LOST IN THE ZOOM: DIAGNOSING AND REBALANCING EVIDENCE USE IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) with zoom-in tools improve visual question answering, yet acquiring useful visual evidence does not ensure correct reasoning over it. To understand this, we propose a new protocol to analyze tool-use trajectories and reasoning. The protocol separates trace curation, tool execution, evidence verification, and answer evaluation, then reuses trajectories for continuation comparisons. Applied to VLMs with zoom-in tools across six datasets, the protocol yields the *EvidenceZoom* benchmark, comprising 35.4K trajectories covering 3.9K questions. Attention diagnosis over trajectories reveals that reasoning consistently dominates attention allocation and that zoomed-in images are overlooked, with attention head patterns showing substantial variation in their focus within the zoomed-in region. Based on these findings, we propose *Head-Gated Attention Reallocation*, a training-free method that corrects the zoom–reasoning attention imbalance, scaling each head's correction by its relative within-zoom focus. Across five VLMs from the GLM, Qwen, and Gemma series, our method improves the six-dataset mean accuracy over the best training-free baseline and even surpasses the training-based method on Qwen3-VL-8B. To summarize, our work highlights the gap between acquiring visual evidence and using it effectively, and suggests attention reallocation as a promising direction for improving visual reasoning with zoom-in tools.

1 INTRODUCTION

Vision-language models (VLMs) with zoom-in tools aim to improve visual question answering on images by acquiring enlarged views of small details at a higher effective resolution and incorporating the returned evidence into their reasoning (Wu & Xie, 2024; Khayatkhoei et al., 2025). However, we observe that even when a model localizes the relevant object and obtains useful visual evidence through zooming, it can still reach an incorrect answer (Figure 1(a)). Prior work similarly finds that useful visual representations can remain underused (Fu et al., 2025; Liu et al., 2025a), recognition and reasoning can fail to compose (Wang et al., 2026a), and generated text can compound errors despite available visual input (Huang et al., 2024; Liu et al., 2025b). Ma et al. (2026) further observe that models often answer incorrectly even when their zoomed-in images contain the relevant visual information. Despite these diagnostic advances, evaluation of VLMs with zoom-in tools commonly summarizes performance through end-to-end answer accuracy (Wu & Xie, 2024; Khayatkhoei et al., 2025; Wei et al., 2026). Answer accuracy, however, cannot distinguish between failures to acquire evidence and failures to utilize it.

To systematically evaluate and diagnose reasoning after the model has accepted tool-returned evidence as useful, we introduce the *EvidenceZoom* benchmark, built from curated tool-use trajectories for evaluating reasoning over self-verified evidence. Its construction protocol separates task curation, tool execution, evidence verification, and answer evaluation, then reuses trajectories for continuation comparisons. We construct this benchmark using VLMs with zoom-in tools, where a model acquires zoomed-in images and performs self-verification to judge whether the returned images contain useful visual evidence for the question. We retain the trajectories accepted by self-verification, including those with incorrect answers, so that their accuracy reflects reasoning over evidence alone, without confounding from acquisition failures. To cover diverse model families and datasets, we construct it using six VLMs from the GLM (Z.ai, 2025), Qwen (Bai et al., 2025; Team,

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

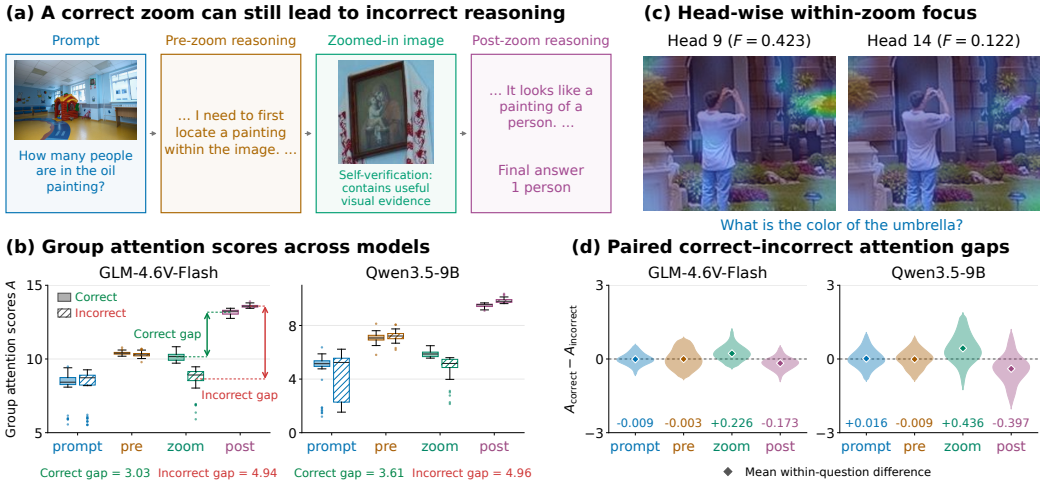


Figure 1: Reasoning–zoom attention, within-question attention differences and head-dependent within-zoom focus. (a) A Qwen3.5-9B trajectory in which the zoomed-in image contains useful visual evidence but the post-zoom reasoning is incorrect. (b) Box plots of group attention scores A and mean post-zoom gaps between correct and incorrect trajectories. (c) Within-zoom attention maps of two Qwen3.5-9B heads on V*Bench. (d) Correct-minus-incorrect attention difference for trajectory pairs from the same model and question.

2026; Qwen Team, 2026), Gemma (Gemma Team, 2026), and DeepSeek (DeepSeek-AI, 2026) series on six datasets. This yields 35,412 trajectories, covering 3,923 questions (Table 1).

Within the self-verified trajectories, we diagnose attention at two levels: how it is allocated between the zoomed-in image and reasoning, and how each head focuses within the zoomed-in image. Across models and datasets, the zoomed-in image receives less attention than both pre-zoom and post-zoom reasoning, revealing a reasoning–zoom attention imbalance (Figure 1(b)). Heads further differ in within-zoom focus, with some concentrating on a few tokens and others spreading attention broadly. This ordering is largely consistent across datasets (Figure 1(c)).

Based on these findings, we propose *Head-Gated Attention Reallocation*, a training-free method that reallocates attention from reasoning to the zoomed-in image during post-zoom decoding. At each decoding step, we adjust the last-layer attention according to the imbalance between reasoning and the zoomed-in image, scaling the adjustment for each head by its within-zoom focus. This reallocation changes how much attention each head assigns to the zoomed-in image without altering its distribution within the image, and requires no additional forward pass.

Across the five GLM, Qwen, and Gemma VLMs and six datasets, our attention reallocation approach improves average accuracy over the strongest training-free decoding baseline by 0.65 to 1.85 percentage points. Moreover, it even surpasses a training-based method on Qwen3-VL-8B. Scaling each head's adjustment by its focus consistently outperforms uniform reallocation at negligible decoding cost, confirming the role of within-zoom focus. Furthermore, we validate self-verification, finding that its judgments reflect target retention in zooms, remain consistent across questions about the same zoom, and align with human assessments.

Our contributions are summarized as follows: (i) We introduce a benchmarking protocol and the *EvidenceZoom* benchmark for evaluating reasoning over self-verified useful visual evidence in tool-use trajectories, with 35,412 zoom-in tool-use trajectories from six VLMs in four model series on six datasets. (ii) Our attention diagnosis mechanism reveals a reasoning-zoom imbalance and head-dependent within-zoom focus. (iii) Moreover, our Head-Gated Attention Reallocation, a training-free intervention approach, yields gains in average accuracy across six datasets on five VLMs. Together, our analyses characterize the gap between acquiring visual evidence and its effective use, motivating attention reallocation as a means of improving visual reasoning with zoom-in tools.

2 RELATED WORK

Visual acquisition and post-zoom evaluation. ViCrop (Khayatkhoei et al., 2025) and CARVE (Ge et al., 2026) use attention to guide region selection, while ZwZ (Wei et al., 2026) distills region-informed supervision into full-image training. MED (Ma et al., 2026) separates intrinsic learning from tool-induced gains and harms, including incorrect answers despite human-confirmed evidence in zoomed-in images. In this work, we develop a protocol to understand the effect of zoom-in tool use on post-zoom reasoning based on trajectories whose acquired evidence the model judges useful.

Visual reasoning failures and evidence use. MMVP (Tong et al., 2024b) and VisOnlyQA (Kamoi et al., 2024) expose perceptual limitations. Other studies identify underused visual representations (Fu et al., 2025; Liu et al., 2025a), failures to compose recognition and reasoning (Wang et al., 2026a), conflicts with language priors (Liu et al., 2025b), and errors reinforced by generated text (Zhong et al., 2024). HARM/VEED (Wang et al., 2026b) finds that similar visual attention can accompany real and hallucinated objects, highlighting the need to distinguish attention patterns from successful evidence use. While these studies focus on the interplay of visual tokens, attention, and reasoning, we target tool use and its role in post-zoom reasoning.

Attention and decoding interventions. Prior methods modify visual attention through sink redistribution (Kang et al., 2025), calibration and layer-wise reallocation (Zhao et al., 2026; Li et al., 2026; Lu et al., 2026), or added cross-attention layers (Mahajan et al., 2025). Other approaches target reasoning heads (Park et al., 2025), residual activations (Li et al., 2025), generated-token over-trust (Huang et al., 2024), or output logits (Chen et al., 2026; Ha et al., 2026). Our training-free method reallocates attention between reasoning and zoomed-in evidence, scaling each head's correction by its relative within-zoom focus.

3 ANALYZING ATTENTION UNDER ZOOM-IN TOOL USE

To investigate the utility of zoom-in tool use in VLMs and their reasoning trajectories, we first outline our *EvidenceZoom* benchmark. We then perform an in-depth analysis of how attention to zoom evidence affects post-zoom reasoning and answer correctness.

3.1 BENCHMARKING EVIDENCE USE WITH SELF-VERIFICATION

We examine whether VLMs fail to use zoom evidence even when a zoom-in tool is invoked to answer a question about the image. Figure 2(a) illustrates this protocol. A VLM first answers each question by reasoning, calling the zoom-in tool, and continuing to reason over the returned zoomed-in image. Self-verification then prompts the same model with the question and its zoomed-in images, but without its reasoning or answer, to judge whether the images contain useful visual evidence for solving the question. The *EvidenceZoom* benchmark excludes rejected trajectories, i.e., those in which zoomed-in images do not contain the required visual information as judged by the VLM itself, retaining the accepted ones with both correct and incorrect answers. We provide the self-verification prompt in Appendix A.3. Performance on these trajectories therefore reflects how well a model reasons over evidence it verifies as useful. To ensure diversity in our benchmark, we curate trajectories from six VLMs, GLM-4.6V-Flash (Z.ai, 2025), Qwen3-VL-8B (Bai et al., 2025), Qwen3.5-9B (Team, 2026), Qwen3.8-27B (Qwen Team, 2026), Gemma-4-31B (Gemma Team, 2026), and DeepSeek-V4-Flash-Vision-Exp (DeepSeek-AI, 2026), on HR-Bench 4K/8K (Wang et al., 2025), V*Bench (Wu & Xie, 2024), CV-Bench-2D (Tong et al., 2024a), MMStar (Chen et al., 2024), and UltraVR (Huang et al., 2026). Table 1 reports retained trajectory and question counts for the full six-model benchmark. For the diagnosis and intervention evaluation, in the following sections, we use the five GLM, Qwen, and Gemma VLMs.

3.2 MEASURING ATTENTION TO ZOOM EVIDENCE

We begin by formulating attention under zoom-in tool use. Here, the model processes a token sequence containing the input image and question, its pre-zoom reasoning, the returned zoomed-in images, and its reasoning and answer after the latest zoom. Let $\mathcal{I}$ denote the index set of positions in this token sequence. We denote these token groups by disjoint subsets of $\mathcal{I}$: $\mathcal{I}_{\text{prompt}}$ (input image and question), $\mathcal{I}^r_{\text{pre}}$ (pre-zoom reasoning), $\mathcal{I}^r_{\text{zoom}}$ (zoomed-in image), and $\mathcal{I}_{\text{post}}$ (post-zoom reason-

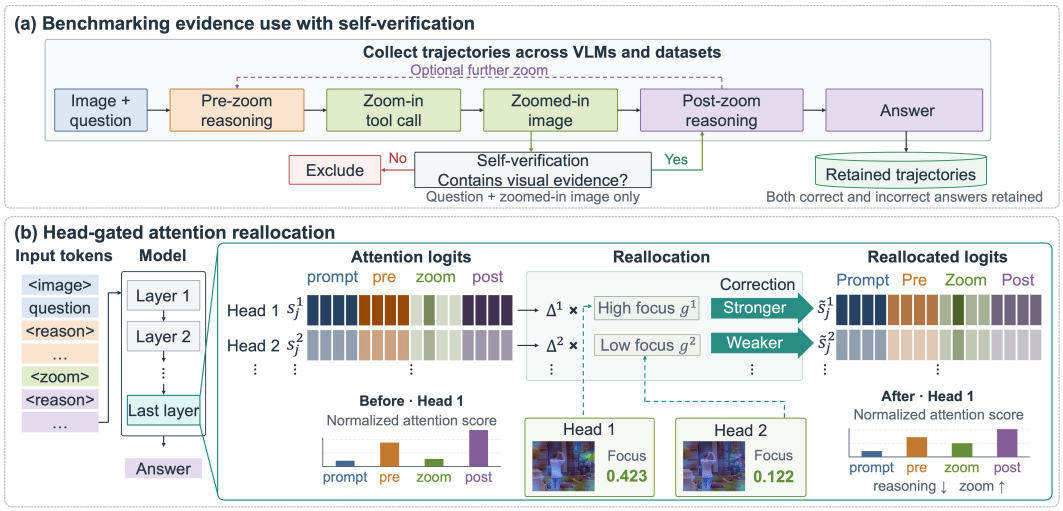


Figure 2: Overview of *EvidenceZoom* benchmarking and our *Head-Gated Attention Reallocation* approach. (a) We benchmark evidence use across VLMs and datasets, retaining trajectories whose zoomed-in images are judged to contain useful visual evidence by self-verification. (b) Our intervention shifts last-layer attention logits based on reasoning–zoom imbalance. (c) Each head's correction is scaled by its relative within-zoom focus.

Table 1: Self-verified trajectories in the *EvidenceZoom* benchmark for each model and dataset. Each cell reports the number of trajectories and the distinct questions they cover.

Model	HR-4K	HR-8K	VStar	CV-2D	MMStar	UltraVR	Total
Dataset questions	200	200	191	1,438	1,500	490	4,019
GLM-4.6V-Flash	707 (147)	744 (141)	405 (126)	4,320 (852)	3,113 (884)	1,578 (293)	10,867 (2,443)
Qwen3-VL-8B	603 (159)	563 (158)	631 (156)	439 (318)	865 (518)	201 (142)	3,302 (1,451)
Qwen3.5-9B	814 (182)	774 (177)	661 (169)	1,645 (837)	1,251 (695)	342 (245)	5,487 (2,305)
Qwen3.8-27B	486 (170)	481 (159)	511 (169)	1,307 (727)	1,595 (823)	128 (92)	4,508 (2,140)
Gemma-4-31B	399 (181)	390 (175)	376 (166)	720 (650)	891 (663)	296 (189)	3,072 (2,024)
DeepSeek-V4-Flash	421 (170)	368 (151)	330 (144)	3,137 (1,191)	3,815 (1,380)	105 (70)	8,176 (3,106)
All	3,430 (198)	3,320 (198)	2,914 (191)	11,568 (1,423)	11,530 (1,481)	2,650 (432)	35,412 (3,923)

ing and answer), where r indexes the tool-use round. Following the order of the tool interaction, these index sets appear as

$$(\mathcal{I}_{\text{prompt}}, \mathcal{I}^1_{\text{pre}}, \mathcal{I}^1_{\text{zoom}}, \ldots, \mathcal{I}^r_{\text{pre}}, \mathcal{I}^r_{\text{zoom}}, \mathcal{I}_{\text{post}}), \tag{1}$$

where system-prompt and special tokens are omitted. For simplicity, we denote the latest-round sets $\mathcal{I}^r_{\text{pre}}$ and $\mathcal{I}^r_{\text{zoom}}$ by $\mathcal{I}_{\text{pre}}$ and $\mathcal{I}_{\text{zoom}}$, omitting the superscript r. Next, we define attention from query position $t \in \mathcal{I}$ to a key position $j \le t$. For layer ℓ and query head h, the attention logit is given by

$$s^{(\ell,h)}_{t,j} = \frac{\langle \mathbf{q}^{(\ell,h)}_t, \mathbf{k}^{(\ell,h)}_j \rangle}{\sqrt{d_k}}, \tag{2}$$

where $\mathbf{q}^{(\ell,h)}_t$ is the query vector at position t, $\mathbf{k}^{(\ell,h)}_j$ is the key vector at position j, and d_k is their dimension. Using these logits, we measure attention at two levels: how it is allocated across token groups and how each head focuses within the zoomed-in image.

3.2.1 How much attention does the zoom receive?

Even when the zoomed-in image contains evidence required by the model to answer (evaluated via self-verification), post-zoom decoding may rely more on the generated reasoning than on the image itself. Following Huang et al. (2024), who examine last-layer attention to study reliance on generated tokens over visual tokens, we analyze how attention is allocated between the returned

zoomed-in image and the generated reasoning at the last layer. As these token groups vary widely in size, their total attention mass reflects both token count and the weights assigned to individual tokens, and thus cannot be compared directly. Although averaging softmax weights within each group removes the effect of token count, absolute differences between these averages can remain small even when their ratio is large. Therefore, we retain a logarithmic scale and aggregate the logits with log-mean-exp (Lowe et al., 2021). This exponentiates the logits as softmax does, averages them within each group, and takes the logarithm, so the result stays on the logit scale. Formally, for each token group γ and query head h at the current decoding step t, we define the group attention score using log-mean-exp as

$$A^h_{\mathcal{I}_\gamma} = \log\left(\frac{1}{|\mathcal{I}_\gamma|}\sum_{j\in\mathcal{I}_\gamma}\exp(s^h_j)\right), \quad \gamma \in \{\text{prompt}, \text{pre}, \text{zoom}, \text{post}\}, \tag{3}$$

where we shorten $s^{(\ell,h)}_{t,j}$ to s^h_j because our analysis considers the last layer ℓ at a fixed decoding step t.

3.2.2 How focused is each head within the zoom?

Beyond how much attention it receives, the zoomed-in image may be attended to diffusely or selectively depending on the head. We measure this selectivity using the entropy of each head's attention within the zoomed-in image. For each zoomed-in image token $j \in \mathcal{I}_{\text{zoom}}$, we compute the conditional attention weight of head h by applying softmax over the zoom logits, $\alpha^h_j = \text{softmax}_{j\in\mathcal{I}_{\text{zoom}}}(s^h_j)$. We then compute the Shannon entropy of these conditional weights, $H^h_{\mathcal{I}_{\text{zoom}}} = -\sum_{j\in\mathcal{I}_{\text{zoom}}}\alpha^h_j \log \alpha^h_j$, with lower entropy indicating stronger concentration. As this entropy is bounded above by $\log|\mathcal{I}_{\text{zoom}}|$ for $|\mathcal{I}_{\text{zoom}}| > 1$, we normalize it by this bound and define the within-zoom focus as one minus the normalized entropy,

$$F^h_{\mathcal{I}_{\text{zoom}}} = 1 - \frac{H^h_{\mathcal{I}_{\text{zoom}}}}{\log|\mathcal{I}_{\text{zoom}}|} \in [0,1]. \tag{4}$$

The focus $F^h_{\mathcal{I}_{\text{zoom}}}$ is zero when attention is spread uniformly over the zoomed-in image and approaches one as it concentrates on fewer tokens.

3.2.3 Findings

We now analyze attention allocation across token groups along with the head-wise focus within the zoomed-in image. We randomly sample 20 questions per setting from retained trajectories with a single tool-use round ($r = 1$) and one zoomed-in image. For the attention allocation analysis, we compute $A^h_{\mathcal{I}_\gamma}$ at every post-zoom query position $t \in \mathcal{I}_{\text{post}}$ and report the average over these positions and all heads. For the within-question attention difference analysis, we subtract the paired incorrect trajectory's score from the correct trajectory's score within each model and question. For the head-focus analysis, we average the focus $F^h_{\mathcal{I}_{\text{zoom}}}$ of each head over the same positions and report its rank across heads, normalized to $[0,1]$ within each dataset.

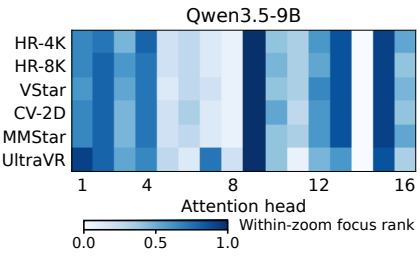


Figure 3: Rank of within-zoom focus for Qwen3.5-9B heads.

Reasoning dominates the zoom. Figure 1(b) shows that reasoning consumes most of the attention budget across GLM-4.6V-Flash and Qwen3.5-9B on HR-Bench 4K. The zoomed-in image receives a lower group attention score than both reasoning groups, even below the pre-zoom reasoning generated before its return. The same pattern holds for Qwen3.5-9B across all six datasets despite their different image resolutions and question types (Figure 11 in Appendix D.2).

Zoom is overlooked. Figure 1(d) compares correct and incorrect trajectories for the same question within each model. Its positive mean zoom differences and negative mean post-zoom differences indicate that the incorrect trajectories assign lower scores to zoom tokens and higher scores to their generated post-zoom reasoning. This contrast indicates that incorrect reasoning can accompany an overlooked zoom and greater reliance on generated text. Together, these results suggest that a

zoomed-in image self-verified as containing visual evidence is not fully leveraged in the last-layer attention readout. We therefore reallocate attention from reasoning tokens to the returned zoomed-in image to improve evidence use.

Heads differ in within-zoom focus. Figure 3 shows that Qwen3.5-9B heads differ in their focus within the zoomed-in image. The relative ordering of high- and low-focus heads remains broadly consistent across all six datasets, suggesting that each head has its own intrinsic degree of focus. The same pattern is observed for GLM-4.6V-Flash, Qwen3-VL-8B, Qwen3.8-27B, and Gemma-4-31B (Figure 12 in Appendix D.2). We therefore use within-zoom focus to determine how strongly each head is adjusted, so that the reallocated attention concentrates on specific details rather than spreading across the image.

Reasoning dominance is not a proximity effect. We examine whether the high group attention score of post-zoom reasoning arises from being adjacent to the answer. To verify this, we swap the zoomed-in image and post-zoom reasoning groups of Qwen3.5-9B, making the zoomed-in image adjacent to the answer instead, while keeping all other tokens unchanged. We define each group by its content rather than its position. Figure 4 shows the group attention scores in Eq. (3) with and without this swap, averaged over matched answer-token queries and heads on the same trajectories. Post-zoom reasoning still receives a higher score than the zoomed-in image after the swap, despite being placed farther from the answer. Therefore, reasoning dominance stems from the model's reliance on its generated reasoning rather than its adjacency to the answer.

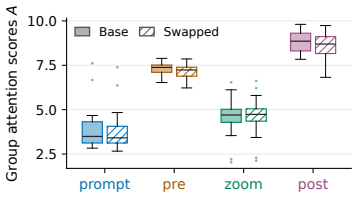


Figure 4: Group attention scores of Qwen3.5-9B with and without swapping the zoom and post-zoom reasoning.

4 Head-Gated Attention Reallocation

We propose *Head-Gated Attention Reallocation* to mitigate the reasoning–zoom imbalance by reallocating attention from reasoning to the zoomed-in image. For each head in the last layer, we measure this imbalance and adjust the attention logits according to the head's within-zoom focus. Since both pre-zoom and post-zoom reasoning receive higher group attention scores than the zoomed-in image, we reallocate attention from both. For each head h at the current decoding step t (suppressed), we define the imbalance as

$$\Delta^h = [\, A^h_{\mathcal{I}_{\text{reason}}} - A^h_{\mathcal{I}_{\text{zoom}}} \,]_+, \tag{5}$$

where $\mathcal{I}_{\text{reason}} = \mathcal{I}_{\text{pre}} \cup \mathcal{I}_{\text{post}}$ denotes the union of pre-zoom and post-zoom reasoning, and the positive part $[\cdot]_+$ restricts reallocation to heads in which reasoning dominates the zoom.

However, adjusting heads by the imbalance alone overlooks each head's intrinsic degree of focus within the zoomed-in image: a less focused head spreads the reallocated attention across the image, whereas a more focused head concentrates it on specific details. We therefore apply stronger reallocation to more focused heads using a head gate based on within-zoom focus,

$$g^h = \frac{F^h_{\mathcal{I}_{\text{zoom}}}}{F^h_{\mathcal{I}_{\text{zoom}}} + \overline{F}_{\mathcal{I}_{\text{zoom}}} + \epsilon} \in [0, 1), \tag{6}$$

where $\overline{F}_{\mathcal{I}_{\text{zoom}}}$ is the mean within-zoom focus over all query heads in the last layer, and ϵ is a constant for numerical stability. Using focus relative to the mean $\overline{F}_{\mathcal{I}_{\text{zoom}}}$ compensates for the absolute level of focus in each sample, so the gate requires neither preselected heads nor a fixed cutoff. Finally, we combine the two to reduce each head's imbalance, raising the zoom logits and lowering the reasoning logits by the same amount:

$$\tilde{s}^h_j = \begin{cases} s^h_j + \frac{1}{2}\, g^h \min(\Delta^h, \Delta_{\text{cap}}), & j \in \mathcal{I}_{\text{zoom}}, \\ s^h_j - \frac{1}{2}\, g^h \min(\Delta^h, \Delta_{\text{cap}}), & j \in \mathcal{I}_{\text{reason}}, \\ s^h_j, & \text{otherwise}, \end{cases} \tag{7}$$

where $\Delta_{\text{cap}} > 0$ is a hyperparameter that caps the imbalance to prevent overcorrection, and the factor $1/2$ splits the correction evenly between the two groups. A head is thus adjusted more strongly

when reasoning dominates more and the head concentrates more within the zoomed-in image. Note that the adjustment changes how much attention the zoomed-in image receives but preserves each head's spatial distribution within it, implying that the concentration on specific details comes from the head gate, not from spatial sharpening. Since the adjustment is applied only to the last layer and reuses its existing attention logits, each decoding step adds only $O\big(N_{\text{head}} \cdot (|\mathcal{I}_{\text{zoom}}| + |\mathcal{I}_{\text{reason}}|)\big)$ operations, where N_{head} denotes the number of query heads in the last layer.

Implementation details. For each intervention, we resume decoding from the original context of a retained trajectory through the first zoomed-in image and apply reallocation only to the newly generated tokens. We set $\Delta_{\text{cap}} = 1.5$ and $\epsilon = 10^{-6}$ for every model and dataset, and provide further implementation details in Appendix A.

5 Experiments

5.1 Experimental settings

Datasets. For intervention evaluation, we reuse trajectories from five of the six VLMs in the *EvidenceZoom* benchmark, *GLM-4.6V-Flash*, *Qwen3-VL-8B*, *Qwen3.5-9B*, *Qwen3.8-27B*, and *Gemma-4-31B*, on *HR-Bench 4K/8K* (Wang et al., 2025), *V*Bench* (Wu & Xie, 2024), *CV-Bench-2D* (Tong et al., 2024a), *MMStar* (Chen et al., 2024), and *UltraVR* (Huang et al., 2026).

Comparison. We denote *Base* as standard decoding without any intervention. We compare our method against training-free baselines including *ResDec* (Chen et al., 2026), which guides decoding using the history of recent output logits, and *FLB* (Ha et al., 2026), which boosts the output logits from the initial decoding steps throughout subsequent generations. Additionally, we compare with the training-based method *ZwZ-8B* (Wei et al., 2026), which is trained from Qwen3-VL-8B via region-to-image distillation. For this comparison, we apply our method to the same backbone.

Evaluation. We average correctness over matched runs within each question and then equally weight questions within each dataset. Deviations report the SD over 10,000 resampled evaluations. We use one-sided McNemar test for statistical analysis.

5.2 Main results

Table 2 shows the accuracy of our approach and the training-free baselines across six datasets. Our proposed *Head-gated Attention Reallocation* method achieves the highest mean accuracy on all five VLMs, outperforming ResDec and FLB by at least 0.65, 1.85, 1.38, 0.81, and 0.70 percentage points, respectively. The gains are consistent across datasets, as our method improves over the base models in 29 of 30 model–dataset pairs, with a slight decrease only on UltraVR for Gemma-4-31B. We provide full and per-category results in Appendix D. Furthermore, Figure 5 compares the accuracy gains of our method and the training-based ZwZ-8B (Wei et al., 2026) over the Qwen3-VL-8B Base model. By consistently improving over the Base model on all six datasets, our method achieves 73.26% mean accuracy, surpassing ZwZ-8B both without tools (71.22%) and with tools (71.00%). Overall, these results show that reallocating attention improves evidence use more effectively than training-free decoding methods and even a training-based model. Qualitatively, in the example of Figure 6, the base model overlooks the zoomed-in evidence, whereas our method attends to it, identifies the yellow hazard-label background, and predicts the correct answer.

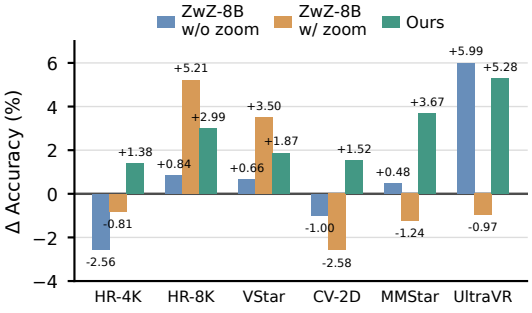


Figure 5: Accuracy gains of ZwZ-8B and Ours over Qwen3-VL-8B Base.

5.3 Ablation study

To isolate the effect of the head gate, we compare three variants on five VLMs. *Base*, *Uniform* applies the reallocation in Eq. (7) with $g^h = 1$ for all heads, and *Ours*. Since Uniform differs from Ours only in the head gate, this comparison isolates the contribution of head gating. Table 3 reports

Table 2: Accuracy of our method and training-free methods across six datasets. Within each model, **bold** and underline denote the best and second-best mean accuracy. Superscripts denote one-sided McNemar test against the second-best method, with $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Model	Method	HR-4K	HR-8K	VStar	CV-2D	MMStar	UltraVR	Average
GLM-4.6V-Flash	Base	$\underline{88.50}_{1.22}$	$87.28_{1.17}$	$87.59_{1.01}$	$75.39_{0.55}$	$72.63_{0.47}$	$20.42_{0.81}$	$71.97_{0.38}$
	ResDec	$87.65_{1.15}$	$87.77_{1.17}$	$87.85_{0.94}$	$75.61_{0.55}$	$72.67_{0.46}$	$\underline{21.47}_{0.84}$	$72.17_{0.37}$
	FLB	$87.76_{1.22}$	$\underline{88.10}_{1.15}$	$\mathbf{88.25}_{0.89}$	$\underline{75.66}_{0.54}$	$\underline{72.70}_{0.46}$	$21.35_{0.84}$	$\underline{72.30}_{0.37}$
	Ours	$\mathbf{89.03}^{*}_{1.21}$	$\mathbf{88.69}^{*}_{1.05}$	$\underline{87.99}_{0.96}$	$\mathbf{76.26}^{***}_{0.54}$	$\mathbf{73.05}^{***}_{0.46}$	$\mathbf{22.69}^{**}_{0.86}$	$\mathbf{72.95}^{***}_{0.37}$
Qwen3-VL-8B	Base	$\underline{88.83}_{0.86}$	$82.07_{1.33}$	$88.23_{0.84}$	$83.02_{0.49}$	$75.45_{0.41}$	$5.28_{0.61}$	$70.48_{0.33}$
	ResDec	$88.35_{0.81}$	$\underline{83.14}_{1.29}$	$87.94_{0.86}$	$80.90_{0.48}$	$76.51_{0.41}$	$\underline{9.15}_{0.50}$	$71.00_{0.32}$
	FLB	$86.75_{0.81}$	$82.10_{1.31}$	$\underline{89.41}_{0.79}$	$\underline{83.05}_{0.40}$	$\underline{78.04}_{0.42}$	$\underline{9.15}_{0.50}$	$\underline{71.42}_{0.32}$
	Ours	$\mathbf{90.21}^{**}_{0.77}$	$\mathbf{85.06}^{*}_{0.98}$	$\mathbf{90.10}_{0.72}$	$\mathbf{84.54}^{*}_{0.31}$	$\mathbf{79.12}_{0.41}$	$\mathbf{10.56}_{0.50}$	$\mathbf{73.26}^{***}_{0.27}$
Qwen3.5-9B	Base	$87.52_{1.22}$	$86.76_{1.23}$	$\underline{90.73}_{0.99}$	$78.53_{0.59}$	$77.86_{0.40}$	$25.95_{0.92}$	$74.56_{0.39}$
	ResDec	$\underline{88.01}_{1.11}$	$\underline{90.06}_{1.10}$	$90.58_{0.91}$	$77.79_{0.59}$	$\underline{78.29}_{0.41}$	$26.90_{0.86}$	$\underline{75.27}_{0.36}$
	FLB	$87.66_{1.07}$	$88.97_{1.15}$	$89.97_{0.86}$	$\underline{79.23}_{0.54}$	$77.30_{0.36}$	$\underline{28.44}_{0.96}$	$75.26_{0.36}$
	Ours	$\mathbf{89.03}^{*}_{1.01}$	$\mathbf{90.10}_{1.10}$	$\mathbf{92.14}_{0.90}$	$\mathbf{79.72}_{0.56}$	$\mathbf{78.95}_{0.42}$	$\mathbf{29.97}_{0.97}$	$\mathbf{76.65}^{***}_{0.35}$
Qwen3.8-27B	Base	$\underline{96.03}_{0.61}$	$\underline{92.92}_{0.49}$	$\underline{98.21}_{0.47}$	$92.80_{0.15}$	$82.79_{0.23}$	$57.52_{0.89}$	$86.71_{0.22}$
	ResDec	$95.34_{0.77}$	$\underline{92.92}_{0.78}$	$97.83_{0.57}$	$92.60_{0.20}$	$82.85_{0.26}$	$\underline{60.78}_{0.89}$	$\underline{87.05}_{0.26}$
	FLB	$95.93_{0.62}$	$92.63_{0.47}$	$97.63_{0.68}$	$\underline{92.87}_{0.13}$	$\underline{83.21}_{0.23}$	$57.52_{0.89}$	$86.63_{0.23}$
	Ours	$\mathbf{96.23}_{0.55}$	$\mathbf{93.89}^{*}_{0.47}$	$\mathbf{98.52}_{0.30}$	$\mathbf{93.01}_{0.13}$	$\mathbf{83.28}_{0.23}$	$\mathbf{62.23}_{1.03}$	$\mathbf{87.86}^{*}_{0.22}$
Gemma-4-31B	Base	$91.53_{0.71}$	$93.43_{0.84}$	$93.37_{0.71}$	$77.31_{0.17}$	$60.23_{0.31}$	$\mathbf{36.95}_{0.82}$	$75.47_{0.26}$
	ResDec	$\underline{91.80}_{0.66}$	$93.14_{0.88}$	$92.97_{0.73}$	$\underline{79.05}_{0.21}$	$\underline{60.68}_{0.33}$	$36.68_{0.86}$	$\underline{75.72}_{0.27}$
	FLB	$91.71_{0.71}$	$\underline{93.81}_{0.75}$	$\underline{93.67}_{0.52}$	$78.59_{0.19}$	$60.38_{0.31}$	$36.07_{0.82}$	$75.71_{0.24}$
	Ours	$\mathbf{92.63}_{0.47}$	$\mathbf{94.29}_{0.69}$	$\mathbf{94.58}_{0.43}$	$\mathbf{79.26}_{0.22}$	$\mathbf{60.99}_{0.33}$	$\underline{36.77}_{0.78}$	$\mathbf{76.42}^{**}_{0.21}$

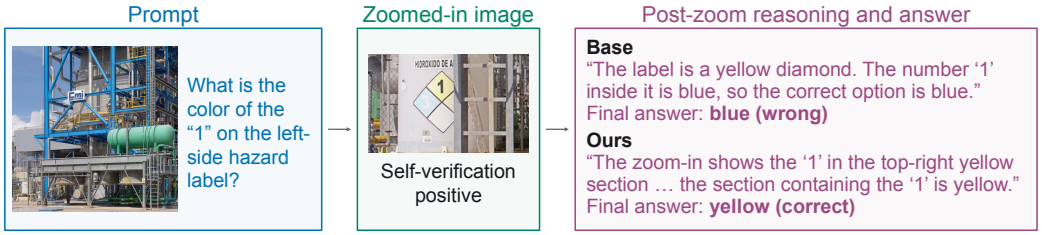


Figure 6: Qualitative example from HR-Bench. Compared with Base, Ours shifts attention to the zoomed-in image and corrects the answer.

the six-dataset mean accuracy of each variant. Uniform improves over Base by 0.42, 1.26, 0.98, and 0.65 percentage points on GLM, Qwen3-VL, Qwen3.5, and Gemma, respectively, but decreases accuracy by 0.06 percentage points on Qwen3.8. Our method achieves the highest accuracy with gains over Uniform be 0.57, 1.53, 1.11, 1.21, and 0.30 percentage points, showing that head gating further strengthens reallocation. We provide the full results in Appendix D.

Decoding efficiency. We compare per-token decoding time against *Base*, *ResDec*, *FLB*, and *Uniform*. We use Qwen3.5-9B on an RTX A6000 GPU with identical continuation tokens for the controlled timing comparison. Table 4 reports per-token decoding time alongside natural output lengths. Our method adds a mean overhead of 3.5% over *Base*, with head gating contributing only 0.16 ms per token over *Uniform*. In terms of output length, our method generates 3.5% more post-zoom tokens than *Base* but 6.3% and 10.9% fewer than ResDec and FLB, respectively. Thus, despite a marginal per-token overhead, the shorter outputs keep the total decoding cost of our method below that of ResDec and FLB. Notably, our method requires neither training nor an additional model pass, keeping its per-token cost close to that of the model without intervention. We provide the measurement protocols and results on longer sequences in Appendices A.5 and D.3.

5.4 Evaluation of Self-Verification

Target retention in zooms. Using V*Bench's evidence annotations, we examine whether self-verification reflects target retention. For 115 single-target direct-attribute questions, we compute the intersection over ground truth (IoG (Wang et al., 2018)) between each zoom and the annotated target. Figure 7(a) shows the mean IoG for each VLM. Across all five VLMs, positive judgments show higher mean IoG with lower standard deviations (full results are in Appendix B.1), indicating

Table 3: Ablation of reallocation and head gating. Values are mean accuracy over six datasets.

Model	Base	Uniform	Ours
GLM-4.6V-Flash	$71.97_{0.38}$	$\underline{72.38}_{0.39}$	$\mathbf{72.95}_{0.37}$
Qwen3-VL-8B	$70.48_{0.33}$	$\underline{71.74}_{0.33}$	$\mathbf{73.26}_{0.27}$
Qwen3.5-9B	$74.56_{0.39}$	$\underline{75.54}_{0.38}$	$\mathbf{76.65}_{0.35}$
Qwen3.8-27B	$\underline{86.71}_{0.22}$	$86.65_{0.24}$	$\mathbf{87.86}_{0.22}$
Gemma-4-31B	$75.47_{0.26}$	$\underline{76.12}_{0.24}$	$\mathbf{76.42}_{0.21}$

Table 4: Decoding speed and average output length. Parentheses show change from Base.

Method	Decode (ms/token)	Generated tokens
Base	40.1	666.6
ResDec	40.4(+0.8%)	736.7(+10.5)
FLB	40.2(+0.4%)	774.4(+16.2)
Uniform	41.3(+3.1%)	767.7(+15.2)
Ours	41.5(+3.5%)	689.9(+3.5)

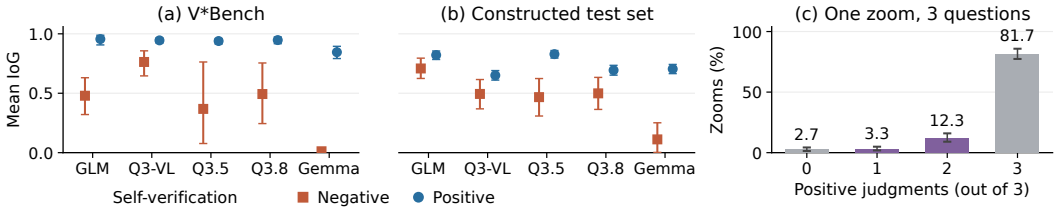


Figure 7: Mean IoG by self-verification on (a) V*Bench and (b) the constructed 200-pair test set. (c) Agreement of self-verification across three attribute questions per zoom.

that self-verified zooms retain more of the target with greater consistency. These results suggest that self-verification provides a reliable signal of whether a zoom contains the target evidence.

Zoom judgments across datasets and attributes. We examine whether self-verification remains consistent across datasets and across questions about different attributes. We sample 200 image–question pairs from V*Bench, HR-Bench 8K, CV-Bench-2D, and MMStar, and generate two additional questions about other attributes of each queried entity, yielding 600 questions. We compute the IoG between each zoom and the annotated target, and the agreement of self-verification across the three questions for each zoom. Figure 7(b) and (c) show the mean IoG and the agreement of judgments on the constructed test set, respectively. Similar to the results on V*Bench, positive judgments have higher mean IoG across all five VLMs. Judgments fully agree across the three questions for 81.7% of zooms, indicating that self-verification is consistent across attributes.

Human evaluation of zooms. We conduct a human evaluation on 100 zoom–question pairs, 25 each from GLM-4.6V-Flash, Qwen3.5-9B, Gemma-4-31B, and DeepSeek-V4-Flash-Vision-Exp, with 50 positive and 50 negative self-verification judgments. Ten raters independently judge whether each zoom provides useful visual evidence. We find that individual raters agree with the self-verification on 80–93% of pairs, with inter-rater agreement of Fleiss' $\kappa = 0.737$ (95% CI $[0.661, 0.807]$) and 87.3% mean pairwise agreement. This indicates that self-verification align with human assessments of visual evidence, supporting the reliability of self-verification for trajectory curation. We provide the full protocols and results in Appendix B.

6 Conclusion

We presented a reusable protocol for curating tool-use trajectories and evaluating reasoning over self-verified evidence. Applied to six VLMs, including DeepSeek, with zoom-in tools across six datasets, it yields the *EvidenceZoom* benchmark, comprising 35,412 trajectories that expose the gap between accepting visual evidence and using it to answer correctly. Furthermore, our attention diagnosis characterizes reasoning–zoom imbalance and differences in spatial focus across heads, motivating *Head-Gated Attention Reallocation*, which shifts last-layer attention toward zoom evidence according to each head's imbalance and relative within-zoom focus, without training or an additional model pass. Intervention evaluation on the five GLM, Qwen, and Gemma VLMs shows gains over training-free baselines, with an additional gain over the training-based baseline on Qwen3-VL-8B. Target-retention analyses, comparisons across question attributes, and human evaluation further characterize the evidence selected by self-verification. The protocol, the diagnosis, and the intervention provide a basis for investigating and improving reasoning with tool-acquired visual evidence.

Limitations and future work. The generalization of our method to other tools and longer interaction sequences remains to be tested. Future work could extend the protocol to other tools and tasks. Relevance-aware head selection and interventions across layers and tool-use rounds are further directions for improving evidence use.

AI USE STATEMENT

We used generative AI tools to polish the writing. The authors take responsibility for the final text, methods, experimental results, claims, and accompanying artifacts.

REPRODUCIBILITY STATEMENT

We provide the anonymous project at `https://anonymous.4open.science/r/EvidenceZoom` and the benchmark in the supplementary material. The accompanying code release includes model-specific implementations of our intervention for the five evaluated backbones, a runnable example trajectory, dependency versions, and execution instructions. Section 3.1 describes trajectory curation and evidence screening, Section 4 specifies the intervention, and Section 5.1 presents the experimental settings and evaluation procedure. Appendix A documents dataset characteristics, model sampling parameters, the self-verification prompt, and implementation details. The protocols for evaluating self-verification, including human assessment, are described in Appendix B.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087, 2024.

Xinrong Chen, Xu Chu, Yingmin Qiu, Hengyuan Zhang, Jing Xiong, Shiyu Tang, Shuai Liu, Shaokang Yang, Cheng Yang, Hayden Kwok-Hay So, et al. Residual decoding: Mitigating hallucinations in large vision-language models via history-aware residual guidance. *arXiv preprint arXiv:2602.01047*, 2026.

DeepSeek-AI. DeepSeek-V4-Flash-Vision-Exp. Hugging Face model repository, 2026. URL `https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash-Vision-Exp`.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: Vlms overlook their visual representations. *arXiv preprint arXiv:2506.08008*, 2025.

Yuyao Ge, Shenghua Liu, Yiwei Wang, Lingrui Mei, Baolong Bi, Xuanshan Zhou, Jiayu Yao, Jiafeng Guo, and Xueqi Cheng. Focusing by contrastive attention: Enhancing vlms' visual reasoning. In *European Conference on Computer Vision*, pp. 555–571, 2026.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Google DeepMind. Gemini 3.7 Flash. Google AI for Developers model documentation, `https://ai.google.dev/gemini-api/docs/models/gemini-3.7-flash`, 2026.

Jiwoo Ha, Jongwoo Baek, and Jinhyun So. First logit boosting: Visual grounding method to mitigate object hallucination in large vision-language models. *arXiv preprint arXiv:2604.00455*, 2026.

Gexin Huang, Yanting Yang, Myeongkyun Kang, Beidi Zhao, Jun Zhou, Chen Zhou, Gang Wang, Zu-hua Gao, and Xiaoxiao Li. Ultravr: A diagnostic ultra-resolution image-vqa benchmark for evidence-grounded reasoning. *arXiv preprint arXiv:2606.05576*, 2026.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Ryo Kamoi, Yusen Zhang, Sarkar Snigdha Sarathi Das, Ranran Haoran Zhang, and Rui Zhang. Visonlyqa: Large vision language models still struggle with visual perception of geometric information. *arXiv preprint arXiv:2412.00947*, 2024.

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

Mahyar Khayatkhoei, Prateek Chhikara, Filip Ilievski, et al. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *International Conference on Learning Representations*, pp. 68194–68213, 2025.

Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, Ming Li, Xiaonan Luo, and Guanbin Li. Cross-modal attention calibration for lvlm hallucination mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40186–40196, 2026.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Proceedings of the International Conference on Machine Learning*, volume 267, pp. 35799–35819, 2025.

Benlin Liu, Amita Kamath, Madeleine Grunde-McLaughlin, Winson Han, and Ranjay Krishna. Visual representations inside the language model. *arXiv preprint arXiv:2510.04819*, 2025a.

Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jen-tse Huang, Qiuzhi Liu, Pinjia He, and Zhaopeng Tu. Insight over sight: Exploring the vision-knowledge conflicts in multimodal LLMs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17825–17846, 2025b.

Scott C. Lowe, Thomas Trappenberg, and Sageev Oore. LogAvgExp provides a principled and performant global pooling operator. *arXiv preprint arXiv:2111.01742*, 2021.

Haolang Lu, Bolun Chu, WeiYe Fu, Guoshun Nan, Junning Liu, Minghui Pan, Qiankun Li, Yi Yu, Hua Wang, and Kun Wang. Reallocating attention across layers to reduce multimodal hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4157–4167, 2026.

Yan Ma, Weiyu Zhang, Tianle Li, Linge Du, Xuyang Shen, and Pengfei Liu. What does vision tool-use reinforcement learning really learn? disentangling tool-induced and intrinsic effects for crop-and-zoom. In *Proceedings of the International Conference on Machine Learning*, 2026.

Shweta Mahajan, Hoang Le, Hyojin Park, Farzad Farhadzadeh, Munawar Hayat, and Fatih Porikli. Attention guided alignment in efficient vision-language models. *arXiv preprint arXiv:2511.17793*, 2025.

Yein Park, Minbyul Jeong, and Jaewoo Kang. Thinking sparks!: Emergent attention heads in reasoning models during post training. *arXiv preprint arXiv:2509.25758*, 2025.

Qwen Team. Qwen3.8-27B model card, 2026. URL `https://huggingface.co/Qwen/Qwen3.8-27B`.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pp. 87310–87356, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024b.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. 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, pp. 7907–7915, 2025.

Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7774–7783, 2018.

Yucheng Wang, Yifan Hou, Aydin Javadov, Mubashara Akhtar, and Mrinmaya Sachan. Compose and fuse: Revisiting the foundational bottlenecks in multimodal reasoning. In *International Conference on Learning Representations*, 2026a.

Zichuan Wang, Songlin Yang, Bo Peng, Zhenchen Tang, Yang Li, Beibei Dong, and Jing Dong. Same attention, different truths: Put logit-lens over visual attention to detect and mitigate lvlm object hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25315–25325, 2026b.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, et al. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.

Z.ai. GLM-4.6V-Flash model card, 2025. URL https://huggingface.co/zai-org/GLM-4.6V-Flash.

Beidi Zhao, Wenlong Deng, Xinting Liao, Yushu Li, Nazim Shaikh, Yao Nie, and Xiaoxiao Li. When rag hurts: Diagnosing and mitigating attention distraction in retrieval-augmented lvlms. *arXiv preprint arXiv:2602.00344*, 2026.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11991–12011, 2024.

A Implementation details

A.1 Datasets

Table 5: Dataset characteristics. Green checks and red crosses indicate whether a dataset provides localized evidence annotations and whether it queries multiple attributes of the same object.

Dataset	Questions	Type	Categories	Evidence Annotation	Multi-Attr.	Focus
HR-Bench 4K	200	MCQ	2	×	×	high-resolution perception
HR-Bench 8K	200	MCQ	2	×	×	high-resolution perception
V*Bench	191	MCQ	2	✓	×	guided visual search
CV-Bench-2D	1,438	MCQ	2	×	×	2D perception (counting, relations)
MMStar	1,500	MCQ	6	×	×	general, vision-indispensable
UltraVR	490	MCQ/VQA	4	✓	×	ultra-resolution evidence-grounded reasoning

A.2 Sampling parameters

Table 6: Sampling parameters of each model.

Model	temperature	top_p	top_k	min_p	presence_penalty	repetition_penalty	max_tokens	reasoning_effort
GLM-4.6V-Flash	0.8	0.6	2	–	–	1.1	32,768	–
Qwen3-VL-8B	0.7	0.8	20	–	1.5	1.0	32,768	–
Qwen3.5-9B	1.0	0.95	20	0.0	1.5	1.0	32,768	–
Qwen3.8-27B	1.0	0.95	20	0.0	0.0	1.0	32,768	xhigh
Gemma-4-31B	1.0	0.95	64	–	–	–	32,768	–
DeepSeek-V4-Flash-Vision-Exp	1.0	0.95	–	–	–	–	32,768	max

A.3 Self-verification prompt

The verifier judges whether the zoomed-in image provides useful, potentially partial visual evidence. The prompt is reproduced below.

```
You are given one zoomed-in image and the question below.

{question}

The answering model also has the original image and may inspect
additional crops.

Judge whether this crop provides useful visual evidence for any part
of solving the question.  Useful evidence may identify a relevant
object, support part of a comparison, check a candidate, or rule out
a candidate region.

The crop may provide only partial evidence.  It does not need to
establish the final answer or preserve its position within the
original image.

Respond "yes" if the crop provides useful evidence.
Respond "no" if it is irrelevant or too unclear to provide useful
evidence.

Do not answer the question.  Respond with exactly one of these two
strings:
yes
no
```

A.4 Intervention implementation

FLB uses $\beta_{\mathrm{FLB}} = 0.1$, $\gamma_{\mathrm{FLB}} = 0.3$, and $\lambda_{\mathrm{FLB}} = 0.05$. ResDec uses $\alpha_{\mathrm{ResDec}} = 0.5$, $\beta_{\mathrm{ResDec}} = 0.1$, an eight-token history, and 64 candidates. Method subscripts distinguish these hyperparameters from the token-group label γ and conditional attention weights α_j^h.

In Eq. (2), the query and key vectors $\mathbf{q}_t^{(\ell,h)}$ and $\mathbf{k}_j^{(\ell,h)}$ are taken after the backbone's native projections, any query/key normalization, and positional transformations. The group scores of Eq. (3) are computed with a numerically stable log-sum-exp reduction, subtracting $\log|\mathcal{I}_\gamma|$ to obtain $A^h_{\mathcal{I}_\gamma}$. In the entropy $H^h_{\mathcal{I}_{\text{zoom}}}$ used by Eq. (4), the argument of $\log \alpha^h_j$ is clamped below at 10^{-8}. The probabilities themselves are not renormalized after this numerical guard. A one-token zoomed-in image is assigned zero normalized entropy and focus $F^h_{\mathcal{I}_{\text{zoom}}} = 1$. All reductions and the bias are computed in fp32 before casting back to the model dtype. Risk gating and additional within-zoom sharpening are disabled. Decoder state for ResDec and FLB is reset for every model `generate` call, including a new call after a tool return. Saved inference and evaluation artifacts are keyed by stage, question, and run index so interrupted posthoc sweeps can resume without replacing completed output.

A.5 Decoding efficiency comparison implementation

Following the distinction between crop-processing overhead and decoding time in Khayatkhoei et al. (2025), we identify the operation being timed rather than treating every source of inference cost as interchangeable. The timing cohort consists of 20 Qwen3.5-9B traces selected in fixed hash order from 75 eligible frozen traces with at least 128 continuation tokens. Selection does not inspect correctness, attention, or runtime. It includes six UltraVR, four CV-Bench-2D, four MMStar, three HR-Bench 4K, two V*Bench, and one HR-Bench 8K question. Prefix lengths range from 870 to 33,297 tokens. This is a timing workload, not an accuracy sample balanced across datasets.

All six methods receive the same full image, crop, serialized prefix, and continuation IDs. Production attention interventions and ResDec/FLB processors execute normally. A final token constraint then forces the saved next token. Greedy selection and disabled EOS/stopping conditions ensure exactly 128 forwards. The common forcing mask, token selection, GPU-to-CPU output collection, and detokenization remain inside the timed interval. Model loading, image preparation, method installation, acquisition, verification, evaluation, and upload are excluded. Each method receives a warmup, and method order is shuffled within each repetition. CUDA synchronization separates prefill/first-token time from the next 127 cached steps. Repetitions are averaged within questions before computing the reported means, SDs, and paired percentage overheads. Saved records verify token identity, forward counts, and complete method/repetition coverage. Historical elapsed times are not used because their hardware and output lengths differ.

B Analysis of self-verification

B.1 Analysis using V*Bench

We examine saved self-verification judgments on V*Bench's 115 direct-attribute questions with a single annotated target. Let P be the effective zoom region after clipping and any minimum-size expansion by the tool, and let T be the annotated target box in the same image coordinates. With $|\cdot|$ denoting image-plane area, we define

$$\text{IoG} = \frac{|P \cap T|}{|T|}.$$

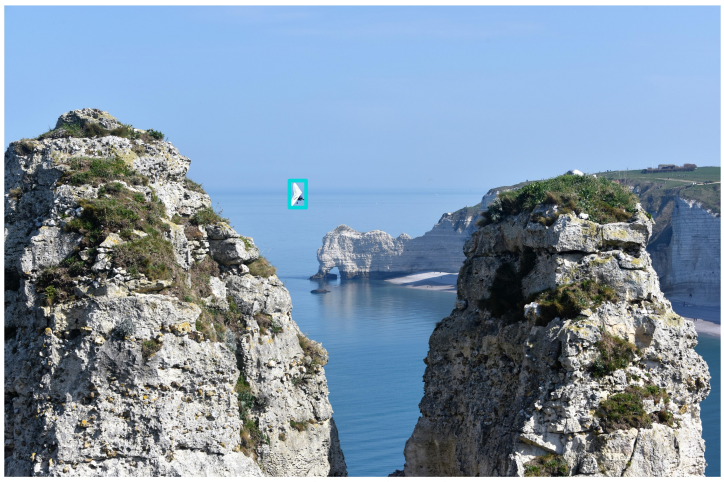
Intersection over ground truth (IoG), as defined by Wang et al. (2018), measures the fraction of the annotated target box retained by the zoom.

The mean IoG differences between positive and negative judgments for GLM, Qwen3-VL, and Qwen3.5 are 0.477, 0.182, and 0.571. The corresponding question-clustered 95% intervals are $[0.321, 0.635]$, $[0.084, 0.304]$, and $[0.095, 0.938]$, respectively. The Qwen3.8 contrast is 0.454, with an interval of $[0.184, 0.708]$. The ZwZ contrast is 0.685, with an interval of $[0.530, 0.792]$. The Gemma-4-31B-it contrast is 0.834, with an interval of $[0.782, 0.885]$.

B.2 Analysis using a constructed test set

We construct a small test set to evaluate self-verification while holding the target object fixed and varying the queried attribute. We randomly sample candidate image–question pairs from HR-Bench 8K, V*Bench, CV-Bench-2D, and MMStar. For each candidate, Gemini 3.7 Flash (Google DeepMind, 2026) identifies an entity relevant to the question, annotates its bounding box, and constructs

three multiple-choice questions about distinct visual attributes of that entity. We retain an eligible original attribute question and add two questions, or rewrite all three when needed. After model checks and human review, we select 200 objects with approximate balance across the four datasets, yielding 600 questions. Each object's three questions share the same image as illustrated in Figure 8.



Source image and target box



Zoom

Attribute	Question	Choices (gold in bold)
Readable marking	What word is written on the hang glider?	**Sunny**, Stormy, Cloudy, Windy
Color	What is the predominant color of the hang glider's wing?	**White**, Red, Blue, Green
Shape/form	What is the general shape of the hang glider's wing?	**Triangular**, Rectangular, Circular, Crescent-shaped

Figure 8: Example from the constructed test set, based on an image from HR-Bench 8K. The object and box remain fixed while the requested attribute changes.

Table 7: Mean IoG with 95% bootstrap intervals by positive or negative self-verification judgment for five backbones.

	V*Bench		Constructed test set	
Backbone	**Negative**	**Positive**	**Negative**	**Positive**
GLM-4.6V-Flash	0.479 [0.321, 0.631]	0.956 [0.907, 0.988]	0.708 [0.624, 0.795]	0.821 [0.784, 0.855]
Qwen3-VL-8B	0.763 [0.646, 0.857]	0.945 [0.923, 0.965]	0.494 [0.369, 0.615]	0.650 [0.611, 0.691]
Qwen3.5-9B	0.368 [0.077, 0.763]	0.939 [0.915, 0.960]	0.467 [0.308, 0.624]	0.829 [0.796, 0.859]
Qwen3.8-27B	0.493 [0.245, 0.755]	0.947 [0.917, 0.974]	0.499 [0.364, 0.633]	0.694 [0.652, 0.735]
Gemma-4-31B-it	0.012 [0.000, 0.025]	0.846 [0.792, 0.895]	0.112 [0.002, 0.251]	0.705 [0.665, 0.742]

B.3 Human evaluation

We analyze ten complete human reviews of the same 100 crop–question pairs (50 with positive and 50 with negative self-verification labels). The cohort contains 25 examples each from Qwen3.5-9B, GLM-4.6V-Flash, Gemma-4-31B-it, and DeepSeek-V4-Flash-Vision-Exp, drawn from HR-Bench 4K (34), HR-Bench 8K (33), and V*Bench (33). Raters see the self-verification prompt, the zoomed-in image, and the question. The full image, model identity, and other information are hidden. They answer yes or no according to whether they judge the zoomed-in image to provide useful visual evidence.

All ten raters answered all 100 items, yielding 1,000 ratings without missing responses. Fleiss' κ among the human raters is 0.737 (95% CI $[0.661, 0.807]$), with mean pairwise agreement of 87.3% and unanimous agreement on 61 items. We compute intervals using 20,000 bootstrap resamples of the 86 distinct dataset–question clusters, retaining all crops and ratings for each sampled question. Figure 9 summarizes all 55 pairwise Cohen's κ comparisons among the ten raters and the VLM's self-verification label. These estimates describe agreement on the balanced, consensus-screened cohort.

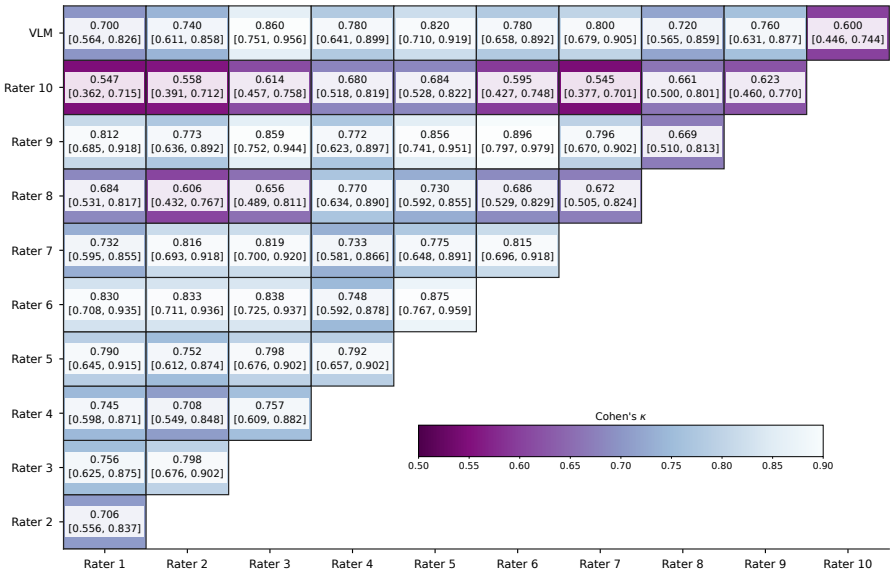


Figure 9: Pairwise Cohen's κ among ten human raters and the VLM on the 100 shared human-review items. Each cell reports κ and its 95% bootstrap interval.

C ATTENTION DIAGNOSTICS

Figure 10 shows all $N_{\text{head}} = 16$ Qwen3.5-9B heads for the V*Bench example summarized in Figure 1(c). For each head h, the map averages the conditional attention weights α_j^h for zoom-token positions $j \in \mathcal{I}_{\text{zoom}}$ over post-zoom query positions $t \in \mathcal{I}_{\text{post}}$. These weights are normalized within $\mathcal{I}_{\text{zoom}}$ at each query, as defined in Section 3.2.2, and all panels share one color range. The dashed box is the evidence annotation from the dataset.



Figure 10: Complete Qwen3.5-9B head atlas.

D ADDITIONAL RESULTS

In this section, we provide additional ablation studies, per-category analyses, extended attention diagnostics, and an analysis of the computational cost of our approach.

Table 8: Full ablation study (%).

Model	Method	HR-4K	HR-8K	VStar	CV-2D	MMStar	UltraVR	Average
GLM-4.6V-Flash	Base	$\underline{88.50}_{1.22}$	$87.28_{1.17}$	$87.59_{1.01}$	$75.39_{0.55}$	$72.63_{0.47}$	$20.42_{0.81}$	$71.97_{0.38}$
	Uniform	$87.54_{1.36}$	$\underline{87.87}_{1.17}$	$\mathbf{88.24}_{0.91}$	$\underline{75.67}_{0.55}$	$\underline{72.87}_{0.46}$	$\underline{22.12}_{0.87}$	$\underline{72.38}_{0.39}$
	Ours	$\mathbf{89.03}_{1.21}$	$\mathbf{88.69}_{1.05}$	$\underline{87.99}_{0.96}$	$\mathbf{76.26}_{0.54}$	$\mathbf{73.05}_{0.46}$	$\mathbf{22.69}_{0.86}$	$\mathbf{72.95}_{0.37}$
Qwen3-VL-8B	Base	$\underline{88.83}_{0.86}$	$82.07_{1.33}$	$88.23_{0.84}$	$\underline{83.02}_{0.49}$	$75.45_{0.41}$	$5.28_{0.61}$	$70.48_{0.33}$
	Uniform	$87.33_{0.91}$	$\underline{84.38}_{1.23}$	$\underline{89.17}_{0.94}$	$82.97_{0.31}$	$\underline{78.47}_{0.43}$	$\underline{8.10}_{0.61}$	$\underline{71.74}_{0.33}$
	Ours	$\mathbf{90.21}_{0.77}$	$\mathbf{85.06}_{0.98}$	$\mathbf{90.10}_{0.72}$	$\mathbf{84.54}_{0.31}$	$\mathbf{79.12}_{0.41}$	$\mathbf{10.56}_{0.50}$	$\mathbf{73.26}_{0.27}$
Qwen3.5-9B	Base	$87.52_{1.22}$	$86.76_{1.23}$	$90.73_{0.99}$	$\underline{78.53}_{0.59}$	$77.86_{0.40}$	$25.95_{0.92}$	$74.56_{0.39}$
	Uniform	$\underline{88.61}_{1.15}$	$\underline{89.65}_{1.25}$	$\underline{91.02}_{0.97}$	$78.24_{0.56}$	$\underline{78.83}_{0.32}$	$\underline{26.90}_{0.95}$	$\underline{75.54}_{0.38}$
	Ours	$\mathbf{89.03}_{1.01}$	$\mathbf{90.10}_{1.10}$	$\mathbf{92.14}_{0.90}$	$\mathbf{79.72}_{0.56}$	$\mathbf{78.95}_{0.42}$	$\mathbf{29.97}_{0.97}$	$\mathbf{76.65}_{0.35}$
Qwen3.8-27B	Base	$\underline{96.03}_{0.61}$	$92.92_{0.49}$	$\underline{98.21}_{0.47}$	$92.80_{0.15}$	$\underline{82.79}_{0.23}$	$\underline{57.52}_{0.89}$	$\underline{86.71}_{0.22}$
	Uniform	$\underline{96.03}_{0.61}$	$\underline{93.05}_{0.56}$	$97.63_{0.67}$	$\underline{92.92}_{0.11}$	$82.76_{0.24}$	$\underline{57.52}_{0.89}$	$86.65_{0.24}$
	Ours	$\mathbf{96.23}_{0.55}$	$\mathbf{93.89}_{0.47}$	$\mathbf{98.52}_{0.30}$	$\mathbf{93.01}_{0.13}$	$\mathbf{83.28}_{0.23}$	$\mathbf{62.23}_{1.03}$	$\mathbf{87.86}_{0.22}$
Gemma-4-31B	Base	$91.53_{0.71}$	$93.43_{0.84}$	$93.37_{0.71}$	$77.31_{0.17}$	$60.23_{0.31}$	$\underline{36.95}_{0.82}$	$75.47_{0.26}$
	Uniform	$\underline{92.08}_{0.47}$	$\underline{94.19}_{0.85}$	$\underline{94.28}_{0.52}$	$\underline{78.67}_{0.20}$	$\underline{60.38}_{0.33}$	$\mathbf{37.13}_{0.82}$	$\underline{76.12}_{0.24}$
	Ours	$\mathbf{92.63}_{0.47}$	$\mathbf{94.29}_{0.69}$	$\mathbf{94.58}_{0.43}$	$\mathbf{79.26}_{0.22}$	$\mathbf{60.99}_{0.33}$	$36.77_{0.78}$	$\mathbf{76.42}_{0.21}$

D.1 PER-CATEGORY ANALYSIS

Table 9: Per-category scores (%).

(a) HR-4K, HR-8K, VStar, and CV-2D

Model	Method	HR-4K		HR-8K		VStar		CV-2D	
		Cross	Single	Cross	Single	Direct attributes	Relative position	Count	Relation
GLM-4.6V-Flash	Base	79.71$_{2.52}$	93.30$_{1.24}$	77.14$_{2.79}$	92.85$_{0.99}$	86.97$_{1.31}$	**89.66**$_{0.00}$	72.07$_{0.65}$	**92.42**$_{0.41}$
	ResDec	78.75$_{2.32}$	92.52$_{1.19}$	77.87$_{2.70}$	93.20$_{1.04}$	87.31$_{1.22}$	**89.66**$_{0.00}$	72.34$_{0.65}$	**92.42**$_{0.41}$
	FLB	79.07$_{2.67}$	92.52$_{1.13}$	78.05$_{2.67}$	93.62$_{1.03}$	**87.83**$_{1.15}$	**89.66**$_{0.00}$	72.40$_{0.63}$	**92.42**$_{0.41}$
	Uniform	78.91$_{2.97}$	92.26$_{1.26}$	77.02$_{2.75}$	93.83$_{1.08}$	87.82$_{1.16}$	**89.66**$_{0.00}$	72.41$_{0.64}$	**92.42**$_{0.41}$
	Ours	**80.58**$_{2.61}$	**93.66**$_{1.17}$	**78.39**$_{2.61}$	**94.35**$_{0.82}$	87.49$_{1.25}$	**89.66**$_{0.00}$	**73.11**$_{0.64}$	**92.42**$_{0.41}$
Qwen3-VL-8B	Base	79.01$_{1.76}$	95.10$_{0.85}$	68.80$_{2.20}$	92.09$_{1.65}$	92.24$_{0.71}$	78.62$_{2.27}$	81.82$_{0.55}$	93.75$_{0.00}$
	ResDec	76.85$_{1.62}$	95.70$_{0.82}$	69.29$_{2.21}$	93.61$_{1.55}$	90.47$_{0.96}$	81.88$_{1.85}$	79.46$_{0.53}$	93.75$_{0.00}$
	FLB	75.30$_{1.58}$	94.07$_{0.85}$	68.11$_{1.89}$	92.67$_{1.82}$	90.44$_{0.93}$	**86.96**$_{1.55}$	81.85$_{0.45}$	93.75$_{0.00}$
	Uniform	73.95$_{1.86}$	95.88$_{0.90}$	**71.30**$_{2.02}$	94.26$_{1.57}$	90.77$_{1.03}$	85.33$_{2.12}$	81.76$_{0.35}$	93.75$_{0.00}$
	Ours	**80.40**$_{1.46}$	**96.48**$_{0.84}$	**71.30**$_{1.41}$	**95.46**$_{1.38}$	**92.32**$_{0.81}$	84.78$_{1.55}$	**83.16**$_{0.35}$	**96.88**$_{0.00}$
Qwen3.5-9B	Base	82.76$_{2.14}$	91.42$_{1.35}$	81.19$_{2.02}$	91.58$_{1.48}$	90.52$_{1.31}$	91.16$_{1.38}$	73.72$_{0.74}$	94.16$_{0.66}$
	ResDec	82.50$_{1.96}$	92.53$_{1.17}$	84.88$_{1.97}$	94.54$_{1.15}$	89.19$_{1.24}$	93.39$_{1.14}$	72.73$_{0.74}$	**94.25**$_{0.74}$
	FLB	82.15$_{2.04}$	92.17$_{0.92}$	84.27$_{1.86}$	93.03$_{1.33}$	89.60$_{1.07}$	90.71$_{1.45}$	75.04$_{0.69}$	92.81$_{0.62}$
	Uniform	82.70$_{2.05}$	93.45$_{1.25}$	**86.30**$_{2.15}$	92.54$_{1.45}$	89.73$_{1.21}$	93.63$_{1.51}$	73.89$_{0.71}$	92.39$_{0.60}$
	Ours	**83.11**$_{1.91}$	**93.89**$_{0.87}$	83.63$_{1.74}$	**95.68**$_{1.38}$	**91.22**$_{1.12}$	**93.99**$_{1.42}$	**75.24**$_{0.71}$	**94.25**$_{0.56}$
Qwen3.8-27B	Base	92.53$_{1.35}$	**98.92**$_{0.00}$	86.60$_{0.87}$	98.43$_{0.56}$	**99.50**$_{0.50}$	96.27$_{0.92}$	88.50$_{0.29}$	97.21$_{0.00}$
	ResDec	92.64$_{1.26}$	97.58$_{0.97}$	87.88$_{1.19}$	97.31$_{1.06}$	98.68$_{0.84}$	96.57$_{0.69}$	88.18$_{0.37}$	97.12$_{0.13}$
	FLB	92.32$_{1.37}$	**98.92**$_{0.00}$	86.87$_{1.03}$	97.65$_{0.00}$	98.02$_{1.13}$	**97.06**$_{0.00}$	88.63$_{0.26}$	97.21$_{0.00}$
	Uniform	92.53$_{1.35}$	**98.92**$_{0.00}$	86.42$_{1.22}$	**98.82**$_{0.00}$	98.02$_{1.11}$	**97.06**$_{0.00}$	88.45$_{0.23}$	**97.49**$_{0.00}$
	Ours	**92.97**$_{1.21}$	**98.92**$_{0.00}$	**88.22**$_{1.03}$	**98.82**$_{0.00}$	**99.50**$_{0.50}$	**97.06**$_{0.00}$	**88.90**$_{0.26}$	97.21$_{0.00}$
Gemma-4-31B	Base	89.22$_{0.98}$	93.58$_{1.02}$	91.98$_{1.04}$	94.68$_{1.28}$	95.61$_{0.76}$	88.99$_{1.48}$	62.61$_{0.33}$	**93.53**$_{0.00}$
	ResDec	90.00$_{0.97}$	93.40$_{0.89}$	91.98$_{1.04}$	94.15$_{1.39}$	95.61$_{0.90}$	87.80$_{1.22}$	65.93$_{0.41}$	**93.53**$_{0.00}$
	FLB	**90.78**$_{0.81}$	92.53$_{1.14}$	91.77$_{1.05}$	95.57$_{1.05}$	**96.82**$_{0.45}$	87.50$_{1.27}$	65.05$_{0.36}$	**93.53**$_{0.00}$
	Uniform	89.61$_{0.81}$	**94.27**$_{0.52}$	**92.80**$_{1.23}$	95.39$_{1.18}$	**96.82**$_{0.45}$	89.29$_{1.27}$	65.20$_{0.38}$	**93.53**$_{0.00}$
	Ours	**90.78**$_{0.81}$	**94.27**$_{0.52}$	92.39$_{0.86}$	**95.92**$_{1.05}$	**96.82**$_{0.45}$	**90.18**$_{0.89}$	**66.32**$_{0.41}$	**93.53**$_{0.00}$

(b) MMStar and UltraVR

Model	Method	MMStar						UltraVR			
		Coarse perception	Fine-grained perception	Instance reasoning	Logical reasoning	Math	Science & technology	AD	CCTV	RS	WSI
GLM-4.6V-Flash	Base	69.93$_{1.03}$	60.77$_{1.21}$	77.28$_{0.85}$	79.10$_{1.00}$	**88.47**$_{0.68}$	62.54$_{2.21}$	2.34$_{1.02}$	32.14$_{1.94}$	33.92$_{1.92}$	**46.40**$_{2.43}$
	ResDec	**70.11**$_{1.00}$	61.10$_{1.19}$	77.12$_{0.90}$	79.10$_{1.00}$	**88.47**$_{0.68}$	61.77$_{2.06}$	2.57$_{1.00}$	32.14$_{1.94}$	38.28$_{2.27}$	**46.40**$_{2.43}$
	FLB	69.97$_{0.97}$	60.96$_{1.22}$	77.50$_{0.95}$	79.29$_{0.98}$	**88.47**$_{0.68}$	61.77$_{2.06}$	2.48$_{1.03}$	32.14$_{1.94}$	37.88$_{2.08}$	**46.40**$_{2.43}$
	Uniform	**70.11**$_{1.00}$	60.98$_{1.23}$	77.41$_{0.85}$	**79.55**$_{0.91}$	**88.47**$_{0.68}$	**63.31**$_{2.06}$	**2.85**$_{1.08}$	**34.31**$_{1.94}$	39.07$_{2.34}$	**46.40**$_{2.43}$
	Ours	**70.11**$_{1.00}$	**61.54**$_{1.18}$	**77.66**$_{0.91}$	**79.55**$_{0.91}$	**88.47**$_{0.68}$	**63.31**$_{2.06}$	2.79$_{1.06}$	**34.31**$_{1.94}$	**41.85**$_{2.22}$	**46.40**$_{2.43}$
Qwen3-VL-8B	Base	79.07$_{0.00}$	64.38$_{1.25}$	82.81$_{0.87}$	78.67$_{0.94}$	88.69$_{1.14}$	56.60$_{0.00}$	0.45$_{0.45}$	15.00$_{5.00}$	20.00$_{0.00}$	31.82$_{4.55}$
	ResDec	**81.40**$_{0.00}$	64.38$_{1.24}$	82.81$_{0.87}$	78.67$_{0.94}$	**89.29**$_{1.12}$	62.26$_{0.00}$	**2.25**$_{0.45}$	**25.00**$_{5.00}$	20.00$_{0.00}$	**54.55**$_{0.00}$
	FLB	79.07$_{0.00}$	69.22$_{1.30}$	82.81$_{0.85}$	**81.33**$_{0.94}$	89.19$_{1.10}$	**66.04**$_{0.00}$	**2.25**$_{0.45}$	**25.00**$_{5.00}$	20.00$_{0.00}$	**54.55**$_{0.00}$
	Uniform	**81.40**$_{0.00}$	70.16$_{1.24}$	83.85$_{1.13}$	80.00$_{0.94}$	**89.29**$_{1.14}$	64.15$_{0.00}$	1.35$_{0.45}$	**25.00**$_{5.00}$	30.00$_{0.00}$	40.91$_{4.55}$
	Ours	**81.40**$_{0.00}$	**72.04**$_{1.23}$	**84.90**$_{0.87}$	80.00$_{0.94}$	**89.29**$_{1.12}$	64.15$_{0.00}$	**2.25**$_{0.45}$	**25.00**$_{5.00}$	**40.00**$_{0.00}$	**54.55**$_{0.00}$
Qwen3.5-9B	Base	72.40$_{0.52}$	69.64$_{0.99}$	80.36$_{1.16}$	84.88$_{0.84}$	88.60$_{0.63}$	73.41$_{1.23}$	8.33$_{2.02}$	24.60$_{2.58}$	40.48$_{2.22}$	31.99$_{1.13}$
	ResDec	71.18$_{0.88}$	67.12$_{0.83}$	81.75$_{1.12}$	**87.11**$_{0.90}$	87.13$_{0.84}$	**77.18**$_{1.18}$	8.21$_{1.79}$	24.60$_{2.58}$	43.37$_{1.91}$	33.14$_{1.13}$
	FLB	71.35$_{0.52}$	68.15$_{0.73}$	80.46$_{1.04}$	84.36$_{0.86}$	**90.20**$_{0.71}$	71.67$_{1.05}$	11.69$_{2.15}$	**29.37**$_{2.61}$	40.14$_{2.26}$	**34.29**$_{1.13}$
	Uniform	**72.92**$_{0.00}$	70.28$_{0.66}$	**82.00**$_{0.99}$	85.22$_{0.49}$	89.33$_{0.71}$	75.03$_{1.00}$	8.58$_{2.24}$	22.22$_{2.01}$	46.94$_{2.45}$	31.99$_{1.13}$
	Ours	71.35$_{0.52}$	**71.51**$_{0.90}$	80.20$_{1.23}$	86.25$_{0.86}$	89.77$_{0.83}$	76.03$_{1.30}$	**12.19**$_{2.04}$	25.79$_{2.84}$	**50.17**$_{2.41}$	**34.29**$_{1.13}$
Qwen3.8-27B	Base	74.99$_{0.66}$	67.19$_{0.66}$	79.90$_{0.58}$	93.39$_{0.37}$	95.96$_{0.41}$	**88.00**$_{0.00}$	79.17$_{2.38}$	44.44$_{0.00}$	58.62$_{0.00}$	**9.09**$_{0.00}$
	ResDec	74.39$_{0.51}$	66.90$_{0.93}$	80.83$_{0.66}$	**94.18**$_{0.37}$	95.66$_{0.50}$	**88.00**$_{0.00}$	79.17$_{2.38}$	**55.56**$_{0.00}$	62.07$_{0.00}$	**9.09**$_{0.00}$
	FLB	74.99$_{0.51}$	**67.60**$_{0.86}$	81.18$_{0.56}$	**94.18**$_{0.37}$	95.96$_{0.41}$	**88.00**$_{0.00}$	79.17$_{2.38}$	50.00$_{0.00}$	55.17$_{0.00}$	**9.09**$_{0.00}$
	Uniform	74.09$_{0.59}$	66.43$_{0.81}$	80.83$_{0.66}$	**94.18**$_{0.37}$	**96.55**$_{0.41}$	86.67$_{0.00}$	79.17$_{2.38}$	50.00$_{0.00}$	55.17$_{0.00}$	**9.09**$_{0.00}$
	Ours	**75.29**$_{0.59}$	67.19$_{0.66}$	**81.64**$_{0.67}$	**94.18**$_{0.37}$	95.96$_{0.41}$	**88.00**$_{0.00}$	**80.15**$_{2.76}$	**55.56**$_{0.00}$	**65.52**$_{0.00}$	**9.09**$_{0.00}$
Gemma-4-31B	Base	**53.15**$_{0.68}$	46.28$_{0.00}$	55.56$_{0.72}$	67.94$_{1.23}$	89.39$_{0.87}$	54.46$_{0.63}$	**34.69**$_{1.45}$	**34.29**$_{0.00}$	46.49$_{3.63}$	37.21$_{1.40}$
	ResDec	**53.15**$_{0.68}$	47.52$_{0.41}$	56.57$_{0.72}$	**68.41**$_{1.32}$	89.39$_{0.87}$	54.46$_{0.63}$	**34.69**$_{1.45}$	30.00$_{1.43}$	46.49$_{3.63}$	**38.37**$_{1.41}$
	FLB	52.76$_{0.56}$	46.69$_{0.41}$	55.56$_{0.72}$	67.94$_{1.23}$	**89.90**$_{1.01}$	54.91$_{0.45}$	**34.69**$_{1.45}$	31.43$_{0.00}$	46.49$_{3.63}$	36.43$_{1.40}$
	Uniform	52.36$_{0.39}$	46.83$_{0.39}$	56.57$_{0.72}$	**68.41**$_{1.32}$	88.38$_{0.87}$	55.21$_{0.88}$	**34.69**$_{1.45}$	31.43$_{0.00}$	**48.25**$_{3.63}$	**38.37**$_{1.40}$
	Ours	52.76$_{0.56}$	**48.35**$_{0.73}$	**57.58**$_{0.72}$	67.94$_{1.23}$	89.39$_{0.87}$	**55.36**$_{0.64}$	**34.69**$_{1.45}$	**34.29**$_{0.00}$	46.49$_{3.63}$	36.82$_{1.29}$

D.2 Additional attention diagnostics

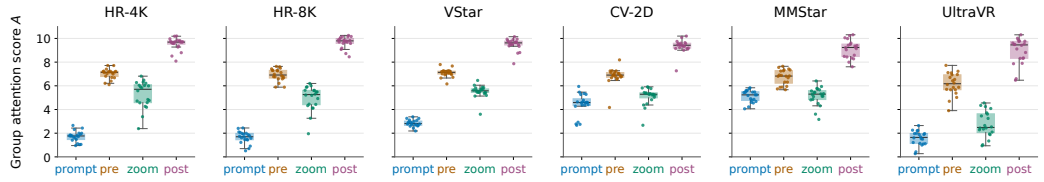


Figure 11: Group attention scores A of Qwen3.5-9B across six datasets. Each dataset includes 20 trajectories with one zoom-in tool call ($r = 1$).

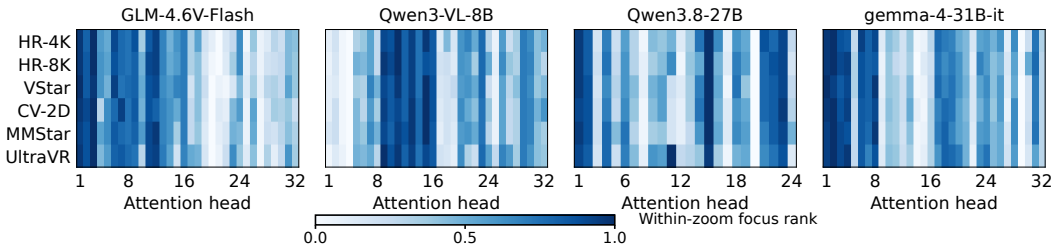


Figure 12: Within-zoom focus rank for four additional models across six datasets.

D.3 Computational cost

Following the distinction between crop-processing overhead and decoding time in Khayatkhoei et al. (2025), we separate fixed-sequence decoding cost from natural continuation length and elapsed time. To extend the 127-step comparison in Table 4, we measure Qwen3.5-9B in BF16 over 1,024 cached steps on RTX A6000 GPUs (48 GB). We sample two trajectories from distinct questions per dataset, with at least 1,050 continuation tokens and at most 32,768 prefix tokens. Base, FLB, ResDec, and our method receive identical images, prefixes, and continuation tokens. For each question, all methods run on the same GPU for three repetitions. After averaging repetitions within questions and weighting questions equally, our method takes 43.01 ms per cached token, compared with 41.57 ms for ResDec and 41.14 ms for FLB. Mean paired overheads are 3.43% and 4.52%, respectively.

Table 10 reports natural continuation lengths and elapsed times for 100 trajectories per backbone. We sample 100 trajectories from each backbone, independently of correctness, attention, output length, and runtime. All four methods use the same sampled trajectories, with matching source identity, reused conversation, and question metadata. Generated-token counts are summed across continuation turns, including reasoning and later tool requests, and exclude acquisition and self-verification. Elapsed time includes source reconstruction, prefills, decoding, and later zoom execution. Model loading is excluded.

Table 10: Generated tokens and continuation elapsed time for 100 sampled trajectories per backbone, matched across all four methods.

		ResDec		FLB		Uniform		Ours	
Backbone	n	Tokens	Seconds	Tokens	Seconds	Tokens	Seconds	Tokens	Seconds
GLM-4.6V-Flash	100	87.1	4.85	84.8	5.19	89.5	5.19	84.0	3.98
Qwen3-VL-8B-Instruct	100	93.6	9.56	82.0	8.99	80.8	9.45	87.9	9.50
Qwen3.5-9B	100	1093.1	53.09	1078.5	57.00	1090.1	55.59	824.5	46.21
Qwen3.8-27B	100	198.3	32.15	198.1	32.24	185.8	31.93	190.3	32.36
Gemma-4-31B-it	100	305.7	27.11	308.8	26.81	287.5	24.93	276.2	25.30