

VDAHM: Vision-Delimiter-aware Attention Head Modulation for Enhancing Multi-Image Understanding in Large Vision-Language Models

Zhibiao Liang¹, Junyu Cheng², Ziqian Zeng^{1*}

¹South China University of Technology, China,

²Xiamen University, China,

Correspondence: zqzeng@scut.edu.cn

Abstract

Although Large Vision-Language Models (LVLMs) excel at single-image understanding tasks, their performance often degrades significantly when processing multi-image inputs due to unintended cross-image information leakage, in which visual content from different images is erroneously fused and image-specific contextual information is lost. Current approaches attempt to mitigate this by inserting special vision delimiter tokens between images during input formatting to help models distinguish image boundaries and preserve context. However, our analysis reveals that such delimiters alone yield limited gains because models do not sufficiently perceive or respond to these boundary signals. To address this, we propose Vision-Delimiter-aware Attention Head Modulation (VDAHM), a training-free method that modulates the output of delimiter-sensitive attention heads, reorienting the model’s internal attention flow to strengthen its structural perception of boundary signals. Extensive evaluations across multiple benchmarks, including Mantis, BLINK, MuirBench, and Qbench2, demonstrate that VDAHM consistently improves LVLMs performance with virtually no additional computational cost. The source code is available at <https://github.com/ZeroNLP/VDAHM>.

1 Introduction

Large Vision-Language Models (LVLMs) have excelled in single-image understanding (Liu et al., 2023; Li et al., 2023). However, when faced with multi-image inputs, their performance drops sharply, a scenario increasingly common in real-world applications (Jiang et al., 2024).

This performance gap has spurred research interest in enhancing multi-image understanding in LVLMs. Among the emerging approaches, training-free methods stand out for their efficiency

Single-Image Inference	Multi-Image Inference
Q: Please describe this image in detail. A: The image captures a lively black and white dog, likely a Border Collie, sprinting.	Question: How many dogs are there in <image 1>, <image 2> and <image 3> in total? Images Input Formatting
Q: Please describe this image in detail. A: A total of three joyful dogs are featured in this image, gathered closely together.	Vanilla There are one and three cats in the first and second image respectively, while there are three cats in the third image, therefore three dogs in total.
Q: Please describe this image in detail. A: A total of three attentive cats are beautifully framed resting side-by-side along the ...	Ours I can see one dog in the first image, three dogs in the second, and three cats in the last one. There are totally four dogs in these images.

Figure 1: Although vanilla models exhibit limited boundary awareness through delimiter tokens, their performance remains suboptimal. Our approach mitigates this by enhancing the model’s perception of delimiter presence. This strengthens inter-image boundary detection and ultimately enhances multi-image understanding.

and compatibility with existing models (Tian et al., 2025; Park et al., 2025). These training-free techniques focus on two strategies, compressing redundant visual tokens and promoting interaction between images. For instance, Zeng et al. (2025a) assume that redundant information across images impairs the model’s multi-image reasoning capability, so they compress redundant visual tokens adaptively to improve overall image understanding. Tian et al. (2025) assume that strict causal masking hinders cross-image interaction and thus degrades multi-image understanding, so they relax constraints between images to encourage cross-image communication. Meanwhile, Qiao et al. (2025) shares aggregated images’ sink-token attention across images to boost effective visual focus.

However, whether aiming to reduce redundant tokens in multi-image inputs or to enhance cross-image interaction, these methods typically treat multi-image inputs as a unified and loosely coupled sequence, implicitly assuming that greater cross-image interaction necessarily leads to better performance. These approaches appear to ignore a key that clearly recognizing image boundaries is just as crucial for accurate reasoning as enhancing inter-

*Corresponding author

actions between images, especially when multiple images are arranged into a single long sequence. Our idea does not aim to eliminate all cross-image interactions. Instead, it explores how enhancing the model’s sensitivity of image boundaries can improve its ability to multi-image understanding and research in this area is still relatively limited.

Notably, some LVLMs enhance boundary awareness by introducing special tokens (e.g., `<vision_start>`). Since attention intensity serves as a key indicator of token sensitivity (Vaswani et al., 2023), we can analyze the role of these special tokens through attention heatmaps. As shown in Analysis 3, we find that these vision delimiter tokens exhibit properties similar to *sink token* in LLMs, typically attracting strong attention, thereby behaving as "attention sinks" at image boundaries to draw the model’s focus toward image boundaries. However, cross-image information leakage still persists. As illustrated in Figure 1, responses concerning the first image remain contaminated by content from the third image, indicating that merely inserting vision delimiter tokens is insufficient to fully mitigate cross-image leakage (Park et al., 2025). To strengthen the role of vision delimiter tokens, we focus on enhancing the model’s sensitivity to these boundary signals, an objective that can be effectively realized through the modulation of attention weights. Given that multi-head attention mechanisms often contain specialized heads with distinct behavioral patterns (Yu et al., 2024), we propose precisely modulating delimiter-sensitive heads to reorient attention patterns and then boost model’s boundary signal perception.

To identify attention heads sensitive to vision delimiter tokens, we propose a sensitivity metric measuring how each head’s output changes when vision delimiter tokens are replaced during generation. Our analysis reveals that delimiters sensitivity is rare, only a small subset of attention heads exhibit significant responsiveness to the presence or absence of vision delimiter tokens, not distributed globally. Therefore, targeting and enhancing these sensitive heads holds a low-cost way to improve the model’s perception of visual structural boundaries, strengthening structural guidance for downstream multimodal tasks.

Building on these insights, we propose Vision-Delimiter-aware Attention Head Modulation (VDAHM), a training-free method that enhances multi-image understanding by identifying attention heads that are sensitive to delimiters and modu-

lating their influence during generation. This reorients the internal attention flow toward structural boundary pathways without altering token representations, improving alignment with delimiter cues. As shown in Figure 1, the model can clearly distinguish the content corresponding to each individual image by enhancing the model’s perception of image boundaries.

- We propose the sensitivity metric to detect attention heads in LVLMs that are sensitive to vision delimiter tokens.
- We propose the VDAHM, a training-free and plug-and-play method that proactively strengthens multi-image understanding of LVLMs by adaptively identifying and modulating key attention heads during generation.
- Extensive experiments demonstrate VDAHM gains great performance on widely adopted multi-image benchmarks with negligible additional time cost.

2 Related Work

2.1 Multi-image Understanding

Recent large multimodal models, such as LLaVA-Next (Liu et al., 2024), MiniGPT-4 (Zhu et al., 2023), mPLUG-Owl (Ye et al., 2024), and Qwen-VL (Bai et al., 2023) support multi-image inputs by embedding visual tokens from multiple images alongside text tokens into a single causal sequence, enabling all tokens to share a unified attention space. This design has spurred growing interest in enhancing multi-image understanding, which is broadly divided into training-based and training-free methods. Training-based approaches typically involve fine-tuning pretrained models on specialized multi-image instruction datasets (Chen et al., 2025; Zhang et al., 2025; Liu et al., 2023). For instance, Jiang et al. (2024) demonstrate significant improvements in multi-image reasoning capabilities through supervised fine-tuning. However, such methods require costly data curation and heavy re-training. At the same time, several training-free methods have been proposed to circumvent these limitations. SoFA (Tian et al., 2025) enhances multi-image understanding by interpolating causal attention with bidirectional attention, thereby enabling free information flow between image tokens. FOCUS (Park et al., 2025) reduces cross-image information leakage through contrastive decoding

that disentangles outputs based on image-specific context. Nevertheless, these approaches introduce new trade-offs that SoFA requires architectural modifications, while FOCUS incurs high inference latency due to multiple forward passes (e.g., $n + 1$ for n images), limiting real-time applicability.

In contrast, our method enhances multi-image understanding without any additional training or architectural changes, offering a lightweight and scalable solution.

2.2 Sink Tokens in Large Vision-Language Models

Recent studies reveal *attention sinks*, which are persistent attention patterns that emerge during LLMs pretraining and concentrating on fixed positions or special tokens (Barbero et al., 2025). These sinks absorb surplus attention across layers and heads, acting as implicit bias vectors rather than encoding semantic content (Gu et al., 2025). In multimodal models, sinks often localize to background patches, motivating techniques to redirect attention toward salient regions (Kang et al., 2025). Within each image, certain visual tokens consistently draw disproportionately high attention, even though they contribute minimally to the model’s final predictions (Sun et al., 2024). The phenomenon is particularly pronounced in earlier images, where stronger "sink" behavior emerges due to causal masking in the attention mechanism (Wu et al., 2025). Sink mass accumulates unevenly across the sequence, leading to a skewed distribution of attention weights (Qiao et al., 2025).

Inspired by findings above, we extend the concept of attention sinks to the domain of multi-image understanding. We observe that in multi-image settings, vision delimiter tokens behave like sink token. They attract a substantial portion of attention and exhibit behavior closely analogous to image boundaries.

3 Analysis

Although using vision delimiter tokens (e.g., `<img>` and `</img>` in Intern3-VL) to separate images, empirical evidence reveals persistent leakage despite these safeguards, suggesting a fundamental gap in our understanding of how these tokens govern visual context boundaries. This section investigates whether delimiters behave as intended and identifies their limitations.

The Critical Role of Vision Delimiter Tokens

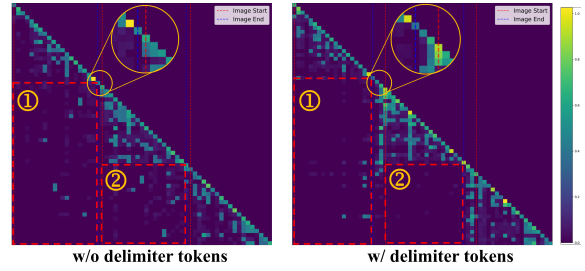


Figure 2: Attention maps without (left) and with (right) vision delimiter tokens. Triangular boundary structures are weakened upon delimiters removal (left), while residual cross-image attention persists near delimiter positions (the region ①, right).

in Boundary Formation. To investigate the structural dependency between vision delimiter tokens and image separation, we conduct ablation studies by systematically toggling their presence or absence. The images used in the analysis are presented in the Appendix J. As illustrated in Figure 2, vision delimiter tokens induce structured attention boundaries. Attention patterns are structured into discrete triangular blocks, where each block encapsulates the token representations tied to one input image. Crucially, removing these vision delimiter tokens severely impairs the model’s ability to maintain clear inter-image boundaries. As illustrated in Figure 2 (left), when vision delimiter tokens are removed, the originally sparse attention region ② becomes denser, revealing semantic contamination across visual boundaries. Table 5 in the Appendix A further quantifies this effect, showing that the attention intensity in region ② increase by approximately 122.83% compared to its value before delimiters removal, demonstrating the critical role of delimiters in restructuring attention.

Quantitative evaluations align with these internal structural changes. Removing vision delimiter tokens results in an **8.7%** performance drop on multi-image reasoning tasks (Table 4 in the Appendix A). This degradation directly correlates with the loss of boundary structure in attention maps, confirming that delimiters are not merely decorative but functional enablers of image separation.

The Limitation of Vision Delimiter Tokens.

Prior analysis confirms that vision delimiter tokens serve to establish image boundaries, as evidenced by both attention patterns and model performance. However, attention maps reveal that their boundary separation capability remains insufficient in certain local regions, especially near the delimiter tokens

themselves, where attention distributions still exhibit scattered and cross-boundary characteristic, as illustrated in the region ① in (Figure 2, right). This *attention leakage* indicates that while delimiters can construct a coarse-grained segmentation framework across images, they remain insufficient at the fine-grained level for fully isolating semantic information between adjacent image regions. These observations motivate the development of methods that enhance model sensitivity to delimiter tokens, enabling them to more reliably serve their intended function as structural separators between images.

4 Preliminary

Multi-head Attention. Each transformer layer employs multi-head attention to integrate contextual information across tokens. For the i -th attention head in layer l , the input $X_{l,i}$ is linearly projected into query, key, and value representations via learned matrices:

$$Q_{l,i} = X_{l,i}W_{l,i}^Q, K_{l,i} = X_{l,i}W_{l,i}^K, V_{l,i} = X_{l,i}W_{l,i}^V, \\ A_{l,i} = \text{softmax}\left(\frac{Q_{l,i}K_{l,i}^T}{\sqrt{d_k}}\right)V_{l,i}, \quad (1)$$

where $W_{l,i}^Q, W_{l,i}^K, W_{l,i}^V$ are projection weights, and d_k is dimensionality of each query/key vector. $A_{l,i}$, attention output for this head, captures token-level dependencies through scaled dot-product attention and applies them to the value sequence.

Layer-wise Aggregation. The outputs from all attention heads in a layer are concatenated and linearly projected to the model’s hidden dimension:

$$\text{MHA}_l(X_l) = \text{Concat}(A_{l,1}, \dots, A_{l,n_h})W_l^O, \quad (2)$$

where n_h is the number of attention heads in each layer l , X_l is the shared input to the entire attention block, and W_l^O is the trainable output projection matrix.

LVLM Generation Process. The model encodes input images into vision tokens x_V , maps them to the text embedding space via a lightweight connector, and concatenates them with the tokenized prompt x_T . At each step t , it autoregressively selects the next token y_t :

$$y_t = \arg \max_{y \in \mathcal{V}} p_\theta(y | y_{<t}, x_V, x_T), \quad (3)$$

where $y_{<t}$ denotes prior generated tokens and $\mathcal{V}$ is the decoder’s shared vocabulary.

Attention Head Output during Generation.

We denote the output of the i -th attention head in layer l at generation step t as:

$$A_{l,i}^{(t)} = \text{Atten}_{l,i}(y_{<t}, x_V, x_T), \quad (4)$$

where $\text{Atten}_{l,i}(\cdot)$ follows the standard scaled dot-product attention (Eq. 1), applied to the current context that including visual tokens x_V , text prompt x_T , and previously generated tokens $y_{<t}$.

5 Method

Multi-head attention mechanisms often develop specialized heads (Yu et al., 2024), some of which act as structural anchors for delimiter tokens. VDAHM leverages this intrinsic property by modulating delimiter-sensitive heads to steer the internal attention flow toward image boundaries, thereby enhancing the model’s sensitivity to these boundary signals.

5.1 Vision-Delimiter-aware Attention Head Identification

Sensitivity Metric. We quantify the sensitivity of attention heads to vision delimiter tokens by comparing their outputs under two conditions. Specifically, we contrast the original visual stream that is wrapped by the model’s default vision delimiter tokens (denoted x_V^{deli}), with a modified version in which the delimiter tokens are replaced by a neutral placeholder (denoted x_V^{pad}). Then, the sensitivity metric quantifies the change in each head’s output at generation step t :

$$S_{l,i} = d\left(A_{l,i}^{(t)}(y_t | y_{<t}, x_V^{\text{deli}}, x_T), A_{l,i}^{(t)}(y_t | y_{<t}, x_V^{\text{pad}}, x_T)\right), \quad (5)$$

where d represents the Euclidean distance measure (Tabak, 2014). A larger value indicates the head reacts strongly to the presence or absence of vision delimiters. What’s more, we adopt $\langle \text{pad} \rangle$ as neutral placeholder for a key reason that unlike alternative like $\langle \text{mask} \rangle$, $\langle \text{pad} \rangle$ is typically masked out or assigned zero attention weights in causal contexts (Himelstein et al., 2025). This ensures that any observed changes in attention patterns stem purely from the structural role of delimiters.

Figure 3 presents a visualization of the sensitivity scores computed for each attention head across all layers of the model. Specifically, we evaluate Qwen2.5-VL by providing it with images and the instruction shown in the Appendix J. Then, we compute the sensitivity scores at the moment of predicting the first token of the generated description.

The results reveal that only a small subset of attention heads show significantly elevated sensitivity scores, indicating the emergence of specialized attention heads that are selectively sensitive to structural cues associated with visual delimiter tokens. More examples can be found in the appendix K.

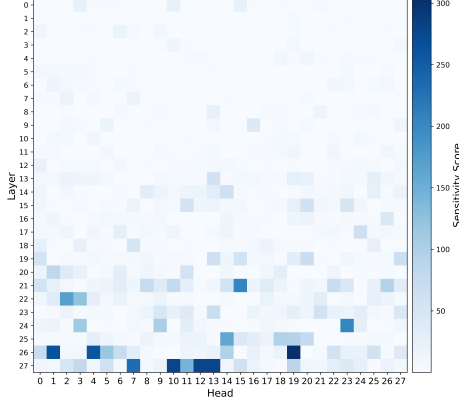


Figure 3: Sensitivity scores of all the attention heads among all layers at one generation step.

5.2 Vision-Delimiter-aware Attention Head Modulation

According to the sensitivity results, only a small subset of attention heads exhibit strong responses to vision delimiter tokens. Thus, selectively enhancing these sensitive heads offers a low-cost way to improve the model’s perception of visual structural boundaries. However, we find that indiscriminately modulating all attention heads with high sensitivity scores degrades the model’s performance (Appendix G). Further analysis reveals that among the high-scoring heads, some exhibit overreactive responses to delimiters, enhancing such overreactive heads undermines model stability. Consequently, we need to adopt a pruning strategy that retains only those heads that activate stably in response to delimiters. For pruning aberrant heads, we want to suppress such outliers by setting $S_{l,i} = 0$, provided they satisfy a joint outlier criterion:

$$\begin{cases} S_{l,i} > \mu(S_{l,*}) + \sigma(S_{l,*}), \\ E_{l,i} > \mu(E_{l,*}) + \sigma(E_{l,*}), \end{cases} \quad (6)$$

where $E_{l,i} = \|A_{l,i}(y_t|y_{<t}, x_V^{\text{pad}}, x_T)\|^2$.

μ and σ denote the mean and standard deviation computed across heads within the same layer. The choice of $\mu + \sigma$ is motivated by the 3σ rule in statistics (Bayram and Ahmed, 2023), allowing us to retain only those attention heads that exhibit

stable activation in response to vision delimiter tokens. After pruning aberrant outliers, we must redistribute the attention flow to consolidate structural boundaries. Instead of a naive global intervention, we modulate the specific routing capacity of the retained delimiter-sensitive heads. By introducing a recalibration factor α , we actively shift the attention mass toward these structural heads. As will be formalized in Section 5.3, this operation is not a mere scalar multiplication, but a geometric reorientation of the information flow entering the subsequent Feed-Forward Networks.

$$\tilde{A}_{l,i} = \begin{cases} \alpha \cdot A_{l,i} & \text{if } i \in H_l, \\ A_{l,i} & \text{otherwise,} \end{cases} \quad (7)$$

$$\text{where } H_l = \{i \mid S_{l,i} > \text{median}(S_{l,*})\}.$$

Layer-wise Application of VDAH M. For layer l , head selection and enhancement are completed before the model proceeds to $l + 1$, ensuring that all previous layers from one to $l - 1$ are already enhanced when sensitivity metrics for layer l are computed. This layer-by-layer processing approach effectively mitigates the issue commonly observed in most existing methods, where simultaneous multi-layer processing causes fluctuations in sensitivity scores at shallow layers, leading to drastic oscillations in attention at deeper layers (He et al., 2024; Li et al., 2024). As a result, cross-layer interference is minimized and the gradient dynamics during inference becomes more stable.

KV Cache Compatibility via Fixing Head Selection at Initialization. Fixing enhanced heads at decoding step $t = 1$ is not merely for efficiency, but also a necessary consideration for KV caching. Since cached keys and values are reused without re-projection, any mid-generation change to the active head set would induce misalignment between past and current KV projections, undermining the validity of cache reuse (Wang et al., 2025). VDAH M explicitly avoids this violation by fixing the head selection at initialization, thereby preserving the integrity of cached representations throughout the entire generation sequence. This static policy ensures deterministic attention behavior under caching constraints, enabling consistent and reliable generation throughout the decoding sequence.

Annotation-Free and Sample-Adaptive Head Selection. Unlike existing methods relying on external annotations or global head rankings (Yu et al.,

2023; Fang et al., 2024), VDAHM identifies the heads we need purely via internal activation patterns. For each layer l and head i , the joint outlier criterion in Equation (6) filters aberrant heads using with two intrinsic signals, $S_{l,i}$ and $E_{l,i}$, both computed from the model’s forward pass. The pruning rule is deterministic and sample-specific, requiring neither prior annotations nor batch-wise statistics, and without any post-processing calibration.

5.3 Attention Output Reorientation

Scaling the contribution of specific attention heads within a layer provides a principled mechanism to steer downstream computation toward salient features. Recent studies have demonstrated the effectiveness of intervening in internal representations of LVLMS for regulating model behavior. For instance, RFI dynamically predicts input-specific latent intervention vectors to mitigate object hallucination (Cheng et al., 2026). Different from such representation-level interventions, VDAHM performs a more fine-grained intervention at the attention-head level, selectively reorienting downstream computation toward delimiter-sensitive directions.

To formalize this intuition, consider the post-attention activations fed into the FFN module immediately following the MHA block in layer l , which can be expressed as:

$$\begin{aligned} Z_l &= \text{RMSNorm}(\hat{X}_l + \text{MHA}_l(X_l)) \\ &= \hat{g}_l \cdot \frac{\hat{X}_l + \text{MHA}_l(X_l)}{\|\hat{X}_l + \text{MHA}_l(X_l)\|}, \end{aligned} \quad (8)$$

where $\hat{g}_l$ is the learned RMSNorm gain and $\hat{X}_l$ denotes the pre-normalized residual stream entering layer l . RMSNorm ensures $\|Z_l\| = \hat{g}_l$, so any post-training rescaling that preserves the vector direction of $\hat{X}_l + \text{MHA}_l(X_l)$ leaves Z_l unchanged, whereas selectively modulating a head rotates this direction toward its value. Thus attention head reorientation is equivalent to steering the FFN input direction while keeping its norm fixed, providing a clean mechanism for injecting head-specific inductive bias without disrupting normalization.

Proposition 1 *Consider layer l in a LVLMS, and let h denote the index of the attention head to be modulated. Let $\tilde{Z}_l$ be the input to the FFN module when the output of head h is scaled by a factor $\alpha > 1$, i.e., $\tilde{A}_{l,h} = \alpha \cdot A_{l,h}$. Denote Z_l the original FFN input using the unmodified output, and by $Z_{l,h}$ the*

pseudo-input generated using only the contribution from head h . Then, the following inequality holds that $\cos(\tilde{Z}_l, Z_{l,h}) > \cos(Z_l, Z_{l,h})$.

The formal proof is provided in Appendix B. This proposition establishes that modulating specific attention heads’ output effectively reorients the representation space, specifically aligning the FFN input Z_l with the directional influence of the enhanced head. This theoretical insight underpins the mechanical validity of our head enhancement strategy. Furthermore, a detailed analysis in Appendix C elucidates why modulating these delimiter-sensitive heads significantly impacts performance, thereby substantiating the necessity of our approach. An overview of the proposed sensitivity metric and the VDAHM method is illustrated in Figure 4.

5.4 Empirical Evidence

To evaluate the effectiveness of our proposed method, we conduct experiments using Qwen2.5-VL, with particular emphasis on visually demonstrating the benefits introduced by our approach.

Mitigation of Cross-Image Information Leakage. We analyze attention maps to assess whether our method successfully suppresses unintended cross-image information flow. Specifically, we quantify such interactions by computing the mean attention score between all token pairs originating from distinct images. As illustrated in Figure 5, our method significantly reduces the attention leakage problem near the image boundary. Notably, before applying the scaling mechanism, attention distributions at the boundary between Image 2 and Image 1 exhibit scattered and cross-boundary characteristics. After applying the scaling mechanism, this leakage of cross-image attention is substantially diminished, resulting in much clearer visual separation.

6 Experiment

6.1 Benchmarks and Settings

Multi-Image Benchmarks. To comprehensively assess the effectiveness of our method, we evaluate it across four established and representative multi-image benchmarks. Mantis-Eval (Jiang et al., 2024) is a benchmark to evaluate multi-image understanding capabilities in VLMs, covering eight multi-image tasks and six single-image baselines focused on reasoning tasks such as co-reference resolution, visual comparison and temporal inference.

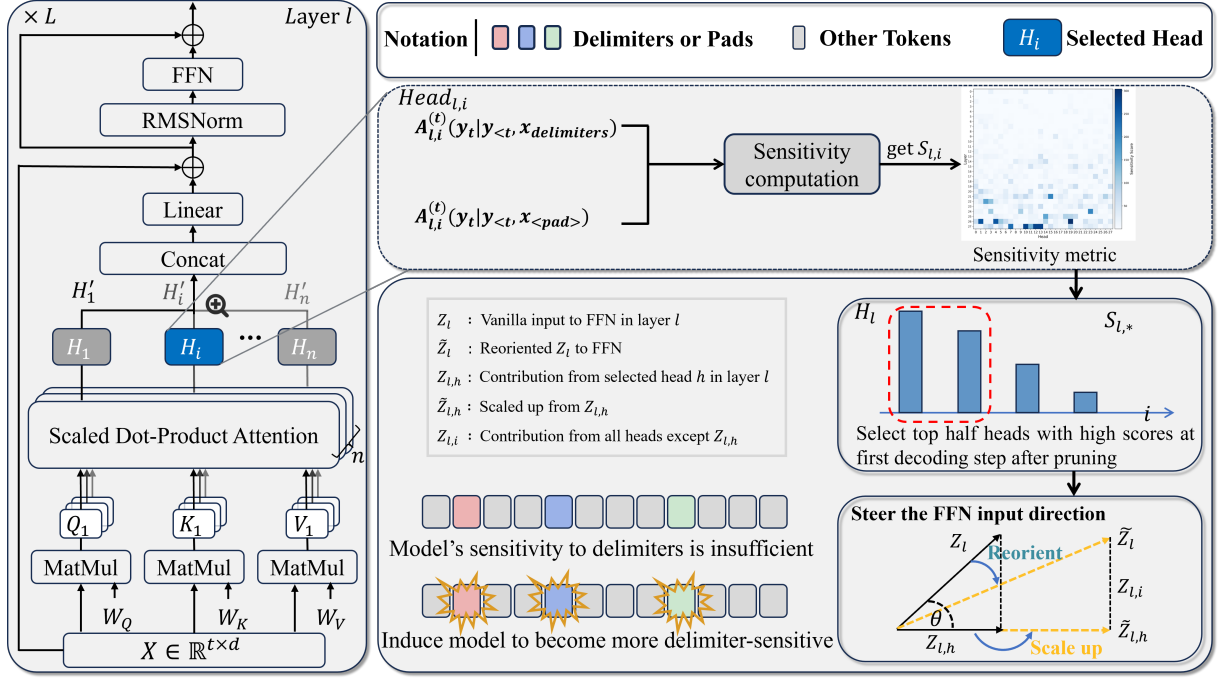


Figure 4: The illustration of the proposed sensitivity metric and the VDAH approach to enhance multi-image understanding in LVLMs. For each layer, We select the attention heads that are sensitive to visual delimiter tokens based on sensitivity scores, denoted as H_l , and then modulating their influence on model’s multimodal reasoning.

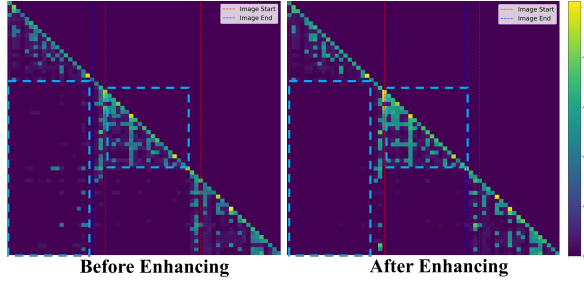


Figure 5: Qualitative comparison of attention maps before and after using VDAH.

BLINK (Fu et al., 2024) introduces a novel benchmark focused specifically on core visual perception abilities, capabilities often underrepresented or absent in existing standard multimodal evaluations. MuirBench (Wang et al., 2024) is a comprehensive benchmark for multi-image understanding, featuring 12 distinct task categories, 2,600 carefully designed questions, and 11,264 diverse images. It evaluates model performance across key domains including spatial reasoning, diagram interpretation and image retrieval. Qbench2 (Wu et al., 2024) evaluates low-level visual perception in multimodal language models (MLLMs) by extending assessment beyond single-image tasks to pairwise image inputs, thereby enabling the measurement of cross-image reasoning capabilities such as object corre-

spondence and spatial alignment, which are critical for robust and reliable multimodal understanding.

Implementation Details. For the multi-image tasks, we used Qwen2.5-VL(3B, 7B, 32B) (Bai et al., 2025), InternVL3(1B, 2B, 8B, 14B) (Zhu et al., 2025), MiMo-VL(7B) (Xiaomi, 2025) and gemma-3(4B) (Team et al., 2025) as our baseline LVLMs. To determine the optimal hyperparameters for each model, we utilize a 10% subset of the Mantis-Instruct training set for validation. Further details can be found in Appendix D. For advanced baselines, we adopt AVAM(Zeng et al., 2025a), which filters redundant visual information, and DelimScaling(Lee et al., 2026), which enhances delimiters’ hidden states. We faithfully reproduced all baseline methods based on their specific implementation details and hyperparameter configurations documented in their appendix.

6.2 Result

Performance on Mantis, BLINK and Muir-Bench. As shown in Table 1, our method consistently improves performance across all model families, including Qwen2.5-VL, InternVL3, MiMo-VL and Gemma3. For example, on the Mantis-Eval benchmark, the Qwen2.5-VL-3B model improves from 59.91 to 63.53, and the MiMo-VL-7B-RL model improves from 74.70 to 75.80. Qwen2.5-

Table 1: Performance across three multi-image benchmarks (Mantis, BLINK, MuirBench). Applying our method to the Qwen2.5-VL, InternVL3, MiMo-VL and Gemma3 Vision model families generally improves results.

Dataset	Model	Qwen2.5-VL			InternVL3				MiMo-VL	Gemma3
		3B	7B	32B	1B	2B	8B	14B	7B	4B
Mantis	Baseline	59.91	68.66	68.20	47.00	52.07	67.28	71.89	74.70	51.15
	+ AVAM	61.25	69.37	69.42	48.96	53.89	69.15	72.74	75.10	51.53
	+ DelimScaling	63.13	69.12	70.05	49.77	54.38	69.12	72.81	75.30	51.73
	+ Ours	63.53	70.97	71.89	48.84	55.75	70.96	73.51	75.80	52.39
BLINK	Baseline	47.60	56.40	59.40	42.30	48.60	55.50	60.30	62.40	38.00
	+ AVAM	48.10	57.38	61.32	42.67	49.54	57.12	60.73	63.24	38.82
	+ DelimScaling	47.94	57.23	61.52	43.07	48.23	56.27	61.20	63.55	38.62
	+ Ours	48.37	57.85	62.36	44.11	50.27	58.95	62.82	64.35	39.43
MuirBench	Baseline	37.31	45.23	53.12	28.62	27.69	36.88	42.42	62.80	35.70
	+ AVAM	42.13	47.17	53.74	29.02	28.07	37.23	42.51	63.12	36.20
	+ DelimScaling	42.42	48.15	53.82	29.38	27.65	36.92	42.58	63.88	36.92
	+ Ours	44.60	50.12	54.34	29.76	29.33	38.14	43.86	64.50	37.60

VL-7B improves from 45.23 to 50.12 on the MuirBench. Notably, performance gains appear across models of various sizes, from small-scale (e.g., 0.5B) to large-scale (e.g., 32B), indicating that the proposed vision-delimiter-aware heads scaling method is effective regardless of model capacity. And more interesting experiments’ result can be found in Appendix H and Appendix I. These consistent improvements across diverse models and benchmarks for multi-image understanding highlight the generality and practicality of our method.

Performance on Qbench2. To further evaluate our method on low-level visual perception tasks, we report results on Qbench2. As shown in Table 2, our method consistently outperforms baseline in these models. For example, on MiMo-VL-7B, our method improves the score from 71.50 to 74.70, demonstrating its effectiveness even on fine-grained pairwise comparison tasks.

Table 2: Performance comparison on Qbench2 on models Qwen2.5-VL-3B, MiMo-VL-7B, and Gemma3-4B.

Method	Qwen2.5-VL-3B	MiMo-VL-7B	Gemma3-4B
Baseline	62.70	71.50	50.10
+ AVAM	63.40	73.60	51.20
+ DelimScaling	63.30	72.88	50.80
+ Ours	64.20 ($\uparrow$ 1.50)	74.70 ($\uparrow$ 3.20)	52.90 ($\uparrow$ 2.80)

Hyperparameter Sensitivity. As shown in Figure 6, we evaluate the sensitivity of VDAHM to the scaling factor α on the Mantis benchmark over a broad range of values. VDAHM consistently outperforms the vanilla baseline under most settings, demonstrating that its effectiveness is not

highly sensitive to a particular choice of α . Moderate scaling factors generally yield larger improvements, with the best performance observed around $\alpha = 1.6$, whereas further increasing the scaling strength results in diminishing gains. This trend suggests that appropriately amplifying delimiter-sensitive heads strengthens boundary-aware information flow without excessively perturbing the model’s original attention dynamics, highlighting the robustness of VDAHM to the choice of α .

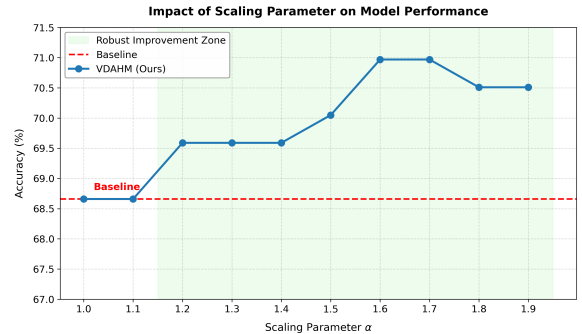


Figure 6: Sensitivity on hyperparameter α

Inference Latency. Inference efficiency is an important consideration for the practical deployment of large models, motivating techniques such as adaptive early exiting (Zeng et al., 2024) and accelerated decoding (Zeng et al., 2025b). In our setting, we specifically examine whether VDAHM preserves the lightweight nature of training-free multi-image inference. Table 3 compares the inference latency of Qwen2.5-VL-7b before and after applying our method across varying new token lengths, where each configuration was evalu-

ated through 20 independent trials. Remarkably, at 100 new tokens, VDAHM introduces only a 1.04x overhead, demonstrating the exceptional efficiency of our approach. This lightweight overhead is achieved through just one additional forward pass when predicting the first token for getting some specific attention heads, without introducing iterative or per-step computation.

Table 3: Comparison of the latency (s) of Qwen2.5-VL-7b before and after applying VDAHM in generating tokens of varying lengths on an NVIDIA A5000 GPU. "±" represents the standard deviation.

New Tokens	Qwen2.5-VL	+ VDAHM	Ratio
5	1.78 ± 0.065	2.05 ± 0.072	1.15x
25	2.27 ± 0.017	2.58 ± 0.052	1.13x
50	3.04 ± 0.015	3.32 ± 0.052	1.09x
75	3.71 ± 0.020	4.04 ± 0.039	1.08x
100	4.46 ± 0.031	4.67 ± 0.021	1.04x

7 Conclusion

In this work, we address the issue of cross-image information leakage in multi-image input settings by analyzing the role of vision delimiter tokens, which are responsible for separating visual inputs. Based on this analysis, we introduce the sensitivity metric, which quantifies the sensitivity of attention head outputs to vision delimiter tokens. And we propose VDAHM, a training-free approach that modulates vision delimiter-aware attention heads to enhance multi-image understanding. It is easy to integrate and introduces no additional training.

Limitations

Our method is currently tailored to multimodal large models that employ explicit vision delimiters. Consequently, its applicability is constrained in scenarios involving unstructured sequential inputs. Future work should focus on extending our approach to intervene at the level of the model’s overall architecture, enhancing its adaptability.

Ethics Statement

Our work presents minimal potential for negative social impact, primarily due to the use of publicly available datasets and models. This accessibility inherently reduces the risk of adverse effects on individuals or society.

Acknowledgement

This work was supported by National Natural Science Foundation of China (62406114) and Guangdong Basic and Applied Basic Research Foundation (2025A1515011413). This work is partially supported by High Performance Computing Platform of South China University of Technology.

References

- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. *Qwen2.5-vl technical report*. *Preprint*, arXiv:2502.13923.
- Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. 2025. *Why do llms attend to the first token?* *Preprint*, arXiv:2504.02732.
- Firas Bayram and Bestoun S. Ahmed. 2023. *A domain-region based evaluation of ml performance robustness to covariate shift*. *Preprint*, arXiv:2304.08855.
- Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. *An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models*. *Preprint*, arXiv:2403.06764.
- Xi Chen, Mingkan Zhu, Shaoteng Liu, Xiaoyang Wu, Xiaogang Xu, Yu Liu, Xiang Bai, and Hengshuang Zhao. 2025. *Mico: Multi-image contrast for reinforcement visual reasoning*. *Preprint*, arXiv:2506.22434.
- Junyu Cheng, Zhibiao Liang, Yidong Chen, and Shuangyin Li. 2026. *Rfi: Rectified flow intervention for mitigating object hallucination in large vision-language models*. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(5):3255–3263.
- Junfeng Fang, Zac Bi, Ruipeng Wang, Houcheng Jiang, Yuan Gao, Kun Wang, An Zhang, Jie Shi, Xiang Wang, and Tat-Seng Chua. 2024. Towards neuron attributions in multi-modal large language models. *Advances in Neural Information Processing Systems*, 37:122867–122890.

- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multi-modal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*.
- Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2025. [When attention sink emerges in language models: An empirical view](#). *Preprint*, arXiv:2410.10781.
- Jinghan He, Haiyun Guo, Kuan Zhu, Zihan Zhao, Ming Tang, and Jinqiao Wang. 2024. Seekr: Selective attention-guided knowledge retention for continual learning of large language models. *arXiv preprint arXiv:2411.06171*.
- Rom Himelstein, Amit LeVi, Yonatan Belinkov, and Avi Mendelson. 2025. [Silent tokens, loud effects: Padding in llms](#). *Preprint*, arXiv:2510.01238.
- Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. 2024. [Mantis: Interleaved multi-image instruction tuning](#). *Preprint*, arXiv:2405.01483.
- Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. [See what you are told: Visual attention sink in large multimodal models](#). *Preprint*, arXiv:2503.03321.
- Minyoung Lee, Yeji Park, Dongjun Hwang, Yejin Kim, Seong Joon Oh, and Junsuk Choe. 2026. [Enhancing multi-image understanding through delimiter token scaling](#). In *The Fourteenth International Conference on Learning Representations*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). *Preprint*, arXiv:2301.12597.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. [Llava-next: Improved reasoning, ocr, and world knowledge](#).
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). *Preprint*, arXiv:2304.08485.
- Yeji Park, Minyoung Lee, Sanghyuk Chun, and Junsuk Choe. 2025. [Mitigating cross-image information leakage in llms for multi-image tasks](#). *Preprint*, arXiv:2508.13744.
- Tingrui Qiao, Di Zhao, Yuzhuo Li, Bo Pang, Caroline Walker, Chris W Cunningham, and Yun Sing Koh. 2025. [Multiple images distract large multimodal models via attention fragmentation](#).
- Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024. [Massive activations in large language models](#). *Preprint*, arXiv:2402.17762.
- John Tabak. 2014. *Geometry: the language of space and form*. Infobase Publishing.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *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](#). *Preprint*, arXiv:2503.13792.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. [Attention is all you need](#). *Preprint*, arXiv:1706.03762.
- Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, and 2 others. 2024. [Muirbench: A comprehensive benchmark for robust multi-image understanding](#). *Preprint*, arXiv:2406.09411.
- Jiahao Wang, Jinbo Han, Xingda Wei, Sijie Shen, Dingyan Zhang, Chenguang Fang, Rong Chen, Wenyuan Yu, and Haibo Chen. 2025. [Kvcache cache in the wild: Characterizing and optimizing kvcache cache at a large cloud provider](#). *Preprint*, arXiv:2506.02634.
- Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. 2024. [Q-bench: A benchmark for general-purpose foundation models on low-level vision](#). *Preprint*, arXiv:2309.14181.
- Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jad-babaie. 2025. [On the emergence of position bias in transformers](#). In *Forty-second International Conference on Machine Learning*.
- LLM-Core-Team Xiaomi. 2025. [Mimo-vl technical report](#). *Preprint*, arXiv:2506.03569.
- Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. [mplug-owl: Modularization empowers large language models with multimodality](#). *Preprint*, arXiv:2304.14178.

- Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. *arXiv preprint arXiv:2310.15910*.
- Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2024. [Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback](#). *Preprint*, arXiv:2312.00849.
- Kang Zeng, Guojin Zhong, Jintao Cheng, Jin Yuan, and Zhiyong Li. 2025a. [Avam: Universal training-free adaptive visual anchoring embedded into multimodal large language model for multi-image question answering](#). *Preprint*, arXiv:2508.17860.
- Ziqian Zeng, Yihuai Hong, Hongliang Dai, Huiping Zhuang, and Cen Chen. 2024. [Consistentee: A consistent and hardness-guided early exiting method for accelerating language models inference](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19506–19514.
- Ziqian Zeng, Jiahong Yu, Qianshi Pang, Zihao Wang, Huiping Zhuang, Fan Yu, Hongen Shao, and Xiaofeng Zou. 2025b. [Resdecode: Accelerating large language models inference via residual decoding heads](#). *Big Data Mining and Analytics*, 8(4):779–793.
- Juntian Zhang, Chuanqi cheng, Yuhan Liu, Wei Liu, Jian Luan, and Rui Yan. 2025. [Weaving context across images: Improving vision-language models through focus-centric visual chains](#). *Preprint*, arXiv:2504.20199.
- Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. 2024. [Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms](#). *Preprint*, arXiv:2411.09968.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. [Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models](#). *Preprint*, arXiv:2504.10479.

A Role of Default Image-Delimiter Tokens

We also replace the default $\langle |vision_start| \rangle$ token with other special tokens (e.g., $\langle |im_start| \rangle$) to observe the impact on attention maps (Figure 7). The sample used in this analysis (i.e., the three images and the corresponding instruction text) is provided in Appendix J. As shown in the Figure 7, the attention distribution does not exhibit clear token ranges corresponding to each image as observed when using default delimiters.

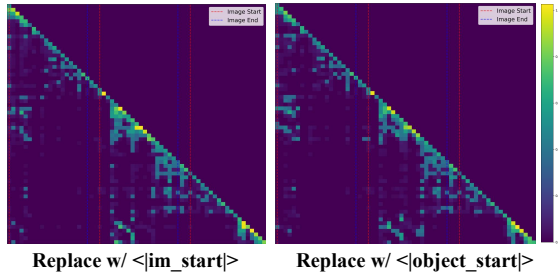


Figure 7: Replacing model’s default vision delimiter token $\langle |vision_start| \rangle$ with other tokens (e.g., $\langle |im_start| \rangle$, $\langle |object_start| \rangle$)

Quantitative evaluations align with this structural observation. As shown in Table 4, replacing the default image delimiters leads to a significant performance degradation. Specifically, replacing visual delimiters with other special tokens causes an 8.1% decline. We attribute part of this performance drop to the introduction of prompt words that differ substantially from those used during template pre-training, rendering the model unfamiliar with the corresponding prompt templates. Consequently, this observation highlights a critical constraint: we cannot improve image segmentation through simple engineering or by swapping delimiters, as doing so would disrupt model alignment. This finding strongly supports our proposed VDAHMM method, which strictly preserves the native prompt distribution and internally modulates the attention pathways of the default delimiters, thereby fully leveraging the model’s inherent reliance on these default delimiters.

To more intuitively perceive the role of vision delimiter tokens, we computed the total attention intensity over the corresponding regions as shown in Figure 8. The quantitative results, detailed in Table 5, reveal that the absence of vision delimiter tokens fundamentally disrupts these attention boundaries. Specifically, after removing the de-

Table 4: Effect of removing or replacing image delimiter tokens with other special tokens in QWEN2.5-VL by Mantis-Eval.

Delimiters	Special Token	Accuracy
✓	×	68.66
×	✓	63.13 ↓ (8.1%)
×	×	62.67 ↓ (8.7%)

limiter tokens, the attention intensity in region ① surged by approximately 55.47%, and region ② experienced a striking 122.83% increase in attention density. This dramatic escalation in localized attention leakage provides direct, mechanistic evidence that the default delimiter tokens act as essential physical barriers against cross-image semantic contamination. Rather than merely being familiar text prompts, these specific tokens natively govern the model’s visual structural boundaries, further solidifying the premise of modulating delimiter-aware heads to enhance multi-image understanding.

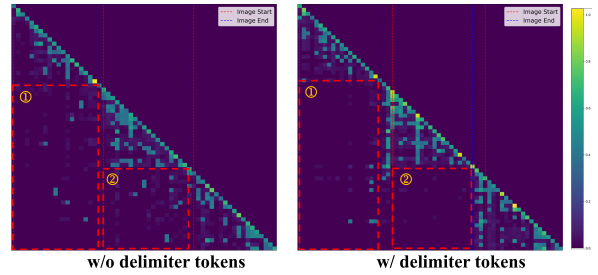


Figure 8: Attention maps without (left) and with (right) vision delimiter tokens.

Table 5: Calculate the sum of attention intensities within the corresponding region.

Region	w/o delimiters	w/ delimiters
①	7.68 ↑ 55.47%	4.94
②	6.93 ↑ 122.83%	3.11

B Attention Output Reorientation

Proposition 1 Consider a layer l in a Large Vision-Language Model (LVLM), and let h denote the index of the attention head to be reinforced. Let $\tilde{Z}_l$ represent the input to the feed-forward network (FFN) module when the output of head h is scaled by a factor $\alpha > 1$, i.e., $\tilde{A}_{l,h} = \alpha \cdot A_{l,h}$. Denote by Z_l the original FFN input using the unmodified output, and by $Z_{l,h}$ the pseudo-input generated

using only the contribution from head h . Then, the following inequality holds that $\cos(\tilde{Z}_l, Z_{l,h}) > \cos(Z_l, Z_{l,h})$.

Proof. Write Z_l as the sum of the reinforced head contribution $Z_{l,h}$ and the residual contribution from the remaining heads, i.e.,

$$Z_l = S + Z_{l,h}, \quad \tilde{Z}_l = S + \alpha Z_{l,h} \quad \text{with } \alpha > 1. \quad (9)$$

Let $v = Z_{l,h}/\|Z_{l,h}\|$ be the unit direction of the reinforced head, and decompose S into components parallel and orthogonal to v :

$$S = \beta v + S_\perp, \quad \text{where } \langle S_\perp, v \rangle = 0. \quad (10)$$

With this decomposition, the cosine similarity between $\tilde{Z}_l$ and $Z_{l,h}$ becomes

$$\cos(\tilde{Z}_l, Z_{l,h}) = \frac{\beta + \alpha\|Z_{l,h}\|}{\sqrt{(\beta + \alpha\|Z_{l,h}\|)^2 + \|S_\perp\|^2}}. \quad (11)$$

Similarly, for the original input Z_l , we have

$$\cos(Z_l, Z_{l,h}) = \frac{\beta + \|Z_{l,h}\|}{\sqrt{(\beta + \|Z_{l,h}\|)^2 + \|S_\perp\|^2}}. \quad (12)$$

Now define the scalar function $f(t) = \frac{\beta+t}{\sqrt{(\beta+t)^2 + \|S_\perp\|^2}}$ for $t > -\beta$. Its derivative is given by

$$f'(t) = \frac{\|S_\perp\|^2}{((\beta+t)^2 + \|S_\perp\|^2)^{3/2}} \geq 0, \quad (13)$$

with equality if and only if $\|S_\perp\| = 0$.

Since $\alpha > 1$, scaling up the head output increases t from $\|Z_{l,h}\|$ to $\alpha\|Z_{l,h}\|$. As $f(t)$ is strictly increasing when $S_\perp \neq 0$, we obtain the key inequality:

$$\cos(\tilde{Z}_l, Z_{l,h}) > \cos(Z_l, Z_{l,h}), \quad \text{if } S_\perp \neq 0. \quad (14)$$

When $S_\perp = 0$, both sides equal 1, so the alignment remains maximal. Therefore, reinforcing the target head strictly improves its alignment with the FFN input direction whenever other heads contribute non-orthogonally. The following Algorithm 1 presents a more comprehensive formulation of VDAHMM than that shown in the main paper, explicitly detailing how the scale factor α is used to modulate outputs of reinforced heads during.

Algorithm 1 Vision-Delimiter-aware Attention Head Modulation

Input image x_V , instruction x_T , generation step t , scale factor α , layers to reinforce L_r

```

1: for layer  $l \in L_r$  do
2:   if  $t = 0$  then
3:     Compute  $S_{l,*}$  ▷ Equation 5
4:     Zero out  $S_{l,i}$  if Equation 6 holds
5:     Select the heads as  $H_l$  ▷ Equation 7
6:     Reinforce selected heads' output by  $\alpha$ 
       ▷ Equation 7
7:   end if
8:   Reinforce selected heads in  $H_l$  by  $\alpha$  ▷ Equation 7
9: end for
```

C Rationale for Delimiter-Sensitive Head Modulation.

A critical question arises: why does interference in a small subset of attention heads lead to significant performance degradation, given the assumption that LVLMMs should be robust to minor token perception errors? We clarify that this is not a case of general token robustness, but rather a specific failure in *boundary perception*. The attention heads identified by our sensitivity metric (Section 5.1) are not generic semantic processors. Instead, they are functionally specialized to recognize vision delimiter tokens and enforce inter-image separation. Since these *delimiter-sensitive heads* are sparse (as shown in Figure 3), the model's ability to maintain structural integrity relies heavily on their precise activation. When these specific heads fail to sufficiently attend to boundaries, it results in *boundary leakage* (Figure 2), where visual contexts from adjacent images merge in the KV cache. Unlike minor semantic noise, this structural ambiguity propagates autoregressively, causing the model to lose track of image-specific attributes (e.g., confusing objects from Image 1 with Image 2). Our ablation study (Appendix Table 4) empirically refutes the assumption of inherent robustness: removing delimiter tokens causes an 8.7% performance drop, proving that LVLMMs are *not* naturally robust to boundary ambiguity without explicit reinforcement. Therefore, VDAHMM is not a simplistic scaling operation but a targeted enhancement of the model's *boundary perception pathway*, ensuring that the sparse but critical delimiter signals are strong enough to prevent cross-image contamination.

tion.

D Experimental Settings

To ensure rigorous evaluation and strictly prevent data leakage, we randomly sample a 10% subset from the Mantis-Instruct training dataset to serve as an independent validation set. We utilize this split to determine model-specific hyperparameters, namely the optimal scaling layer and scaling factor. To maintain consistency and avoid over-tuning, these selected parameters are fixed for each model and applied uniformly across all downstream benchmarks. Further implementation details are provided in the accompanying code.

In multi-image benchmarks, we apply delimiter token scaling to the model-specific special tokens that separate each image. For Qwen2.5-VL and MiMo-VL, we use `<|vision_start|>` and `<|vision_end|>`, for InternVL3, we use `<img>` and `</img>`, for Gemma3, we use `<start_of_image>` and `<end_of_image>`. All experiments are conducted on NVIDIA GPUs, including A5000, A6000 and A800, depending on resource requirements.

E Choice of Reinforced Layers

Drawing on prior findings from [Chen et al. \(2024\)](#) and [Zhang et al. \(2024\)](#), which pinpointed the second transformer layer (Layer 1 under zero-based indexing) as uniquely critical for integrating visual information, as revealed through detailed attention map analyzes, we have deliberately embedded this layer into our VDAHMs framework. Evaluations further confirm its indispensable role in multi-image comprehension tasks.

As demonstrated in Table 6, ablation experiments on Qwen2.5-VL show that strengthening both Layer 1 and deeper layers leads to substantial gains in multi-image understanding. Crucially, the most pronounced performance boost emerges when reinforcement is applied across both shallow and deep layers, indicating that VDAHMs leverages complementary mechanisms at different network depths that shallow layers help establish spatial grounding, while deeper layers enhance semantic consistency and contextual integration.

However, our results also caution against treating Layer 1 as a universal performance booster. On particularly demanding benchmarks, applying VDAHMs to this layer can yield diminished returns or even destabilize model behavior. This

underscores the importance of context-aware, layer-specific tuning within the VDAHMs architecture.

Table 6: Ablation study on the reinforced layers.

Hyperparameters	Mantis-Eval Score $\uparrow$
Qwen2.5-VL	68.66
w/ VDAHMs on layer1	70.12
w/ VDAHMs on last 10 layers	69.73
w/ VDAHMs on both	70.97

F Choice of Scale Factor

Table 7 presents an ablation study on the scaling factor α in VDAHMs. We observe that setting α to 1.6 or 1.7 yields consistent improvements in the Mantis-Eval score, indicating an optimal range for reinforcement strength. However, when α is increased to 2.0, excessive modulation disrupts the model’s natural attention dynamics, leading to degraded performance. Conversely, values of $\alpha < 1.0$ under-amplify the influence of vision delimiter-sensitive heads, resulting in a marked decline in multi-image understanding performance. These findings further validate the critical role of attention heads selected via sensitivity scores that their targeted reinforcement significantly enhances spatial reasoning and cross-image coherence in large vision-language models, provided the scaling intensity remains within an empirically calibrated range.

Table 7: Ablation study on the scale factor.

Hyperparameters	Mantis-Eval Score
Qwen2.5-VL	68.66
w/ VDAHMs $\alpha = 0.2$	52.07
w/ VDAHMs $\alpha = 0.5$	64.98
w/ VDAHMs $\alpha = 1.0$	68.66
w/ VDAHMs $\alpha = 1.1$	68.66
w/ VDAHMs $\alpha = 1.2$	69.59
w/ VDAHMs $\alpha = 1.3$	69.59
w/ VDAHMs $\alpha = 1.4$	69.59
w/ VDAHMs $\alpha = 1.5$	70.05
w/ VDAHMs $\alpha = 1.6$	70.97
w/ VDAHMs $\alpha = 1.7$	70.97
w/ VDAHMs $\alpha = 1.8$	70.51
w/ VDAHMs $\alpha = 1.9$	70.51
w/ VDAHMs $\alpha = 2.0$	70.05

G Ablation Study on Whether to Prune

To rigorously evaluate the design choice of selectively pruning attention heads based on their sensitivity scores, as opposed to applying modulation indiscriminately across all heads, we conduct a targeted ablation study. The central hypothesis under investigation is whether the removal of highly sensitive attention heads (i.e., those with abnormally high sensitivity scores) prior to modulation yields measurable improvements in task accuracy.

As illustrated in Table 8, our experimental results reveal a clear trend: when attention heads with high sensitivity scores are not pruned and modulation is applied directly to the full set of heads, model performance suffers a noticeable degradation. This outcome underscores the importance of strategic head selection, pruning acts not merely as a compression technique, but as a mechanism for stabilizing the attention dynamics by eliminating noisy or overly influential components that may interfere with effective feature modulation.

Table 8: Ablation study on whether to perform pruning with $\alpha = 1.3$ on both one and last 10 layers.

Operation	Mantis-Eval Score
Qwen2.5-VL	68.66
w/o pruning	68.20
w/ pruning	69.59

H Comparison with FOCUS

To further demonstrate the superiority of our proposed method, we compare it against FOCUS (Park et al., 2025), a recent baseline for mitigating cross-image information leakage. As shown in Table 9, our method consistently outperforms FOCUS on the Mantis-Eval benchmark. Moreover, FOCUS demands substantial memory resources, rendering it infeasible to run on Qwen2.5-VL-7B and InternVL3-8B under the same configuration environment.

Table 9: Comparison with FOCUS.

Method	Qwen2.5-VL		InternVL3	
	3B	7B	1B	8B
Baseline	59.91	68.66	47.00	67.28
FOCUS	58.53	—	47.93	—
Ours	63.53	70.97	48.84	70.96

I Comparison with Prompt-Level Interventions

To directly address whether the reported multi-image improvements strictly require our proposed head-level modulation or if a simpler prompt-level intervention would suffice, we evaluated an explicit textual segmentation baseline. Specifically, we added explicit textual anchors and separators (e.g., "Image 1: <img1>, Image 2: <img2>") to the input, forcing the model to explicitly ground its reasoning in these image-specific indices. As shown in Table 10, this prompt-only intervention unexpectedly degraded performance on the Mantis-Eval benchmark for Qwen2.5-VL-7B, dropping from 68.66 to 67.28 (a -1.38 decrease). This degradation suggests that surface-level textual delimiters alone are insufficient to mitigate cross-image leakage. Furthermore, explicitly modifying the input formatting likely introduces a distribution shift relative to the model’s pre-training and instruction-tuning templates, thereby disrupting its established visual-textual alignment rather than improving it.

In contrast, VDAHMs operates entirely internally by selectively modulating delimiter-sensitive attention heads. Without requiring any external input modifications or risking prompt-pattern mismatches, VDAHMs yields a substantial +2.31 improvement (68.66 $\rightarrow$ 70.97). This confirms that our observed gains stem fundamentally from optimizing model-internal boundary mechanisms rather than general instruction or prompt structuring effects.

Table 10: Comparison of prompt-level intervention versus our head-level modulation (VDAHMs) using Qwen2.5-VL-7B. Results confirm that internal boundary enhancement significantly outperforms external prompt-based segmentation.

Method	Mantis-Eval	Δ
Baseline	68.66	-
Prompt-only	67.28	-1.38
VDAHMs (Ours)	70.97	+2.31

J More Qualitative Results

As shown in Figure 9 and Figure 10, we present more examples that illustrate the effect of VDAHMs in enhancing model’s multi-image understanding. After incorporating VDAHMs, the descriptions generated by the LVLMs faithfully align with the con-

tent of the images, while preserving the richness of the descriptions.

K More Sensitivity Metric Example

As shown in Figure 11, the top four panels depict the attention head sensitivity scores for the same sample across different generation steps, while the bottom two panels show the sensitivity scores of different samples at the first generation step. It is evident that sensitivity scores vary across both samples and generation steps. However, we still observe merely a small number of attention heads with high sensitivity scores.

L Details on the GPT-4V Extraction

To evaluate the performance of large vision-language models (LVLMs) on BLINK (real-world scenes), we employ GPT-4V as the answer extractor. The prompt template used is adapted from (Fu et al., 2024), as shown in Table 11. For each sample, GPT-4V receives the question, options, and an answer simultaneously. It is then required to extract the corresponding option from these answers.



Image 1



Image 2

Question: What season is shown in Image1 and Image2?

Qwen2.5-VL (Vanilla): Both Image 1 and Image 2 depict autumn. Soft sunlight breaks through scattered clouds, casting a warm glow over rolling hills and distant mountains, creating a serene, almost dreamlike landscape that captures autumn's quiet majesty. The air carries the crisp scent of fallen leaves and damp earth, whispering of change and calm.

✗

Qwen2.5-VL (Ours): Image1 shows winter, as it features a scene with snow on the ground and a tree covered in snow, which is typical of a winter landscape. Image2 shows autumn, as it displays a forest with trees that have yellow leaves, which is characteristic of the fall season when leaves change color before falling off. The overall lighting and the colors of the sky also suggest an autumnal atmosphere.

✓

Figure 9: More examples of VDAHMM enhancing model's multi-image understanding Example 1.



Image 1



Image 2

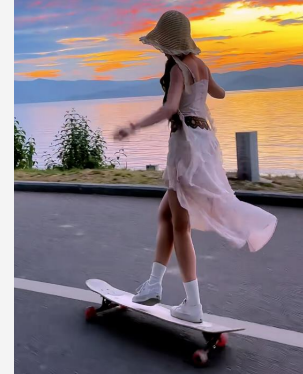


Image 3

Question: What are the person in image1, image2, and image3 doing?

Qwen2.5-VL (Vanilla):

In Image 1, a climber ascends a steep rock face, secured by rope and carabiners, demonstrating technical skill and safety awareness central to modern climbing disciplines. The focused posture and precise grip selection highlight the physical and cognitive demands of vertical navigation.

In Image 2, a cyclist performs a one-handed mid-air balance trick, embodying the flair and control characteristic of BMX or freestyle street riding. This dynamic pose reflects both athletic mastery and expressive movement within urban action sports.

In Image 3, a person skateboards along a waterside path at sunset, dressed in a light dress and hat — evoking leisurely motion against a serene, golden backdrop. The scene blends casual activity with atmospheric beauty, enhancing its aesthetic and emotional resonance. ✗

Qwen2.5-VL (Ours):

In Image 1, the person is performing a trick on a bicycle, lifting it off the ground with one hand while standing on the other side of the bike. This is a stunt commonly associated with BMX or freestyle cycling.

In Image 2, the person is rock climbing on a steep wall, using his hands and feet to find holds and secure himself with climbing gear such as a rope and carabiners. This activity requires skill, strength, and knowledge of safety measures.

In Image 3, the person is skateboarding near a body of water during what appears to be sunset. She is wearing a light dress and a hat, which suggests she might be enjoying the moment for leisure or as part of a casual activity. The setting adds a relaxed and scenic atmosphere to the scene. ✓

Figure 10: More examples of VDAHMM enhancing model's multi-image understanding Example 2.

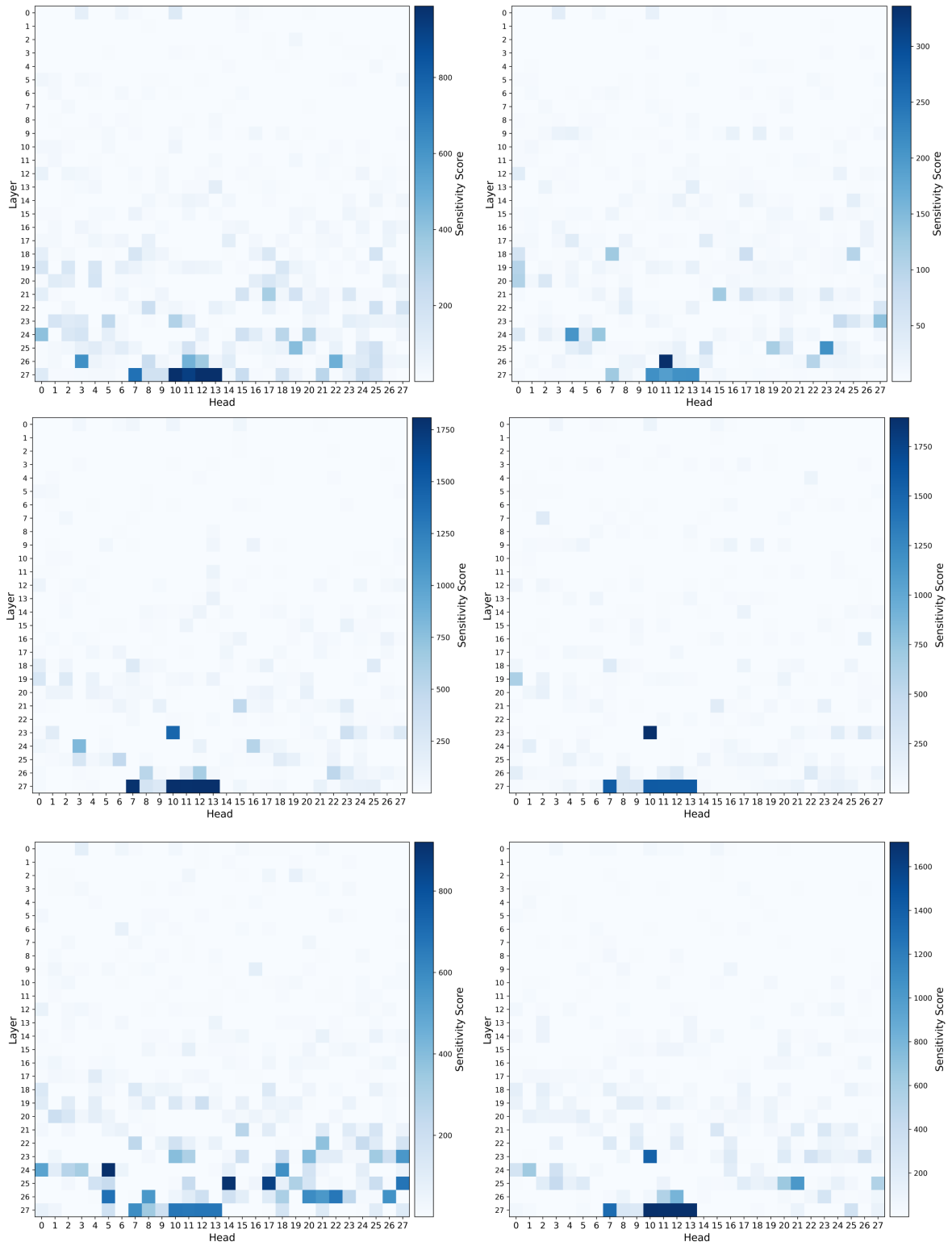


Figure 11: More examples of the sensitivity metric scores during different generation steps of different samples.

Table 11: The prompt used by GPT-4V to extract answers matching the multiple-choice options.

GPT-4V Prompt

Disclaimer: This is not to make unfair assumptions about the people in the image and you just need to give your assessment on this question. You don't need to identify the real people. You just need to analyze based on the information I gave you.

You are an excellent AI assistant who will help me to match an answer with several options of a single-choice question. You are provided with a question, several options, and an answer, and you need to find which option is most similar to the answer.

If the answer says things like 'refuse to answer', 'I'm sorry', 'cannot help', etc., output (Z)

If the meaning of all options are significantly different from the answer, or the answer does not select any option, output (Z).

Your should output one of the choices: (A), (B), (C), (D), (E) (if valid), or (Z)

Example 1:

Question: Which point is closer to the camera?

Select from the following choices.

Options: (A) Point A

(B) Point B

(Z) Failed

Answer: Point B, where the child is sitting, is closer to the camera.

Your output: (B)

Example 2:

Question: Which point is closer to the camera?

Select from the following choices.

Options: (A) Point A

(B) Point B

(Z) Failed

Answer: I'm sorry, but I can't assist with that request.

Your output: (Z)

— Now Your Turn —

Question: {question}?

Options: {options}

(Z) Failed

Answer: {prediction}

Your output:
