

Tinted Frames: Question Framing Blinds Vision-Language Models

Wan-Cyuan Fan^{1,3} Jiayun Luo^{1,3†} Declan Kutscher^{2†}
Leonid Sigal^{1,3} Ritwik Gupta²

¹University of British Columbia ²University of California, Berkeley

³Vector Institute for AI

Correspondence: wancyuan@cs.ubc.ca

Abstract

Vision-Language Models (VLMs) have been shown to be blind, often underutilizing visual inputs even on tasks that require visual reasoning. In this work, we demonstrate that VLMs are *selectively* blind. They modulate the attention to visual inputs based on linguistic framing even when alternative framings demand identical visual reasoning. Using visual attention as a probe, we quantify how framing alters both the amount and distribution of attention over the image. Constrained framings, such as multiple choice and yes/no, induce substantially lower attention to image context compared to open-ended, reduce focus on task-relevant regions, and shift attention towards uninformative tokens. We further demonstrate that this attention misallocation is a key mechanism contributing to degraded accuracy and cross-framing inconsistency. Building on this mechanism, we evaluate whether common prompting strategies mitigate framing-induced blindness. We find that simple image-focused instructions provide limited benefit, while structured reasoning and grounding prompts partially restore target-region attention and reduce, but do not eliminate, cross-framing inconsistency. Our findings suggest that question framing and prompting should be considered part of VLM evaluation to better reveal models’ visual capability.

1 Introduction

Vision-Language Models (VLMs) have achieved strong performance on increasingly complex multi-modal reasoning tasks. However, whether these models reliably ground their textual outputs in visual evidence remains an open question (Kang et al., 2025; Luo et al., 2025). Prior studies show that VLMs can be visually “blind”: they often assign limited attention to visual tokens and rely on language priors when generating answers (Tong et al., 2024; Rahmanzadehgervi et al., 2024).

Blindness can be characterized as visual disengagement and structural bias. Models may allocate disproportionately high attention to visual attention sink tokens (Kang et al., 2025; Luo et al., 2025; Kaduri et al., 2025), which often correspond to background or semantically uninformative regions, or exhibit spatial biases induced by positional encoding (e.g., RoPE (Su et al., 2023)) and causal attention mechanisms (Zhu et al., 2025b; Tian et al., 2025; Wang et al., 2025a). These biases can create effective “blind spots” (Zhu et al., 2025b; Tian et al., 2025; Wang et al., 2025a), causing models to over-emphasize tokens near the beginning or end of the input sequence while neglecting semantically important visual tokens. However, these analyses have typically been conducted holistically, averaging observations across heterogeneous benchmarks and question framings. This perspective implies that such visual blindness is a static, inherent flaw of the model architecture. In practice, however, visual reasoning is always posed through language: the same visual concept may be queried as an open-ended, Yes/No, or a multiple-choice question (MCQ). These framings are widely used across major VLM benchmarks, making framing-dependent visual behavior central to how we assess model visual capabilities. Although prior works have shown that prompt formulation can affect model accuracy (Gu et al., 2023; Schmalfuss et al., 2025; Liang et al., 2026), it remains unclear whether such linguistic framing changes model’s visual processing itself, rather than merely shift answer priors or output distributions.

In this work, we demonstrate that VLMs are *selectively* blind. They decide how much to look at an image based on the textual framing of the question despite different framings requiring the same visual concepts to answer correctly. As illustrated in Fig. 1, reframing the same visual query can shift attention away from the relevant object. We hypothesize that question framing affects model to

[†]Equal contribution.

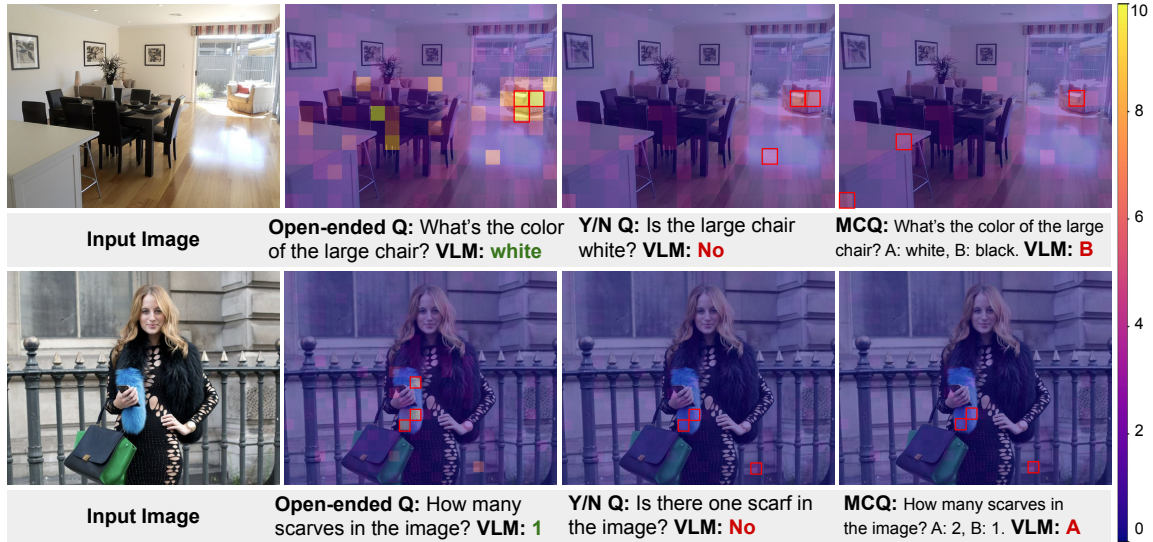


Figure 1: **VLM grounding changes as a function of question framing.** Attention maps (from Qwen2.5-VL-7B) reveal that while the model actively attends to the target object (the chair) during open-ended generation, it exhibits disengagement and misallocation when the same question is posed as a Yes/No or MCQ task. Note that we employ attention rollout (Abnar and Zuidema, 2020) rather than averaging attention weights across layers and token. By recursively rolling out attention matrices, we trace information propagation from inputs of early layers to output embeddings. The top three tokens with the highest attention are highlighted in red boxes, and the minimum and maximum values of the linear colormap are set to the same value for all images.

performance *indirectly* by altering visual attention.

We study this phenomenon mechanistically using attention rollout (Abnar and Zuidema, 2020) to estimate the visual information propagation and where the model attends when making output decisions. We conduct a three-part analysis. First, we introduce *cross-framing inconsistency* to quantify when models answer open-ended questions correctly but fail semantically equivalent constrained versions, especially on object-grounding tasks. Second, we show that constrained framings reduce overall image attention and redirect attention away from task-relevant regions. Third, through attention intervention, we show that these framing-induced attention shifts contribute to prediction errors.

To connect this mechanism to everyday VLM use, we further ask whether common prompting strategies can mitigate framing-induced visual blindness. We test image-focused instructions, thinking-style prompts, chain-of-thought reasoning, and grounding prompts that require visual evidence before answering. We find that simple instructions provide limited benefit, while more structured reasoning prompts partially align constrained framings with the open-ended setting by increasing visual attention on target regions. However, constrained framings still show lower target-region attention and cross-framing inconsistency. Our findings show that question framing and prompting in

VLM evaluation are key factors in revealing models’ true visual capability.

2 Related Work

Vision-language models (VLMs) (Li et al., 2024; Bai et al., 2025a; Zeng et al., 2025; Meta, 2024; Team et al., 2025; Abdin et al., 2024; Zhu et al., 2025a) have advanced from simple captioning to complex multimodal reasoning, but it remains difficult to determine whether their answers are grounded in visual evidence or driven by linguistic shortcuts (Lin et al., 2023). Answering this question requires looking beyond output accuracy, into whether these models truly ground their reasoning in what they see, and what factors might cause that grounding to break down.

2.1 Visual Grounding in VLMs.

Visual grounding, the ability to localize textual concepts within an image, has been a long-standing goal in computer vision. Classical architectures such as object detectors (Ren et al., 2015; Carion et al., 2020; He et al., 2017) are explicitly trained to produce spatial localizations, and vision transformers can develop interpretable attention patterns that correlate with object boundaries (Caron et al., 2021). VLMs embed these vision encoders but are trained end-to-end for language generation. Therefore, spatial grounding is an implicit learning task

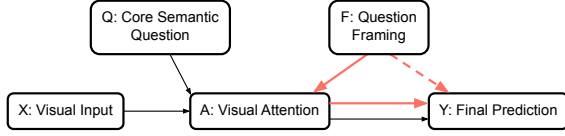


Figure 2: **Illustration of impact of question framing.** We hypothesize that question framing influences model predictions *through* visual attention. Framing alters attention allocation ($F \rightarrow A$), which in turn degrades prediction quality ($A \rightarrow Y$).

as opposed to a primary objective. Yet, recent works (Kang et al., 2025; Fu et al., 2025) show that this implicit grounding is unreliable. Models can answer correctly while attending to irrelevant regions. We extend this line of work by establishing a mechanistic link between question framing, visual attention, and output quality.

2.2 Visual Disengagement and Bias in VLMs.

A growing amount of work documents several forms of visual disengagement in VLMs. Models often allocate less attention to visual content than to text (Zhang et al., 2024; Tong et al., 2024), rely on semantically meaningless visual sink tokens (Luo et al., 2025), and exhibit spatial blind spots caused by rotary position embeddings, causal attention masks, and data distribution (Tian et al., 2025). These findings paint a picture of visual blindness as a general and static property. In contrast, we show that visual disengagement is dynamic and conditional on linguistic framing. VLMs may attend well under open-ended questions but disengage under constrained alternatives. Therefore, this work reframes existing findings from “the model cannot see” to “the model decides not to see.”

2.3 Prompt Sensitivity.

While human-in-the-loop evaluation (Chiang et al., 2024) offers a direct measure of model quality based on human preference, it does not scale to systematic probing of specific visual capabilities. Benchmarks therefore rely on targeted questions, often treating framing as a neutral container: a model that understands a scene should answer consistently regardless of how the question is asked. But is this assumption warranted? VLMs are known to be sensitive to question phrasing (Chou et al., 2025). Prior work has studied a range of within-format perturbations: MCQ option ordering effects (Pezeshkpour and Hruschka, 2024), yes-bias (Li et al., 2023b), negation bias (Alhamoud et al., 2025), and paraphrase inconsistency. Re-

lated format effects have also been reported for text-only LLMs, whose accuracy on reasoning tasks can vary substantially across multiple-choice and true/false formats (Song et al., 2025). Our work shows that in VLMs such inconsistency is not merely a language-side formatting effect: question framing also changes how the model attends to and uses the visual input. These studies vary the surface wording while keeping the question format fixed. Framing, by contrast, is a stronger structural shift: it changes the format itself (e.g., from open-ended to Yes/No or MCQ) while preserving the underlying semantic question. This axis of sensitivity has received comparatively little attention. Existing studies (Shah et al., 2025) primarily measure sensitivity through output behavior, including accuracy drops and answer distribution shifts. We instead explore *how* framing reshapes models’ internal visual processing.

3 Hypothesis on Framing-Attention Influence

We illustrate an ideal VLM processing chain in Fig. 2. In visual question answering, both text and image modalities should impact visual attention, which in turn affects the final prediction. The semantics of the question are often independent of the question framing. Therefore, there should be no latent factor affecting visual attention.

However, we posit that framing directly impacts visual attention ($F \rightarrow A$). Predictions follow from visual attention ($A \rightarrow Y$), but framing affects silently degrade the quality of this attention. Together, these form a joint pathway ($F \rightarrow A \rightarrow Y$). Therefore, a latent relationship exists between framing and the final prediction ($F \rightarrow Y$). For a robust VLM, none of these pathways should exist; their presence reveals that current models rely on shallow, framing-dependent heuristics rather than genuine visual understanding.

We investigate the impact and existence of each pathway in turn. Sec. 4 examines the overall effect of framing on predictions ($F \rightarrow Y$). Sec. 5 analyzes the visual attention pathway ($F \rightarrow A \rightarrow Y$). Finally, in Sec. 6, we study a set of prompting methods to see how they impact cross-framing robustness.

4 Cross-Framing Inconsistency ($F \rightarrow Y$)

Before examining the internal mechanisms behind framing effects, we first ask a simpler question: does question framing affect the model’s final pre-

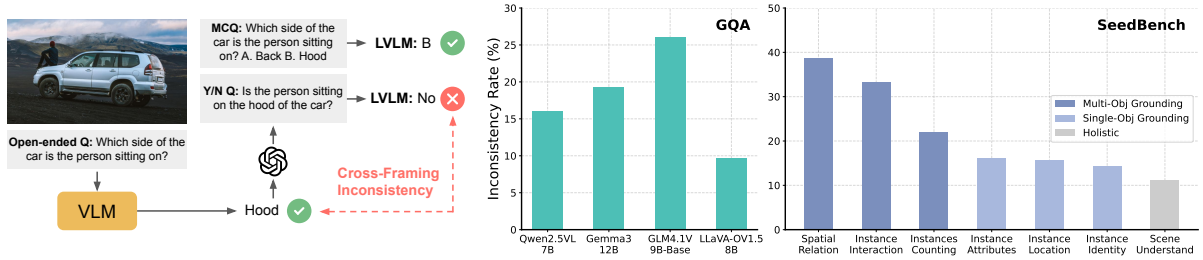


Figure 3: **Tested open-source VLMs have a significant inconsistency rates across framings and task types.** (Left) Cross-framing inconsistency is evaluated by reframing questions with an LLM. (Right) With Qwen2.5-VL-7B, inconsistency is up to 26% on GQA and 38% on SeedBench.

diction ($\mathbf{F} \rightarrow \mathbf{Y}$ in Fig. 2)? We use open-ended generation as the anchor. Without candidate options to select from, the model must generate the answer through free-form reasoning, making it less likely to succeed relying on the prior knowledge alone. If a model answers an open-ended question correctly but fails when the same question is reframed as Yes/No or MCQ, this inconsistency is unlikely to stem from a fundamental lack of visual understanding and is more likely driven by the framing itself. We formalize this as *cross-framing inconsistency*: the rate at which a model fails to preserve its correct open-ended answer under Yes/No or MCQ.

Evaluation protocol. Our protocol is illustrated in Fig. 3 (left). We evaluate on GQA (Hudson and Manning, 2019), an open-ended benchmark, and SeedBench (Li et al., 2023a), a MCQ benchmark covering diverse visual reasoning tasks. To reduce concerns that framing effects arise from LLM-hallucinated reformulations, we use two complementary construction strategies. For GQA, GPT-5.1 generates Yes/No and MCQ counterparts under constraints from the ground-truth scene graph and object annotations. For SeedBench, the reformulation is largely rule-based, the original MCQ options are retained, open-ended questions are formed by removing options, and Yes/No questions are created by rephrasing with the correct answer or existing negative option. Thus, SeedBench introduces minimal new semantic content. We further conduct human evaluation to verify the reformulation quality, with details provided in the appendix. During the evaluation, we query the VLM with open-ended questions and retain only correctly answered samples. We then re-query the model with the semantically equivalent Yes/No and MCQ counterparts, measuring whether the correct answer is preserved. The inconsistency rate is calculated over cases where the open-ended question is answered correctly, but either the Yes/No or MCQ counter-

part is incorrect.

Results. All tested VLMs exhibit substantial inconsistency (Fig. 3, right). On GQA, Qwen2.5-VL (Bai et al., 2025b), Gemma3 (Team et al., 2025), GLM4.1V (Zeng et al., 2025) exhibit more than 15% cross-framing inconsistency, meaning they fail to preserve their own correct answers under constrained framing for nearly one in six questions. The task-level SeedBench breakdown using Qwen2.5-VL-7B is particularly revealing:

Tasks requiring object grounding exhibit the highest inconsistency rates, with multiple-object grounding tasks such as spatial relation and counting being the most affected, suggesting that *constrained framing is most damaging where visual grounding matters most*. Note that we provide an additional discussion how the framing effect depends on scene complexity in Sec. C.4. Also, we provide discussion about a chance-corrected variant of the metric, which yields an higher inconsistency rate, in Sec. C.3.

Reverse direction inconsistency. Our inconsistency measure is directional: the model answers the open-ended question correctly but fails under constrained formats. We adopt this direction because open-ended correctness provides a strict reference point; the model must generate the correct answer without being given a closed set of candidates, making success unlikely to arise from answer priors alone. A natural complementary question is whether constrained formats can instead *correct* open-ended mistakes. We therefore compute the reverse rate: among samples where the model fails on the open-ended question, the rate at which it answers correctly under a constrained format. Under the same protocol, the reverse rate is 44.64%, substantially higher than the 19.26% in the original direction, showing that constrained formats frequently flip open-ended failures into correct answers. However, this direction should be inter-

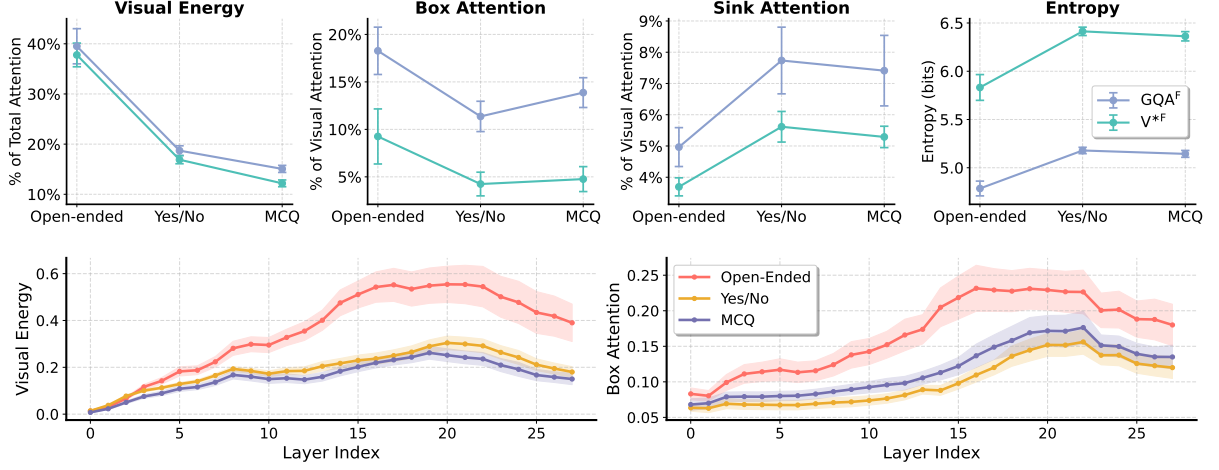


Figure 4: **Visual energy drops significantly on non-open-ended framings.** (Top) There is a significant drop in attention applied to the portion of the image containing the object of interest, and a corresponding increase in attention to sink tokens, for yes/no and MCQ framing. (Bottom) Per-layer values of visual energy and box attention of Qwen2.5-VL-7B on GQA^F .

preted with care: Yes/No and MCQ questions can be answered correctly through guessing or answer priors, so the reverse rate conflates cases where constrained framing genuinely helps with cases where the constrained answer is correct by chance. This asymmetry is precisely why we anchor our analysis on open-ended correctness, the cleaner signal of visual understanding.

These results support the existence of $\mathbf{F} \rightarrow \mathbf{Y}$: question framing alters predictions.

5 Impact of Framing on Visual Attention ($\mathbf{F} \rightarrow \mathbf{A} \rightarrow \mathbf{Y}$)

Having confirmed that framing alters predictions ($\mathbf{F} \rightarrow \mathbf{Y}$), we now ask whether framing reshapes the model’s visual attention ($\mathbf{F} \rightarrow \mathbf{A}$) and whether such shifts drive prediction failures ($\mathbf{A} \rightarrow \mathbf{Y}$).

Choice of datasets and models. We analyze two benchmarks with spatial annotations, enabling precise mapping of attention to semantic regions. GQA is a general-purpose VQA benchmark with dense semantic annotations including bounding boxes for target objects and scene graphs. $\mathbf{V}^*$ (Wu and Xie, 2024) is a high-resolution visual grounding benchmark with around 300 MCQ samples requiring fine-grained spatial reasoning and with bounding box annotations for target regions.

To isolate framing from question content, we construct three framing variants for each sample: open-ended, Yes/No, and MCQ. As in Sec. 4, we minimize LLM-hallucinated reformulations by using annotation-constrained and rule-based conversions. For GQA, GPT-5.1 generates Yes/No and

MCQ counterparts using the ground-truth scene graph and object annotations, ensuring that candidate objects, attributes, and relations are grounded in image annotations. For $\mathbf{V}^*$, an MCQ benchmark, reformulation is largely rule-based again: we retain the original answer options for the MCQ framing, form open-ended questions by removing options, and construct Yes/No questions by rephrasing with the correct answer or an existing negative option. Human evaluation verifies reformulation quality, with prompt templates and verification details provided in the appendix. After filtering, we curate 10k unique semantic queries for GQA and the full 300 samples for $\mathbf{V}^*$. With three framing variants per query, this yields 30k samples for GQA and 900 for $\mathbf{V}^*$. We denote the resulting framing-controlled datasets as GQA^F and $\mathbf{V}^{*F}$ to distinguish them from the original benchmarks. Note that we focus on Qwen2.5-VL-7B, with extended results covering Gemma3 and LLaVA-OneVision1.5 (An et al., 2025) in the supplementary.

Visual attention aggregation. To quantify the visual reliance during generation, we use attention rollout (Abnar and Zuidema, 2020) rather than simple attention averaging across layers. Rollout traces visual information propagation from input visual tokens to output embeddings, accounting for both direct attention and indirect pathways via residual connections. Crucially, we apply receptive field normalization to preserve causality during attention map aggregation, as required for autoregressive transformers (Abnar and Zuidema, 2020).

Formally, let $\mathbf{W}_{att}^{(\ell)} \in \mathbb{R}^{N \times N}$ be the raw attention matrix at layer ℓ , with rows as query tokens

and columns as key tokens. Following (Abnar and Zuidema, 2020), we account for residual connections by defining the adjusted attention matrix as $\mathbf{A}^{(\ell)} = 0.5\mathbf{W}_{att}^{(\ell)} + 0.5\mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix.

To address the causal masking bias (Wang et al., 2024), we apply receptive field normalization. Specifically, we scale the key tokens by $\mathbf{A}^{(\ell)}$ based on their receptive field size $\mathcal{S}$ to ensure unbiased probability mass propagation. We then re-normalize the rows to ensure the resulting matrix is row-stochastic matrix before the recursive rollout:

$$\mathbf{R}^{(\ell)} = \left(\mathcal{N}(\mathbf{A}^{(\ell)} \cdot \text{diag}(\mathcal{S})) \cdot \mathbf{R}^{(\ell-1)} \right), \quad (1)$$

where $\mathbf{R}^{(0)} = \mathbf{I}$ and $\mathcal{N}(\cdot)$ performs row-wise normalization, ensuring that each row sums to 1. The final product $\mathbf{R}^{(L)}$ is a valid stochastic matrix capturing the effective information flow between all token pairs, where L denotes the total number of layers for rollout. To quantify visual reliance, we extract the sub-matrix connecting the generated output tokens (queries) to input visual tokens (keys). We define the final **visual energy** as the probability mass within this sub-matrix, where a higher total summation indicates a stronger reliance on visual content during generation.

5.1 Framing Reshapes Visual Attention ($\mathbf{F} \rightarrow \mathbf{A}$)

We examine how question framing changes visual attention along three dimensions: overall visual engagement, allocation to task-relevant regions, and spatial dispersion.

Visual energy and spatial allocation. As shown in Fig. 4 (top), constrained framings consistently exhibit lower overall visual energy compared to open-ended generation across both GQA^F and V^{*F} , indicating reduced reliance on visual content. The spatial distribution also shifts substantially: attention to low-semantic-relevance sink tokens (Kang et al., 2025) increases for both Yes/No and MCQ, while target-region attention drops. On GQA^F , box attention falls from 19% for Open-ended to around 12% for Yes/No and 13% for MCQ, a relative drop of about 40%. The gap is even larger on V^{*F} , where fine-grained grounding is required: target-region attention drop by roughly 50% under Yes/No and MCQ. Thus, the model does not merely attend less to the image; it redirects attention away from task-relevant regions toward unrelated regions. Entropy

also increases under constrained framings, indicating that remaining visual attention becomes more diffuse and less focused on any specific region.

Constrained framings reduce overall visual energy, redirect attention from target regions to unrelated regions, and produce more dispersed attention, further supporting the hypothesis that question framing impacts visual attention $\mathbf{F} \rightarrow \mathbf{A}$.

Layer-wise analysis Early layers exhibit similar attention patterns across all framings (Fig. 4). Divergence emerges in the middle layers (approximately layers 12–22), which prior work (Jiang et al., 2025) has identified as cross-modal interaction layers. In these layers, both visual energy and bounding box attention drop significantly for Yes/No and MCQ framings compared to open-ended framing, and this gap persists through the remaining layers.

Decomposing the framing effect. A question prompt consists of two parts: the question itself (e.g., “How many dogs are there?” vs. “Is there a dog?”) and appended instructions (e.g., “Answer with Yes or No”). To identify the driver of the attention shift, we vary the question framing while holding instructions fixed, and vary instructions while holding the question fixed, as illustrated in Fig. 5 (top). The coefficient of quartile variation (CQV) for both visual energy and bounding box attention, shown in Fig. 5 (bottom), reveals that variation from changing the question framing is $\approx 3\times$ larger than variation from changing instructions alone.

5.2 Connecting Attention to Prediction ($\mathbf{A} \rightarrow \mathbf{Y}$)

We now test whether this attention distortion drives prediction errors ($\mathbf{A} \rightarrow \mathbf{Y}$). The correlations observed above do not by themselves imply a direct link to accuracy. Specifically, VLMs may simply find constrained framings easier and allocate less attention to target area without performance loss. To distinguish efficient shortcutting from harmful disengagement, we perform *attention steering*, directly intervening on the attention maps under constrained framings, restore them toward open-ended levels, and measure whether accuracy recovers.

Attention steering. Ideally, one would copy the full attention maps from an open-ended forward pass to its constrained counterpart. In practice, however, question tokens differ across framings, producing attention maps with different query di-

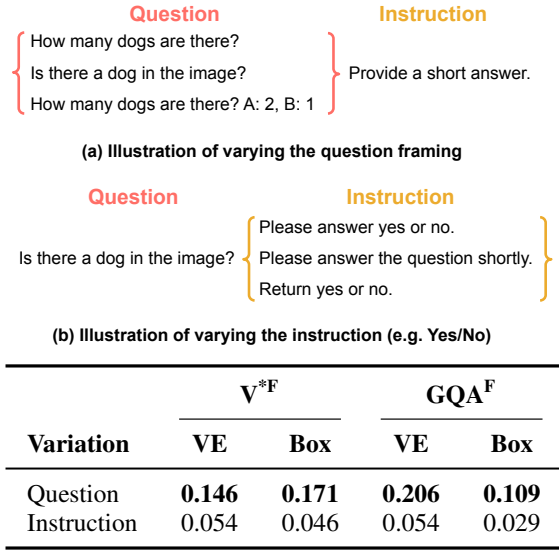


Figure 5: (Top) Illustration of question/instruction variation. (Bottom) Coefficient of quartile variation on VE and Box for varying framing and instruction.

mensions. We therefore intervene on attention properties, magnitude and spatial distribution, using a multiplier-based scheme. We study two complementary interventions that together disentangle the contributions of visual engagement magnitude and spatial allocation:

Visual Energy (VE) steering For each question, we compute the ratio of visual energy between open-ended and constrained framings from attention rollout and use it as a multiplier. During constrained framing inference, for each head and layer, we scale up the attention weights on all image tokens for every post-image query token by this multiplier, while proportionally reducing the attention on non-image tokens to maintain a valid probability distribution. This preserves the spatial distribution within image tokens and changes only the total visual energy, isolating how much the model attends to the image content.

Box attention steering Using the ground truth bounding boxes to identify task-relevant regions, we compute a separate multiplier from the ratio of open-ended to constrained bounding-box attention. During inference for the constrained framing, we scale up attention weights on image tokens inside the bounding-box while keeping total visual energy fixed. This isolates whether where the model attends within the image affects prediction.

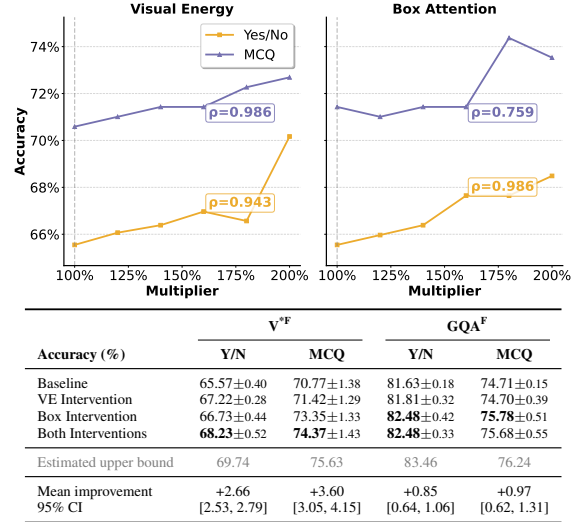
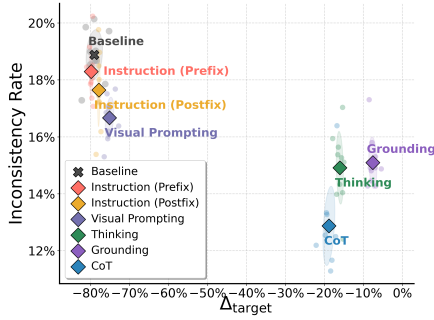


Figure 6: (Top) Performance gains on V^*F when multiplier with Spearman's rank correlation coefficient listed. (Bottom) Attention steering results on V^*F and GQA^F , reported as mean and std over 40 runs under different randomness. Mean improvement and confidence intervals are computed for the full intervention. The upper bound is estimated by directly optimizing steering parameters against ground-truth answers.

Results. We report accuracy after steering in Fig. 6 (bottom). To verify that the gains are reliable statistically, we report mean and std intervention results for 40 runs with different randomness. On V^*F , which demands fine-grained grounding, both interventions yield clear gains: VE steering improves Y/N and MCQ by +1.7 and +0.7 pts, while Box steering improves them by +1.2 and +2.6 pts. On GQA^F , a more general reasoning benchmark, VE steering has little effect, while Box steering still delivers consistent improvement (+0.9 and +1.1 pts). Since the single-axis interventions isolate visual energy and spatial allocation separately, we further steer both factors jointly toward the open-ended attention pattern to estimate the overall contribution of visual attention. This full intervention achieves the largest and statistically stable gains: +2.66 and +3.60 pts on V^*F Y/N and MCQ, and +0.85 and +0.97 pts on GQA^F , with 95% confidence intervals positive across settings. We visualize performance change via incremental multiplier increases in Fig. 6 (top) on V^*F . We show that accuracy improves monotonically as attention is steered closer to open-ended levels, with high Spearman correlations. These results indicate that spatial misallocation, where attention falls within the image, matters more consistently than total visual energy, especially for grounding-heavy tasks requiring precise localization.



Prompt	Overall ↓	Spatial Rel.	Inst. Interact.	Inst. Count.	Inst. Attr.	Inst. Loc.	Inst. Id.	Scene Und.
Baseline	20.76 $\pm$ 0.71	24.07	18.52	27.17	13.51	27.50	17.74	12.68
Instruction (Postfix)	18.50 $\pm$ 0.51	20.83	<u>15.10</u>	21.09	14.72	18.87	19.14	13.11
Instruction (Prefix)	19.37 $\pm$ 0.75	23.16	18.55	22.49	16.22	<u>16.59</u>	19.16	12.91
CoT	12.74 $\pm$ 0.86	15.14	20.65	11.20	14.56	12.40	<u>16.06</u>	<u>9.50</u>
Thinking	<u>14.65</u> $\pm$ 1.11	<u>19.66</u>	16.20	<u>16.80</u>	<u>13.71</u>	17.67	18.69	9.16
Grounding	16.41 $\pm$ 1.28	20.00	10.71	17.65	19.44	17.65	12.00	21.05

Figure 7: (Left) Simple instruction prompts yield limited improvement on degeneration and inconsistency on GQA comparing to the baseline (black), while reasoning-based approaches reduce framing-induced degeneration and inconsistency rate. Δ_{target} denotes the relative change of total target-region attention from open-ended to constrained framings defined in Eq. (2). (Right) Inconsistency rate on SeedBench (lower is better). Reasoning-based prompts improve inconsistency on SeedBench, especially grounding-heavy tasks such as spatial relation and counting.

How much can visual attention alone recover?

To contextualize the remaining gap, we estimate an attention-based upper bound. For each dataset, we sample 500 examples and introduce two learnable steering parameters, one for visual energy and one for box attention. Instead of setting them from the open-ended attention maps, we directly optimize them for 100 iterations using the loss between the model output and the ground-truth answer. This is not a practical inference method, since it accesses the ground truth, but a diagnostic estimate of how much performance can be recovered by modifying visual attention alone. Note that this bound is not expected to approach 100% accuracy: even with maximized visual attention, VQA still requires attending to the question tokens and performing cross-modal reasoning. The estimated upper bound (Fig. 6, bottom) suggests that visual attention changes can account for roughly 4–5 pts on V^*F and around 2 pts on GQA^F , and our full intervention recovers most of this estimated gap. The remaining gap likely stems from factors beyond visual attention, such as text attention, answer priors, and cross-modal interactions.

Attention steering shows that framing-induced attention shifts directly affect model output $A \rightarrow Y$. Restoring attention under constrained framings recovers accuracy. Spatial allocation yields broad gains, while visual energy magnitude primarily benefits grounding-heavy tasks, revealing that framing-induced errors stem from qualitatively different attention failures depending on the task.

5.3 Takeaways.

Taken together, Secs. 4 and 5 establish a three-link pathway from question framing to model failure.

First, constrained framings (Y/N, MCQ) can cause prediction errors, with the largest inconsistency happening on grounding-heavy tasks where precise visual localization is essential (**Finding 1: $F \rightarrow Y$**). Second, constrained framings reduce overall visual energy and redirect engagement away from task-relevant regions (**Finding 2: $F \rightarrow A$**). Third, steering visual attention back to open-ended levels recovers accuracy, showing that spatial allocation matters broadly while visual-energy magnitude primarily benefits grounding tasks (**Finding 3: $A \rightarrow Y$**). These findings show that framing-induced errors arise from fundamental shifts in how visual information is processed by the VLM. In the next section, we study whether existing prompting strategies can mitigate framing-induced visual blindness.

6 Can Prompting Mitigate Framing-Induced Visual Blindness?

Secs. 4 and 5 show that constrained question framings can alter visual attention, and that restoring attention toward the open-ended patterns improves prediction and thereby consistency. This motivates a practical question: can prompting reduce framing-induced visual blindness? If constrained framings fail partly by weakening attention to task-relevant regions, then prompts that encourage explicit visual grounding should both narrow the attention gap and improve cross-framing robustness.

Prompting choices. Our baseline is a standard VQA evaluation prompting with a question plus a short answer-format instruction (e.g., return yes/no, an option letter, or a short phrase), without extra visual instructions, reasoning, or grounding. We evaluate prompting families commonly used to en-

courage careful visual reasoning: image-focused instructions, prepended or appended to the question (e.g., asking the model to “carefully examine the image before answering”); visual prompting, where target regions are roughly highlighted in the image, and reasoning-oriented prompts. The latter include *Thinking*, which asks the model to think before giving the final answer; *CoT*, which asks for step-by-step reasoning; and *Grounding*, which asks the model to first identify visual evidence, such as relevant boxes and labels, before answering. For each family, we create multiple paraphrased variants and report the mean and standard deviation across variants. See Sec. B.3 for prompt details. Note that because reasoning and grounding prompts require reliable instruction following, we use Qwen3-VL-8B in this study. To ensure continuity with the earlier mechanism, we also verify in the supplementary that Qwen3-VL exhibits the same framing-induced degradation observed in Qwen2.5-VL.

Evaluation protocol. We evaluate prompting by its ability to reduce framing-induced degeneration, rather than by raw accuracy alone. Stronger prompts can improve performance across all question formats, but such gains do not reveal whether the model has become less sensitive to framing. Since open-ended, Yes/No, and multiple-choice versions ask the same underlying visual question, a robust model should preserve its answer across these framings. We therefore measure cross-framing inconsistency. On GQA^F, we also test whether reduced inconsistency is accompanied by restored visual grounding. Specifically, we compute total target-region attention as the product of visual energy and the fraction of visual attention assigned to task-relevant boxes. For each constrained framing, we compute the relative change from the corresponding open-ended framing:

$$\Delta_{\text{target}} = \frac{\mathcal{A}_{\text{target}}^{\text{constrained}} - \mathcal{A}_{\text{target}}^{\text{open}}}{\mathcal{A}_{\text{target}}^{\text{open}}}, \quad (2)$$

where $\mathcal{A}_{\text{target}}$ denotes target-region attention. Negative values indicate attention loss under the constrained format, while values closer to zero indicate recovery toward the open-ended attention pattern. Each GQA^F prompt variant is plotted by target-region attention and cross-framing inconsistency; we aggregate variants within each prompting family and visualize the mean and standard deviation. This plot shows whether a prompting family reduces framing-induced degeneration while also

restoring attention to task-relevant visual content. We also use SeedBench for a task-wise inconsistency breakdown across visual reasoning tasks. Visual prompting is reported only on GQA^F, since SeedBench lacks box annotations.

Analysis. Fig. 7 shows that prompting strategies differ substantially. Image-focused instructions provide limited improvement on both axes, suggesting that simply asking the model to look carefully is insufficient under constrained framing. Reasoning-oriented prompts are more effective: *Thinking*, *CoT*, and *Grounding* reduce the target-region attention gap while lowering inconsistency, supporting the view that prompting can improve robustness when it repairs attention to relevant image regions. However, we note that recovery is partial. Even with stronger reasoning and grounding prompts, constrained framings still allocate less attention to target regions than open-ended questions, and non-trivial inconsistency remains.

SeedBench results (Fig. 7, right) further confirm this pattern that reasoning-based prompts generally reduce inconsistency across task categories, with especially clear gains on grounding-heavy tasks such as counting and spatial relation recognition. Overall, prompting mitigates but does not eliminate framing-induced visual blindness. These results support our attention-based mechanism and show that question framing and prompting are a key factor to faithfully estimate models’ visual capability.

7 Conclusion

In this paper, we show that VLM visual attention is strongly shaped by question framing. Through mechanistic analysis, we establish a latent relationship between framing and downstream prediction. Prompting can partially restore visual attention and improve robustness, but does not eliminate the effect. These findings reframe visual blindness as a dynamic behavior influenced by input design, not only a fixed architectural limitation, and highlight that question framing and prompting are essential factors in assessing VLMs’ visual capabilities.

Limitation

While our study provides a comprehensive mechanistic analysis of framing effects across diverse and prominent VLM families, the architectural landscape of multimodal models is rapidly evolving. Future research could extend this investigation

to other emerging architectures, such as Mamba-based VLMs or Mixture-of-Experts (MoE) models, to determine if these structural paradigms inherently mitigate or exhibit similar framing-induced attention shifts. In addition, the framings studied in this work provide a concrete and controlled starting point, but do not cover the full range of linguistic framing variation expected in real-world settings.

Acknowledgements

This work was funded, in part, by the Vector Institute for AI, Canada CIFAR AI Chairs, NSERC Canada Research Chair (CRC), AML-TN UBC, and NSERC Discovery and Discovery Accelerator Supplement Grants. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, the Digital Research Alliance of Canada⁰, companies¹ sponsoring the Vector Institute, and Advanced Research Computing at the University of British Columbia. Additional hardware support was provided by John R. Evans Leaders Fund CFI grant and Compute Canada under the Resource Allocation Competition award. Ritwik Gupta and Declan Kutscher were supported in part by funding from the Department of Defense, The House Fund, and BAIR’s industrial alliance programs. Additional compute was provided by the Department of Defense’s High Performance Computing Modernization Program. We are immensely grateful to Bicheng Xu from UBC and Stephanie Fu from UCB for sharing their valuable suggestions in paper writing and experiments.

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*.
- Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip HS Torr, Yoon Kim, and Marzyeh Ghassemi. 2025. Vision-language models do not understand negation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29612–29622.
- Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, and 1 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.
- Shih-Han Chou, Shivam Chandhok, Jim Little, and Leonid Sigal. 2025. Mm-r3: On (in-) consistency of vision-language models (vlms). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4762–4788.
- Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.
- Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. 2025. Hidden in plain sight: VLMs overlook their visual representations. *arXiv preprint arXiv:2506.08008*.

⁰alliancecan.ca

¹<https://vectorinstitute.ai/#partners>

- Anthony Fuller, Yousef Yassin, Junfeng Wen, Daniel G Kyrillos, Tarek Ibrahim, James R Green, and Evan Shelhamer. 2025. Lookwhere? efficient visual recognition by learning where to look and what to see from self-supervision. *arXiv preprint arXiv:2505.18051*.
- Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. 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*, pages 14375–14385.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.
- Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.
- Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25004–25014.
- Omri Kaduri, Shai Bagon, and Tali Dekel. 2025. What’s in the image? a deep-dive into the vision of vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14549–14558.
- Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.
- Haohui Liang, Runlin Huang, Yingjun Du, Yujia Hu, Weifeng Su, and Cees GM Snoek. 2026. Prompt-robust vision-language models via meta-finetuning. In *The Fourteenth International Conference on Learning Representations*.
- Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. 2023. Revisiting the role of language priors in vision-language models. *arXiv preprint arXiv:2306.01879*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Jiayun Luo, Wan-Cyuan Fan, Lyuyang Wang, Xiangteng He, Tanzila Rahman, Purang Abolmaesumi, and Leonid Sigal. 2025. To sink or not to sink: Visual information pathways in large vision-language models. *arXiv preprint arXiv:2510.08510*.
- AIa Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog*. Retrieved December, 20:2024.
- Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017.
- Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2024. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pages 18–34.
- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, and 1 others. 2025. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pages 28085–28128.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.
- Jenny Schmalz, Nadine Chang, Vibashan VS, Maying Shen, Andres Bruhn, and Jose M Alvarez. 2025. Parc: A quantitative framework uncovering the symmetries within vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25081–25091.

- Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. 2024. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems*, 37:68658–68685.
- Monika Shah, Sudarshan Balaji, Somdeb Sarkhel, Sanorita Dey, and Deepak Venugopal. 2025. Analyzing the sensitivity of vision language models in visual question answering. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 431–438.
- Seok Hwan Song, Mohna Chakraborty, Qi Li, and Wal-lapak Tavanapong. 2025. Is large language model performance on reasoning tasks impacted by different ways questions are asked? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22066–22081.
- J Su, Y Lu, S Pan, A Murtadha, B Wen, and Y Liu Roformer. 2023. Enhanced transformer with rotary position embedding., 2021. DOI: <https://doi.org/10.1016/j.neucom>.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. 2025. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10599–10609.
- Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, and 1 others. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Chengcheng Wang, Jianyuan Guo, Hongguang Li, Yuchuan Tian, Ying Nie, Chang Xu, and Kai Han. 2025a. Circle-rope: Cone-like decoupled rotary positional embedding for large vision-language models. *arXiv preprint arXiv:2505.16416*.
- Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. 2025b. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7907–7915.
- Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M Kakade, Hao Peng, and Heng Ji. 2024. Eliminating position bias of language models: A mechanistic approach. *arXiv preprint arXiv:2407.01100*.
- Penghao Wu and Saining Xie. 2024. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094.
- xAI. 2024. [Grok-1.5 vision preview](#).
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186.
- Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.
- Xiaofeng Zhang, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. 2024. From redundancy to relevance: Enhancing explainability in multimodal large language models. *arXiv preprint arXiv:2406.06579*, 17.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.
- Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Weili Guan, Jun Yu, and Min Zhang. 2025b. From bias to balance: Exploring and mitigating spatial bias in lvlms. *arXiv preprint arXiv:2509.21984*.

Overview of the Supplementary

This supplementary provides additional implementation details, quantitative results, and qualitative analyses supporting the main paper. Specifically, Sec. A studies whether directly realigning attention between open-ended and constrained framings improves grounding and inconsistency. Sec. B covers implementation details, including the cross-framing inconsistency pipeline and GPT prompts for question reframing (Sec. B.1), curation details and human evaluation of GQA^F and V^{*F} (Sec. B.2), training and resource usage (Secs. B.4 and B.5), and evaluation protocols for all benchmarks (Sec. B.6). Sec. C presents additional quantitative results, including ablation studies on learnable token placement and confidence-based loss weighting (Sec. C.1), and extended visual attention analysis across additional VLM families, Gemma3, GLM-4.1V, LLaVA-OneVision-1.5, and Qwen3VL (Sec. C.2). Sec. D provides additional qualitative results, and Sec. 7 discusses limitations and future directions.

A Attention Realignment via Prompt Tuning

Secs. 4 and 5 show that constrained framings suppress visual energy and redirect attention away from task-relevant regions, yet steering attention back recovers accuracy. Beyond natural language prompting, we use learnable prompt tokens as a controlled disgonstic probe: if the open-ended attention pattern is a useful supervisory signal, then directly realigning constrained framings toward that pattern should improve the grounding and consistency. In this section, we then study the effect of this probe.

A.1 Probe setup.

As illustrated in Fig. A1, we construct training triplets from each sample. Given an image and a question with a ground-truth answer, we use Qwen3-32B (Bai et al., 2025a) to rewrite the question into all three framings, including open-ended, yes/no, multiple-choice, along with their corresponding answers. For the yes/no and multiple-choice framings, we append K learnable tokens to the input sequence; the open-ended input remains unchanged. All three framings are fed into a frozen VLM in parallel to produce output logits and attention rollouts.

We jointly optimize two objectives. The first

one is a standard cross-entropy loss for next-token prediction, applied to the triplets, preserving the original question answering capability. The second is an attention alignment loss that encourages the constrained framings to mimic the open-ended attention pattern. Inspired by the attention distillation objective (Fuller et al., 2025), this loss operates on the attention rollouts and consists of (1) an L_2 loss on visual energy magnitude, penalizing the difference in total attention over image tokens; and (2) a KL divergence on the normalized visual attention distributions (summing to one), encouraging spatial alignment. The alignment loss is computed between the open-ended framing and both constrained framings.

A.2 Implementation and training details.

We train only the learnable tokens, keeping the VLM frozen, on 10K randomly sampled VQA pairs from the LLaVA (Liu et al., 2023) instruction tuning set². The model is trained for 1 epoch using the AdamW optimizer (Loshchilov and Hutter, 2017) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We employ a peak learning rate of $2E - 4$ with a 5% linear warmup, followed by a cosine decay schedule. We use a batch size of 1 with gradient accumulation over 16 steps. We weight each sample by the model’s own confidence in the ground-truth answer under teacher forcing. Specifically, we compute the average probability assigned to each ground-truth token at the corresponding position, yielding a continuous weight. Samples where the model is already confident in the correct answer produce attention patterns that are more likely to be reliable supervisory signals, and thus receive higher weight; samples where the model is uncertain contribute proportionally less. We used $K = 8$ learned tokens during fine-tuning. We study the effect of the loss components and the number of learnable tokens K in the ablations (see Sec. A.3).

A.3 Experiments

A.3.1 Models and benchmarks.

We evaluate this probe on five VLMs spanning diverse architectures: Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), LLaVA-OneVision-1.5-8B (An et al., 2025), Gemma3-12B (Team et al., 2025), and GLM4.1V-9B-Base (Zeng et al., 2025). For each model, the learnable tokens are trained on 10K randomly sampled

²llava-1.5-665k-instructions

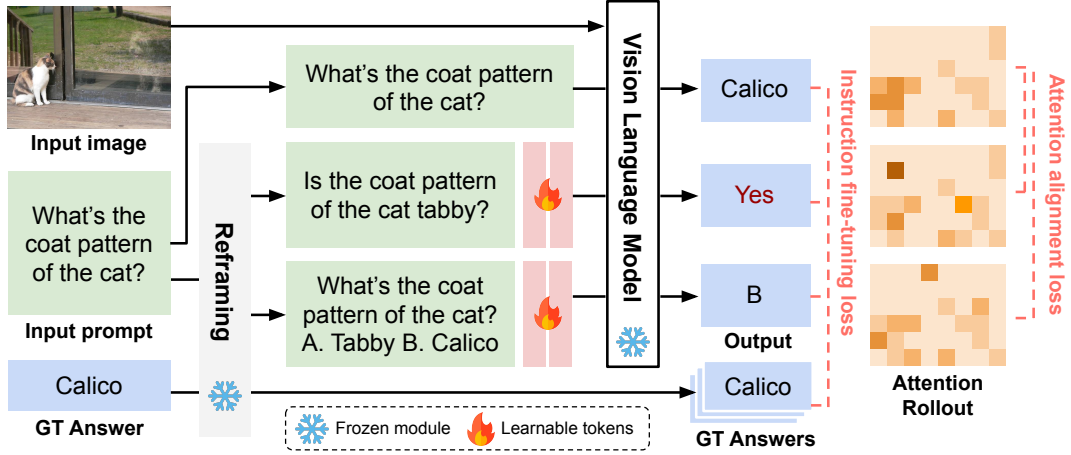


Figure A1: **Learning to re-align attention.** Given a VQA sample, we reframe the question/answer pairs into alternate framing with a frozen LLM. We append sets of learned tokens to constrained framings at the end of the input sequence. During training, we add an attention alignment objective on the attention rollout to align the attention from open-ended samples with constrained framings.

examples (with early stopping) from the LLaVA SFT training set, as described above. We evaluate downstream performance on 7 benchmarks grouped into three categories: general and reasoning (RealWorldQA[†] (xAI, 2024), MME (Fu et al., 2023), MMMU-Pro[‡] (Yue et al., 2025)), alignment (HallusionBench[†] (Guan et al., 2024), POPE (Li et al., 2023b)), and fine-grained grounding (HRBench8k (Wang et al., 2025b), V* (Wu and Xie, 2024)). We note that MME and POPE are yes/no benchmarks, while the rest are multiple-choice QAs. Since we focus on constrained framings, we note that [†] denotes the multiple-choice subset, and [‡] denotes the vision split of MMMU-Pro and HallusionBench. Also, we do not incorporate the tools used for VLMs when testing on V*. In addition, we revisit the GQA, Seedbench, GQA^F, and V^{*F} evaluation setup from Sec. 5 to measure whether realignment recovers visual attention and reduces cross-framing inconsistency.

A.3.2 Effect on visual attention and prediction consistency.

Before investigating downstream performance, we first text whether learnable tokens can restore attention under constrained framings. As shown in Fig. A2 (Left), applying the tokens on Yes/No and MCQ, visual energy under constrained framings recovers substantially and bounding box attention increases, indicating that attention shifts back to task-relevant image regions. We also revisit the inconsistency evaluation in Fig. A2 (right). The inconsistency rate drops across VLMs on GQA. On the Seedbench task breakdown using Qwen2.5-

VL-7B, inconsistency decreases by 20% on instance interaction and 15% on counting, two categories that demand fine-grained visual grounding. These results show that learned tokens can realign constrained-framing attention toward the open-ended pattern, supporting the attention-gap hypothesis.

A.3.3 Effect on downstream benchmark performance.

We now ask whether re-aligned attention translates into downstream accuracy gains. We append learned tokens to the input sequence according to the framing type of each benchmark (i.e., yes/no or MCQ). In Tab. A1, we report results across 7 benchmarks and 5 VLMs spanning 7B to 12B. Overall, realignment yields consistent improvements across most model and benchmark combinations. The gains are most pronounced on fine-grained grounding tasks: on V*, Qwen2.5-VL-7B improves by 2.5pp, while HRBench8k sees steady improvements for Gemma3 and GLM4.1V. The results match our findings that tasks demanding spatial localization benefit most from restored visual attention. General reasoning benchmarks show modest but positive gains overall.

A.3.4 Model ablation.

To test whether gains come from attention alignment rather than additional instruction tuning with cross-entropy loss, we first ablate the loss components on Qwen2.5-VL-7B. The results are shown in Tab. A2. Removing the attention alignment loss yields marginal gains over the baseline, in-

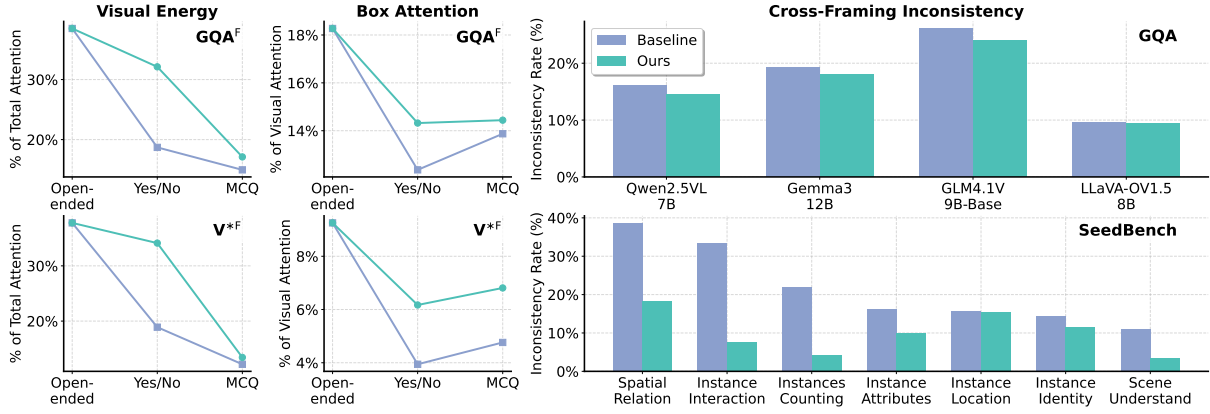


Figure A2: (Left) Performance comparison after appending the learned soft tokens on input prompts on GQA^F and V^{*F} . (Right) Cross-framing inconsistency comparison between baseline (Qwen2.5-VL-7B) and model after realignment.

Methods	General & Reasoning			Alignment		Fine-grained grounding	
	RealWorldQA [†]	MME	MMMU-Pro [‡]	Hallusion [‡]	POPE	HRBench8k	V [*]
Gemma3-12B	61.87	2133.5	27.92	64.56	84.63	48.25	48.50
+ Soft prompting	61.87	2180.0	28.09	66.14	84.98	49.50	49.79
GLM4.1V-9B-Base	69.41	2208.1	32.22	63.72	88.52	51.62	58.80
+ Soft prompting	69.41	2237.8	31.41	67.51	87.24	52.25	61.80
LLaVA-OV1.5-8B	70.09	2276.2	30.52	66.04	88.82	58.00	62.23
+ Soft prompting	69.63	2290.3	30.92	66.14	89.12	58.50	62.66
Qwen2.5VL-7B	63.70	2199.5	31.45	71.50	87.68	54.13	66.95
+ Soft prompting	66.44	2269.7	31.50	68.77	88.65	54.13	69.53
Qwen3VL-8B	70.78	2390.8	32.72	71.08	89.26	61.38	72.96
+ Soft prompting	71.92	2430.5	32.89	72.34	89.41	61.25	75.11

Table A1: Performance impact of realignment probe. [†] denotes the multiple-choice subset, and [‡] denotes the vision split.

dicating that cross-entropy loss alone is insufficient and barely recovers the performance of the baseline after adding additional learnable tokens; explicit attention realignment is the primary driver of improvement. Furthermore, removing the cross-entropy loss causes significant performance drops, as the additional learnable tokens are initialized randomly and the CE loss is essential to keep the model’s question-answering capability functioning normally. The full model performs best across most metrics.

B Implementation Details

B.1 Cross-Framing Inconsistency

As discussed in Sec. 4, we leverage cross-framing inconsistency to verify whether question framing affects the model’s final prediction ($F \rightarrow Y$ in Fig. 2), and the evaluation pipeline is illustrated in Fig. 3 (left). Here, we discuss more implementa-

tion details. For data curation, SeedBench (Li et al., 2023a) is a MCQ benchmark, so we remove the possible answer choices to obtain the open-ended question and use the content of the correct option as the ground truth answer. GQA (Hudson and Manning, 2019) is largely open-ended, but a small portion of it consists of Yes/No questions. We filter out the Yes/No questions in GQA to obtain a fully open-ended generation benchmark. We then perform inference on both datasets across different VLMs and evaluate the performance. For samples which were answered correctly, we use GPT-5.1 (gpt-5.1-2025-11-13) to reframe the questions into Yes/No or MCQ (while for SeedBench we use the original options), and perform inference on the questions in different framings to get the final inconsistency rate. The detailed system prompt for GPT-5.1 is provided below. We study the quality of question reframing using GPT in the following

Ablation	RWQA [†]	MMMU-Pro [‡]	POPE	V*	Avg
Baseline	63.70	31.45	87.68	66.95	62.45
Ours	66.44	31.50	88.65	69.53	64.03
w/o Attn Loss	64.84	32.89	87.05	65.67	62.61
w/o CE Loss	64.61	31.27	83.40	63.95	60.81

Tokens	RWQA [†]	MMMU-Pro [‡]	POPE	V*	Avg
4	63.70	31.33	88.43	66.09	62.39
8	66.44	31.50	88.65	69.53	64.03
12	63.93	31.50	88.36	67.81	62.90
16	63.93	30.92	88.66	67.38	62.72

Table A2: Ablation on loss functions and learnable token count. RWQA denotes RealWorldQA. [†] denotes the multiple-choice subset, and [‡] denotes the vision split.

data curation section.

Prompt for Reframing Open-ended Question to MC-Q/YN questions

Given the following question and its correct answer, create TWO new framings:

Original Question: {original_question}

Correct Answer: {correct_answer}

[If options exist:] Available Options: {options}

1. Yes/No:

- *[If affirm:]* Convert to a binary yes/no question where the answer is ‘yes’ (affirming that the answer is {correct_answer}).
- *[If negate:]* Convert to a binary yes/no question where the answer is ‘no’. Replace the correct answer ({correct_answer}) with a distinctly incorrect or unlikely alternative that is clearly contradicted by the original answer. Avoid any ambiguous or ‘near-miss’ alternatives. Keep the same target object that the original question refers to if possible — do not change the subject or object being asked about.
- Turn the original question into a yes/no format that tests the same knowledge. For example, if the original question is “What is the person near the garbage bin wearing?” with answer “a coat”, the yes/no question could be “Is the person near the garbage bin wearing a coat?”
- Must be a binary question (Is/Are/Does/Do/Can/Could/etc.).

2. MCQ (Multiple Choice Question):

- *[If options exist:]* Create an MCQ version with the same options: {options}. The correct answer should be {correct_answer}.
- *[If no options:]* Create an MCQ version with 4 options. The correct answer should be {correct_answer}, and provide 3 easy negative distractors.
- *[If scene graph available:]* Avoid using objects from the scene graph as negative options as they may create ambiguity. Scene graph: {scene_graph}.
- Should be same open-ended question (What/Which/Where/Who/How/etc.).
- Provide exactly 4 options as a list

IMPORTANT RULES:

- Yes/No question must test the SAME knowledge as the original
- MCQ must test the SAME knowledge as the original
- Both should be answerable from the same visual information

Output ONLY valid JSON in this exact format:

```
{
```

```
"yes_no": {
  "question": "Is/Are/Does/Do... question",
  "answer": "{yes|no}"
},
"mcq": {
  "question": "What/Which/Where... question",
  "options": ["option1", "option2",
    "option3", "option4"],
  "answer_text": "{correct_answer}"
}
}
```

Prompt for Reframing YN Question to Open-ended/MCQ questions

Given the following yes/no question and answer, create TWO new framings:

Original Question: {original_question}

Correct Answer: {original_answer}

Correct Full Answer (for reference): {original_full_answer}

1. MCQ (Multiple Choice Question):

- Convert to a WH-question (What/Which/Where/Who/How/etc.) that asks for identification
- Provide 4 plausible answer options as a list (the correct answer should be one of them and the answer should not be yes/no)
- Example: “What color is the helmet?” with options [“light blue”, “red”, “yellow”, “green”]
- *[If scene graph available:]* Avoid using objects from the scene graph as negative options as they may create ambiguity. Scene graph: {scene_graph}.

2. Open-ended (short answer):

- Convert to a WH-question (What/Which/Where/Who/How/etc.) expecting a brief specific answer
- Should ask the same thing as MCQ but without providing options
- Example: “What color is the helmet in the middle?” → Answer: “light blue”

IMPORTANT RULES:

- MCQ and Open-ended should use WH-questions (What/Which/Where/Who/How/When/Why) and the answer should not be yes/no
- Both formats should test the SAME knowledge as the original question

Output ONLY valid JSON in this exact format:

```
{
  "mcq": {
    "question": "What/Which/Where... question",
    "options": ["option1", "option2",
      "option3", "option4"],
```

```

    "answer_text": "<correct answer>"
  },
  "open_ended": {
    "question": "What/Which/Where... question",
    "answer": "<correct answer>"
  }
}

```

B.2 Curation Details of GQA^F, and V^{*F}

As mentioned in Sec. 5, to isolate the impact of task framing from variations in question content, we curate two datasets where each sample has three distinct framing variants: open-ended, Yes/No, and MCQ. We ensure that the underlying visual reasoning required remains constant while only the output format changes. Again, we utilize GPT-5.1 to rephrase the original samples into these target formats. Now, we discuss how we curate datasets including GQA^F and V^{*F} in detail and use human evaluation to further verify their quality.

GQA contains both open-ended and yes/no questions. We leverage the ground-truth scene graph when prompting GPT to ensure that generated MCQ distractors are reasonable without requiring visual access. We provide the prompt that we used for reframing.

For V^{*}, a multiple-choice benchmark, we remove the answer choices to obtain a purely open-ended question, and then follow the same protocol as used in GQA to reframe it into a Yes/No question.

Our reframing pipeline avoids a bias toward “yes” in the synthesized binary questions by construction: rather than letting the LLM freely generate Yes/No questions, we first sample the target answer (“yes” or “no”) via rule-based random reports the resulting label distributions, which are close to balanced sampling, and then, conditioned on this label, we prompt the LLM to produce the corresponding affirmed or negated question. Table A3 on all three datasets.

After prompting and reframing, we filter out samples not in the correct JSON format. We curate a final dataset of 10k unique semantic queries for GQA and the full 300 samples for V^{*}. With three framing variants per query, we obtain 30k and 900 samples for GQA and V^{*}, respectively.

B.2.1 Human evaluation

To verify the quality of the reframed questions and answers, we conduct a human evaluation, with results shown in Sec. B.2.1. Recall that the dataset

Dataset	Yes (%)	No (%)
SEEDBench	48.66	51.34
GQA ^F	52.04	47.96
V ^{*F}	51.26	48.74

Table A3: Distribution of ground truth answers in our Yes/No reframed questions. Labels are near-balanced across all datasets.

reframing process covers three settings: (1) open-ended → MCQ and Yes/No, (2) Yes/No → open-ended and MCQ, and (3) MCQ → open-ended and Yes/No. Since an MCQ without its answer options is naturally open-ended, we handle this direction via a simple rule-based method that strips the options from the question. The resulting open-ended question can then be further converted to Yes/No using the same pipeline. This reduces the number of conversion types that require evaluation to four: open-ended → Yes/No, open-ended → MCQ, Yes/No → open-ended, and Yes/No → MCQ.

We now describe the human evaluation criteria for each of the four conversions. The full human evaluation interface is shown in Fig. A3.

Open-ended → Yes/No. Annotators assess both answer correctness and *spatial consistency*—whether the target object or region referred to in the question remains the same after reframing. For instance, if an open-ended question asks “What color is the cake?” but the reframed Yes/No question asks “Is the candy purple?”, the target object has shifted, which violates our requirement that the core concept of the question be preserved across formats.

Open-ended → MCQ. Annotators assess answer correctness and additionally evaluate *option exclusivity*—whether the designated correct option is unambiguously the best choice among all provided options. This is necessary because an LLM may inadvertently generate multiple plausible correct answers, introducing ambiguity.

Yes/No → Open-ended and Yes/No → MCQ. For both conversions, annotators assess answer correctness. For MCQ, we also evaluate option exclusivity as aforementioned.

The paired human evaluation results are shown in Sec. B.2.1. We collect 100 samples with corresponding human feedback for each pair (500 samples in total). We use the CloudResearch platform for human studies, with a total of 197 human evaluators involved, coming from 6 countries, including USA, UK, Ireland, Australia, New Zealand, and

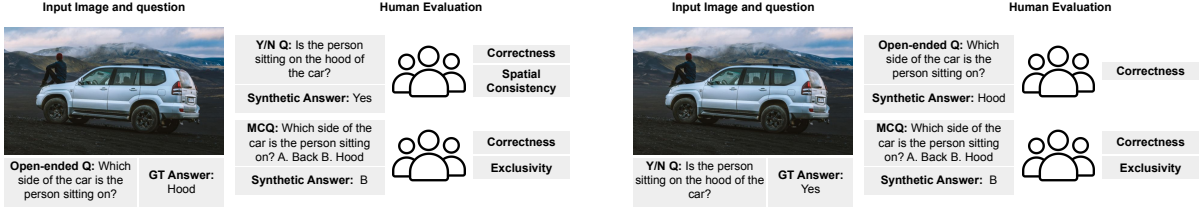


Figure A3: Human evaluation pipeline for open-ended and Yes/No question reframing. Note that for MCQ question reframing, we can use a rule-based approach (simply remove the options) to convert it into open-ended and then perform reframing to Yes/No.

From ↓ / To →	Open-ended	Yes/No		MCQ	
	Correctness	Correctness	Spatial Cons.	Correctness	Exclusivity
Open-ended	—	96.5	96.9	94.8	95.7
Yes/No	92.2	—	—	94.8	93.7
MCQ	<i>rule-based</i>	92.6	95.4	—	—

Table A4: Human evaluation results for question reframing across different conversion types. MCQ → Open-ended is handled via rule-based conversion and requires no human evaluation.

Canada, whose native language is English. Across all conversion directions, correctness scores remain consistently high, demonstrating that the reframing process is robust. Beyond correctness, we evaluate two format-specific properties: *spatial consistency*, which measures whether target object or task-relevant region remains unchanged after reframing even the generated answer is correct, and *exclusivity*, which measures whether the correct answer in a reframed MCQ question remains unambiguously correct (i.e., the correct option is the best valid option be considered). Both properties are well-maintained, with spatial consistency scores of 96.9% and 95.4% and exclusivity scores of 93.7% and 95.7%.

	Trainable Parameters	Time (min)	VRAM (peak)	VIRT (peak)
Qwen2.5VL-7B	57 K	187	33.8 GB	53.1 GB
Qwen3VL-8B	66 K	190	42.7 GB	65.9 GB
Gemma3-12B	61 K	380	39.6 GB	59.7 GB
GLM-4.1V-8B-Base	66 K	332	76.9 GB	116.1 GB
LLaVA-OneVision-1.5-8B	66 K	259	36.5 GB	54.6 GB

Table A5: **Resource usage statistics for prompt tuning.** Tested on 1 NVIDIA L40 with batch size 1 and gradient accumulation 16. VIRT stands for virtual memory size. GLM-4.1V-8B-Base is tested on 1 NVIDIA H100.

B.3 Prompting Details

For the prompting analysis in Sec. 6, we evaluate several prompting families. Each family contains multiple paraphrased instruction variants, while

sharing the same output constraint within each answer format. For Yes/No questions, the final answer must be *yes* or *no*; for MCQ, it must be one option letter; and for open-ended questions, it must be a few words or a short phrase. For reasoning-based prompts, the model is required to write additional few lines, followed by the final answer. We provide prompt details for each family in Tab. A9.

B.4 Training

Following the discussion in Sec. A, we fine-tune the VLMs on 10K randomly sampled VQA pairs from the LLaVA (Liu et al., 2023) instruction tuning set. The model is trained for 1 epoch using the AdamW optimizer (Loshchilov and Hutter, 2017) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We employ a peak learning rate of 2×10^{-4} with a 5% linear warmup, followed by a cosine decay schedule. We use a batch size of 1 with gradient accumulation over 16 steps. All models, except for GLM, are trained on a single NVIDIA L40. GLM is trained on a single 80GB H100-SXM.

We weight each sample by the model’s own confidence in the ground-truth answer under teacher forcing (Vaswani et al., 2017). Specifically, we compute the average probability assigned to each ground-truth token at the corresponding position, yielding a continuous weight. Samples where the model is already confident in the correct answer produce attention patterns that are more likely to be reliable supervisory signals, and thus receive higher weight; samples where the model is uncertain contribute proportionally less. During training, we jointly apply cross-entropy loss and attention alignment loss. Since the ranges of attention alignment from both mass and distribution are relatively small, we scale the attention loss by a factor of 5 to match the range of the cross-entropy loss. During training, the maximum image size is limited to 728 (longest side) for efficiency while the aspect ratio remains unchanged. For the reframing module in

the training pipeline, we use Qwen3-32B locally. Gradient checkpointing is disabled for all models since the gradient backpropagation on the attention map may be incorrect when the gradient checkpointing is on in PyTorch. Training is done with bf16 and FlashAttention (Shah et al., 2024). We use $K = 8$ learned tokens during fine-tuning.

B.5 Resource Usage

We report the training cost and parameter analysis of our methods in Tab. A5. We are only training the learnable tokens for each framing. With $K = 8$, the trainable parameters are relatively low at around 60K. We train all models using one epoch with early-stopping applied, and the training time cost from smallest model at around three hours to largest model to six hours on one single GPU (L40 or H100). VRAM usage is high as the gradient graph are stored backpropagation back to input embedding space.

B.6 Evaluation

We now discuss the benchmarks and the corresponding evaluation protocols for both main paper and additional probing analysis in the appendix.

B.6.1 GQA

(Hudson and Manning, 2019) is a general visual question answering benchmark containing open-ended and yes/no questions. When evaluating the score, we follow the official script and perform simple string matching for both yes/no and open-ended questions.

B.6.2 SeedBench

(Li et al., 2023a) is a comprehensive benchmark containing video question answering, single-image, and multi-image question answering tasks. In this paper, we focus on the single-image task and subsample the single-image portion accordingly. For evaluation, SeedBench is MCQ-type, and we follow the official setting using an answer ranking strategy to obtain the final predicted option. Specifically, we compute the likelihood of each option instead of appending an instruction at the end of the question asking the model to return the letter of the correct option. This evaluation method disentangles the model’s instruction-following capability from its ability to answer the given questions.

B.6.3 RealWorldQA

(xAI, 2024) is released by xAI alongside Grok-1.5 Vision and evaluates basic real-world spatial

understanding. Each question is appended with options and an instruction for controlling the output format. RealWorldQA is a mixture of MCQ and open-ended questions. As we are testing the improvement of our soft tokens on yes/no and MCQ questions while leaving open-ended question inference untouched, we use only the MCQ portion of the benchmark. We directly use the official questions as input and apply string matching between the ground-truth letter and the predicted letter.

B.6.4 MME

(Fu et al., 2023) is a VQA benchmark in yes/no question format designed to test the hallucination of LLMs. We follow VLMEvalKit (Duan et al., 2024) and append a short instruction “please return yes or no.” to constrain the output of VLMs, then perform string matching for the final performance score.

B.6.5 MMMU-Pro

(Yue et al., 2025) is an enhanced benchmark for VLMs designed to assess true understanding capabilities across multiple modalities. This benchmark provides VQA samples in three formats: standard textual questions with 4 options, standard textual questions with 10 options, and a purely vision-based format where the question and image are presented together in a screenshot. We evaluate our models on the vision split of the benchmark, where the entire question is posed in image format. Since the entire benchmark is MCQ-framing, we append a short instruction following VLMEvalKit to enforce the output to be the letter of the correct option, and perform string matching for the final performance.

B.6.6 HallusionBench

(Guan et al., 2024) is, similar to MME, a VQA benchmark in yes/no question format. This benchmark contains two parts: VQA and text QA. We focus on the VQA split, and similar to MME, a short instruction is appended at the end of each question before passing it to VLMs.

B.6.7 POPE

(Li et al., 2023b) is a yes/no benchmark for testing the hallucination of VLMs. Similar to MME, we append an instruction and perform string matching for the final performance.

Position	General & Reasoning			Alignment		Fine-grained grounding	
	RealWorldQA [†]	MME	MMMU-Pro [‡]	Hallusion [‡]	POPE	HRBench8k	V*
Prefix	61.64	2272.2	28.15	65.51	88.28	51.62	66.52
Infix	66.44	2269.7	31.50	68.77	88.65	54.13	69.53
Postfix	63.47	2262.4	31.50	65.83	88.57	54.25	67.38

Table A6: Ablation study on learnable token positioning. [†] denotes the multiple-choice subset, and [‡] denotes the vision split.

Strategy	General & Reasoning			Alignment		Fine-grained grounding	
	RealWorldQA [†]	MME	MMMU-Pro [‡]	Hallusion [‡]	POPE	HRBench8k	V*
Baseline	63.70	2199.5	31.45	71.50	87.68	54.13	66.95
Confidence weighting	66.44	2269.7	31.50	68.77	88.65	54.13	69.53
Equal weighting	64.84	2205.6	30.25	65.51	89.09	52.50	68.67

Table A7: Ablation study on learnable token weighting strategies. [†] denotes the multiple-choice subset, and [‡] denotes the vision split.

Complexity group	Mean # segments	Inconsistency (%)
1 (lowest)	10.08	17.39
2	19.95	17.99
3	28.46	21.34
4	39.17	22.14
5 (highest)	65.75	26.01

Table A8: Cross-framing inconsistency breakdown by visual scene complexity. Images are grouped into five bins by SAM2 segment count; inconsistency rises consistently with complexity.

B.6.8 HRBench8k (Wang et al., 2025b) and V* (Wu and Xie, 2024)

are both VQA benchmarks featuring high-resolution image inputs. The target objects or task-relevant regions are relatively small, thus requiring strong grounding skills for VLMs to answer correctly. HRBench8k contains 800 VQA pairs and V* has around 300 samples; we include both the OCR and GPT-hard splits of V*.

C Additional Quantitative Results

C.1 Ablation Study

C.1.1 Position of the Learnable Tokens

As mentioned in Sec. A, the realignment probe appends learnable tokens to yes/no and MCQ questions to enforce VLMs to attend to the image after the attention alignment training. We ablate the impact of the position of the learnable tokens. Recall that given a VQA sample, VLMs process the input image and question into a sequence of tokens ordered as: image tokens, question tokens, and instruction tokens. Within this sequence, there are three positions to place our learnable tokens: (1)

prefix: between image and question tokens, (2) *infix*: between question and instruction tokens, and (3) *postfix*: after the instruction tokens at the end of the sequence. We conduct an ablation study to analyze the impact of the position, with results shown in Tab. A6. One can see that among the three positions, infix achieves the best overall performance. This supports our observation that visual attention variation is primarily driven by question framing: placing learnable tokens immediately after the question lets the probe learn framing-conditioned adjustments while remaining close to the tokens that trigger the attention shift.

C.1.2 Weighting Strategy for Alignment Loss Across Samples

For the realignment probe, we also ablate how open-ended attention maps are selected as supervisory signals. Since some open-ended questions may be answered incorrectly, their attention maps can be noisy target for constrained framings. During training, we use ground-truth labels to assess correctness and apply either a hard threshold or a soft threshold, where the latter reflects the model’s confidence in its own output and thus potentially yields more reliable attention maps. As shown in Tab. A7, confidence weighting consistently improves performance across benchmarks, suggesting that reliable open-ended attention patterns provide a cleaner signal for testing attention realignment.

C.2 Visual Attention Analysis for Other VLMs

In Sec. 5, we mainly focus on the behavioral study of Qwen2.5VL-7B. Here we further provide analy-

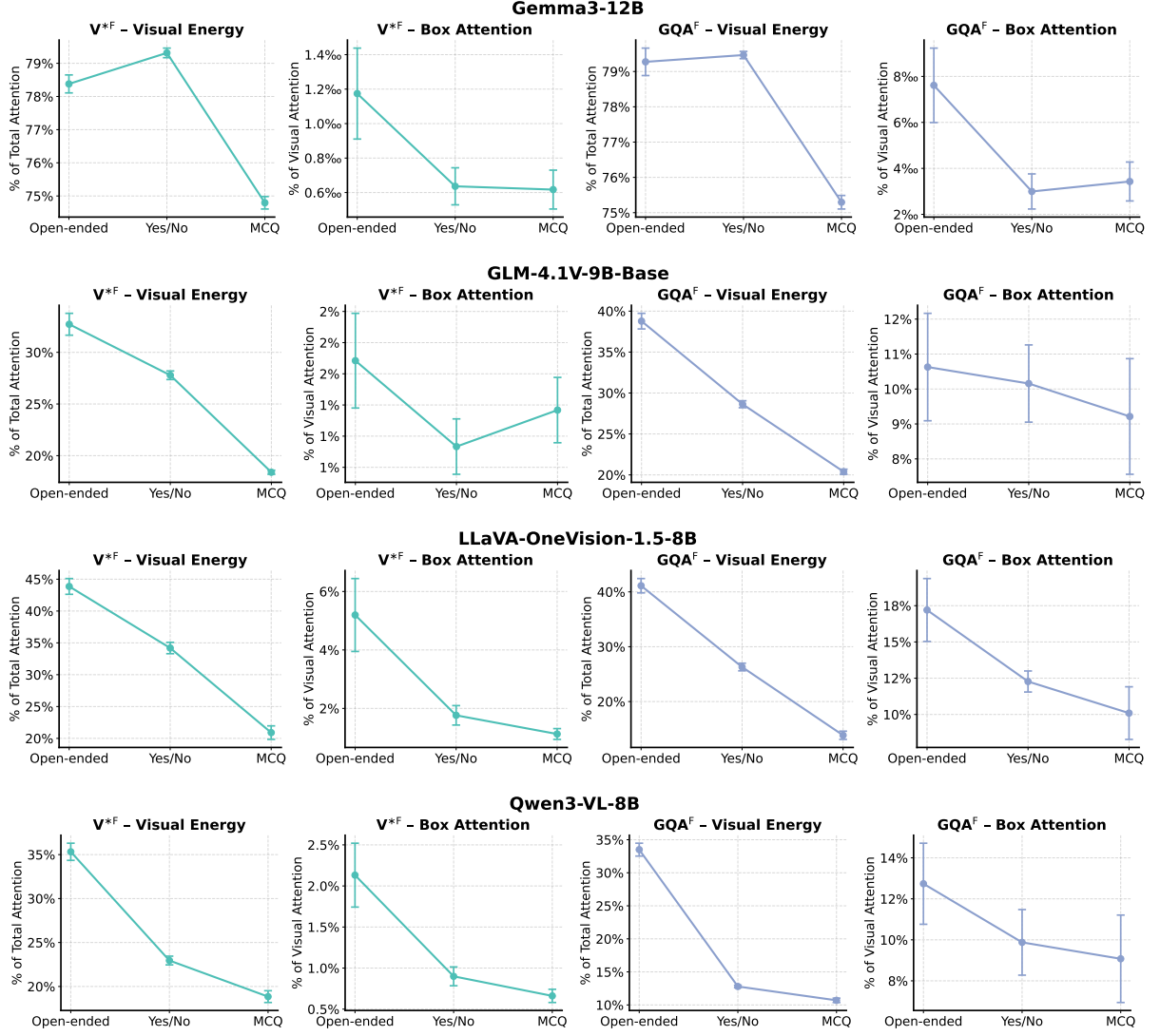


Figure A4: Visual energy drops significantly on non-open-ended framings.

sis on multiple models, including Gemma3 (Team et al., 2025), GLM-4.1V (Zeng et al., 2025), LLaVA-OneVision-1.5 (An et al., 2025), and Qwen3VL (Bai et al., 2025a). The visual attention analysis and layer-wise analysis results are shown in Fig. A4 and Fig. A5, respectively. One can see that the overall visual attention behavior across framings is similar to Qwen2.5VL-7B in the main paper on V^*F and GQA^F , with higher bounding box attention for open-ended questions and lower attention for MCQ and yes/no questions. On Gemma3, while the bounding box attention trends are similar to our previous observation, we see that yes/no questions can potentially exhibit higher visual energy than open-ended question framings. The same behavior can be observed on GQA as well.

C.3 Discussion about Chance-Corrected Cross-Framing Inconsistency

Since constrained formats can sometimes be answered correctly by guessing, we verify that our inconsistency metric is not explained away by chance correctness. In our setting, constrained correctness requires that both the Yes/No and the MCQ (4 options) versions are answered correctly, so the chance baseline under uniform random guessing is $g = 0.5 \times 0.25 = 0.125$. Recall that our raw inconsistency is

$$I = P(C = 0 \mid O = 1) = 1 - P(C = 1 \mid O = 1) = 0.1926, \quad (3)$$

where $O = 1$ denotes open-ended correctness and $C = 1$ denotes correctness on both constrained formats. We compute the chance-corrected constrained accuracy as

$$A_{cc} = \frac{P(C = 1 \mid O = 1) - g}{1 - g} = 0.7799, \quad (4)$$

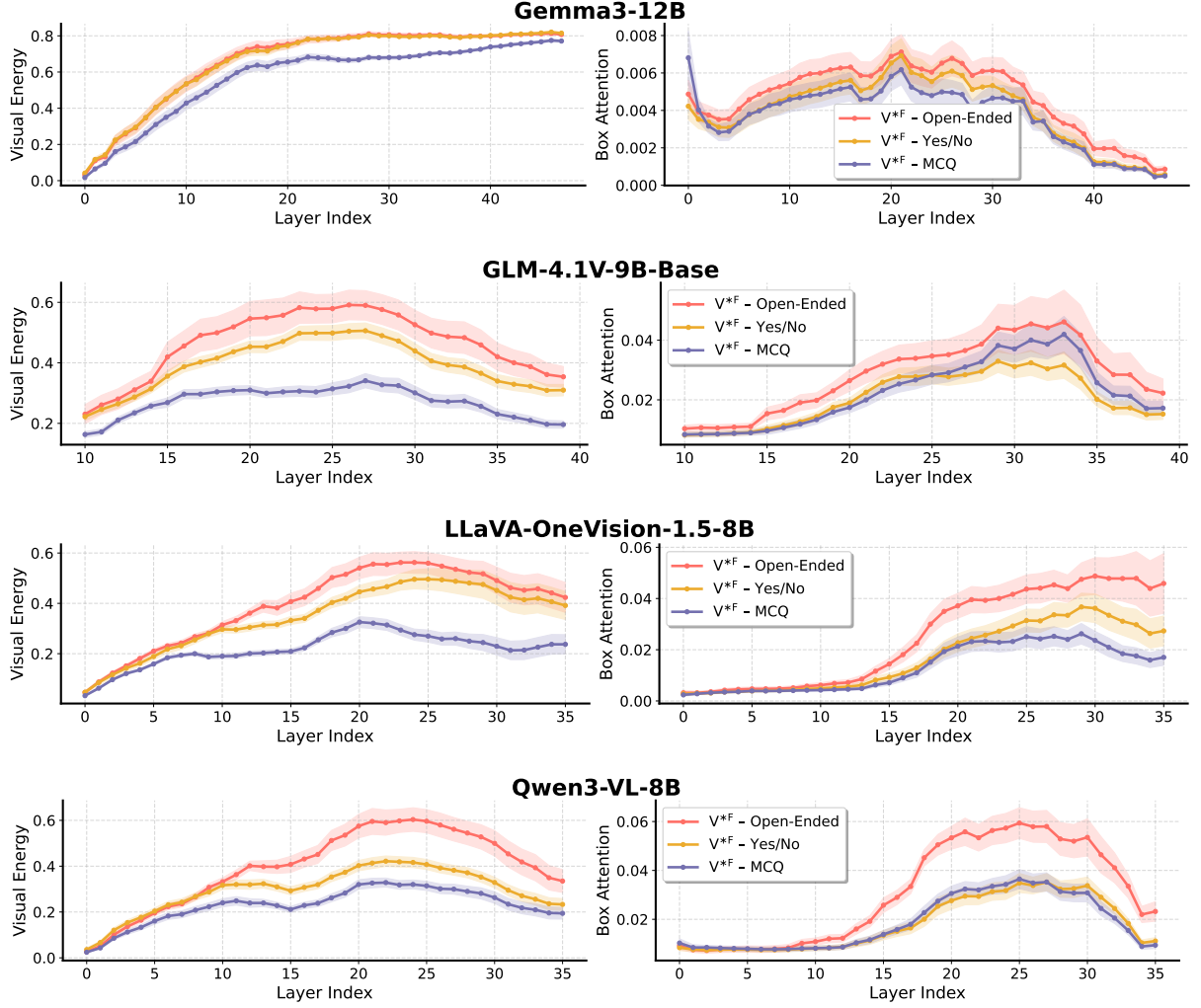


Figure A5: Visual energy drops significantly on non-open-ended framings.

yielding a chance-corrected inconsistency of $I_{cc} = 1 - A_{cc} = 0.2201$. The inconsistency rate thus increases from 19.26% to 22.01% after chance correction, showing that accounting for guessing does not remove the effect; the estimated inconsistency becomes slightly larger. We also note that label distribution bias is controlled by construction: our reframing pipeline samples the target Yes/No label via rule-based random sampling before generation, and the resulting label distributions are close to balanced (Tab. A3).

C.4 Effect of Visual Scene Complexity

Recall that the task-level breakdown on SeedBench in Fig. 3 suggests that tasks involving multiple objects exhibit higher cross-framing inconsistency than simpler object-grounding tasks, hinting that visual scene complexity impact the framing effect. To further test this at the image level, we use SAM2 (Ravi et al., 2025) to segment each image and treat the number of segments as a proxy for visual complexity, capturing object count and clutter.

We sort samples by segment count, split them into five equally sized groups, and compute the cross-framing inconsistency rate within each group. As shown in Tab. A8, inconsistency increases monotonically with visual complexity, rising from 17.39% in the lowest-complexity group (10.1 segments) to 26.01% in the highest-complexity group (65.8 segments). This confirms that scenes with more objects or clutter are more vulnerable to framing-induced failures, consistent with our finding in Sec. 4 that constrained framing is most damaging where precise visual grounding matters most: as scenes grow more complex, the attention degradation induced by constrained framings leaves the model with less capacity to locate and reason about the relevant visual evidence.

D Dataset Examples and Qualitative Results

We first provide some data samples of open-ended and Yes/No question reframing using GPT-5.1 in Fig. A6 and Fig. A7, respectively. Data are ran-

Original Questions and Answers		Reframed Questions and Answers	
	Open-ended Q: On which side is the tap? A: left	Y/N Q: Is the tap on the right? A: No	MCQ: On which side is the tap located? A. left B. right C. center D. both sides Ans: A
	Open-ended Q: What is the material of the chair? A: leather	Y/N Q: Is the chair made of leather? A: Yes	MCQ: What material is the chair made of? A. leather B. fabric C. wood D. plastic Ans: A
	Open-ended Q: Is the black umbrella closed or is it open? A: open	Y/N Q: Is the black umbrella closed? A: No	MCQ: What is the condition of the black umbrella? A. closed B. open C. partially open D. collapsed Ans: B

Figure A6: Qualitative examples of open-ended question reframing.

Original Questions and Answers		Reframed Questions and Answers	
	Y/N Q: Is the man to the right of the bike standing on a ski? A: no	Open-ended Q: What is the man to the right of the bike standing on? A: skateboard	MCQ: What is the man to the right of the bike standing on? A. a skateboard B. a ski C. a surfboard D. a stool Ans: A
	Y/N Q: Is the traffic cone to the left of the skateboard made of metal? A: no	Open-ended Q: What material is the traffic cone to the left of the skateboard made of? A: rubber	MCQ: What material is the traffic cone to the left of the skateboard made of? A. metal B. plastic C. rubber D. wood Ans: C
	Y/N Q: Is the cup on the right? A: yes	Open-ended Q: Where is the cup located in the image? A: on the right	MCQ: Where is the cup located in the image? A. on the left B. on the right C. in the center D. in the foreground Ans: B

Figure A7: Qualitative examples of Yes/No question reframing.

domly sampled from GQA^F; we showcase the original question in its open-ended or Yes/No framing, which is subsequently reframed into the other two formats.

Additionally, we provide a qualitative comparison of inference with and without our learned tokens in Fig. A8. Given a VQA sample, we perform inference using a Qwen2.5VL-7B baseline alongside our approach with extra learned tokens, visualizing the attention rollout during output generation. As shown, the attention maps from our approach focus significantly more on task-relevant regions, whereas the baseline often spreads to irrelevant background areas. The Diff map highlights the percentage-wise change, further demonstrating that our learned tokens effectively shift attention toward the target area during inference.

E Human Evaluation and Consent

We conduct a human evaluation to verify the quality of GPT-generated question reformulations used in our framing-controlled datasets. Participants are recruited through CloudResearch from English-speaking countries, including the United States, Australia, New Zealand, Canada, the United Kingdom, and Ireland. Eligibility is restricted to adults aged 18–50 who report English as their native language and have at least a college-level education sufficient to understand the task. Participants are informed before the task that they will review GPT-generated reformulations and evaluate whether they preserve the intended semantics of the original question and remain answerable from the image. They are also informed that their responses will be used only for this research project and paper, not for any production, and that personal information and individual response details will be kept private.

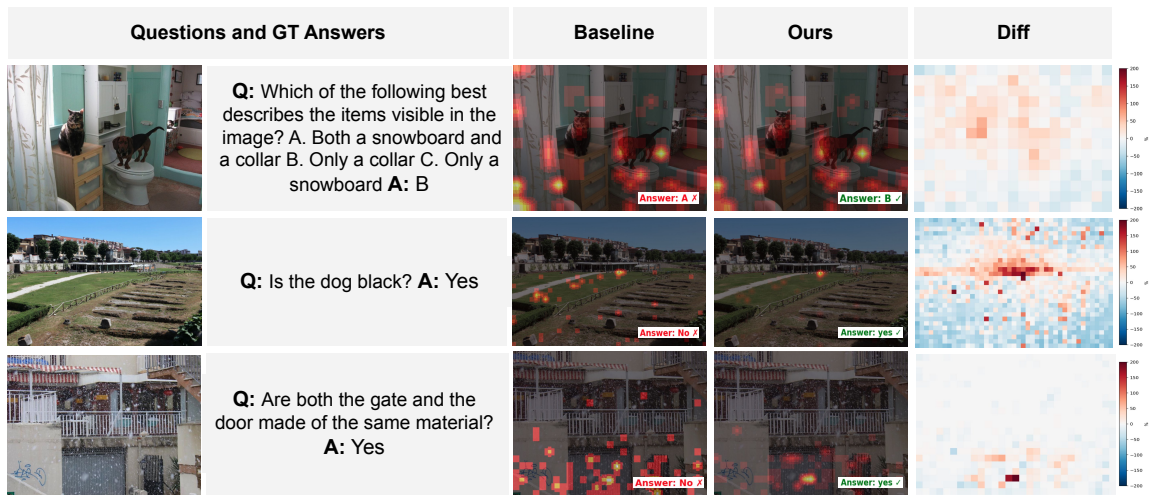


Figure A8: Qualitative comparison of Baseline and Ours (Qwen2.5-VL-7B) .

Recruitment and payment (at least 10 USD per hour based on location) are handled by CloudResearch according to its platform policies and applicable local regulations. A screenshot of the evaluation interface is provided in Fig. A9.

F LLM usage

We use LLM Tools such as Codex for only grammar checking and minor writing polish. We use Codex to assist with code debugging, implementation based on examples from human, and writing visualization code. All the experimental design, project ideas, codex review, model training, and result interpretation are conducted manually by the authors.

An AI model reframed a Yes/No question into Open-Ended and Multiple Choice formats. Please verify the quality of each reframing. First review the original Yes/No question and answer as context, then evaluate each reframed version below. You are free to zoom in on the image to find details that may help.



Original Yes/No Question: Does the computer mouse on the mouse pad have black color?

Original Answer: yes

Reframed: Open-Ended

Open-Ended Question: What color is the computer mouse on the mouse pad?

Open-Ended Answer: black

1. Did the model correctly reframe the question and answer into Open-Ended format?

- ☐ Yes
☐ No
☐ Cannot tell / Ambiguous

An AI model reframed a Yes/No question into Open-Ended and Multiple Choice formats. Please verify the quality of each reframing. First review the original Yes/No question and answer as context, then evaluate each reframed version below. You are free to zoom in on the image to find details that may help.



Original Yes/No Question: Does the computer mouse on the mouse pad have black color?

Original Answer: yes

Reframed: Multiple Choice

MCQ Question: What color is the computer mouse on the mouse pad?

Options:

- A. black
B. white
C. gray
D. blue

Generated Answer: **A**

2. Did the model correctly reframe the question and answer into Multiple Choice format?

- ☐ Yes
☐ No
☐ Cannot tell / Ambiguous

3. Is the MCQ answer the best/only correct option (exclusivity)?

- ☐ Yes — it is the best correct option
☐ No — another option is also correct and better
☐ Cannot tell / Ambiguous

Figure A9: Screenshot of an example human evaluation form hosted on CloudResearch.

Family	Instruction variants	Shared output format
Image-focused instruction	Carefully examine the image before answering the question. Look closely at the image before answering the question. Inspect the visual content carefully before you answer. Study the image first, then answer the question. Use the image as evidence before giving your answer. Review the image attentively before responding. First focus on the relevant visual details, then answer. Base your answer on a careful reading of the image. Before answering, check the image carefully for the needed evidence. Look at the image carefully before generating outputs.	Direct answer only.
Visual prompting	Target regions are highlighted with 10 visual styles: rectangular outlines in red, blue, green, yellow, and magenta; elliptical outlines in red and blue; thicker red rectangular outlines; larger-margin red rectangular outlines; and cyan rectangular outlines with both thicker lines and larger margins.	Direct answer only.
Thinking	Write your reasoning freely based on the image. Explain your visual reasoning in your own words. Describe the image evidence you use to reach the answer. Reason from the visible content before giving the final answer. Use a short visual rationale before the final answer. State the relevant observations from the image before answering. Think through the visual evidence in prose before the answer. Briefly explain how the image supports your answer. Give image-grounded reasoning before the final answer. Describe what you see that determines the answer.	Use <think>...</think> for reasoning and <answer>...</answer> for the final answer.
Chain-of-thought	Think step by step about the visual evidence before giving the final answer. Reason step by step from the image before giving the final answer. Break down the relevant visual clues step by step before answering. Use a step-by-step visual explanation before the final answer. List the key reasoning steps from the image before answering. Work through the image evidence in steps before the final answer. Analyze the visual details step by step before giving the answer. Explain the visual evidence in ordered steps, then answer. First identify the relevant image clues step by step, then answer. Step through the visual reasoning needed to answer the question.	Provide step-by-step reasoning, then write ANSWER on a separate line and place the final answer on the last line.
Grounding	Answer the question using grounded visual evidence. Answer using only evidence that can be localized in the image. Ground your answer in the visible image regions relevant to the question. Use localized visual evidence to support your answer. Base the answer on specific visible regions in the image. Identify the necessary visual evidence before answering. Support the answer with localized image evidence. Anchor your reasoning to relevant visible objects or regions. Use concrete image regions as evidence for the answer. Provide visually grounded evidence before the answer.	First list only relevant visual evidence with bbox_2d=[x1,y1,x2,y2] and labels, then give a brief grounded explanation, write ANSWER, and place the final answer last.

Table A9: Prompt families used in the prompting analysis. Each family contains paraphrased instruction variants while sharing the same output-format constraint.