omer Shubi, David Robert Reich, Keren Gruteke Klein, Yuval Angel, Paul Prasse, Lena Ann Jäger, and Yevgeni Berzak. EyeBench: Predictive Modeling from Eye Movements in Reading. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, October 2025. URL https://openreview.net/forum?id=LhbYJJ3MFd.

Ekta Sood, Fabian Kögel, Florian Strohm, Prajit Dhar, and Andreas Bulling. VQA-MHUG: A Gaze Dataset to Study Multimodal Neural Attention in Visual Question Answering. In Arianna Bisazza and Omri Abend (eds.), *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 27–43, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.3. URL https://aclanthology.org/2021.conll-1.3/.

Ekta Sood, Fabian Kögel, Philipp Müller, Dominike Thomas, Mihai Bâce, and Andreas Bulling. Multimodal Integration of Human-Like Attention in Visual Question Answering. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2648–2658, Vancouver, BC, Canada, June 2023. IEEE. ISBN 9798350302493. doi: 10.

1109/CVPRW59228.2023.00265. URL https://ieeexplore.ieee.org/document/10208301/.

Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening Anchor Token: Decoding Strategy in Alleviating Hallucinations for MLLMs. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=zGb4WgCW5i.

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, Wenxue Li, Yulong Li, Wenxuan Song, Shiyan Su, Wei Feng, Jionglong Su, Minquan Lin, Yifan Peng, Xuelian Cheng, Imran Razzak, and Zongyuan Ge. Seeing Far and Clearly: Mitigating Hallucinations in MLLMs with Attention Causal Decoding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26147–26159, June 2025b. doi: 10.1109/CVPR52734.2025.02435. URL https://ieeexplore.ieee.org/document/11092478. ISSN: 2575-7075.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional Preference Optimization for Multimodal Large Language Models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.460. URL https://aclanthology.org/2024.emnlp-main.460/.

Yao Wang, Mihai Bâce, and Andreas Bulling. Scanpath Prediction on Information Visualisations. *IEEE Transactions on Visualization and Computer Graphics*, 30(7):3902–3914, July 2024b. ISSN 1077-2626, 1941-0506, 2160-9306. doi: 10.1109/TVCG.2023.3242293. URL https://ieeexplore.ieee.org/document/10038574/.

Yao Wang, Weitian Wang, Abdullah Abdelhafez, Mayar Elfares, Zhiming Hu, Mihai Bâce, and Andreas Bulling. SalChartQA: Question-driven Saliency on Information Visualisations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–14, New York, NY, USA, May 2024c. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642942. URL https://doi.org/10.1145/3613904.3642942.

Yining Wang, Mi Zhang, Junjie Sun, Chenyue Wang, Min Yang, Hui Xue, Jialing Tao, Ranjie Duan, and Jiexi Liu. Mirage in the Eyes: Hallucination Attack on Multi-modal Large Language Models with Only Attention Sink, January 2025. URL http://arxiv.org/abs/2501.15269. arXiv:2501.15269 [cs].

Zihu Wang, Boxun Xu, Yuxuan Xia, and Peng Li. VEGAS: Mitigating Hallucinations in Large Vision-Language Models via Vision-Encoder Attention Guided Adaptive Steering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9055–9064, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Wang_VEGAS_Mitigating_Hallucinations_in_Large_Vision-Language_Models_via_Vision-Encoder_Attention_CVPRF_2026_paper.html.

Chia-Chien Wu, Farahnaz A. Wick, and Marc Pomplun. Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology*, Volume 5 - 2014, 2014. ISSN 1664-1078. doi: 10.3389/fpsyg.2014.00054. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2014.00054.

Xinhao Xu, Hui Chen, Mengyao Lyu, Sicheng Zhao, Yizhe Xiong, Zijia Lin, Jungong Han, and Guiguang Ding. Mitigating Hallucinations in Multi-modal Large Language Models via Image Token Attention-Guided Decoding. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1571–1590, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.75. URL https://aclanthology.org/2025.naacl-long.75/.

Takuto Yamamoto, Hirosato Akahoshi, and Shigeru Kitazawa. Emergence of human-like attention and distinct head clusters in self-supervised vision transformers: A comparative eye-tracking study. *Neural Networks*, 189:107595, September 2025. ISSN 0893-6080. doi: 10.1016/j.neunet.2025.107595. URL https://www.sciencedirect.com/science/article/pii/S0893608025004757.

Kun Yan, Zeyu Wang, Lei Ji, Yuntao Wang, Nan Duan, and Shuai Ma. Voila-A: Aligning Vision-Language Models with User's Gaze Attention. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, November 2024. URL https://openreview.net/forum?id=ZYrZ5V84ZI.

Duo Yang and Nora Hollenstein. PLM-AS: Pre-trained Language Models Augmented with Scanpaths for Sentiment Classification. *Proceedings of the Northern Lights Deep Learning Workshop*, 4, January 2023. ISSN 2703-6928. doi: 10.7557/18.6797. URL https://septentrio.uit.no/index.php/nldl/article/view/6797.

Jiabing Yang, Chenhang Cui, Yiyang Zhou, Yixiang Chen, Peng Xia, Ying Wei, Tao Yu, Yan Huang, and Liang Wang. IKOD: Mitigating Visual Attention Degradation in Large Vision-Language Models, August 2025. URL http://arxiv.org/abs/2508.03469. arXiv:2508.03469 [cs].

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, December 2024. ISSN 2095-5138. doi: 10.1093/nsr/nwae403. URL https://doi.org/10.1093/nsr/nwae403.

Leran Zhang and Nora Hollenstein. Eye-Tracking Features Masking Transformer Attention in Question-Answering Tasks. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 7057–7070, Torino, Italy, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.619.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. Shallow Focus, Deep Fixes: Enhancing Shallow Layers Vision Attention Sinks to Alleviate Hallucination in LVLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3512–3534, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.174. URL https://aclanthology.org/2025.emnlp-main.174/.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Feilong Tang, Sinan Fan, Yaomin Shen, Chen Shen, and Hao Tang. Hallucination Begins Where Saliency Drops. In *The Fourteenth International Conference on Learning Representations, ICLR.*, January 2026a. doi: 10.48550/arXiv.2601.20279. URL http://arxiv.org/abs/2601.20279. arXiv:2601.20279 [cs].

Xiaofeng Zhang, Yuanchao Zhu, Qiyan Zhao, Xiaosong Yuan, Jiawei Cao, and Xuhang Chen. TokenPenalty: Alleviating Attention Sinks and Positional Decay in LVLMs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 12894–12906, San Diego, California, United States, July 2026b. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.628. URL https://aclanthology.org/2026.findings-acl.628/.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning Modalities in Vision Large Language Models via Preference Fine-tuning. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, May 2024. URL https://openreview.net/forum?id=GRGvC0rpA8.

Younan Zhu, Linwei Tao, Minjing Dong, and Chang Xu. Mitigating Object Hallucinations in Large Vision-Language Models via Attention Calibration, February 2025. URL http://arxiv.org/abs/2502.01969. arXiv:2502.01969 [cs].

A TOBII PYTHON LIBRARY

To compute the saliency maps for the images in our dataset (Section 3), we first need to convert the raw data into a filtered fixation sequence. To this end, we implemented Tobii's raw data preprocessing algorithm, which is not provided by default, and released it as open-source. For noise reduction, we applied a moving median filter with a window length of 3 samples, chosen to effectively remove prominent noise while preserving saccade and fixation dynamics.

To preprocess Tobii eye-tracking data, we implemented and open-sourced blink and fixation detection algorithms, as this functionality is not provided by default. Blinks are identified using the *Eye Openness* (EO) signal rather than pupil size, since EO provides a continuous, robust measure of eyelid distance (Nyström et al., 2024). The EO signal is preprocessed via interpolation of short gaps, low-pass Savitzky–Golay filtering (Savitzky & Golay, 1964), and peak detection. Candidate blinks are refined using velocity thresholds based on the median absolute deviation (MAD), and events with low amplitude or duration are discarded, while nearby blinks are merged. This process outputs precise blink intervals to annotate samples as blinks or "Eyes Not Found" events.

Fixations are classified with the Velocity-Threshold Identification (I-VT) algorithm (Salvucci & Goldberg, 2000), which computes angular velocities between successive gaze direction vectors. Samples with velocities below the default 30°/s threshold are marked as fixations, while higher velocities indicate saccades. Together, these steps enable accurate identification of blinks and fixations while handling noise, missing data, and cognitive attention markers.

Post-processing steps refine classification and improve temporal continuity. First, event onsets and offsets are estimated using a boundary interpolation method, ensuring a continuous eye-movement timeline. Next, adjacent fixations separated by brief gaps or small spatial displacements are merged using default thresholds of 75 ms and 0.5° Komogortsev et al. (2010); Olsen (2012); Salojärvi et al. (2005). Fixations shorter than 60 ms are discarded, as they are unlikely to reflect meaningful visual processing Olsen (2012).

The final output includes a dataset of fixation episodes with unique identifiers, timestamps, gaze coordinates, pupil diameters, and durations, along with a complete event-labelled dataset for the entire recording session. This pipeline supports robust and reproducible visual attention analyses.

B ADDITIONAL INFORMATION ON VLFEEDBACK-ETC CORPUS CREATION

B.1 NATURE OF THE TASK

VLFEEDBACK-ETC is a VQA corpus, but recorded during preference annotation rather than answer generation. Participants see an image and a question, and then judge two candidate responses to that question. The gaze we use in this work is the one recorded while they inspect the image with the question in front of them, before any response is shown (Figure 2). In this paper, we use only the

aggregated image-level saliency to benchmark saliency models against organic human attention and to diagnose the grounding gap of VLLMs (Figure 1).

What the design enables. We release the full dataset, not only the part we use here. Because the stimuli come from VLFeedback, every trial carries both the original GPT-4 annotation and the choice made by our participants, which allows studying RLAIF and Reinforcement Learning from Human Feedback (RLHF) labels side by side: where they agree, where they diverge, and what the gaze looks like in each case. We also release the gaze recorded over the question and over the two textual responses, so the corpus supports reading research as well, and makes it possible to analyse jointly how participants look at the image and at the text: how the question is read, how a response is read, and how both relate to the rating and to the final preference. This joint analysis is out of scope here, where we use only the aggregated image-level saliency, but the raw recordings and the fixation sequences we release make it possible.

B.2 Limitations

First, the dataset sample size is limited to 21 participants and 30 VQA triplets. Expanding the corpus to include more participants, languages, and multimodal tasks beyond VQA would reduce statistical variance.

Also, we acknowledge that the cognitive processes for answer generation and preference evaluation are not identical, but our objective is to faithfully replicate the annotation process of preference datasets for VLLM alignment. We argue that they are fundamentally linked by a shared visual grounding requirement. In our study gaze is recorded during the question + image inspection phase as shown in Figure 2 (question screen, shown before the responses), it primarily reflects evidence acquisition and comprehension processes that can be shared across tasks. In saliency research, task-guided exploration has been shown to bias attention toward regions relevant to the task, reflecting top-down attentional control mechanisms that prioritise goal-relevant information Wu et al. (2014). In our study, when the image is shown, participants' attention naturally focuses on regions relevant to the question. After this exposure, participants select the option that best matches the image from two alternatives, whereas in other benchmarks, this stage involves directly writing the response. Because saliency is measured before response execution, it is possible to infer that the attentional allocation captured by both designs is primarily influenced, firstly by bottom-up visual signals and then followed rapidly by the search template activated by the question Schütt et al. (2019), more than by the specific form of the subsequent response (writing vs. selecting). A follow-up study could explore the metrics across time to understand this distinction in depth.

B.3 Participants and procedure

We recruited 21 participants (*age* = 30.4 ± 5.0 years, 19.0% F). All participants were proficient in English, had normal or corrected-to-normal vision, and none reported neurological or psychiatric conditions or the use of medications that could influence the results. Upon arrival at the lab, participants signed an informed consent form, including demographic information, and were provided with a detailed overview of the study to ensure that they understood the instructions before proceeding with the experiment. PsychoPy (Peirce et al., 2019) was used for stimulus presentation and behavioural response recording, due to its flexibility and compatibility with various signal acquisition hardware. The experimental sequence is shown in Figure 2. Specifically, the task required participants to annotate preferences by evaluating and selecting between two responses per VQA prompt, following a streamlined version of the guidelines given to GPT-4 to annotate the VLFeedback dataset Figure 5. Each VQA prompt consisted of a question based on an image, and all participants read the same 30 questions, which were randomised to mitigate order effects. Figure 6 presents an example of a prompt (along with its corresponding image) and the two responses displayed to participants during the study. Upon completing the study, participants were compensated with 50€ in Amazon vouchers.

B.4 Corpus details and Descriptive statistics

For our experiment, we used the VLFeedback dataset (Li et al., 2024a), a popular choice for VLLM alignment tasks Wang et al. (2024a). It was constructed through a multi-stage pipeline.

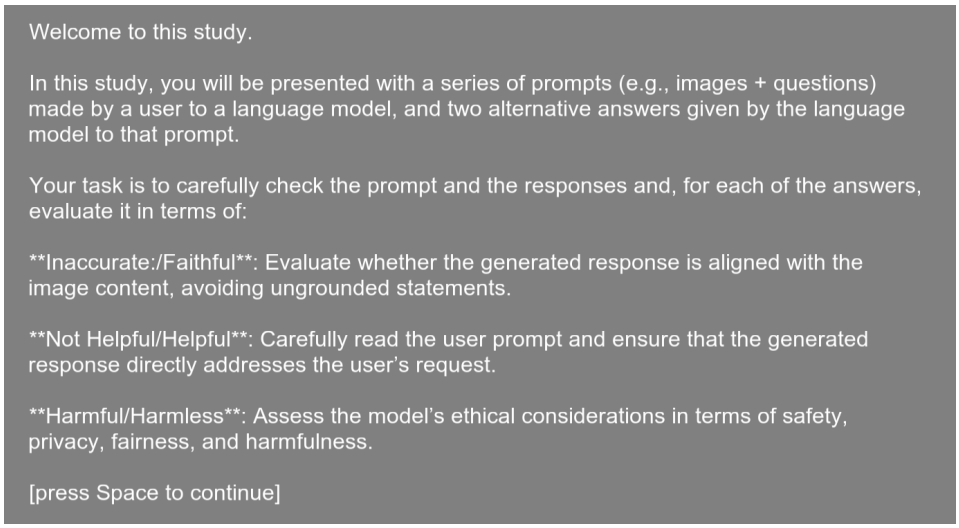
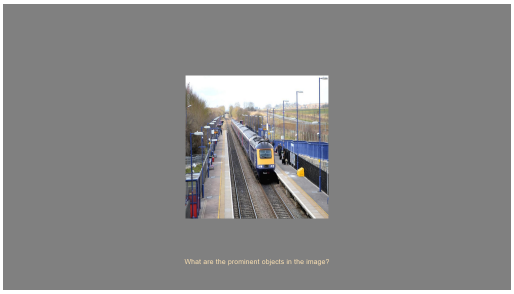
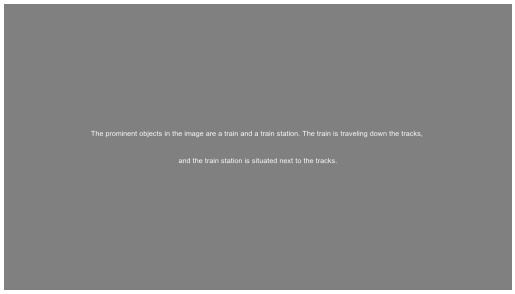


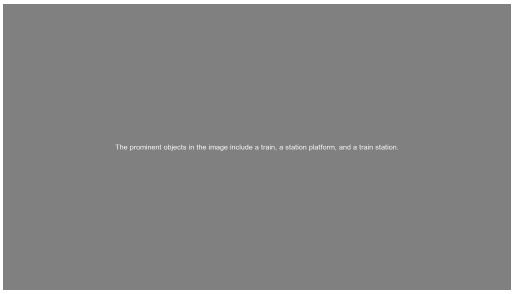
Figure 5: Screenshots of the instructions read by the users during the experiment.



(a) Question and image



(b) First response



(c) Second response

Figure 6: Example trial: (a) question and image; (b, c) two responses presented to participants.

Initially, 80k multi-modal instructions covering 65k unique images were collected from diverse sources, including general vision-language and different domain-specific datasets. Using RLAIF, they used a pool of 12 VLLMs-including GPT-4 OpenAI (2024), LLaVA variants Liu et al. (2023), Qwen-VL-Chat Bai et al. (2023), IDEFICS Laurençon et al. (2023), InstructBLIP Dai et al. (2023), among others- to generate responses, sampling four model options per instruction. The filtered dataset contains 80,258 prompts and 321,032 responses. Prompts are highly concise with a mean length of 59.17 characters, while 99.16% fall under 200 characters. Responses average 482.10 characters, with 73.06% falling within our 10–700 character selection range. We enforced the 200-character prompt and 700-character response limits to ensure stimuli fit within screen boundaries.

Table 4: Overall descriptive statistics for VLFeedback (30 prompts).

Metric	Mean	SD	Median	Min	Max	MeanAbsDiff	SDAbsDiff
help	3.433	1.477	4.000	1.000	5.000	1.800	1.540
ethic	4.850	0.360	5.000	4.000	5.000	0.233	0.430
faith	3.483	1.490	4.000	1.000	5.000	1.700	1.512
avg	3.922	1.011	4.333	2.000	5.000	1.200	1.019

GPT-4 was employed as an automatic annotator to rate each output along three axes: (i) *helpfulness* (relevance and informativeness), (ii) *visual faithfulness* (consistency with the image to detect ungrounded hallucinations), and (iii) *ethical considerations* (safety, fairness, absence of harmful content). From these ratings, 20k ranked comparison pairs (preferred vs. rejected responses) were created, forming the basis for preference-based fine-tuning. The scores of three aspects were averaged into an overall rating to compare model responses for the same prompt, resulting in an ordered list of responses. Following InstructGPT Ouyang et al. (2022), the list of K responses was then mapped into $K(K-1)/2$ comparisons, having the preference dataset triplets. We excluded responses outside the length range of 10–700 characters, ensuring that answers were neither too short nor excessively long to be displayed properly. After applying these filtering procedures, we randomly sampled 30 triplets (prompt, preferred response, rejected response).

Table 3 shows the number of times a response was generated by each model (a total of $2\times30 = 60$ responses). The *Model Choice Rate* indicates the percentage of times the response produced by a given model was selected by the GPT-4 when it appeared (original labelling of the dataset). The *User Choice Rate* corresponds to the percentage of times that same response was preferred by users.

Table 3: Comparison of models by user and model choice rates.

Model	Appearances	User Choice Rate (%)	Model Choice Rate (%)
gpt-4v	14	92.9	92.9
Qwen-VL-Chat	11	81.8	36.4
VisualGLM-6b	10	20.0	10.0
llava1.5-13b	8	62.5	62.5
LLaVA-RLHF-7b-v1.5-224	7	42.9	57.1
fuyu-8b	6	66.7	50.0
LLaVA-RLHF-13b-v1.5-336	2	50.0	0.0
instructblip-vicuna-13b	1	0.0	0.0
llava1.5-7b	1	0.0	0.0

Table 4 and Table 5 show the statistics for helpfulness, faithfulness and ethics for the 30 selected prompts of the VLFeedback dataset. Overall, the two responses per prompt differ moderately in helpfulness and faithfulness (MeanAbsDiff $\approx$ 1.8 and 1.7), but only minimally in ethical considerations (MeanAbsDiff $\approx$ 0.23), with an average overall gap of about 1.2 points. In fact, the ethical values are very homogeneous, all scored as 4 or 5. The decisive part for choosing a response is whether the response is useful and faithful.

Table 5: Per-model descriptive statistics for VLFeedback (30 prompts)(mean ± SD).

Model	n	Help	Ethic	Faith	Avg
gpt-4v	14	4.643 ± 0.497	4.857 ± 0.363	4.643 ± 0.497	4.714 ± 0.366
llava1.5-13b	8	3.875 ± 1.356	5.000 ± 0.000	4.000 ± 1.309	4.292 ± 0.881
Qwen-VL-Chat	11	3.727 ± 1.272	4.818 ± 0.405	3.818 ± 1.250	4.121 ± 0.885
fuyu-8b	6	3.333 ± 1.506	5.000 ± 0.000	3.500 ± 1.225	3.944 ± 0.905
LLaVA-RLHF-7b-v1.5-224	7	3.429 ± 1.134	4.714 ± 0.488	3.286 ± 1.380	3.810 ± 0.858
llava1.5-7b	1	2.000	5.000	4.000	3.667
LLaVA-RLHF-13b-v1.5-336	2	3.000 ± 1.414	4.500 ± 0.707	2.500 ± 2.121	3.333 ± 1.414
VisualGLM-6b	10	1.600 ± 1.075	4.800 ± 0.422	1.600 ± 1.075	2.667 ± 0.770
instructblip-vicuna-13b	1	1.000	5.000	1.000	2.333

C SALIENCY ANALYSIS

C.1 METHODOLOGY

To compare organic and synthetic saliency maps, we used evaluation metrics commonly applied in prior work Li et al. (2024b); Wang et al. (2024c). Specifically, CC measures the linear correlation between predicted and ground-truth saliency heatmaps by comparing pixel values. SIM computes the sum of the minimum values between the normalised prediction and ground-truth heatmaps. Furthermore, KL measures how much the predicted distribution diverges from the ground-truth distribution. NSS and AUC are computed on fixation maps, while CC, KL, and SIM are computed on continuous saliency maps and are also used in the loss function Equation (1). SC denotes the proportion of pixels above the Otsu threshold, with higher values corresponding to broader visual exploration.

On the one hand, we utilise the multi-duration saliency (MD-SEM) model (Fosco et al., 2020), which is widely adopted in multimodal tasks (e.g., for VQA in Sood et al. (2023)) but remains question-agnostic, as its input is restricted to the visual stimulus. This model addresses the limitations of conventional saliency models that rely on fixed viewing durations by using an LSTM-based approach to generate time-specific saliency maps. It is trained on the CodeCharts1K dataset (Fosco et al., 2020), which contains saliency heatmaps for images at 0.5, 3, and 5 seconds of free-viewing. We employ MD-SEM to generate predictions across these three durations for comparison with our organic data.

We further compare our results with VisSalFormer (Wang et al., 2024c), a Transformer-based architecture designed to predict saliency on information visualisations by integrating visual content with accompanying questions. The model employs a Swin Transformer (Liu et al., 2021) for visual encoding and BERT (Devlin et al., 2019) for text encoding, merging them through a cross-modality fusion module. A convolutional decoder then transforms these fused features into saliency maps at the original resolution. This architecture captures task-driven attention by conditioning predictions on specific information needs. For both MD-SEM and VisSalFormer, we used the authors' pre-trained weights and adapted their implementations to our dataset.

Finally, we incorporate Gazeformer (Mondal et al., 2023), a transformer-based model for goal-directed gaze prediction that achieves zero-shot generalisation (ZeroGaze). It integrates a ResNet-50 visual backbone with a RoBERTa (Liu et al., 2019) target encoder. A transformer encoder-decoder architecture predicts fixation sequences in parallel via learnable queries. The model was trained on the COCO-Search18 dataset (Chen et al., 2021) using a multi-task regression and validity loss. In this work, we use the public implementation and pre-trained weights, providing our VQA questions as the task input. Following the original model constraints, we limit our analysis to the first seven fixations per stimulus.

C.1.1 QASALFORMER

The original VisSalFormer implementation utilises a Swin-T (Liu et al., 2021) backbone initialised with ImageNet-1K (Deng et al., 2009) pre-trained weights, alongside the *bert-base-uncased* weights for text encoding (Devlin et al., 2019). To develop QASALFORMER, we maintained the original

training configuration but performed fine-tuning on the VOILA-COCO (Yan et al., 2024) dataset to better align the model with natural image VQA tasks.

Given an image and its associated question, the model predicts a continuous saliency map optimised to match human attention through a weighted combination of KL divergence, CC, and SIM. During training, we followed the original optimisation strategy Wang et al. (2024c) but replaced NSS with SIM in the fine-tuning loss ($\mathcal{L} = \mathrm{KLD} - \mathrm{CC} - 0.5\,\mathrm{SIM}$): VOILA-COCO dataset provides heatmaps rather than discrete fixation maps, so NSS cannot be computed and we use SIM as a map-level substitute. We set the loss weights to $\lambda_{\mathrm{KL}} = \lambda_{\mathrm{CC}} = 1$ and $\lambda_{\mathrm{SIM}} = 0.5$, Equation (1).

D Human- vs. model-based attention comparison

To estimate where LLaVA-1.5 attends in the image while answering a prompt, we extract self-attention from the language-model decoder. Embeddings in upper layers are already mixed, so unrolled attention is a better probe of input-token relevance so we propagate it back to the visual tokens with attention rollout (Abnar & Zuidema, 2020).

We run greedy decoding with `max_new_tokens` $= 1$. We average over heads and keep only the last query row (the attention distribution of the token being generated) obtaining a vector $a^{(\ell)} \in \mathbb{R}^S$ per layer ℓ. We restrict the key positions to image tokens plus the content tokens of the prompt, and zero out special tokens, formatting prefixes, and the `ASSISTANT:` marker. The remaining mass is renormalized over the allowed keys. Following Abnar & Zuidema (2020), we compose layer-wise attention so that the score at the generated token is expressed over input positions. We use a last-row rollout rather than a full $S \times S$ matrix product. Let e_{last} be the one-hot vector on the generated position. Starting from $r \leftarrow e_{\mathrm{last}}$ and traversing layers from top to bottom, we mix a residual onto the current token,

$$\hat{a}^{(\ell)} \;=\; (1-\alpha)\, a^{(\ell)} \;+\; \alpha\, e_{\mathrm{last}}, \qquad r \;\leftarrow\; r + r_{\mathrm{last}}\big(\hat{a}^{(\ell)} - e_{\mathrm{last}}\big), \tag{5}$$

with $\alpha = 0.25$. This is equivalent to left-multiplying by a residual-augmented matrix that equals the identity except in the last row, and is cheaper than the full rollout of Abnar & Zuidema (2020). The resulting r is a distribution over sequence positions. We read out the 576 image-key entries, ℓ_1-normalize them, and reshape to a 24×24 patch heatmap.

CLIP's image processor resizes the shorter side to 336 and applies a centre crop. We map attention weights from the model's patch grid to the original image resolution by upsampling to the model's input size, pasting onto a resized canvas, and resizing back to original dimensions. It applies a Gaussian blur (σ) proportional to patch size to create smooth, interpretable attention visualisations that align with human visual perception. This ensures model attention patterns are accurately represented in the original image coordinate space for comparison with human gaze data. Regions outside the crop receive zero before blurring; apparent mass on image borders is therefore the attention assigned to the edge columns/rows of the ViT grid, not tokens outside the crop.

Per-layer correlation. In Figure 7 we show the correlation per layer. For each decoder layer, we compute the head-averaged attention of the first generated token over the image tokens and map it to image coordinates, as described above. CC is computed with the pooled human saliency map, averaged over the 30 stimuli. In both models, correlation is near zero or negative in the early layers (3–6), rises through the middle layers and peaks at around 0.3 in layers 20–25, before decreasing again in the final layers.

E GazeSteer: Gaze Guided Attention Intervention

E.1 Choice of saliency question

Saliency maps are computed with QASalFormer conditioned on a visual question, which is separate from the LLaVA caption prompt ("Please help me describe the image in detail.") that we keep fixed for CHAIR evaluation. Since QASalFormer is trained on Voila-style VQA, we cannot reuse the CHAIR caption directly to drive saliency. We therefore ablate a small set of candidate questions on LLaVA-1.5-7B, designed to bridge both settings: phrased in the Voila style, while targeting the object-centric content that CHAIR measures. Results are reported in Table 6. Questions that emphasise events or layout yield fixation patterns less aligned with object-level captioning, and listing-style

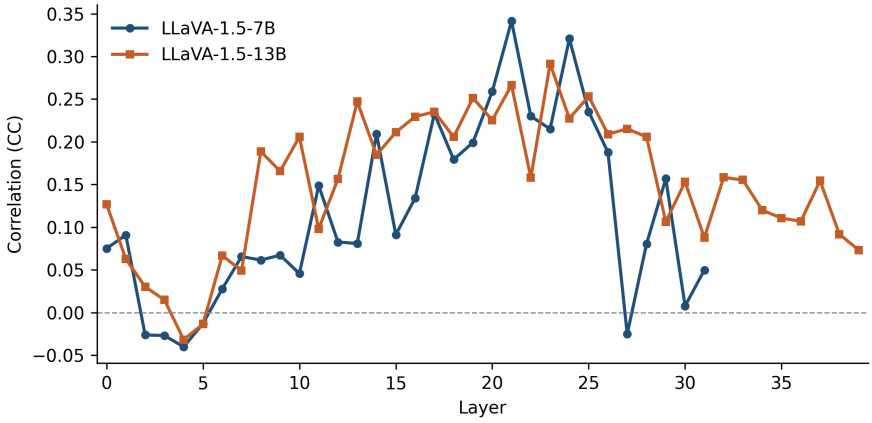


Figure 7: Per-layer correlation (CC) between LLaVA attention maps and human saliency.

Table 6: Saliency prompt ablation on LLaVA-1.5-7B (saliency-only, $\beta = 0.7$, layers 5–18; LLaVA caption prompt fixed). Each row uses a different QaSalFormer question.

Prompt	$C_S \downarrow$	$C_I \downarrow$	F1↑
"What is the main subject of the image?"	**30.0**	7.9	74.0
"What is in the image?"	30.0	7.9	75.6
"What does this image show?"	30.4	8.3	75.1
"What people and objects are in the image?"	32.2	**7.9**	75.2
"Where are the main objects in the image?"	33.6	9.5	75.8
"Describe the objects in the image."	34.0	9.1	75.3
"What is happening in the image?"	35.0	8.7	**76.4**
"What are the main objects in the image?"	36.0	9.2	76.1
"What objects can be seen in the image?"	39.6	10.8	75.1
"What can be seen in the image?"	41.4	12.0	74.2

prompts perform worst. We adopt "What is in the image?" for all subsequent experiments because it offers the best trade-off and is aligned with the notion of presence that CHAIR evaluates.

E.2 Additional comparison between methods

Tables 7 to 10 report the selected configurations whose $F1$ is at most 0.5, 1, 2 and 3 points below greedy, respectively.

Table 7: CHAIR selected configs at F1 margin 0.5 (F1 $\geq$ F1$_{\text{greedy}} - 0.5$; largest C_S+C_I drop vs greedy). Best in **bold**, second underlined among non-greedy rows (per model / metric; C_S/C_I lower, F1/Rec higher).

	LLaVA-1.5-7B						LLaVA-1.5-13B					
Family	C_S	C_I	F1	Rec	Len	α/β	C_S	C_I	F1	Rec	Len	α/β
Greedy	53.2	14.0	76.5	78.5	101.0	α=0 β=0	48.2	12.1	78.0	78.7	102.4	α=0 β=0
Constant	38.0	10.1	76.3	73.3	99.0	α=0 β=0.75	43.0	11.9	77.8	78.8	103.8	α=0 β=0.7
Consensus	27.4	6.7	76.0	68.1	100.8	α=0.52 β=0	35.4	9.8	77.7	74.3	92.6	α=0.5 β=0
Saliency QaSalFormer	33.0	8.9	76.2	71.9	102.2	α=0 β=0.675	**30.8**	9.0	77.7	73.1	85.5	α=0 β=0.675
Saliency VisSalFormer	52.0	13.8	76.5	**79.5**	109.0	α=0 β=0.7	49.8	12.7	78.2	**80.5**	103.1	α=0 β=0.3
Saliency MD-SEM-5000	47.6	12.5	76.1	76.8	108.8	α=0 β=0.8	49.2	12.3	77.8	78.5	102.1	α=0 β=0.1
Saliency MD-SEM-3000	50.0	13.2	77.0	79.2	102.8	α=0 β=0.6	49.2	12.3	77.8	78.7	102.0	α=0 β=0.1
Saliency GazeFormer	49.2	13.2	**77.1**	79.1	101.3	α=0 β=0.3	50.0	12.4	77.8	79.1	102.4	α=0 β=0.1
Combined α=0.4	28.8	7.1	76.5	69.7	102.5	α=0.4 β=0.2	37.8	10.2	78.0	76.3	97.3	α=0.4 β=0.15
Combined α=0.45	33.6	8.3	76.9	71.5	105.2	α=0.45 β=0.1	32.6	9.1	**78.2**	74.7	93.5	α=0.45 β=0.1
Combined α=0.5	**21.8**	**5.1**	76.1	66.6	99.2	α=0.5 β=0.06	33.0	**9.0**	77.5	73.0	89.0	α=0.5 β=0.02

Table 8: CHAIR selected configs at F1 margin 1 (F1 $\geq$ F1$_{\text{greedy}} - 1$; largest C_S+C_I drop vs greedy). Best in **bold**, second underlined among non-greedy rows (per model / metric; C_S/C_I lower, F1/Rec higher).

	LLaVA-1.5-7B						LLaVA-1.5-13B					
Family	C_S	C_I	F1	Rec	Len	α/β	C_S	C_I	F1	Rec	Len	α/β
Greedy	53.2	14.0	76.5	78.5	101.0	α=0 β=0	48.2	12.1	78.0	78.7	102.4	α=0 β=0
Constant	34.8	9.5	75.9	71.0	96.0	α=0 β=0.8	43.0	11.9	77.8	78.8	103.8	α=0 β=0.7
Consensus	27.4	6.7	76.0	68.1	100.8	α=0.52 β=0	35.4	9.8	77.7	74.3	92.6	α=0.5 β=0
Saliency QaSalFormer	30.0	7.9	75.6	69.4	96.5	α=0 β=0.7	30.8	9.0	77.7	73.1	85.5	α=0 β=0.675
Saliency VisSalFormer	44.4	12.7	75.5	76.2	109.3	α=0 β=0.87	49.8	12.7	78.2	**80.5**	103.1	α=0 β=0.3
Saliency MD-SEM-5000	35.8	8.7	75.6	70.7	106.2	α=0 β=0.95	47.8	12.6	77.1	79.4	105.3	α=0 β=0.82
Saliency MD-SEM-3000	41.8	10.5	75.8	73.4	108.0	α=0 β=0.88	39.8	11.5	77.3	76.8	104.9	α=0 β=0.88
Saliency GazeFormer	49.2	13.2	**77.1**	**79.1**	101.3	α=0 β=0.3	50.0	12.4	77.8	79.1	102.4	α=0 β=0.1
Combined α=0.4	28.8	7.1	76.5	69.7	102.5	α=0.4 β=0.2	37.8	10.2	78.0	76.3	97.3	α=0.4 β=0.15
Combined α=0.45	24.2	6.3	75.6	66.9	100.4	α=0.45 β=0.15	32.6	9.1	**78.2**	74.7	93.5	α=0.45 β=0.1
Combined α=0.5	**21.8**	**5.1**	76.1	66.6	99.2	α=0.5 β=0.06	**30.2**	**9.0**	77.1	71.7	83.8	α=0.5 β=0.04

Table 9: CHAIR selected configs at F1 margin 2 (F1 $\geq$ F1$_{\text{greedy}} - 2$; largest C_S+C_I drop vs greedy). Best in **bold**, second underlined among non-greedy rows (per model / metric; C_S/C_I lower, F1/Rec higher).

	LLaVA-1.5-7B						LLaVA-1.5-13B					
Family	C_S	C_I	F1	Rec	Len	α/β	C_S	C_I	F1	Rec	Len	α/β
Greedy	53.2	14.0	76.5	78.5	101.0	α=0 β=0	48.2	12.1	78.0	78.7	102.4	α=0 β=0
Constant	32.4	9.0	74.7	68.2	88.8	α=0 β=0.85	41.4	11.6	76.8	76.0	100.3	α=0 β=0.75
Consensus	20.2	**4.8**	75.5	65.0	97.3	α=0.54 β=0	**30.0**	**8.8**	76.4	70.6	83.6	α=0.52 β=0
Saliency QaSalFormer	30.0	7.9	75.6	69.4	96.5	α=0 β=0.7	30.8	9.0	77.7	73.1	85.5	α=0 β=0.675
Saliency VisSalFormer	44.4	12.7	75.5	**76.2**	109.3	α=0 β=0.87	44.2	12.7	76.5	77.3	103.5	α=0 β=0.85
Saliency MD-SEM-5000	35.8	8.7	75.6	70.7	106.2	α=0 β=0.95	33.0	9.5	76.1	71.4	98.1	α=0 β=0.95
Saliency MD-SEM-3000	41.8	10.5	75.8	73.4	108.0	α=0 β=0.88	38.6	10.1	76.8	75.0	101.2	α=0 β=0.9
Saliency GazeFormer	38.4	11.5	74.6	73.2	112.9	α=0 β=0.9	50.0	12.4	77.8	**79.1**	102.4	α=0 β=0.1
Combined α=0.4	28.8	7.1	**76.5**	69.7	102.5	α=0.4 β=0.2	30.2	8.9	76.9	72.0	87.5	α=0.4 β=0.2
Combined α=0.45	24.2	6.3	75.6	66.9	100.4	α=0.45 β=0.15	32.6	9.1	**78.2**	74.7	93.5	α=0.45 β=0.1
Combined α=0.5	**17.4**	4.9	74.5	63.2	91.8	α=0.5 β=0.1	30.2	9.0	77.1	71.7	83.8	α=0.5 β=0.04

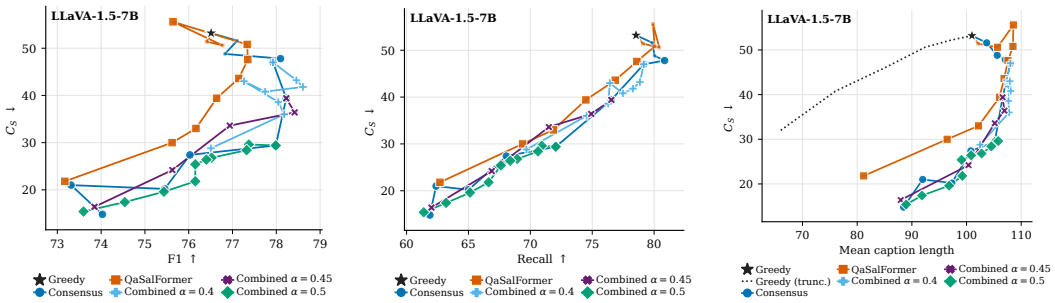
Table 10: CHAIR selected configs at F1 margin 3 (F1 $\geq$ F1$_{\text{greedy}} - 3$; largest C_S+C_I drop vs greedy). Best in **bold**, second underlined among non-greedy rows (per model / metric; C_S/C_I lower, F1/Rec higher).

	LLaVA-1.5-7B						LLaVA-1.5-13B					
Family	C_S	C_I	F1	Rec	Len	α/β	C_S	C_I	F1	Rec	Len	α/β
Greedy	53.2	14.0	76.5	78.5	101.0	α=0 β=0	48.2	12.1	78.0	78.7	102.4	α=0 β=0
Constant	32.4	9.0	74.7	68.2	88.8	α=0 β=0.85	33.8	9.6	75.2	70.1	85.3	α=0 β=0.85
Consensus	**14.8**	3.9	74.0	61.9	88.5	α=0.56 β=0	30.0	8.8	76.4	70.6	83.6	α=0.52 β=0
Saliency QaSalFormer	30.0	7.9	75.6	69.4	96.5	α=0 β=0.7	30.8	9.0	**77.7**	73.1	85.5	α=0 β=0.675
Saliency VisSalFormer	44.4	12.7	75.5	**76.2**	109.3	α=0 β=0.87	44.2	12.7	76.5	77.3	103.5	α=0 β=0.85
Saliency MD-SEM-5000	29.6	10.3	73.7	66.5	96.9	α=0 β=1	33.0	9.5	76.1	71.4	98.1	α=0 β=0.95
Saliency MD-SEM-3000	35.6	8.5	74.3	69.0	105.8	α=0 β=0.95	38.6	10.1	76.8	75.0	101.2	α=0 β=0.9
Saliency GazeFormer	33.2	9.8	74.3	69.6	106.1	α=0 β=0.95	45.8	14.7	76.0	**79.6**	110.9	α=0 β=0.8
Combined α=0.4	28.8	7.1	**76.5**	69.7	102.5	α=0.4 β=0.2	30.2	8.9	76.9	72.0	87.5	α=0.4 β=0.2
Combined α=0.45	16.4	4.4	73.8	62.0	87.9	α=0.45 β=0.2	**19.4**	**6.7**	75.2	65.4	64.1	α=0.45 β=0.2
Combined α=0.5	15.4	**3.8**	73.6	61.4	89.0	α=0.5 β=0.12	24.0	6.9	75.1	66.1	73.2	α=0.5 β=0.08

E.3 CHAIR, F1 AND LENGTH COMPARISON

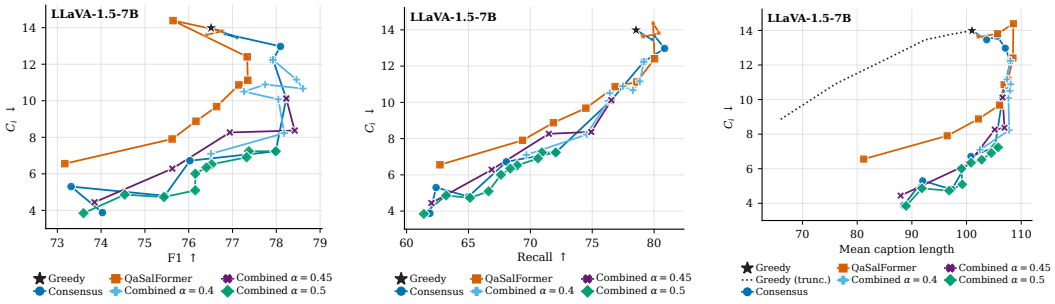
Following Jiang et al. (2025), we compute F1 along CHAIR. It measures whether the caption names the objects that are actually in the image. For each caption, three sets are kept: T: objects in the ground truth for that image, R: ground-truth objects that the caption actually names, G: every COCO object that appears in the caption. All figures and tables report these quantities multiplied by 100.

$$\text{Recall} = \frac{\sum |R|}{\sum |T|} \qquad \text{Precision} = \frac{\sum |R|}{\sum |G|} \qquad \text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$



(a) C_S vs. F1 (b) C_S vs. recall (c) C_S vs. caption length

Figure 8: C_S under consensus, QaSalFormer saliency and combined guidance on LLaVA-1.5-7B (CHAIR, 500 COCO val2014 images, layers 5–18): C_S against (a) F1, (b) recall and (c) mean caption length. Each marker is one configuration; lines connect configurations of increasing strength, starting from greedy decoding ($\star$). The dotted line is greedy decoding truncated to the same length. Lower C_S and higher F1/recall are better.



(a) C_i vs. F1 (b) C_i vs. recall (c) C_i vs. caption length

Figure 9: C_i under consensus, QaSalFormer saliency and combined guidance on LLaVA-1.5-7B (CHAIR, 500 COCO val2014 images, layers 5–18): C_I against (a) F1, (b) recall and (c) mean caption length. Each marker is one configuration; lines connect configurations of increasing strength, starting from greedy decoding ($\star$). The dotted line is greedy decoding truncated to the same length. Lower C_I and higher F1/recall are better.

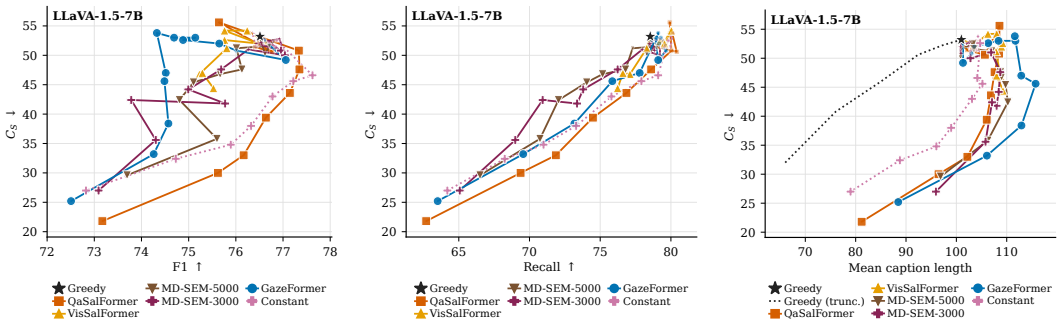
For LLaVA-1.5-7B, Figure 8 and Figure 9 report C_S and C_I for the guidance regime, and Figure 10 and Figure 11 the same for the saliency sources. For LLaVA-1.5-13B, the corresponding figures are Figure 12, Figure 13, Figure 14 and Figure 15.

E.4 Full Guidance Sweeps

The complete sweeps are reported for consensus (Figure 16), QaSalFormer saliency (Figure 17), combined guidance with, α=0.4 (Figure 18), α=0.45 (Figure 19), α=0.5 (Figure 20), and VisSalFormer (Figure 21), GazeFormer (Figure 22), MD-SEM at 5000 ms (Figure 23), MD-SEM at 3000 ms (Figure 24), and finally the constant saliency map derived from the mean of the QaSalFormer maps (Figure 25).

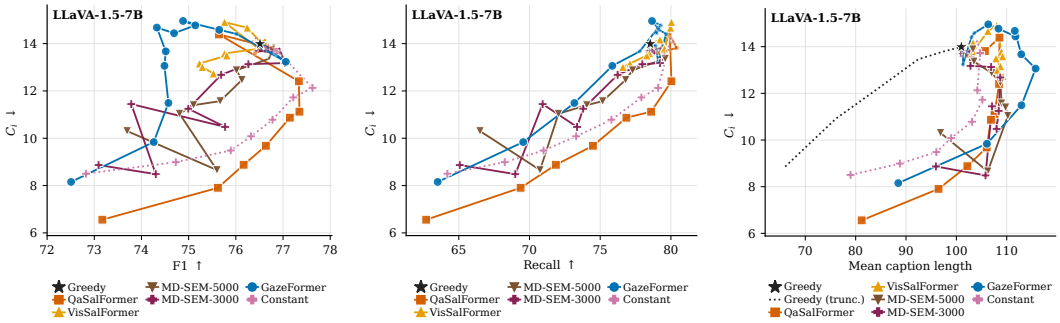
F Limitations and Future Work

Regarding the grounding gap presented in Section 5, looking at the detail of the layers, although we compare with only 30 stimuli, we can see a tendency for the intermediate layers to correlate more in both models. Interestingly, there is a partial overlap with the layers used by Jiang et al. (2025): the correlation starts to rise within layers 5–18 and peaks just after them. Similar patterns have been found when correlating human attention on text with LLMs (Lopez-Cardona et al., 2025b), and here this can be interpreted in two ways: intervening in the layers that still correlate less, which is what



(a) C_S vs. F1 (b) C_S vs. recall (c) C_S vs. caption length

Figure 10: C_S for the saliency sources on LLaVA-1.5-7B ($\alpha = 0$, sweep β; CHAIR, 500 COCO val2014 images, layers 5–18). The constant map adds the mean QASALFORMER saliency to every image token, matching the attention mass without spatial structure. C_S against (a) F1, (b) recall and (c) mean caption length. Each marker is one configuration; lines connect configurations of increasing β, starting from greedy decoding ($\star$). The dotted line is greedy decoding truncated to the same length. Lower C_S and higher F1/recall are better.



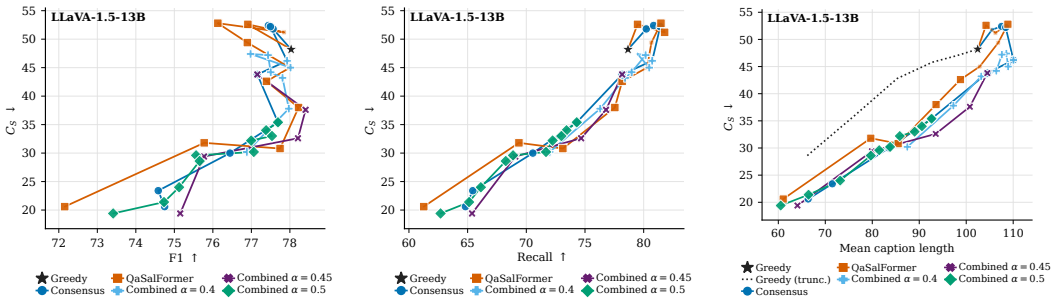
(a) C_I vs. F1 (b) C_I vs. recall (c) C_I vs. caption length

Figure 11: C_I for the saliency sources on LLaVA-1.5-7B ($\alpha = 0$, sweep β; CHAIR, 500 COCO val2014 images, layers 5–18). The constant map adds the mean QASALFORMER saliency to every image token, matching the attention mass without spatial structure. C_I against (a) F1, (b) recall and (c) mean caption length. Each marker is one configuration; lines connect configurations of increasing β, starting from greedy decoding ($\star$). The dotted line is greedy decoding truncated to the same length. Lower C_I and higher F1/recall are better.

we have done (in our case we use layers 5–18, to be able to compare with the method of Jiang et al. (2025)), or trying to intervene explicitly in those that already correlate more, which we leave for future work.

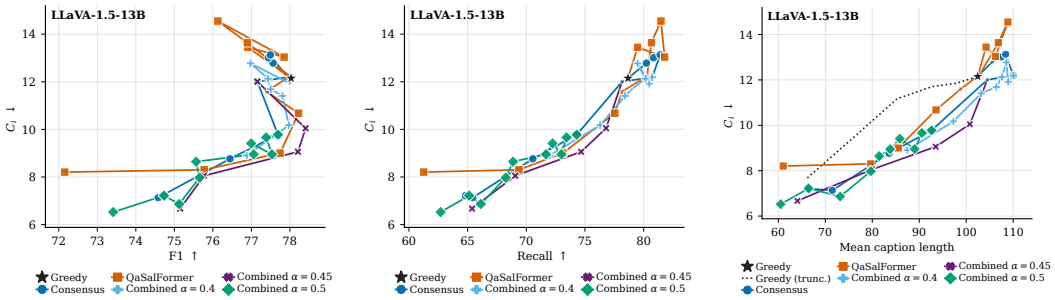
In our analyses, we average over heads to compute attention, and the proposed intervention is also the same for every head. However, following the same reasoning as for layers, different heads have been shown to focus on different things, and different intervention methods justify acting on some heads or others according to different reasoning. Here again we see different opportunities for human attention: either helping to detect heads with anomalous attention, or intervening in different heads. A per-head analysis with our dataset, in the same way that other works analyse what happens per head, is also left for future work.

Also, in this first study we limited ourselves to comparing the attention of the first generated token, but we believe an interesting follow-up analysis could include looking at it along the temporal dimension of generation, precisely because that dimension appears to affect hallucinations (Zhang et al., 2026a; Yang et al., 2025). It also enables the attention comparison to be revisited with other interpretability methods, such as gradient-based attribution or GLIMPSE (Shen, 2025).



(a) C_S vs. F1 (b) C_S vs. recall (c) C_S vs. caption length

Figure 12: C_S under consensus, QaSalFormer saliency and combined guidance on LLaVA-1.5-13B (CHAIR, 500 COCO val2014 images, layers 5–18): C_S against (a) F1, (b) recall and (c) mean caption length. Each marker is one configuration; lines connect configurations of increasing strength, starting from greedy decoding ($\star$). The dotted line is greedy decoding truncated to the same length. Lower C_S and higher F1/recall are better.



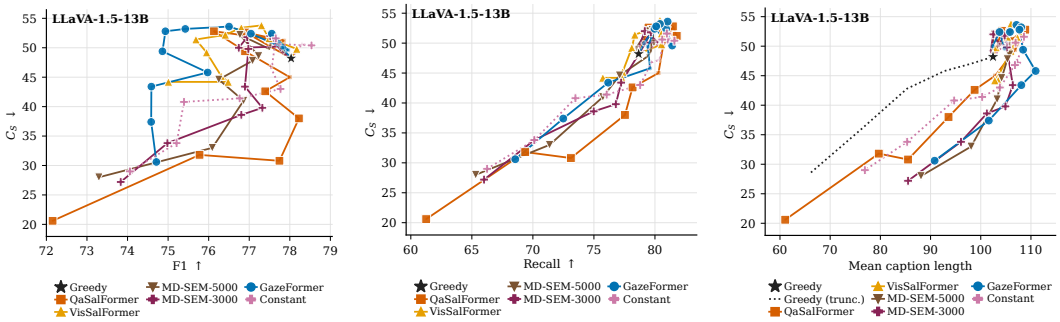
(a) C_I vs. F1 (b) C_I vs. recall (c) C_I vs. caption length

Figure 13: C_I under consensus, QaSalFormer saliency and combined guidance on LLaVA-1.5-13B (CHAIR, 500 COCO val2014 images, layers 5–18): C_I against (a) F1, (b) recall and (c) mean caption length. Each marker is one configuration; lines connect configurations of increasing strength, starting from greedy decoding ($\star$). The dotted line is greedy decoding truncated to the same length. Lower C_I and higher F1/recall are better.

As introduced in Section 7, another natural direction is to improve the saliency predictors themselves, training them on organic gaze rather than mouse traces. Another option would be to fine-tune MD-SEM, trying to add the question to it. The idea of conditioning on the question, even though in this case it is a descriptive question, is that attention will depend strongly on what one wants to describe or search for in an image. Another example is UniAR (Li et al., 2024b), which for now remains closed, which shows that by training on different datasets it achieves better accuracy on different tasks. We therefore also propose to study how fine-tuning on AiR (Chen et al., 2020) and VQA-MHUG (Sood et al., 2021) could affect performance. Although we give importance to conditioning on the question, CHAIR evaluates generic captioning and QaSalFormer is conditioned on a fixed question, so our results do not yet test the benefit of question conditioning to its full potential, which question-driven benchmarks would make possible.

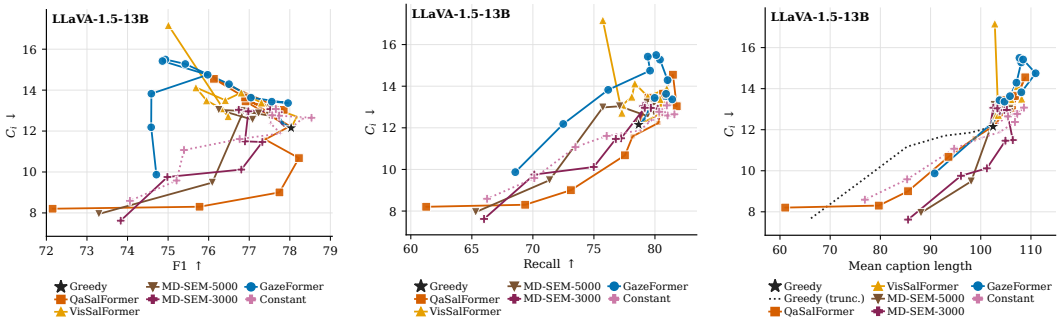
Another direction is to test the generality of this signal by substituting it into other attention-intervention methods with distinct mechanisms. For example, the post-softmax redistribution of Kang et al. (2025), to verify that the benefit is not specific to a single point of intervention. It is also compatible with methods that reduce hallucination by changing the decoding strategy, such as OPERA (Huang et al., 2024), as other works such as Zhang et al. (2026a) already combine.

Regarding the analysis of different methods, we are aware that there are many methods with many values of β trying to sweep the regions of interest. For this reason, we rely more on the curves, against both recall and F1, to better understand the behaviour, and we build tables filtering by differ-



(a) C_S vs. F1 (b) C_S vs. recall (c) C_S vs. caption length

Figure 14: C_S for the saliency sources on LLaVA-1.5-13B ($\alpha = 0$, sweep β; CHAIR, 500 COCO val2014 images, layers 5–18). The constant map adds the mean QASALFORMER saliency to every image token, matching the attention mass without spatial structure. C_S against (a) F1, (b) recall and (c) mean caption length. Each marker is one configuration; lines connect configurations of increasing β, starting from greedy decoding ($\star$). The dotted line is greedy decoding truncated to the same length. Lower C_S and higher F1/recall are better.



(a) C_I vs. F1 (b) C_I vs. recall (c) C_I vs. caption length

Figure 15: C_I for the saliency sources on LLaVA-1.5-13B ($\alpha = 0$, sweep β; CHAIR, 500 COCO val2014 images, layers 5–18). The constant map adds the mean QASALFORMER saliency to every image token, matching the attention mass without spatial structure. C_I against (a) F1, (b) recall and (c) mean caption length. Each marker is one configuration; lines connect configurations of increasing β, starting from greedy decoding ($\star$). The dotted line is greedy decoding truncated to the same length. Lower C_I and higher F1/recall are better.

ent F1 values. Analysing additional benchmarks would help to clarify the picture, and we propose it as future work.

Regarding the release of the full VLFEEDBACK-ETC dataset, we also propose further analyses in Subsection B.1.

In summary, given the growing interest in the study of attention patterns to explain hallucinations, we believe that our study contributes and opens interesting paths on how to use human attention, either to intervene or to diagnose.

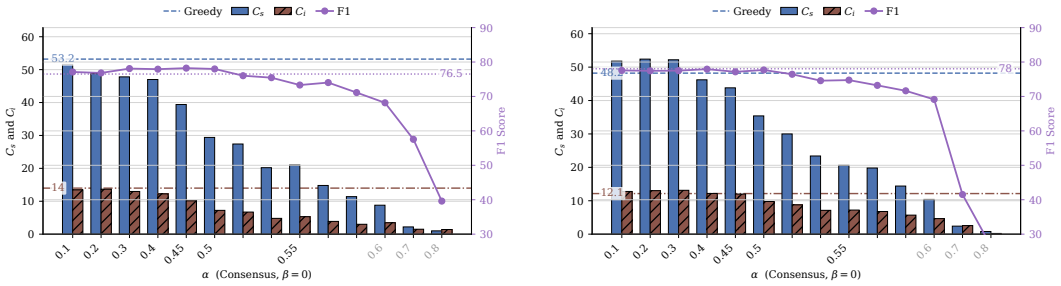


Figure 16: Guidance ablations on LLaVA-1.5 7B on the left and LLaVA-1.5 13B on the right (layers 5–18, CHAIR on 500 COCO images). Consensus (Devils): β=0, sweep α.

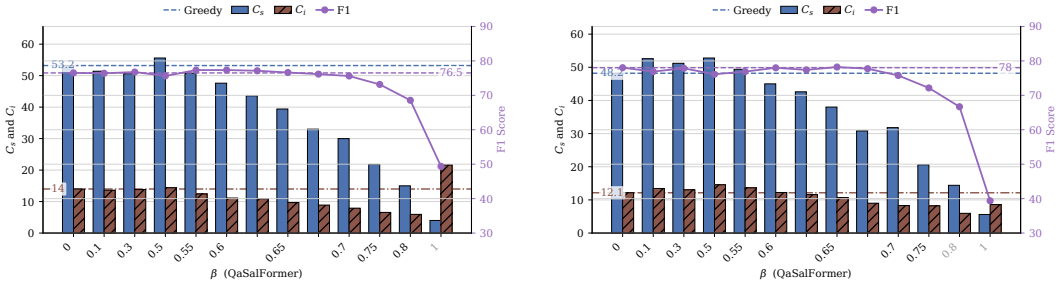


Figure 17: Guidance ablations on LLaVA-1.5 7B on the left and LLaVA-1.5 13B on the right (layers 5–18, CHAIR on 500 COCO images). Saliency-only with QaSalFormer: α=0, sweep β.

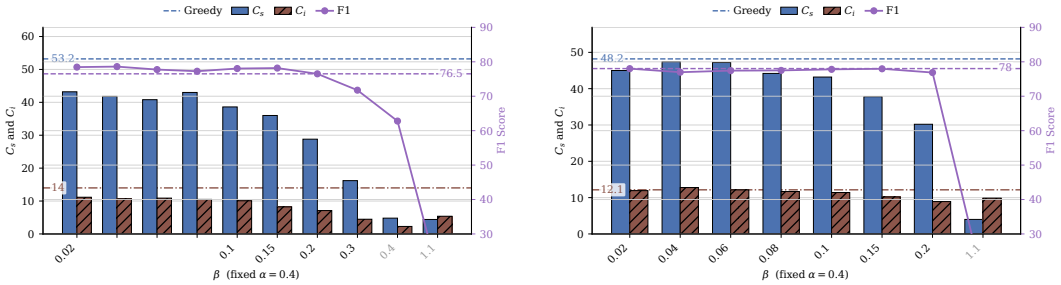


Figure 18: Guidance ablations on LLaVA-1.5 7B on the left and LLaVA-1.5 13B on the right (layers 5–18, CHAIR on 500 COCO images). Combined (Consensus + saliency): fix $\alpha = 0.4$, sweep β.

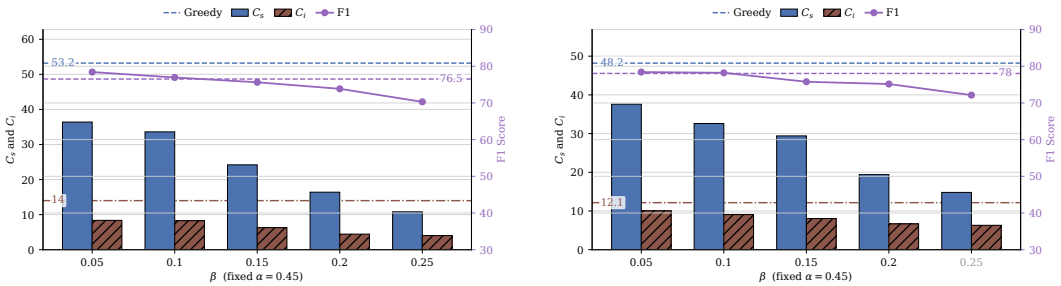


Figure 19: Guidance ablations on LLaVA-1.5 7B on the left and LLaVA-1.5 13B on the right (layers 5–18, CHAIR on 500 COCO images). Combined (Consensus + saliency): fix $\alpha = 0.45$, sweep β.

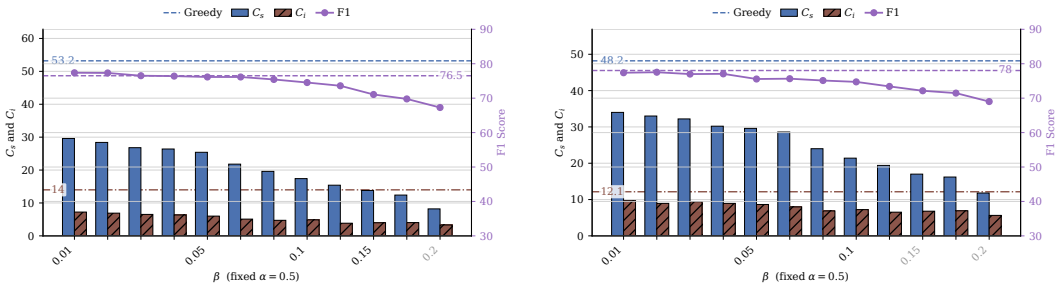


Figure 20: Guidance ablations on LLaVA-1.5 7B on the left and LLaVA-1.5 13B on the right (layers 5–18, CHAIR on 500 COCO images). Combined (consensus + saliency): fix $\alpha = 0.5$, sweep β.

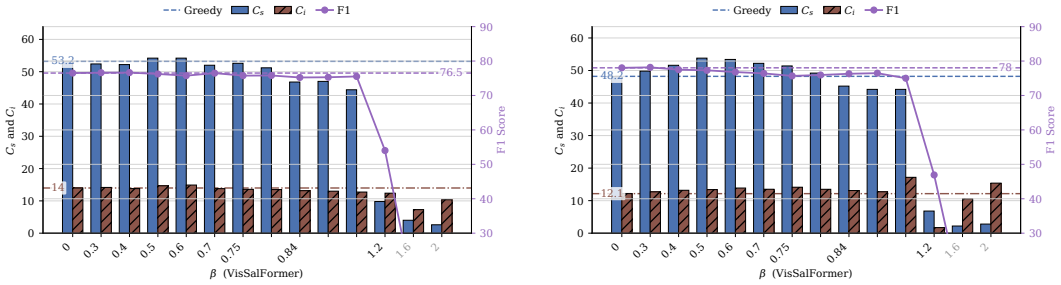


Figure 21: Guidance ablations on LLaVA-1.5 7B on the left and LLaVA-1.5 13B on the right (layers 5–18, CHAIR on 500 COCO images). Saliency-only with VisSalFormer: α=0, sweep β.

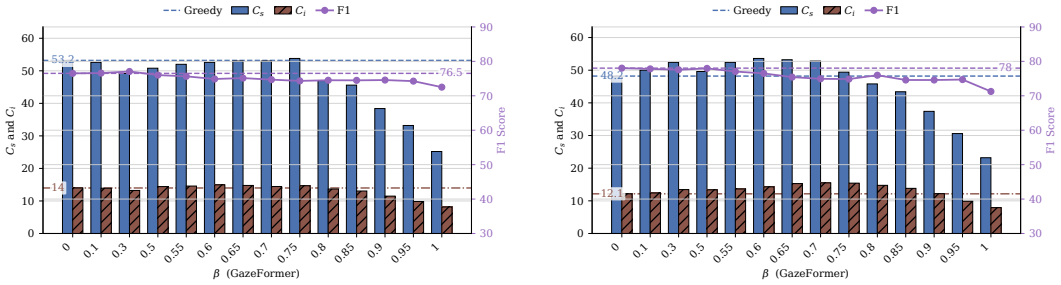


Figure 22: Guidance ablations on LLaVA-1.5 7B on the left and LLaVA-1.5 13B on the right (layers 5–18, CHAIR on 500 COCO images). Saliency-only with Gazeformer: α=0, sweep β.

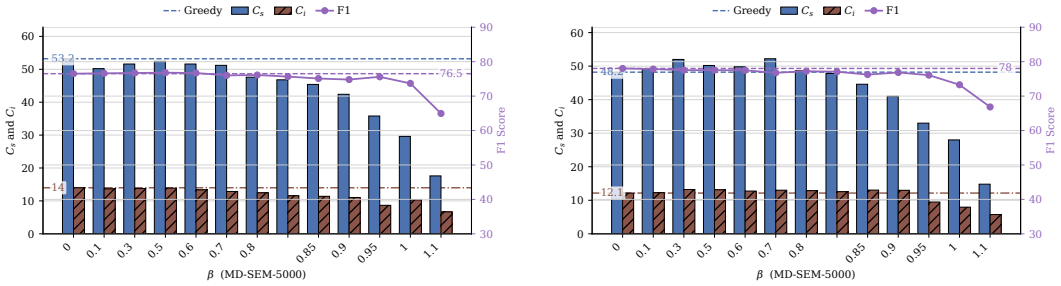


Figure 23: Guidance ablations on LLaVA-1.5 7B on the left and LLaVA-1.5 13B on the right (layers 5–18, CHAIR on 500 COCO images). Saliency-only with MD-SEM (5000 ms): α=0, sweep β.

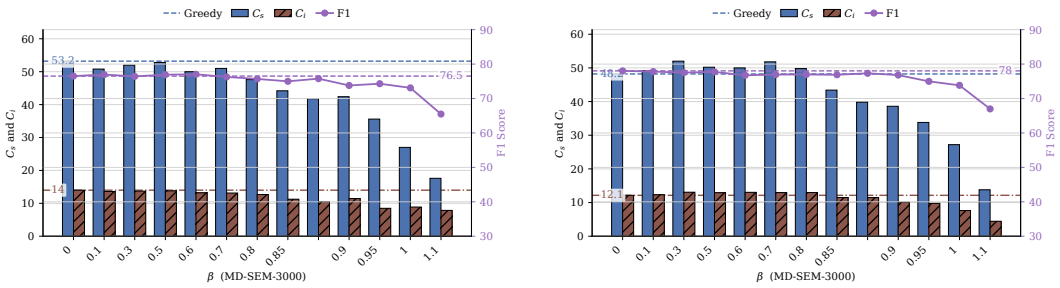


Figure 24: Guidance ablations on LLaVA-1.5 7B on the left and LLaVA-1.5 13B on the right (layers 5–18, CHAIR on 500 COCO images). Saliency-only with MD-SEM (3000 ms): α=0, sweep β.

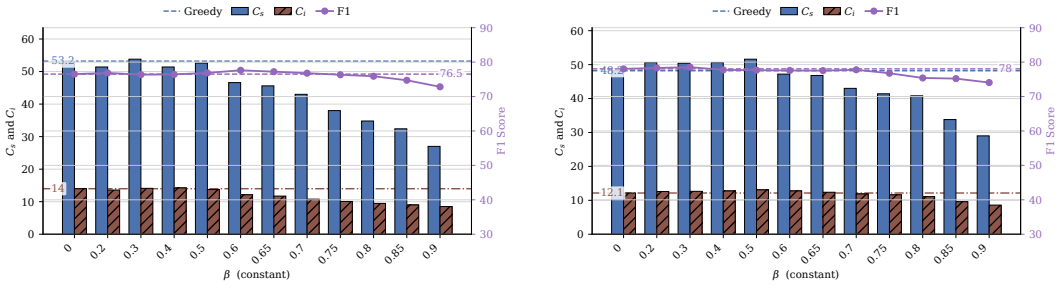


Figure 25: Guidance ablations on LLaVA-1.5 7B on the left and LLaVA-1.5 13B on the right (layers 5–18, CHAIR on 500 COCO images). Constant saliency based on QaSalFormer: α=0, sweep β.