FOCUS WHERE IT MATTERS: TRAINING-FREE SEMANTIC RE-FOCUS FOR VISION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) often fall back on linguistic priors even when the visual input contains the evidence needed to answer a query. We systematically investigate this failure through controlled input perturbations and decoder attention analysis. While visual context is essential for accurate generation, visual tokens receive disproportionately low attention. Moreover, this attention is concentrated in a small subset of heads and layers and is often poorly aligned with query-relevant image regions. Motivated by these insights, we introduce Semantic Re-Focus (SRF), an inference-time method that intervenes at two complementary stages. A semantic relevance map is estimated and is used to apply foveation before visual encoding and a selective, query-conditioned enhancement of visual attention is performed during decoding. Across five benchmark suites and two VLM architectures, SRF improves performance in fine-grained perception, counterfactual settings, object hallucination, and open-ended generation. On Qwen2.5-VL-3B and LLaVA-1.5-7B, it improves MMVP pair accuracy from 40.0% to 45.3%, and lowers the MMHal-Bench hallucination rate from 67.0% to 57.2% respectively. Our results suggest that query-conditioned control of visual evidence is a promising direction for building more reliably grounded multi-modal models.

1 INTRODUCTION

Vision language models (VLMs) have become increasingly capable at combining visual perception with natural language reasoning (Liu et al., 2023; Chen et al., 2024), yet they remain unreliable when the visual input conflicts with strong linguistic priors. In such scenarios, a model may correctly recognize the relevant entities in an image yet generate an answer that reflects statistically frequent language patterns rather than the actual visual context (Vo et al., 2026; Zhao et al., 2025). Such failure modes arise across fine-grained visual discrimination (Tong et al., 2024), counterfactual scenes (Vo et al., 2026; Zhang et al., 2025), object-presence verification (Li et al., 2023), and open-ended generation (Rohrbach et al., 2018; Sun et al., 2024). This suggests that the key challenge is not simply whether visual information is present, but whether query-relevant evidence is sufficiently prioritized when the model computes its output distribution (Chen et al., 2025; Bi et al., 2025). Understanding and resolving this attentional bottleneck is therefore critical to build reliable multi-modal systems.

Recent training-free methods attempt to improve visual grounding through contrastive decoding (Leng et al., 2024; Wang et al., 2024b) or visual-attention amplification (Yin et al., 2025; Liu et al., 2024b; Tu et al., 2026). While promising, many of these approaches intervene at a single stage of the multi-modal pipeline or amplify visual information indiscriminately, without accounting for which specific regions are semantically relevant to the user's prompt. Furthermore, existing evaluations have largely focused on object-hallucination benchmarks and binary question formats (Li et al., 2023; Guan et al., 2024) and it remains unclear whether improvements on these settings transfer to broader failure modes involving fine-grained perception, counterfactual reasoning, and open-ended generation.

To address this gap, we begin with a diagnostic investigation of VLM failure modes using controlled vision-language perturbations and decoder attention analysis. Removing or spatially disrupting the image substantially degrades performance, while moderate degradation is better tolerated. Further, within the decoder, visual attention is highly nonuniform across heads and layers, and qualitative analysis shows that decoder attention is more diffuse and does not necessarily identify the regions

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.

most relevant to the current query. Together, these observations suggest that effective visual grounding requires both preserving query relevant evidence before encoding and selectively emphasizing it during decoding.

Motivated by these findings, we introduce Semantic Re-Focus (SRF), a training-free inference-time approach that operates across two complementary stages of the VLM pipeline. Rather than uniformly sharpening images or indiscriminately boosting visual token weights, SRF estimates task-relevant image regions by computing semantic alignment between the input instruction and visual patch embeddings. Before visual encoding, SRF applies semantic foveation, a process inspired by human foveated vision (Itti & Koch, 2001) that preserves query-relevant regions at high visual fidelity while reducing detail in non-essential areas. During decoding, SRF performs selective attention re-focusing, dynamically re-weighting attention toward query-relevant visual tokens within calibrated vision-responsive heads and layers. This two stage intervention preserves query relevant visual detail before fusion and increases the contribution of corresponding visual tokens during decoding.

We evaluate SRF across two VLM architectures and five complementary benchmarks covering multiple generalization challenges. On Qwen2.5-VL-3B, SRF improves MMVP (Tong et al., 2024) pair accuracy from 40.0% to 45.3%, and on LLaVA-1.5-7B it raises the four subtask MME (Fu et al., 2026) total from 631.66 to 693.33, and reduces MMHal Bench hallucination from 67.0% to 57.2%. Our broader evaluation also reveals substantial benchmark specific variation. For example, contrastive methods that improve hallucination metrics can degrade sharply on the counterfactual VLMBias benchmark. We also identify apparent annotation inconsistencies in the widely used POPE benchmark (Li et al., 2023), highlighting the need to interpret benchmark level gains with care and to evaluate visual grounding across complementary settings. Overall, SRF provides a simple way to intervene on visual evidence at inference time without retraining the VLM. Our contributions are summarized as follows:

- We present a systematic analysis of VLM attention dynamics and input perturbations, demonstrating that visual information processing is constrained by a severe internal attention imbalance and misaligned with query-relevant regions.
- We introduce Semantic Re-Focus (SRF), an inference-time method that combines query-conditioned semantic foveation with selective visual attention enhancement during decoding.
- We evaluate SRF across five benchmark suites covering fine-grained perception, counterfactual bias, object hallucination, and open-ended generation and also report on benchmark annotation discrepancies.

2 Related works

Recent interpretability and diagnostic research has sought to isolate the internal root causes of visual grounding failures in VLMs. A primary driver is language prior dominance, where auto-regressive decoders default to strong linguistic correlations acquired during pre-training rather than conditioning on visual evidence (Lee et al., 2025). This behavior might be driven by a severe attentional imbalance as visual tokens receive a disproportionately small fraction of total decoder attention despite constituting the majority of the input sequence (Chen et al., 2025) and cross-modal integration being highly localized (Yin et al., 2025; Xie et al., 2026). This pattern is consistent with a localized use of visual information within the decoder, where only a subset of heads and layers assigns substantial attention to visual tokens.

To quantify these failures across different visual contexts, a broad spectrum of benchmarks has been developed. POPE (Li et al., 2023) evaluates object hallucination through binary queries; MMVP (Tong et al., 2024) targets fine-grained visual distinctions where linguistic intuition fails; VLMBias (Vo et al., 2026) measures resilience when visual evidence contradicts common priors; MME (Fu et al., 2026) evaluates general perception and reasoning; and MMHal-Bench (Sun et al., 2024) tests hallucination in open-ended generation. However, because these benchmarks vary in prompt structures, answer spaces, and targeted failure types, localized performance gains on a single dataset do not necessarily reflect generalized grounding improvements. Furthermore, annotation inconsistencies, such as ground-truth errors observed in established benchmarks like POPE (Neuhaus & Hein, 2025), highlight the risks of single-benchmark evaluation and motivate assessments across complementary perceptual and reasoning tasks.

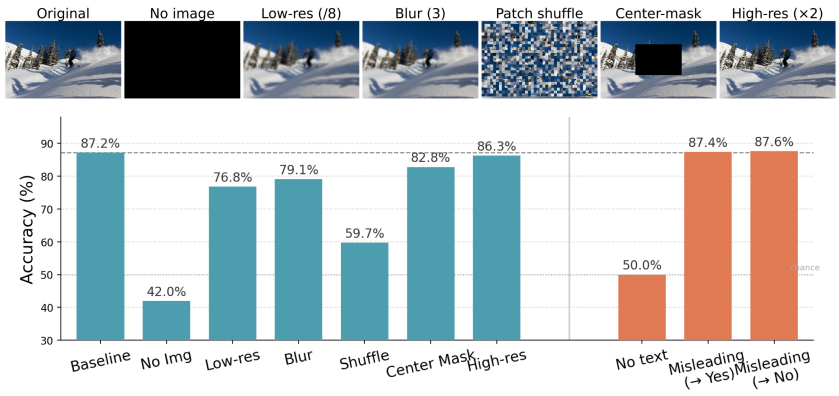

Figure 1: Controlled visual and language perturbations using Qwen2.5-VL-3B on POPE. The top row illustrates the visual perturbations on one example, while the bars report accuracy.

To mitigate these internal failure modes without parameter updates, training-free methods attempt to strengthen the influence of visual evidence during inference. Contrastive decoding methods (Leng et al., 2024; Kim et al., 2024) penalize language priors by contrasting predictions from original and modified contexts. Concurrently, attention-focused interventions seek to re-balance decoder representations. These methods attempt to amplify visual signals across vision-responsive heads (He et al., 2025; Yin et al., 2025), recalibrate attention distributions across decoding steps (Jung et al., 2025; Wan et al., 2025), resolve inter-layer attention discrepancies (Xie et al., 2026), or address structural bottlenecks such as attention sinks (Kang et al., 2025) and localized feature integration (An et al., 2025). While effective, these techniques typically apply unconditioned, global boosts or intervene at a single stage of the pipeline.

3 Empirical Analysis

We first analyze how VLM predictions depend on visual and textual inputs through controlled perturbations. We then examine decoder attention to understand how visual information is distributed across tokens, layers, and attention heads.

3.1 Perturbing the visual and language channels

We begin by examining how VLM predictions depend on visual and textual inputs using controlled perturbations on the POPE dataset and report results in Figure 1. Removing the image entirely drops accuracy from 87.2% to 42.0%, which confirms that the model does draw on visual input and is not answering purely from text priors. Moderate visual degradation through low resolution, Gaussian blur, or center masking causes substantially smaller changes, while spatial patch shuffling reduces accuracy to 59.7%. A marginal positive impact is seen with high-resolution up-sampling.On the language side, removing the question reduces performance to chance level, while adding misleading yes or no statements leaves accuracy close to the baseline. Additional prompt based analyses on VLMBias show that explicit instructions to inspect the image provide little improvement, with detailed results in Appendix Figures 8 and 9. Taken together, the visual evidence is present and used, but its influence is susceptible to being overridden by the language prior, which motivates examining how visual evidence is used internally.

3.2 Attention allocation

We examine how much decoder attention is allocated to visual information. As shown in Figure 2, visual tokens constitute 93% of the input tokens in our Qwen2.5-VL-3B setting, while receiving approximately 8% of the measured decoder attention. The large disparity reveals a strong asymmetry between visual and textual context during decoding. A substantial fraction of nonvisual attention is assigned to system and template tokens. The layer and head analysis further shows that visual

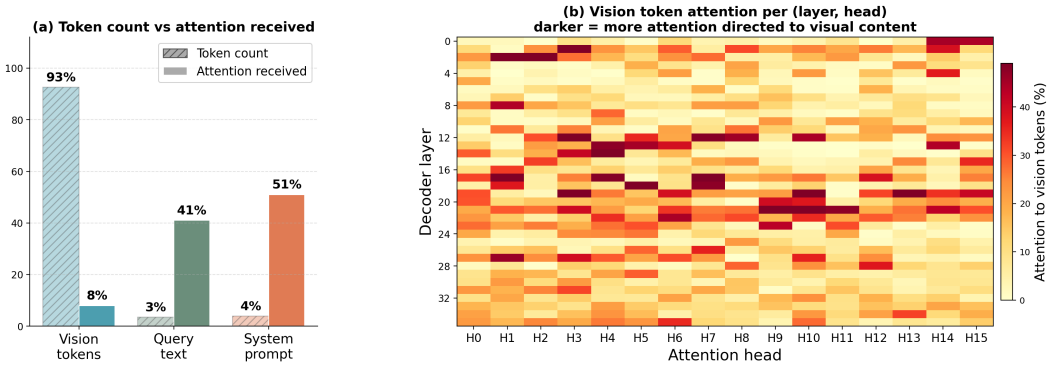

Figure 2: Visual attention imbalance and head specialization in Qwen2.5-VL-3B-Instruct. Left: visual tokens constitute 93% of the input sequence but receive only 8% of the attention. Right: Visual attention is also strongly nonuniform across decoder layers and heads.

attention is highly nonuniform, with a limited subset of heads and layers assigning substantially greater attention to image tokens than the rest. A similar pattern is observed on LLaVA-1.5-7B in Appendix Figure 10. Uniformly increasing visual attention would therefore modify many heads that currently assign little attention to image tokens. These observations motivate a selective decoder intervention that acts only on heads and layers with stronger visual attention.

3.3 Query-grounded attention

The analysis above shows that visual attention is concentrated in a limited subset of heads and layers, but it does not indicate whether that attention falls on image regions relevant to the current query. Figure 3 shows two representative examples where decoder attention remains spatially diffuse and assigns substantial weight outside the region associated with the answer. These examples are qualitative, but they illustrate a limitation of using decoder attention itself to decide which visual tokens should be amplified. Strengthening already attended regions could reinforce visual content that is not relevant to the current query. Hence, we explore an external alignment model to compare the query concept with localized image regions and produce a spatial semantic relevance map. As illustrated in Figure 3 (and Figures 6, 12), the resulting semantic maps place greater emphasis on regions associated with the queried evidence. This motivates an external query-conditioned intervention to direct attention toward the relevant image regions.

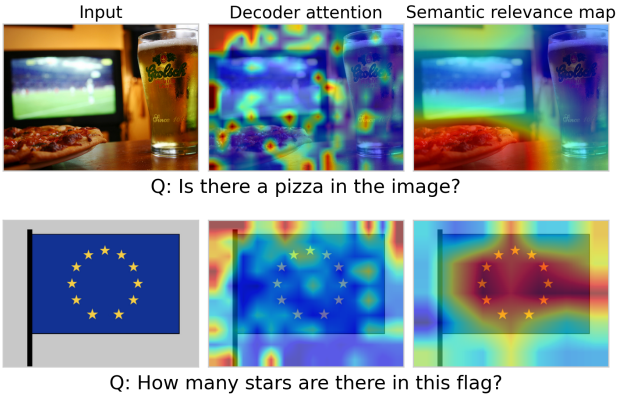

Figure 3: Decoder attention from question to image tokens vs an external query-conditioned semantic relevance map.

4 Method

Inspired by the findings of the empirical analysis, we propose Semantic Re-Focus (SRF), a training-free inference-time method that improves visual reliance by intervening at two points in a frozen vision-language model. SRF first estimates a query-conditioned semantic relevance map from a frozen alignment model (Section 4.1). This signal guides semantic foveation before visual encoding

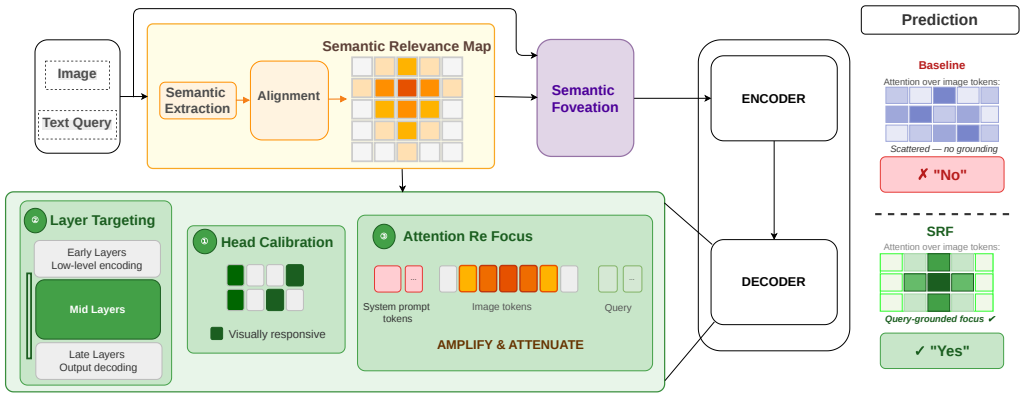

Figure 4: Overview of Semantic Re-Focus (SRF). Given an image and query, a semantic relevance map is estimated and semantic foveation is applied by attenuating less relevant visual detail before encoding. During decoding, vision-responsive heads and layers are identified and attention toward the corresponding visual tokens are selectively modified. Both interventions are applied at inference time while keeping the VLM parameters fixed.

(Section 4.2) and selectively modifies visual attention during decoding within calibrated vision responsive heads and layers (Section 4.3). The schematic is shown in Figure 4 and the visualization of each component is shown in Figure 5.

Let an input consist of an image I and an instruction x. A VLM first maps the image into visual tokens through a vision encoder E_v, then concatenates these visual tokens with text tokens before decoding:

$$\mathbf{v} = E_v(I), \qquad \mathbf{z}_0 = [\mathbf{s}, \mathbf{v}, \mathbf{x}], \tag{1}$$

where $\mathbf{s}$ denotes system or template tokens, $\mathbf{v} = \{v_j\}_{j=1}^{N}$ denotes image tokens, and $\mathbf{x}$ denotes instruction tokens. SRF keeps all model parameters fixed and changes only the input image and selected attention logits during inference.

4.1 Semantic Relevance Map

SRF begins by estimating which image regions are semantically relevant to the current query. Given an instruction x, we extract a compact textual grounding target $t = g(x)$ from the salient nouns or noun phrases in the query. The target is used only to guide visual localization and does not modify the original instruction given to the VLM. We estimate relevance using a frozen image text alignment model. To account for objects appearing at different spatial scales, we cover the image using crops from three spatial kernels. Let $\{I_p^{(k)}\}_{p=1}^{P_k}$ denote the crops at scale k. Their semantic relevance is computed as

$$r_p^{(k)} = \left\langle \phi_v(I_p^{(k)}),\, \phi_t(t) \right\rangle, \tag{2}$$

where ϕ_v and ϕ_t denote normalized image and text embeddings.

The scores from all scales are normalized to obtain a soft spatial mask:

$$m_p = \frac{r_p - \min_{p'} r_{p'}}{\max_{p'} r_{p'} - \min_{p'} r_{p'} + \epsilon}, \qquad m_p \in [0, 1]. \tag{3}$$

We then upsample this patch-level mask to image resolution, giving $M \in [0,1]^{H \times W}$. High values in M correspond to regions semantically aligned with the instruction.

To avoid imposing a false target when the extracted concept is absent or uncertain, SRF computes a confidence score using the strongest alignment observed either at the full-image level or among the local patches,

$$c = \min\left(1, \frac{\max\left(\langle \phi_v(I), \phi_t(t) \rangle, \max_p r_p\right)}{\tau}\right), \tag{4}$$

where τ is a presence threshold. The final mask is softened by this confidence:

$$\bar{M} = cM + (1 - c)\mathbf{1}. \tag{5}$$

When the alignment confidence is high, $\bar{M}$ preserves the spatial relevance estimate. When confidence is low, it approaches a uniform map, limiting the effect of an unreliable grounding target.

4.2 Encoder Intervention - Semantic Foveation

Human vision is spatially non-uniform and visual acuity is highest around the fovea and decreases toward the periphery, allowing fine local detail to be combined with coarser contextual information (Rosenholtz, 2016; Itti & Koch, 2001). This principle has also motivated foveated representations in deep vision models (Pramod et al., 2022; Gizdov et al., 2025). In SRF, the region of high fidelity is not determined by gaze or image center, but by its semantic relevance to the current query. Attenuating detail outside that region shifts the budget toward the evidence the query depends on, and unlike a crop it preserves global scene context, which the model still needs to interpret what it sees.

Let $B_\sigma(I)$ denote a Gaussian-blurred version of the image with blur scale σ. The foveated image is computed as

$$I_{\mathrm{SRF}} = \bar{M} \odot I + (1 - \bar{M}) \odot B_\sigma(I), \tag{6}$$

where $\odot$ denotes elementwise multiplication after broadcasting the mask over color channels. This operation keeps semantically relevant regions sharp while attenuating background detail. The visual tokens used by the language model are then

$$\mathbf{v}^{\mathrm{SRF}} = E_v(I_{\mathrm{SRF}}). \tag{7}$$

This stage does not remove the background entirely. It reduces high-frequency detail outside the relevant region so that the visual encoder preserves coarse context while reducing fine visual detail outside query relevant regions. Visualizations can be seen in Figures 6 and12.

4.3 Decoder Intervention

4.3.1 Vision-responsive Layers and Heads

Cross-modal fusion is not uniformly distributed across the decoder. Based on the empirical analysis, our intervention is restricted to the middle layers where visual attention is more pronounced, $\mathcal{L}^* = \{\ell_s, \ell_s + 1, \ldots, \ell_e\}$.

Within those layers the intervention is further restricted to heads that assign greater attention to visual tokens. We identify them using a small unlabeled calibration set. For each calibration example we run the frozen model without intervention and record attention from instruction or generation positions to image-token positions. Let $A^{\ell,h}$ denote the attention matrix in layer ℓ and head h, let $\mathcal{V}$ be the image-token positions and let $\mathcal{Q}$ be the instruction or generation-query positions. The visual routing score is

$$\rho_{\ell,h} = \frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \sum_{j \in \mathcal{V}} A^{\ell,h}_{ij}, \tag{8}$$

averaged over calibration examples. The selection is made independently within each layer.

$$\mathcal{H}^*_\ell = \mathrm{TopK}_h \left(\mathbb{E}_c[\rho^{(c)}_{\ell,h}] \right), \tag{9}$$

The selected heads are fixed during evaluation. This calibration is label-free and requires only standard forward passes.

4.3.2 Semantic re-focus

The decoder computes pre-softmax attention logits

$$Z^{\ell,h} = \frac{Q^{\ell,h}(K^{\ell,h})^\top}{\sqrt{d_k}}. \tag{10}$$

SRF modifies these logits only for $\ell \in \mathcal{L}^*$ and $h \in \mathcal{H}^*_\ell$, and distinguishes three groups of keys. Let $\mathcal{V}$ index the image-token positions and $\mathcal{S}$ the system and template positions that precede the image in

the prompt. Within $\mathcal{V}$, *semantic* keys are the image tokens the relevance map marks as related to the query and *background* keys are the remainder, while the keys in $\mathcal{S}$ carry fixed text that competes for the same attention budget. The instruction tokens, which follow the image, are never modified.

Let $s_j \in [0, 1]$ denote the semantic relevance of image token j, read from $\bar{M}$ on the visual-token grid. For every query position i,

$$\widetilde{Z}_{ij}^{\ell,h} = \begin{cases} Z_{ij}^{\ell,h} + \lambda_{\text{sem}}\, s_j - \lambda_{\text{bg}}(1 - s_j), & j \in \mathcal{V}, \\ Z_{ij}^{\ell,h} - \lambda_{\text{sys}}, & j \in \mathcal{S}, \\ Z_{ij}^{\ell,h}, & \text{otherwise}, \end{cases} \tag{11}$$

SRF amplifies image-token logits in proportion to their relevance and the two visual terms form a single ramp, from $-\lambda_{\text{bg}}$ at the least relevant token to $+\lambda_{\text{sem}}$ at the most relevant, so the intervention redistributes attention within the image rather than uniformly adding to it. The third term further reduces competition from the system-prompts. The modified attention distribution is then computed normally,

$$\widetilde{A}^{\ell,h} = \text{softmax}\left(\widetilde{Z}^{\ell,h}\right). \tag{12}$$

This preserves the architecture and leaves all learned weights unchanged. Because the softmax normalises over the row, raising the logits of relevant visual tokens also lowers the relative weight of the query-text positions without modifying them. The intervention is selective along three axes, image token position, attention head and decoder layer, so SRF avoids indiscriminately increasing visual attention, which can amplify background noise or disrupt language processing.

We additionally study whether SRF can be combined with contrastive decoding and introduce SRF-C, an optional extension that adds a no-image contrastive pass in Appendix A

Analytical properties of SRF. SRF admits two simple analytical interpretations. Semantic foveation preserves coarse image structure while selectively retaining finer detail according to semantic relevance. In the decoder, relevance based logit amplification monotonically increases the relative attention assigned to more relevant visual tokens within the selected heads and layers. We provide formal derivations in Appendix F.

5 Experimental Setup

We evaluate SRF on two architectures, Qwen2.5-VL-3B-Instruct (Wang et al., 2024a) and LLaVA-1.5-7B (Liu et al., 2024a), with greedy decoding throughout. The relevance map is computed by a frozen CLIP ViT-B/32 (Radford et al., 2021) that is never updated and is shared across both models. We evaluate on five benchmarks spanning different grounding challenges. **MMVP** (Tong et al., 2024) with fine-grained discrimination, **VLMBias** (Vo et al., 2026) with counterfactual visual reasoning. **MME** (Fu et al., 2026) with perception and cognition questions and **MMHal-Bench** (Sun et al., 2024) with hallucination in open-ended generation and **POPE** (Li et al., 2023) with yes/no questions for hallucination evaluation.

On Qwen2.5-VL we compare against VAF (Yin et al., 2025), VCD (Leng et al., 2024), VHR (He et al., 2025) and ILVAD (Xie et al., 2026). On LLaVA-1.5-7B we follow the protocol of Xie et al. (2026) adding CODE (Kim et al., 2024), AGLA (An et al., 2025), VAR (Kang et al., 2025), SPARC (Jung et al., 2025) and ONLY (Wan et al., 2025). More details on hyper-parameters and methods are provided in the Appendix (Section C).

6 Results

Fine-grained perception. MMVP pairs two images that differ in a single visual detail and asks the same question of both. Pair accuracy credits a pair only when both images are answered correctly, so a model that answers only from language priors scores near zero. SRF reaches the highest pair accuracy (45.3) of all compared methods (Table 1), including those that require one or two additional decoding passes, and improves per-image accuracy alongside it. Semantic map and foveation can be attributed to the gain because the discriminating evidence occupies a small region, and the encoder

Table 1: Comparison of training-free visual grounding methods on Qwen2.5-VL-3B-Instruct. "+pass" denotes additional VLM decoding passes beyond standard generation.

		MMVP		VLMBias			
Method	+pass	Pair	Img	Optical	Logos	GameBoard	Avg
Baseline	—	40.0	67.7	49.7	5.6	4.2	18.1
VAF (Yin et al., 2025)	0	40.0	69.0	49.7	6.8	11.7	17.9
VCD (Leng et al., 2024)	+1	37.3	65.7	24.1	0.5	8.3	9.6
ILVAD (Xie et al., 2026)	+1	38.7	68.0	48.5	4.1	6.5	17.7
VHR (He et al., 2025)	+2	38.7	68.3	**51.1**	2.2	11.7	17.9
SRF	0	**45.3**	**70.0**	50.9	**11.4**	**11.9**	**20.2**

Table 2: Performance on MME and MMHal-Bench for LLaVA-1.5-7B.

	MME ↑					MMHal-Bench	
Method	Exist.	Count	Pos.	Color	Total	Score ↑	Hal% ↓
Greedy / Vanilla	195.00	158.33	123.33	155.00	631.66	2.01	67.0
VCD (Leng et al., 2024)[CVPR'24]	190.00	163.33	118.33	143.33	614.99	2.20	61.8
CODE (Kim et al., 2024)[NeurIPS'24]	195.00	131.66	133.33	150.00	609.99	2.05	66.2
AGLA (An et al., 2025)[CVPR'25]	190.00	153.33	118.33	160.00	621.66	2.14	63.8
VAF (Yin et al., 2025)[CVPR'25]	185.00	151.66	133.33	155.00	624.99	1.97	68.2
VAR (Kang et al., 2025)[ICLR'25]	185.00	143.33	143.33	165.00	636.66	2.18	62.2
SPARC (Jung et al., 2025)[ICML'25]	195.00	158.33	123.33	153.33	629.99	2.09	64.6
ONLY (Wan et al., 2025)[ICCV'25]	190.00	150.00	121.66	165.00	626.66	1.88	69.8
VHR (He et al., 2025)[ACL'25]	195.00	158.33	123.33	153.33	629.99	2.10	65.6
ILVAD (Xie et al., 2026)[ICML'26]	195.00	153.33	133.33	160.00	641.66	2.22	61.5
SRF	**196.67**	**173.33**	**146.67**	**176.67**	**693.33**	**2.46**	**57.2**

otherwise spends the budget on the rest of the scene, attenuating the irrelevant surround preserves the detail on which the pair can be separated.

Counterfactual bias. VLMBias modifies familiar subjects so that the memorized answer and the visible answer disagree (Figures 9, 12). The benchmark therefore measures whether visual evidence can override a strong language prior rather than merely agree with it. Table 1 reports three representative categories and the average across all seven and SRF improve the overall average performance. Contrastive methods such as VCD degrade sharply on this benchmark (9.6%), indicating that improvements on hallucination focused settings do not necessarily generalize to counterfactual visual reasoning.

Open ended generation and hallucination. We test whether our intervention remains useful on a different output format, where errors accumulate across freely generated tokens rather than a single decision. On LLaVA-1.5-7B, SRF improves performance across both the MME evaluation and MMHal Bench (Table 2). SRF increases the MME total from 631.66 to 693.33 and the largest gains appear in the *Position* and *Color* subtasks, which require spatially localizing specific visual attributes. It outperforms the strongest compared baseline, ILVAD (which requires an extra forward pass), by 51.67 points. On MMHal-Bench, SRF achieves the best hallucination score 2.46 and the lowest hallucination rate (57.2%). These results show that our dual intervention is effective on a second VLM architecture and extends beyond short form visual discrimination to open ended generation.

Annotation noise in benchmarks. POPE (Li et al., 2023) provides a complementary evaluation focused specifically on object hallucination. SRF improves on the most challenging adversarial split (Table 5). However, during error inspection we found cases where the annotation is inconsistent with the visible content, shown in Figure 7, an issue also documented in a recent manual audit (Neuhaus & Hein, 2025). These results reinforce the need to interpret small differences on benchmarks together with annotation quality and provide more evaluations across complementary visual grounding tasks.

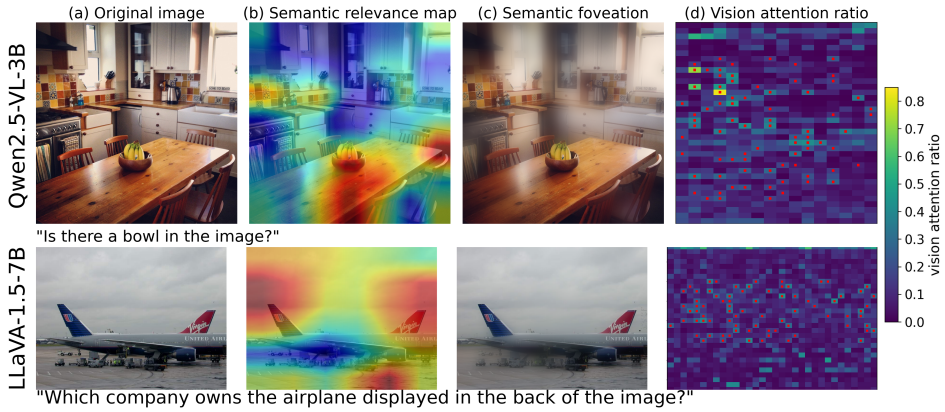

Figure 5: SRF pipeline on Qwen and LLaVA - Semantic relevance map, foveated image and the decoder vision attention ratio, with the heads SRF selects marked in red.

Ablation and Inference Cost. The ablation in Table 3 shows a progressive benefit as the SRF components are combined. Relevance based amplification and attenuation of the background together provide an initial gain, system token suppression adds further improvement, and the strongest performance is obtained with semantic foveation. Replacing the semantic relevance map with a random spatial map largely removes these gains, showing that the benefit does not come from generic amplification or random image modification. SRF requires no additional VLM decoding pass, although semantic relevance estimation introduces additional preprocessing. On Qwen2.5 VL 3B, SRF requires 613ms per sample, compared to 668 to 802 ms for the evaluated multi pass methods. Full runtime details are provided in Appendix C.

Table 3: Ablation study on MMVP with Qwen2.5-VL-3B.

	Dec Attn			Enc Fov	MMVP	
Configuration	$\lambda_{\text{sem}} \uparrow$	$\lambda_{\text{bg}} \downarrow$	$\lambda_{\text{sys}} \downarrow$	σ	**Pair**	**Img**
No intervention (baseline)	–	–	–	–	40.0	67.7
Semantic Relevance Map						
amplification only	✓	–	–	–	41.9	68.3
+ background attenuation	✓	✓	–	–	42.2	68.3
+ system suppression	✓	✓	✓	–	43.0	69.0
+ foveation **(SRF)**	✓	✓	✓	✓	**45.3**	**70.0**
Random map						
+ all	✓	✓	✓	✓	40.7	67.7

7 Conclusion

We studied how visual evidence is used during VLM inference and showed that reliable grounding depends not only on access to the image, but also on preserving and selectively emphasizing the evidence relevant to the current query. We showed that a VLM's own decoder attention does not necessarily identify the image regions most relevant to the query, while a small frozen image text model can provide an independent and reliable semantic relevance signal. Semantic Re-Focus uses this signal at two complementary stages, foveating the input before encoding and selectively reweighting attention within vision responsive heads and layers during decoding. This improves fine grained perception, counterfactual reasoning, and open ended generation and hallucination behavior across two architectures without an additional VLM decoding pass. Moreover, our ablations show that the gains arise from meaningful semantic localization together with complementary encoder and decoder interventions, rather than generic amplification.

Limitations and Future Work. Our study is limited to two VLM architectures and depends in part on the localization quality of the external alignment model. Future work can explore stronger grounding models as the relevance source, finer grained per token relevance for long form generation, and extensions to more architectures, tasks, and modalities. More generally, we see SRF as a starting point for inference time mechanisms that use external signals to selectively shape multi-modal reasoning in frozen foundation models.

AI USE STATEMENT

In this work, we used generative AI tools for improving the writing and as an aid for ideas for the visualizations. Additionally, we used generative AI tools for finding related works and drafting sections of the paper. All AI-assisted content was subject to author review and verification. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI. All research ideas, methodology design, analysis reported in this paper are done by authors.

REPRODUCIBILITY STATEMENT

We have provided the explanation of our methods plus all parameters for reproducing the results. We will share the code upon acceptance.

REFERENCES

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29915–29926. IEEE, 2025.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. Unveiling visual perception in language models: An attention head analysis approach. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4135–4144, 2025.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. In *International Conference on Machine Learning*, pp. 9910–9932. PMLR, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=DgH9YCsqWm.

Andrey Gizdov, Shimon Ullman, and Daniel Harari. Seeing more with less: Human-like representations in vision models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4408–4417. IEEE, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, June 2024.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. Cracking the code of hallucination in lvlms with vision-aware head divergence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3488–3501, 2025.

Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

Mingi Jung, Saehyung Lee, Eunji Kim, and Sungroh Yoon. Visual attention never fades: Selective progressive attention recalibration for detailed image captioning in multimodal large language models. In *International Conference on Machine Learning*, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 87676–87703, 2025.

Junho Kim, Hyun J Kim, Yeon J Kim, and Yong M Ro. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *Advances in Neural Information Processing Systems*, 37:133571–133599, 2024.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. Vlind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4129–4144, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pp. 125–140. Springer, 2024b.

Yannic Neuhaus and Matthias Hein. Repope: Impact of annotation errors on the pope benchmark. *arXiv preprint arXiv:2504.15707*, 2025.

RT Pramod, Harish Katti, and SP Arun. Human peripheral blur is optimal for object recognition. *Vision research*, 200:108083, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Conference on Empirical Methods in Natural Language Processing*, 2018.

Ruth Rosenholtz. Capabilities and limitations of peripheral vision. *Annual review of vision science*, 2 (1):437–457, 2016.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the association for computational linguistics: ACL 2024*, pp. 13088–13110, 2024.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9568–9578, 2024.

Chongjun Tu, Peng Ye, Dongzhan Zhou, Lei Bai, Gang Yu, Tao Chen, and Wanli Ouyang. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *International Journal of Computer Vision*, 134(1):22, 2026.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zifu Wan, Ce Zhang, Silong Yong, Martin Q Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3225–3234. IEEE, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15840–15853, 2024b.

Yutong Xie, Zhenglin Hua, Ran Wang, Wing WY Ng, Xizhao Wang, and Yuheng Jia. Finding the correct visual evidence without forgetting: Mitigating hallucination in lvlms via inter-layer visual attention discrepancy. In *International Conference on Machine Learning*, 2026.

Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14625–14634, 2025.

Jusheng Zhang, Kaitong Cai, Yijia Fan, Jian Wang, and Keze Wang. Cf-vlm: Counterfactual vision-language fine-tuning. *arXiv preprint arXiv:2506.17267*, 2025.

Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Baobao Chang, and Minjia Zhang. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19677–19701, 2025.

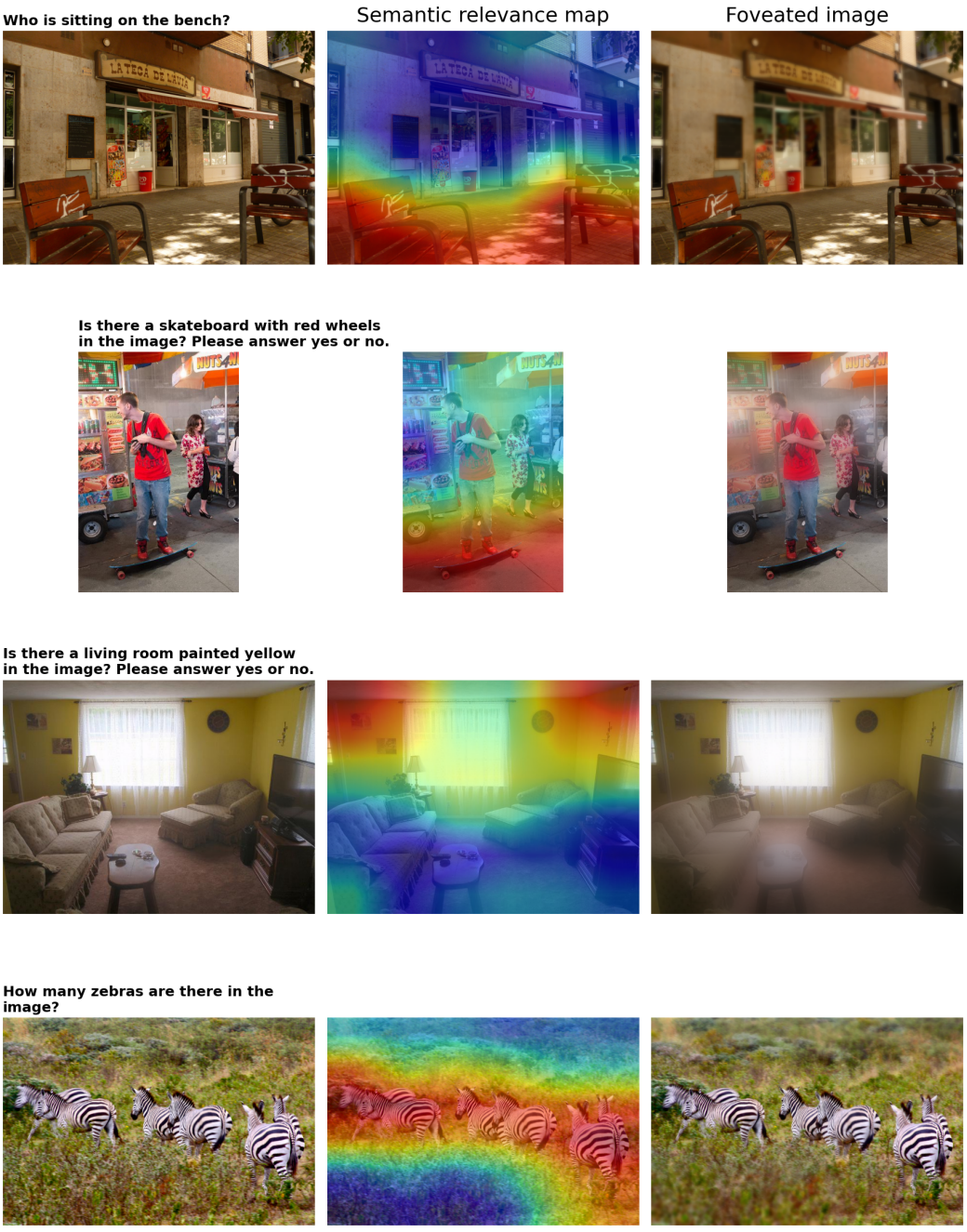

Figure 6: Semantic relevance maps and semantically foveated images on LLaVA, shown for two examples each from the MME and MMHal Bench benchmarks.

A SRF-C - A CONTRASTIVE EXTENSION

SRF improves visual grounding within a single VLM decoding pass by increasing the contribution of semantically relevant visual evidence. We additionally study whether this signal can be combined with contrastive decoding. We refer to this extension as SRF-C. The goal is to preserve the semantic visual emphasis introduced by SRF while further reducing predictions that can be explained primarily by the language context. At each decoding step, let ℓ_{SRF} denote the output logits produced by SRF

Table 4: SRF-C on Qwen2.5-VL-3B-Instruct. The +forward pass column counts additional VLM decoding passes beyond standard generation.

		MMVP		VLind	
Method	**+forward pass**	**Pair**	**Img**	**Q-Acc**	**Pair**
Baseline	-	40.0	67.7	59.6	47.0
VAF (Yin et al., 2025)	0	40.0	69.0	58.9	45.7
VCD (Leng et al., 2024)	+1	37.3	65.7	62.3	47.4
ILVAD (Xie et al., 2026)	+1	38.7	68.0	59.3	46.0
VHR (He et al., 2025)	+2	38.7	68.3	61.6	44.0
SRF	0	45.3	70.0	59.5	46.4
SRF-C	+1	**46.6**	**70.7**	**73.7**	**52.7**

and let $\ell_\varnothing$ denote the logits from a second pass in which the visual input is suppressed while the textual context is kept unchanged. SRF-C modifies the output logits according to

$$\ell_{\mathrm{SRF-C}} = \ell_{\mathrm{SRF}} + \gamma\,(\ell_{\mathrm{SRF}} - \ell_\varnothing)\,, \tag{13}$$

where $\gamma \geq 0$ controls the contrastive strength and $\gamma = 0$ recovers standard SRF. The contrastive term increases the relative score of tokens whose predictions are better supported when visual evidence is available and reduces the influence of tokens that remain strongly supported without the image. Unlike SRF, this extension requires one additional VLM decoding pass.

Results. We additionally evaluate SRF-C on VLind-Bench (Lee et al., 2025), which is designed to measure reliance on language priors using counterfactual visual contexts. VLind-Bench separates language prior effects from simpler failures in commonsense knowledge and visual perception through a sequence of controlled tests. This setting complements MMVP by testing whether the model follows visual evidence when it conflicts with a strong learned prior. Table 4 shows that the contrastive extension improves over SRF on both evaluation settings. On MMVP, pair accuracy increases from 45.3% with SRF to 46.6% with SRF-C. The difference is substantially larger on VLind-Bench, where question accuracy increases to a significant 73.7% and pair accuracy to 52.7%. These results suggest that semantic re-focusing and contrastive suppression provide complementary signals, with the contrastive term being particularly useful when correct predictions require overcoming strong language priors.

B Annotation Noise in Hallucination benchmark

POPE is widely used to evaluate object hallucination in VLMs, but its scores depend on the quality of the underlying object annotations inherited from MSCOCO. A recent manual audit reports incorrect labels in both positive and negative examples, and further shows that such annotation errors can affect measured F1 scores and model rankings (Neuhaus & Hein, 2025). This indicates that small performance differences on POPE should be interpreted with care.

Table 5 reports SRF on the original POPE evaluation. SRF improves on the most challenging adversarial split, while performance on the popular and random splits remains close to the baseline. During qualitative inspection, we observed multiple cases in which the provided label appears inconsistent with the visible image content. Figure 7 shows three representative examples where SRF follows the visible evidence while the baseline agrees with the benchmark label.

These cases highlight a broader challenge in evaluating VLM grounding. A model prediction can be visually correct yet be counted as an error when benchmark annotations are incomplete or inconsistent. Since POPE and similar benchmarks are widely used to compare hallucination mitigation methods, annotation quality can influence both absolute scores and conclusions drawn from small performance differences.

Table 5: Object hallucination performance on POPE using Qwen2.5-VL-3B-Instruct. We report accuracy and F1 (%) on adversarial, popular, and random splits, each containing 3,000 questions.

	Adversarial		Popular		Random	
Method	Acc	F1	Acc	F1	Acc	F1
Baseline	86.37	85.12	87.57	86.25	88.50	87.15
SRF	87.33	86.07	87.40	86.06	88.43	87.06

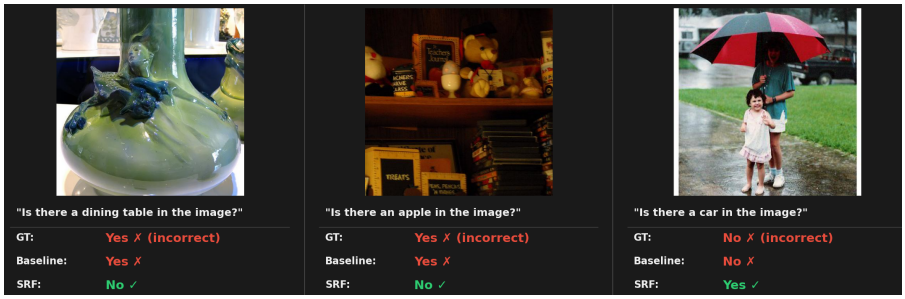

Figure 7: Examples of apparent annotation inconsistencies in POPE. The baseline prediction agrees with the annotation, whereas the SRF prediction agrees with the visual evidence.

C Experimental Setup

Qwen2.5-VL-3B-Instruct has 36 decoder layers with 16 attention heads each and uses dynamic image resolution, which we cap at $512 \times 28 \times 28$ pixels. LLaVA-1.5-7B has 32 layers with 32 heads each and encodes every image at 336×336 pixels into 576 visual tokens. Both models are run in bfloat16 with eager attention so that attention logits can be modified before the softmax.

All results use greedy decoding. For the Qwen2.5 VL 3B comparison in Table 1, we reproduced all baseline methods using their publicly available implementations and reported settings. For the LLaVA 1.5 7B comparison in Table 2, we report the published results from the corresponding papers under the same evaluation setting.

Semantic Relevance Map. The semantic relevance map is computed by a frozen CLIP ViT-B/32. For each image we embed the whole image and crops from a 3×3, a 5×5 and a 7×7 grid in one batched pass, score each crop against the extracted noun phrase and project the scores back to their locations. The presence threshold is $\tau = 0.20$ on Qwen2.5-VL and LLaVA-1.5.

Benchmarks. **MMVP** (Tong et al., 2024) probes fine-grained visual discrimination with 150 image pairs that differ in a single visual detail, with one binary question per image. We report pair accuracy, which credits a pair only when both of its images are answered correctly, together with per-image accuracy. **VLMBias** (Vo et al., 2026) evaluates counterfactual visual reasoning over seven categories. Each image modifies a familiar subject, for example, an animal with an unusual number of limbs or a flag with an unusual number of stars, so the memorized answer and the visible answer disagree. We report exact-match accuracy. **MME** (Fu et al., 2026) evaluates object-level and attribute-level perception with yes-or-no questions. Following the training-free methods we compare against, we use the Existence, Count, Position and Color subtasks and report their total score. **MMHal-Bench** (Sun et al., 2024) evaluates hallucination in open-ended generation over eight question categories. Responses are scored by GPT model used in SOTA works, and we report the mean score and the hallucination rate. **POPE** (Li et al., 2023) measures object hallucination with binary questions of the form "Is there a ⟨object⟩ in the image?". Its random, popular and adversarial splits differ in how negative objects are sampled, the adversarial split drawing on co-occurrence priors, and each contributes 3,000 questions. We report accuracy and F1.

Table 6: Inference cost or time per sample on MMVP images with Qwen2.5-VL-3B-Instruct, measured on a single RTX 2080 Ti.

Method	ms / sample	Acc
VAF	367	40.0
ILVAD	668	38.7
VCD	680	37.3
VHR	802	38.7
SRF	613	**45.3**

Hyper-parameters. SRF uses a small set of parameters, most of which are fixed across all experiments. The head fraction k is 0.20, so the intervention reaches the top fifth of heads in each layer of the band. The foveation scale $\sigma = 20$, presence threshold $\tau = 0.20$, and system suppression $\lambda_{\text{sys}} = 0.30$ for both architectures and all benchmarks.

The decoder layer interval and visual logit strengths are fixed per architecture. The layer band spans the central layers of the decoder, where our analysis in Section 3 shows cross-modal fusion to be concentrated. For Qwen2.5-VL-3B, we use layers 6–31 with $\lambda_{\text{sem}} = 2.0$ and $\lambda_{\text{bg}} = 0.2$. For LLaVA-1.5-7B, we use layers 8–20 with $\lambda_{\text{sem}} = 0.5$ and $\lambda_{\text{bg}} = 0.2$.

Vision-responsive heads are identified using 20 unlabeled calibration examples by ranking heads according to the fraction of attention assigned to visual tokens. This calibration uses neither task labels nor performance feedback and determines only which heads are modified.

D Additional Analysis

Ablation Study The ablation in Table 3 examines how the individual SRF components contribute to the final prediction. Using the semantic relevance map for visual token amplification provides an initial improvement over the baseline, while adding selective system suppression further improves performance. The largest gain appears when semantic foveation is included, indicating that modifying the visual input and reweighting decoder attention provide complementary benefits. In contrast, replacing the semantic map with a random spatial map produces little change even when the remaining interventions are kept identical. These results support the importance of query relevant spatial information rather than generic amplification or spatial perturbation alone.

Inference Cost Table 6 reports wall-clock cost on MMVP with Qwen2.5-VL-3B, measured over all 300 images on a single RTX 2080 Ti. SRF adds no additional VLM decoding pass. The contrastive baselines run the language model a second or third time on a modified input, which roughly doubles their cost, whereas SRF modifies attention inside the single pass the model already performs. Its overhead comes instead from the alignment model and from re-encoding the foveated image.

E Extended Empirical Analysis

The main paper analyzes how visual evidence is used under controlled perturbations and through decoder attention. We extend this analysis in two directions by examining language prior behavior on VLMBias and by repeating the attention analysis on LLaVA-1.5-7B. These experiments test whether the observed behavior persists across a different type of visual conflict and a second VLM architecture.

E.1 Input Perturbations and Bias Rate on VLMBias

VLMBias provides a useful setting for studying how predictions change when visual evidence conflicts with common language priors. We evaluate the full benchmark under the same controlled visual perturbations used in the main analysis and report both accuracy and bias rate, where bias rate measures the fraction of predictions that agree with the answer favored by the language prior. Figure 8 shows that perturbing the visual input changes both quantities substantially, with spatial disruption

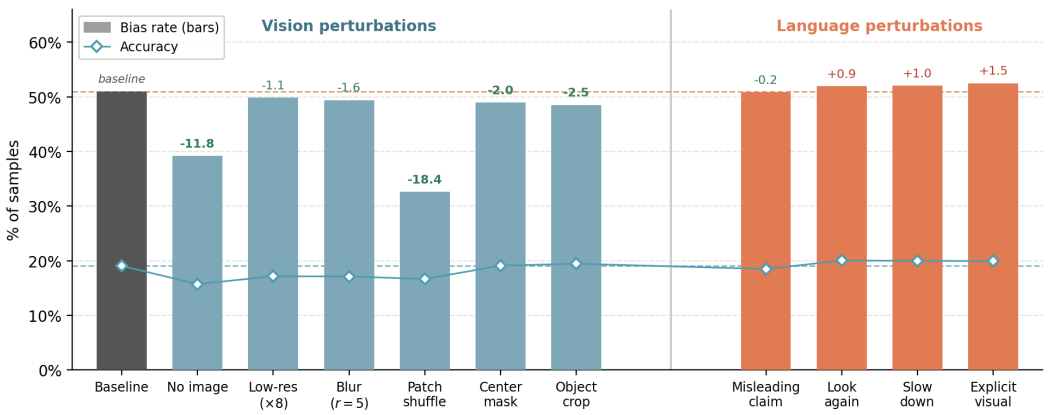

Figure 8: Bias rate (bars) and accuracy (diamonds) across visual and language perturbations on VLMBias.

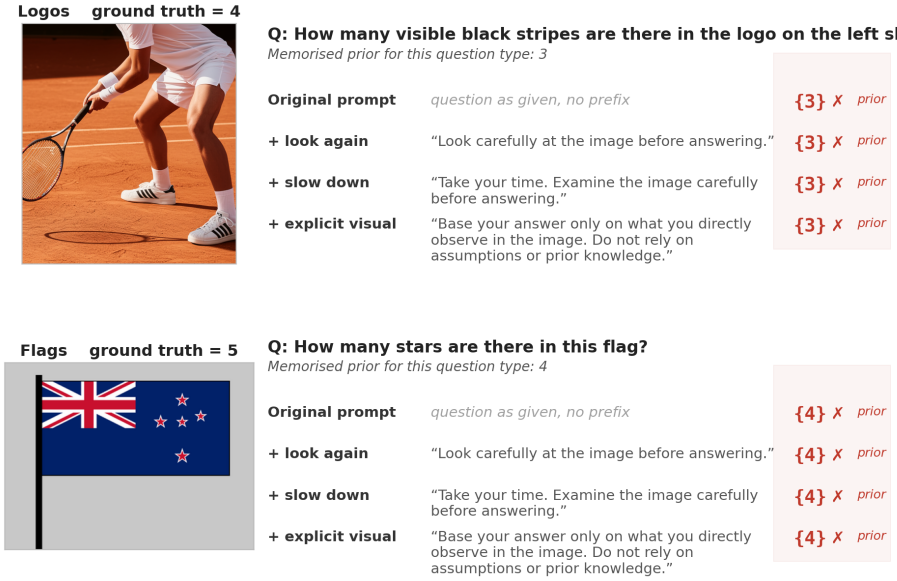

Figure 9: Representative VLMBias examples under prompt based interventions. Despite explicit instructions to inspect the visual evidence, the model continues to produce the response favored by the language prior.

and image removal producing some of the largest changes. Since these perturbations also remove useful visual information, a lower bias rate does not necessarily imply better grounding. Instead, the results show that prior consistent predictions are strongly affected by the form and availability of visual evidence.

E.2 Prompt Based Interventions on VLMBias

A natural alternative to modifying the model internally is to explicitly ask it to rely more strongly on the image. We therefore evaluate several prompt variants that instruct the model to look again, reason carefully, or answer only from directly observed visual evidence. Across the evaluated counting tasks, these instructions produce negligible changes in accuracy and do not reduce the tendency to follow the language prior. Figure 9 shows representative cases where the same prior consistent response is produced despite repeated instructions to inspect the image.

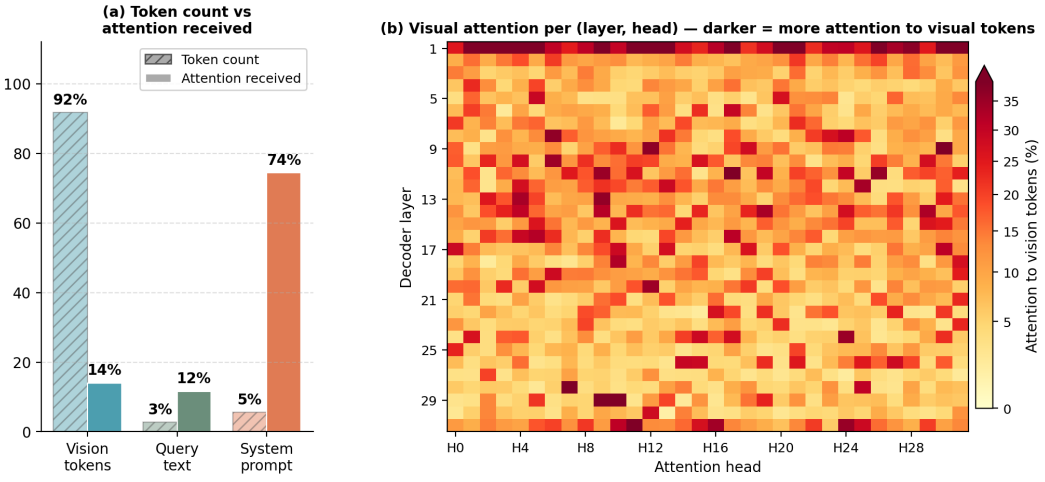

Figure 10: Visual attention imbalance and head specialization in LLaVA-1.5-7B



Figure 11: Attention heatmap and semantic relvance map from external alignment model on LLaVA-1.5-7B

E.3 Attention Analysis on LLaVA

The attention analysis in the main paper is performed on Qwen2.5-VL-3B, leaving open whether the observed concentration of visual attention is specific to that architecture. We therefore repeat the same analysis on LLaVA-1.5-7B by measuring visual attention across decoder layers and attention heads. Figure 10 shows that visual attention is again distributed nonuniformly, with a limited subset of heads contributing much more strongly than others. We additionally inspect the spatial attention maps produced by LLaVA in Figure 11.Together with the head level analysis, these results show a qualitatively similar pattern to Qwen2.5-VL=3B and support applying SRF selectively across heads, layers, and visual locations rather than modifying all decoder attention uniformly.

F Analytical Properties of SRF

F.1 Semantic Foveation as Relevance Gated Detail Preservation

The semantic foveation operation is defined as

$$I_{\mathrm{SRF}} = \bar{M} \odot I + (1 - \bar{M}) \odot B_\sigma(I). \tag{14}$$

Rearranging gives

$$I_{\mathrm{SRF}} = B_\sigma(I) + \bar{M} \odot (I - B_\sigma(I)) \,. \tag{15}$$

A Gaussian blur suppresses higher spatial frequencies while retaining much of the coarse spatial structure of the image. Thus, $I - B_\sigma(I)$ approximately represents the visual detail removed by the blur. The relevance map $\bar{M}$ controls where this detail is retained. When $\bar{M}_{u,v} = 1$, the original image value is recovered at location (u, v), while $\bar{M}_{u,v} = 0$ retains its blurred representation. Intermediate values interpolate between these cases.

This decomposition shows that semantic foveation preserves coarse image structure globally while selectively retaining finer visual detail in regions with higher semantic relevance.

F.2 Monotonic Semantic Reweighting of Visual Attention

Consider a selected decoder layer ℓ, head h, and query position i for which the visual tokens are available in the causal context. For a visual token j, SRF modifies the pre-softmax attention logit as

$$\widetilde{Z}_{ij}^{\ell,h} = Z_{ij}^{\ell,h} + \lambda_{\mathrm{sem}} s_j - \lambda_{\mathrm{bg}}(1 - s_j). \tag{16}$$

For analysis, define

$$\eta = \lambda_{\mathrm{sem}} + \lambda_{\mathrm{bg}}. \tag{17}$$

The modified visual-token logit can then be written as

$$\widetilde{Z}_{ij}^{\ell,h} = Z_{ij}^{\ell,h} - \lambda_{\mathrm{bg}} + \eta s_j. \tag{18}$$

Define the original attention distribution conditioned on visual-token positions as

$$p_j = \frac{\exp(Z_{ij}^{\ell,h})}{\sum_{k \in \mathcal{V}} \exp(Z_{ik}^{\ell,h})}, \qquad j \in \mathcal{V}. \tag{19}$$

The corresponding conditional distribution after SRF is

$$\widetilde{p}_j = \frac{\exp(\widetilde{Z}_{ij}^{\ell,h})}{\sum_{k \in \mathcal{V}} \exp(\widetilde{Z}_{ik}^{\ell,h})} \tag{20}$$

$$= \frac{p_j \exp(\eta s_j)}{\sum_{k \in \mathcal{V}} p_k \exp(\eta s_k)}. \tag{21}$$

The common term $-\lambda_{\mathrm{bg}}$ cancels because the distribution is conditioned on visual-token positions. Therefore, for any two visual tokens a and b,

$$\frac{\widetilde{p}_a}{\widetilde{p}_b} = \frac{p_a}{p_b} \exp\left(\eta(s_a - s_b)\right). \tag{22}$$

When $\eta > 0$ and $s_a > s_b$,

$$\frac{\widetilde{p}_a}{\widetilde{p}_b} > \frac{p_a}{p_b}, \tag{23}$$

showing that SRF increases the relative attention assigned to the visual token with higher semantic relevance.

We can further characterize the expected semantic relevance under $\widetilde{p}$. Let

$$\mu(\eta) = \sum_{j \in \mathcal{V}} \widetilde{p}_j s_j. \tag{24}$$

Differentiating with respect to η gives

$$\frac{d\mu(\eta)}{d\eta} = \sum_{j \in \mathcal{V}} \widetilde{p}_j s_j^2 - \left(\sum_{j \in \mathcal{V}} \widetilde{p}_j s_j \right)^2 \tag{25}$$

$$= \mathrm{Var}_{\widetilde{p}}(s) \geq 0. \tag{26}$$

Hence, within the visual-token subset, the expected semantic relevance of attended tokens is monotonically nondecreasing with the combined relevance strength η. Equality holds when the attended visual tokens have identical semantic relevance.

This result characterizes only the redistribution of attention among visual tokens. It does not imply that the total attention assigned to visual tokens increases, since the full attention distribution also depends on the original logits and on the absolute shifts introduced by λ_{sem}, λ_{bg}, and λ_{sys}. It just establishes that SRF produces the intended monotonic reweighting toward visual tokens with higher semantic relevance.

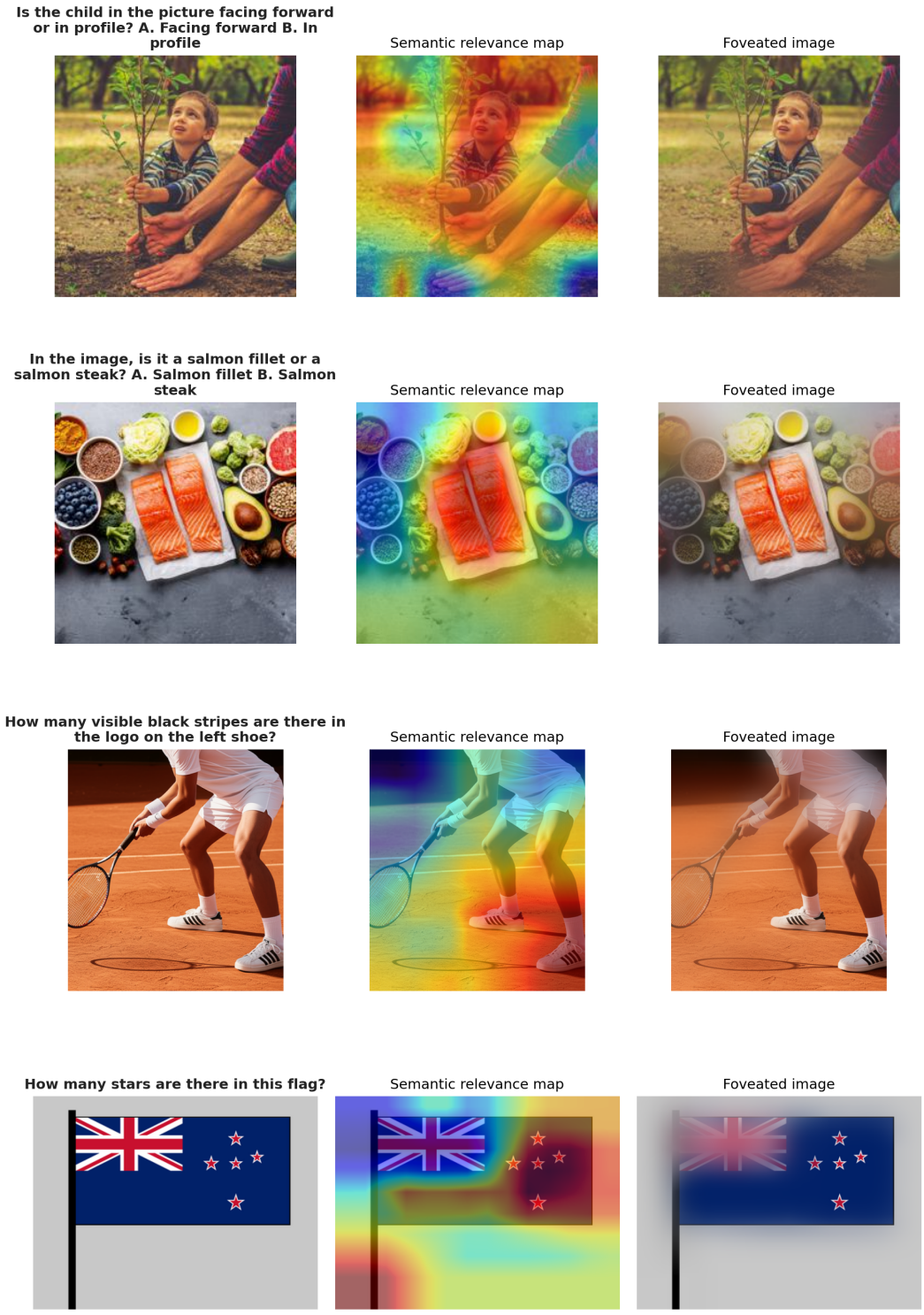

Figure 12: Semantic Relevance Map and Semantic Foveated Image on Qwen shown for two examples each from MMVP and VLMBias datasets respectively