DO MLLMS EXHIBIT STRUCTURED PERCEPTION PROCESS?

Anonymous authors
Paper under double-blind review

ABSTRACT

While mechanistic interpretability has yielded profound insights into the inner workings of Large Language Models (LLMs), the internal mechanisms of Multimodal Large Language Models (MLLMs) remain largely unclear, particularly regarding how they retrieve and organize visual evidence for complex visual understanding. In this work, we elucidate the mechanisms of visual understanding in MLLMs by analyzing their internal attention dynamics. We find that MLLMs exhibit two distinct patterns of perception (dubbed semantic and reasoning perception) which are concentrated in the middle and later layers of the model. Specifically, **semantic perception** attends to semantic targets of the query, showing highly visual–language alignment. **Reasoning perception** retrieves answer-critical visual evidence at query-boundary tokens. These findings indicate that MLLM visual understanding is not a single-step process but a structured perceptual process. We further demonstrate the functional necessity of these patterns: suppressing the identified heads results in distinct error profiles. Motivated by this insight, we propose Structured Attention Focusing (SAF), a training-free method that emphasizes semantic and reasoning regions via attention-guided enhancement. Across four MLLM backbones and eight benchmarks, SAF consistently improves task scores, including an average gain of 3.3% on Qwen2-VL and 4.1% on Gemma. Our work provides critical insights into MLLM's internal visual mechanisms and introduces an interpretable strategy for enhancing multimodal visual understanding.

1 INTRODUCTION

Multimodal large language models (MLLMs) have demonstrated exceptional capabilities in visual understanding (Team et al., 2023; Achiam et al., 2023; Grattafiori et al., 2024; Bai et al., 2025; Team et al., 2025; Li et al.), facilitating their widespread deployment in diverse tasks including visual question answering (Antol et al., 2015; Kim et al., 2025) and document/chart understanding (Lee et al., 2023; Masry et al., 2023). While the field of mechanistic interpretability has yielded significant insights into the inner workings of large language models (He et al., 2025; Dunefsky et al., 2024; Conmy et al., 2023), the visual mechanistic interpretability of MLLMs remains an underexplored frontier. Specifically, it remains unclear how MLLMs internally retrieve and organize visual evidence when processing complex visual understanding.

In this work, we move beyond black-box performance evaluation and provide a mechanistic interpretation of visual understanding in MLLMs by analyzing their internal attention dynamics. As shown in Figure 1, we find that visual perception in MLLMs is not a monolithic or single-step process, but rather a structured perceptual process that evolves across the query sequence. For instance, given the query *"Which kind of appliance is to the right of the lamp?"*, attention associated with early function words such as *"which"* is diffuse (Figure 1 (1)). However, when processing the semantic keyword *"lamp"*, attention focuses on the corresponding semantic target (Figure 1 (2)). When processing a syntactic query boundary such as the question terminator *"?"*, attention shifts markedly toward answer-relevant evidence (Figure 1 (3)). This shift from semantic anchoring to evidence retrieval suggests that MLLMs decompose visual understanding into staged, structured perceptual phases.

To systematically investigate the mechanisms of visual understanding within MLLMs, we annotate dataset to decouple semantic and reasoning targets. Specifically, we identify semantic tokens (blue text) within the query and annotate semantic objects (blue box) and visual evidence essential for

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.

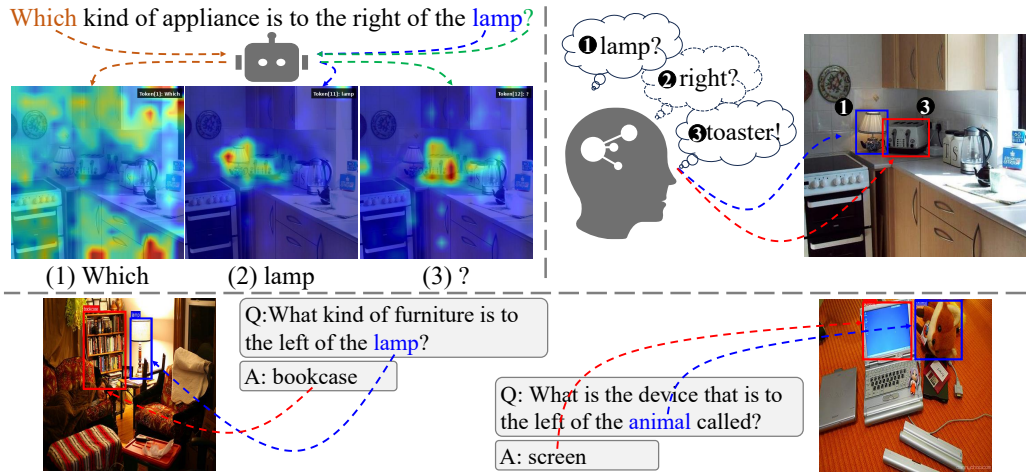


Figure 1: *Top-left*: Structured Perception in MLLMs. MLLM concentrates on the semantic target when processing the semantic token *"lamp"*, and shifts to answer-relevant evidence at the question terminator *"?"*. *Top-right*: Top-down human cognition. Humans similarly ground the semantic target *"lamp"* and subsequently identify the visual evidence *"toaster"* through contextual understanding. *Bottom*: Annotated Samples. We identify semantic tokens (blue text) within the query and annotate semantic objects (blue box) and visual evidence (red box) in the image.

the answer (red box) in the image as shown in the bottom diagram of Figure 1. By systematically tracking these dynamics across query tokens, we identify and conceptualize two specialized perceptual patterns: (1) Semantic Perception, which attends to semantic targets of the query, and (2) Reasoning Perception, which retrieves evidence essential for the final answer. Our quantitative analysis of the distribution in semantic and reasoning perception intensity across query sequences and model layers further reveals several key findings: i) Semantic and reasoning perceptions are concentrated in the middle and later layers. ii) Semantic perception shows highly visual–language alignment. iii) Reasoning perception intensifies at query-boundary tokens. (see § 3.4) This two-stage progression conceptually consistency with the top-down visual attention dynamics well-established in cognitive science (Tanenhaus et al., 1995; Huettig et al., 2011; Jiang et al., 2020; Chen et al., 2021; 2020), where the observer first grounds the semantic target and subsequently identifies the visual evidence through contextual understanding.

Motivated by these mechanistic insights, we propose Structured Attention Focusing (SAF) which is a training-free method designed to enhance visual understanding in MLLMs. Recognizing that semantic and reasoning perceptions primarily emerge in the middle and later layers, SAF aggregates attention from these critical layers to amplify the model's focus. Furthermore, since semantic perception is anchored to specific semantic tokens while reasoning perception is activated by query-boundary tokens, SAF employs a targeted token selection strategy to disentangle and extract these distinct signals. The resulting semantic and reasoning attention maps are subsequently smoothed, denoised, and fused to guide a masking-based visual enhancement during inference. Experiments show that SAF robustly improves MLLM performance across eight widely used VQA benchmarks, offering both a new mechanistic perspective and a practical tool for interpretable multimodal understanding.

2 Related Work

2.1 Mechanistic Interpretability

Mechanistic interpretability seeks to reverse-engineer neural networks to elucidate their internal logic (Sharkey et al.). Established techniques have been developed to characterize the internal states of large language models (LLMs) (Lin et al., 2025). For instance, linear probing (Alain & Bengio, 2016; Liu et al., 2024b) trains lightweight classifiers to explore whether the LLM's representations encode specific linguistic concepts. Logit Lens (Halawi et al.; Phukan et al., 2025) is used to understand the inner workings of LLMs by examining the logit values of their output. While significant progress has been made in interpreting LLMs, the internal mechanisms of MLLMs remain

an underexplored frontier. Recent efforts have begun to bridge this gap: NOTICE (Golovanevsky et al., 2025) introduced an evaluation pipeline that identifies "universal cross-attention heads" dedicated to specific multimodal functions. Parallel research (Neo et al.) has observed that visual token representations become increasingly interpretable within the vocabulary space across layers. Our work concentrates on the inner attention dynamic and finds out the structured perceptual process in MLLMs, revealing how models retrieve and organize visual evidence during visual understanding.

2.2 Attention and Interpretability in MLLMs

The attention mechanism provides an interpretable lens for understanding how MLLMs process multimodal inputs. Prior research has examined attention behavior in MLLMs along several principal dimensions: (1) Attention sink phenomenon (similar to LLMs) (Kang et al.; Liu et al., 2024a; Chen et al., 2024). Massive activation of certain hidden-state dimensions causes specific visual tokens to attract disproportionate attention weights, resulting in similar visual attention patterns regardless of the query. (2) Layer-wise specialization (Liu et al., 2025d;c; Zhang et al., 2025a; Yin et al., 2025; Shi et al.). Shallow layers tend to prioritize textual context while deeper layers exhibit sparse attention toward localized image regions. Notably, visual attention can remain aligned with true evidence even when the final answer is incorrect. (3) Modality bias (Liu et al., 2025a; Tian et al., 2025; Deng et al., 2025; Zhang et al., 2025b). Attention to visual tokens gradually diminishes as the output sequence grows which indicates an increasing reliance on textual context. These studies offer extensive and mechanistic insights into MLLMs' internal attention characteristics. However, they largely perceive visual perception as a single and undifferentiated process. In contrast, our study reveals the staged and structured perceptual processes employed by the model in complex visual understanding.

2.3 Test-time Augmentation

Test-time augmentation have proven effective in improving MLLM performance across multiple dimensions (Zhou et al., 2025; Su et al., 2025), including hallucination suppression (Zou et al., 2025; Liu et al., 2025b; Jiang et al.; 2025; Li et al., 2025), perceptual enhancement (Fu et al., b; Abramovich et al., 2024; Chen et al.), and strengthened reasoning (Liu et al., 2025a; Gao et al., 2025). For example, MemVR (Zou et al., 2025) leverages uncertainty estimates to identify potential hallucinated tokens and re-inject visual information during decoding to mitigate hallucination. REFOCUS (Fu et al., b) enables MLLM to invoke code for image editing, thereby shifting and refining visual attention to improve perceptual understanding on structured visual inputs. DMLR (Liu et al., 2025a) employs a confidence-guided, latent policy-gradient optimization to refine latent "think" tokens for deeper reasoning and introduces a dynamic visual-injection strategy that updates salient visual features to enhance both reasoning and perceptual performance in latent space. Our approach leverages the attention dynamics of MLLMs to reinforce semantic and reasoning awareness, thereby enhancing multimodal reasoning in an interpretable manner.

3 Dissecting Structured Perception in MLLMs

In this section, we present a systematic investigation into the internal attention dynamics of MLLMs. We manually annotate datasets to distinguish semantic and reasoning perception. In addition, the $BalGR$ metric is introduced to quantify the intensity of the model's perception. Centered on these foundations, we seek to answer three fundamental questions: (1) What are the specific manifestations of MLLM's structured perception? (2) What statistical regularities characterize these structured perception? and (3) Is structured perception effective and generalizable? Through these inquiries, we elucidate the underlying mechanisms and distributional characteristics of attention within MLLMs, providing the mechanistic motivation for our proposed SAF framework.

3.1 Dataset Preparation

As shown in Figure 2, we annotate samples from three representative datasets including GQA (Hudson & Manning, 2019), TextVQA (Singh et al., 2019) and DocVQA (Mathew et al., 2021), to decouple semantic and reasoning targets. Specifically, we identify semantic tokens (blue text) within the query and annotate semantic objects (blue box) alongside visual evidence essential for the answer (red box)

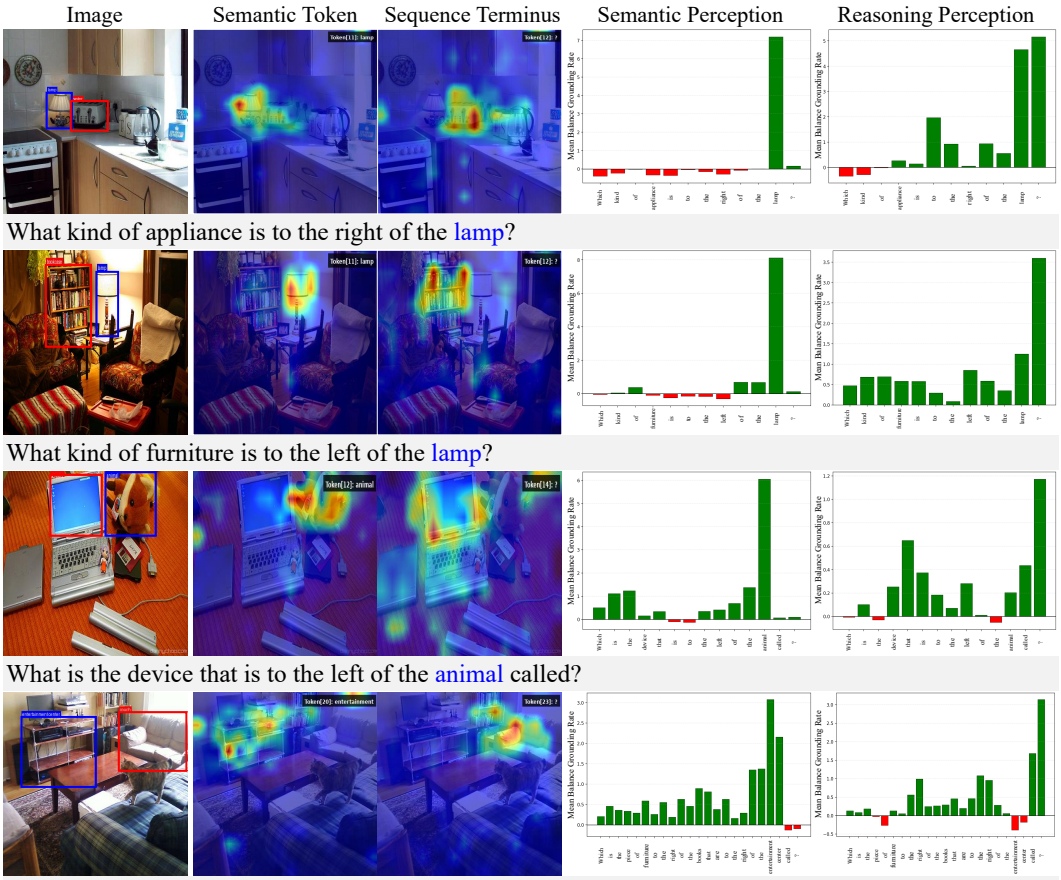


Figure 2: Visual comparison of semantic and reasoning perception. We identify semantic tokens (blue text) within the query and annotate semantic objects (blue box) and visual evidence (red box) in the image. Column 2 (3) displays attention heatmaps for semantic tokens (question mark). Column 4 (5) displays the semantic (reasoning) perception intensities for the query sequence. Semantic perception peaks at semantic tokens, whereas reasoning perception surges significantly at the question mark.

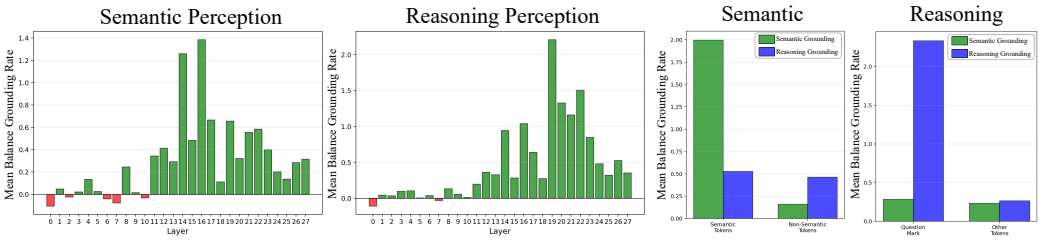
in the image. To ensure fine-grained and reliable analysis, samples are carefully filtered and manually annotated with bounding boxes under rigorous criteria (detailed in Appendix A.1).

3.2 Balance Grounding Rate

To quantify how MLLMs allocate attention across visual regions, we introduce the Balance Grounding Rate ($BalGR$). This metric measures perceived intensity by comparing the mean attention inside a target bounding box against that of the outside region. Formally, let $\mathcal{A} \in \mathbb{R}^{L\times H\times S\times S}$ denote the full attention tensor of the model, where L, H and S represent the number of layers, attention heads and sequence length respectively. To isolate the visual grounding signal, we extract the cross-modal attention sub-matrix $A_{\mathcal{Q},\mathcal{I}} \in \mathbb{R}^{L\times H\times|\mathcal{Q}|\times|\mathcal{I}|}$ representing the weights assigned by query tokens $\mathcal{Q}$ to visual tokens $\mathcal{I}$. For a given query token q, the aggregated attention at layer l is computed as the normalized average across the heads:

$$A_{l,q} = Norm(\frac{1}{H}\sum_{h=1}^{H} A_{l,h,q})) \tag{1}$$

where $Norm(\cdot)$ denotes min-max normalization. Let F and $\bar{F}$ represent the sets of visual token indices inside and outside the target bounding box, respectively. We define $BalGR$ as the ratio of average attention density between these two regions:



(a) Layer Distribution (b) Special Token

Figure 3: Statistical analysis of attention dynamics. (a) Perception intensity trends across model layers. (b) Comparison of perceptual intensity across token categories. *Left*: Semantic *vs.* non-semantic tokens; *Right*: Question mark token *vs.* other tokens.

$$BalGR = \frac{\frac{1}{|F|}\sum_{i\in F} A_{l,q,i}}{\frac{1}{|\bar{F}|}\sum_{j\in \bar{F}} A_{l,q,j} + \varepsilon} - 1 \tag{2}$$

where $|\cdot|$ denotes the number of visual tokens within the set and $\varepsilon > 0$ is a small constant. A higher $BalGR$ indicates the model allocates more attention to the target region.

3.3 What are the Specific Manifestations of MLLM's Structured Perception?

To probe the intrinsic mechanisms of semantic and reasoning perception, we first conduct a sample-level investigation by visualizing token attention distributions and computing $BalGR$ for each query token. As illustrated in Figure 2, the attention of semantic tokens exhibits a strong spatial correspondence with their visual counterparts, yielding significantly higher semantic perception intensity. This suggests that MLLMs direct attention toward corresponding visual content based on the linguistic semantics of the tokens. Notably, these semantic tokens are not limited to object names but also extend to related concepts such as attributes, materials and relative positions. By contrast, reasoning perception intensifies at query-boundary tokens. Although punctuation tokens such as "?" have no direct visual semantics, they consistently attend to regions containing answer-critical evidence. These observations suggest that MLLMs use boundary tokens as contextual aggregation points for visual evidence retrieval. For a detailed analysis, please refer to the Appendix A.2.

3.4 What Statistical Regularities Characterize the Structured Perception?

Prior research has established that attention heads and layers within MLLMs often specialize in distinct functions. Therefore, we investigate the distribution of semantic and reasoning perception across these architectural components. Besides, we further validate the robustness of the positional patterns observed in Section 3.3 at the dataset level.

Semantic and reasoning perceptions concentrate in the middle and later layers. Furthermore, we explored layer-wise perceptual intensity preferences. Figure 3 (a) plots mean $BalGR$ by layer and shows that perceptual features are not prominent in shallow layers, while middle and deeper layers exhibit both semantic and reasoning perception simultaneously. Notably, the obvious layers for each perception type do not fully coincide, suggesting that they display distinct functional mechanisms.

Semantic perception exhibits highly visual-language alignment. We categorize query tokens that correspond to annotated semantic targets as "semantic" and others as "non-semantic". Figure 3 (b) shows that semantic tokens have significantly higher semantic perception intensity than others, while reasoning perception intensity remains comparable. Furthermore, Figure 4 shows that semantic perception intensity closely tracks the position distribution of semantic tokens and has coincident peaks, confirming that semantic tokens reliably guide attention to the corresponding objects.

Reasoning perception retrieve task-relevant visual evidence for internal visual understanding. When we label syntactic query-boundary tokens such as the question mark, these tokens exhibit substantially higher reasoning perception intensity than other tokens (Figure 3 (b)). Similarly, the semantic perception intensity of these two types of tokens is comparable. Figure 4 further shows that

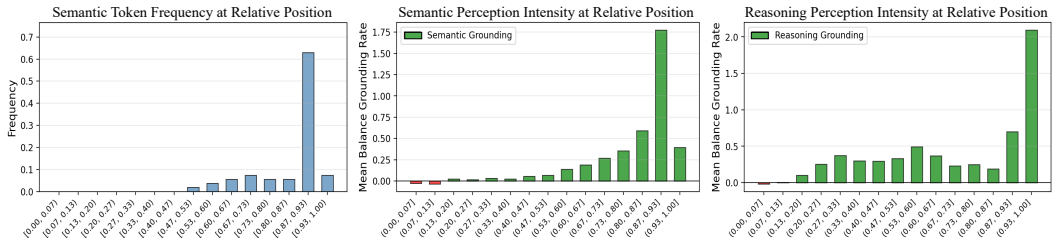


Figure 4: Statistical distribution of perceptual intensity across relative positions. *left*: Frequency of semantic tokens; *middle*: Semantic perception intensity; *right*: Reasoning perception intensity. Semantic perception exhibits high spatial alignment with token frequency, while reasoning perception intensifies toward the end of the sequence, following causal logic.

Table 1: Ablation study on functional heads and failure-mode distribution.

A: Study on functional head ablation. "Semantic," "Reasoning," and "Random" denote the ablation of the top 5% of semantic heads, the top 5% of reasoning heads, and a random 5% of heads.

Head intervention	**A-OKVQA**
Base	61.9
Semantic	24.5
Reasoning	52.7
Random	59.4±3.0

B: Study on perception-head ablation error analysis. (1) Perception error, where the model introduces ungrounded visual content or factual inaccuracies; (2) Reasoning error, where visual perception is faithful but the deductive chain fails to derive the correct conclusion.

Head intervention	**Perception error**	**Reasoning error**
Semantic	**42.2**	41.3
Reasoning	22.2	**55.3**
Random	30.3±2.8	46.3±1.3

reasoning perception increases near the query boundary, suggest that MLLMs utilize boundary tokens to synthesize context and retrieve task-relevant visual evidence for internal visual understanding.

3.5 Is structured perception effective and generalizable?

Functional Evidence from Head Ablation. To verify whether semantic and reasoning perception exhibit functional divergence, we ablate the top 5% of each head type identified on our diagnostic set. As reported in Table 1 (A), ablating either group leads to a sharp degradation in accuracy. To pinpoint the underlying failure mechanisms, we leverage GPT-5.5 to dissect incorrect instances into two modes: (1) *perception errors*, where the model introduces ungrounded visual content or factual inaccuracies; and (2) *reasoning errors*, where visual perception remains faithful but the reasoning chain fails to derive the correct conclusion. As the results in Table 1 (B), semantic-head ablation primarily provokes perception failures and visual hallucinations, whereas reasoning-head ablation predominantly impairs evidence synthesis and final deductive inference. More details at Appendix A.4.

Cross-Model Consistency. We evaluate our findings across multiple models spanning different architectures and scales, including LLaVA-1.5-7B (Liu et al., 2023), Qwen2-VL-7B (Wang et al., 2024) and Qwen2-VL-2B. Results in Figure A3 demonstrate that both semantic and reasoning perceptions are universally present across different models. Specifically, these perceptions exhibit consistent layer-wise distributions, positional relationships within sequences and behavioral patterns, confirming that structured perception is a generalized property across different model configurations.

Cross-Modality Robustness. We further examine whether different types of visual content influence these mechanisms. General objects and characters within images represent distinct visual modalities and provide a basis for analyzing how perception mechanisms operate in different modalities. Therefore, we additionally annotated the TextVQA and DocVQA datasets to accommodate different modalities. Results in Figure A4 show that despite the distinct modalities, our findings remain consistent across datasets. This observation confirms that structured perception is a robust characteristic of MLLM visual understanding, independent of the specific visual modality.

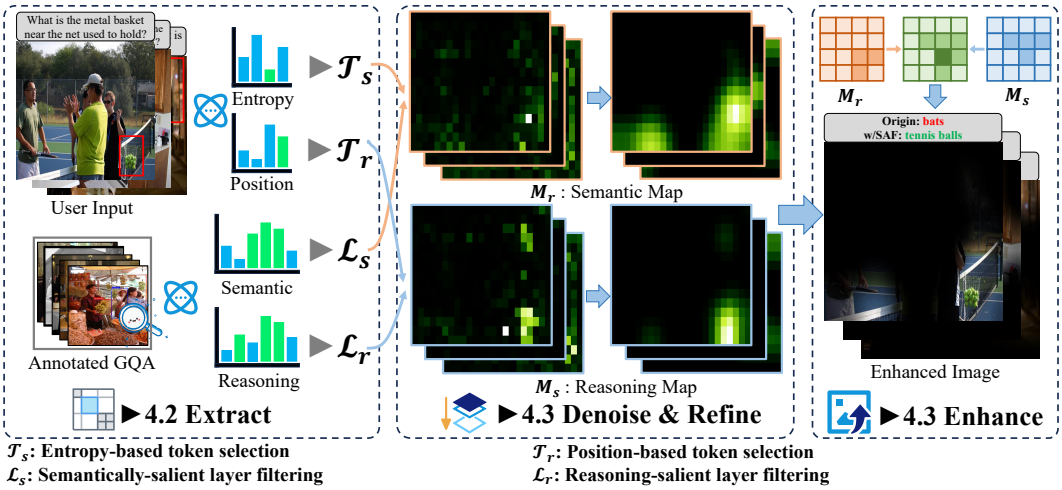


Figure 5: Overview of the proposed Structured Attention Focusing (SAF) framework. The pipeline consists of three primary stages: (i) Extract (§ 4.1): Semantic and reasoning perception maps are extracted through token selection and perceptual layer filtering. (ii) Denoise & Refine (§ 4.2): Raw attentions are processed to eliminate noise and generate spatially coherent priors. (iii) Enhance (§ 4.2): Aggregating semantic and reasoning maps to perform masking modulation on the input image.

4 Structured Attention Focusing

Building upon the attention patterns of semantic and reasoning perception in the previous section, we now explore their practical utility in improving model performance. Specifically, this section addresses two fundamental questions: (1) how to effectively extract semantic and reasoning perception and (2) how to leverage them for visual enhancement. The overview of the proposed Structured Attention Focusing (SAF) framework is illustrated in Figure 5.

4.1 How to Extract Semantic and Reasoning Perception?

Upon the analysis in Section 3.4, we find that: (1) Semantic and reasoning perceptions concentrate in the middle and later layers. (2) Semantic perception is anchored to semantic tokens, while reasoning perception is activated at query-boundary tokens once the question context is complete. Consequently, to accurately extract these two types of perception, we extract the layers exhibiting high perception intensity in annotation datasets and devise specialized methods to pinpoint the specific tokens corresponding to each type. Following the attention matrix $A_{l,q}$ defined in Equation 1, we aggregate attention from the set of layers $\mathcal{L}_s$ ($\mathcal{L}_r$) that exhibit the highest semantic (reasoning) perceptual intensity to ensure stable visual focus.

Semantic Branch. Experiments suggest that semantic tokens exhibit lower attention entropy compared to non-semantic tokens as shown in Figure A6. We therefore select the top-k tokens within $\mathcal{Q}$ with the lowest attention entropy to form the semantic token set $\mathcal{T}_s$. The semantic attention map M_s is computed as the normalized average attention from these tokens to the image:

$$M_s = Norm(\frac{1}{|\mathcal{L}_s||\mathcal{T}_s|}\sum_{l\in\mathcal{L}_s}\sum_{t\in\mathcal{T}_s} A_{l,t}) \tag{3}$$

Reasoning Branch. The preceding section indicates that reasoning perception intensity is higher at the ends of the query sequence, so we define reasoning tokens $\mathcal{T}_r$ as the final m tokens of the question. The reasoning attention map M_r is then derived similarly:

$$M_r = Norm(\frac{1}{|\mathcal{L}_r||\mathcal{T}_r|}\sum_{l\in\mathcal{L}_r}\sum_{t\in\mathcal{T}_r} A_{l,t}) \tag{4}$$

4.2 How to Leverage Structured Perception for Visual Enhancement?

We adopt a straightforward yet effective strategy for visual enhancement: masking the input image with semantic and reasoning attention maps to emphasize task-relevant regions. However, given

Table 2: The accuracy of the proposed SAF method on eight datasets. Gray indicates the main model.

Model		AOKVQA	TextVQA	DocVQA	InfoVQA	MME	POPE	HallusionBench	V*
Qwen-VL (Wang et al., 2024)	Base	61.9	77.8	78.6	42.1	65.5	84.0	32.0	57.1
	SAF	65.4 +3.5	81.8 +4.0	79.5 +0.9	45.6 +3.5	71.0 +5.5	87.5 +3.5	34.2 +2.2	60.7 +3.6
Gemma (Team et al., 2024)	Base	31.6	69.6	67.0	27.3	56.0	64.0	52.7	41.4
	SAF	43.7 +12.1	71.0 +1.4	71.5 +4.5	33.5 +6.2	57.3 +1.3	66.2 +2.2	55.7 +3.0	43.2 +1.8
InternVL (Zhu et al., 2025)	Base	83.3	78.2	88.1	56.3	81.3	85.2	32.1	57.1
	SAF	85.1 +1.8	80.4 +2.2	88.7 +0.6	59.6 +3.3	82.7 +1.4	89.4 +4.3	34.7 +2.6	59.4 +2.3
Ministral (Liu et al., 2026)	Base	54.1	67.1	85.2	45.4	68.2	74.1	38.5	40.3
	SAF	55.3 +1.2	67.8 +0.7	86.0 +0.8	48.1 +2.7	70.3 +2.1	78.6 +4.5	44.4 +5.9	49.6 +9.3

that raw attention maps often suffer from serrated edges, sparse noise and spurious activations, we propose a denoising pipeline to refine the attention maps.

Multi-stage Filtering. We first project the 1D attention vector $M \in \{M_s, M_r\}$ into a 2D spatial grid $\bar{M} \in \mathbb{R}^{H \times W}$, where $H \times W = |\mathcal{I}|$. Subsequently, to suppress sparse noise and ensure spatial continuity, we sequentially employ a Gaussian kernel G_σ to smooth local intensity spikes and a median filter $Median(\cdot)$ with kernel size κ to eliminate salt-and-pepper noise:

$$M_{smooth} = Median(G_\sigma * \bar{M}, \kappa) \tag{5}$$

Adaptive Morphological Filtering. To further isolate meaningful perceptual regions, we apply an adaptive threshold $\tau = \mu + \alpha\sigma$, where μ and σ denote the mean and standard deviation of M_{smooth} respectively. We then perform Connected Component Analysis $CCA(\cdot)$ to enforce morphological constraints, discarding any isolated regions whose area $\mathcal{A}$ falls below a minimum size γ:

$$\tilde{M} = Norm\left(M_{smooth} \odot CCA(M_{smooth} > \tau, \mathcal{A} \geq \gamma)\right) \tag{6}$$

Attention-Guided Enhancement. To leverage the refined perceptual signals, we modulate the input image I via the fused attention prior. The enhanced image I' is derived by computing the clamped summation of both perceptual maps, ensuring a soft union of semantic and reasoning cues:

$$I' = I \odot \mathcal{C}(\tilde{M}_s + \tilde{M}_r) \tag{7}$$

where $\mathcal{C}(\cdot)$ denotes a point-wise saturation and normalization operator that rescales the fused attention into the range $[0, 1]$, defined as $Norm(\min(1, \max(0, x)))$. By suppressing task-irrelevant regions while preserving the fidelity of semantic and reasoning anchors, SAF effectively guides the model to focus on the most informative visual cues during multimodal inference.

5 Experiment

5.1 Experimental Setting

Benchmarks and Backbones. We evaluate the proposed SAF across eight benchmarks: A-OKVQA (Schwenk et al., 2022), TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), MME (Fu et al., a), POPE (Li et al., 2023), HallusionBench (Guan et al., 2024) and V* (Wu & Xie, 2024) (details in Appendix A.3). To assess cross-model generalizability, we evaluate our method over four MLLMs, including Qwen-VL (Wang et al., 2024), Gemma (Team et al., 2024), InternVL (Zhu et al., 2025) and Ministral (Liu et al., 2026).

5.2 Result and Analysis

SAF Improves VQA Accuracy. In Figure A7, we show a few qualitative examples of the SAF helping the MLLM perform self-correction. Comprehensive quantitative evaluations across multiple VQA benchmarks in Table 2 demonstrate that our training-free approach yields performance gains in various models. Specifically, when applied to the Qwen-VL model, SAF achieves a 5.5% improvement on MME and a 3.6% boost on V*, which emphasize visual hallucination and fine-grained perception respectively. These results substantiate the effectiveness of SAF in mitigating hallucinations and sharpening visual perception. More detailed analyses in Appendix A.3.

Ablation Study on the Masking Strategies. We investigate the performance of different masking strategy variants. Table 3 (A) isolates the two perceptual components. Both semantic-only and

Table 3: Ablation study on different masking strategies. strategies.

A: Semantic/reasoning ablation. "Semantic" indicates enhancement is performed using only semantic perception, while "Reasoning" indicates reasoning perception only.

Method	AOKVQA	InfoVQA	V*
Base	61.9	42.1	57.1
Semantic	64.8	43.6	54.5
Reasoning	64.5	44.2	60.3
SAF	**65.4**	**45.6**	**60.7**

B: Generic attention-mask baselines. "All-query" averages visual cross-attention across all query tokens to form the mask, whereas "Last-token" derives the mask from the final token.

Method	AOKVQA	InfoVQA	V*
Base	61.9	42.1	57.1
All-query	63.0	41.8	58.6
Last-token	63.4	43.4	58.1
SAF	**65.4**	**45.6**	**60.7**

Table 4: Ablation Study on the token number and layer selection.

A: Ablation Study on the Quantity of Perceptual Tokens. k indicates the number of semantic tokens. m indicates the number of reasoning tokens. Configuration "E" indicates selecting all query token as semantic and reasoning tokens.

Paradigm	k	m	Avg Acc
Base	/	/	62.4
A	1	1	65.1
B	2	2	65.7
C	3	3	64.7
D	5	5	64.8
E	All	All	64.6

B: Ablation Study on the Perception Layer Selection. "A" indicates selecting first half layers. "B" and "C" follow the same logic. "D" indicates selecting all layers. "E" indicates selecting half of layers exhibiting highest perceptual intensity.

Paradigm	Layers	Avg Acc
Base	/	62.4
A	Early Layers	64.5
B	Middle Layers	65.2
C	Late Layers	65.4
D	Full Layers	65.3
E	Selected Layers	65.7

reasoning-only enhancement improve over the baseline, showing that the two signals contribute complementary information. SAF achieves the best performance by jointly preserving query-grounded semantic context and answer-critical evidence. Table 3 (B) compares SAF with two generic attention-mask baselines, all-query attention and last-token attention. Although these baselines improve over the original image in some settings, they remain weaker than SAF, indicating that the semantic/reasoning decomposition provides benefits beyond simple attention masking.

Ablation Study on the Quantity of Perceptual Tokens. We investigate the impact of the number of perception tokens on model performance as summarized in Table 4 (A). Results show that performance begins to degrade as the token count exceeds an optimal threshold. This indicates that the emergence of structured perception is primarily driven by a localized, sparse subset of tokens within the sequence, rather than a global distribution. This result is consistent with the sequence-wise attention dynamics characterized in Section 3.3, further validating the necessity of precise token selection.

Ablation Study on the Perception Layer Selection. We investigate the sensitivity of model performance to different layer selection strategies as shown in Table 4 (B). The results indicate that utilizing early layers leads to the lowest performance, whereas middle, late or all-layer configurations yield competitive results. This is consistent with our findings in Section 3.4 that semantic and reasoning perceptions concentrate in the middle and later layers. SAF employs a filtered subset of perception layers and achieves the optimal results, validating the effectiveness of layers filtering.

6 Conclusion

In this work, we present a systematic investigation of the internal attention dynamics within MLLMs. Our study reveals a structured perceptual process during visual understanding that can be functionally categorized into semantic and reasoning perception. Through comprehensive qualitative and quantitative analyses, we demonstrate that semantic perception exhibits high vision-language alignment, while reasoning perception predominantly activates at the query boundary. Crucially, these perceptions are proved to be intrinsic functions of MLLMs, pervasive across models and data modalities. Building on these insights, we introduce Structured Attention Focusing (SAF), a training-free method that enhances visual inputs by aggregating dual-perceptual attention maps. Experiment validates that SAF improves performance across eight benchmarks. Ultimately, our findings unveil a sophisticated functional hierarchy within MLLM attention, offering a promising foundation for future research into mechanism-driven model enhancement.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing, improving the clarity of manuscript presentation, and AI-assisted coding. Generative AI tools were not used for research ideation, generating experimental results, fabricating data, drafting figures, or making scientific conclusions. All AI-assisted text and code were carefully reviewed, revised, and verified by the authors. In particular, AI-assisted code was manually inspected and tested for correctness, and all scientific claims, experimental results, and conclusions were independently checked by the authors. We take full responsibility for the final content of this work, including any text, code, or other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We are not aware of any ethical concerns or conflicts of interest associated with this work.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of our work. The proposed SAF method, perception layer selection, and mask strategies are described in detail in the main paper. The experimental setup, including datasets, model configurations and hyperparameters is provided in the Appendix. We also report additional implementation details and ablation studies to facilitate verification of our findings in the Appendix. All experiments are conducted on publicly available datasets and models. The code and necessary scripts for reproducing the experiments will be released upon publication.

REFERENCES

Ofir Abramovich, Niv Nayman, Sharon Fogel, Inbal Lavi, Ron Litman, Shahar Tsiper, Royee Tichauer, Srikar Appalaraju, Shai Mazor, and R Manmatha. Visfocus: Prompt-guided vision encoders for ocr-free dense document understanding. In *European Conference on Computer Vision*, pp. 241–259. Springer, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

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.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. Air: Attention with reasoning capability. In *European Conference on Computer Vision*, pp. 91–107. Springer, 2020.

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

Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10876–10885, 2021.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3867–3876, 2025.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, a.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Richard Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. In *ICLR 2025 Workshop on Foundation Models in the Wild*, b.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19520–19529, 2025.

Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. What do vlms notice? a mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11462–11482, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. Overthinking the truth: Understanding how language models process false demonstrations. In *The Twelfth International Conference on Learning Representations*.

Yinhan He, Wendy Zheng, Yushun Dong, Yaochen Zhu, Chen Chen, and Jundong Li. Towards global-level mechanistic interpretability: A perspective of modular circuits of large language models. In *Forty-second International Conference on Machine Learning*, 2025.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Falk Huettig, Joost Rommers, and Antje S Meyer. Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta psychologica*, 137(2):151–171, 2011.

Ming Jiang, Shi Chen, Jinhui Yang, and Qi Zhao. Fantastic answers and where to find them: Immersive question-directed visual attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2980–2989, 2020.

Nicholas Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *The Thirteenth International Conference on Learning Representations*.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25004–25014, 2025.

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

Byeong Su Kim, Jieun Kim, Deokwoo Lee, and Beakcheol Jang. Visual question answering: A survey of methods, datasets, evaluation, and challenges. *ACM Computing Surveys*, 57(10):1–35, 2025.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pp. 18893–18912. PMLR, 2023.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*.

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

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. *Proceedings of Machine Learning Research*, 267:35799–35819, 2025.

Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, et al. A survey on mechanistic interpretability for multi-modal foundation models. *arXiv preprint arXiv:2502.17516*, 2025.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. Reasoning within the mind: Dynamic multimodal interleaving in latent space. *arXiv preprint arXiv:2512.12623*, 2025a.

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

Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*, 2025b.

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

Yexin Liu, Zhengyang Liang, Yueze Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, et al. Unveiling the ignorance of mllms: Seeing clearly, answering incorrectly. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9087–9097, 2025c.

Zhenhua Liu, Tong Zhu, Chuanyuan Tan, Bing Liu, Haonan Lu, and Wenliang Chen. Probing language models for pre-training data detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1576–1587, 2024b.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, et al. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. *arXiv preprint arXiv:2510.17771*, 2025d.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 14662–14684, 2023.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations*.

Anirudh Phukan, Divyansh Divyansh, Harshit Kumar Morj, Vaishnavi Vaishnavi, Apoorv Saxena, and Koustava Goswami. Beyond logit lens: Contextual embeddings for robust hallucination detection & grounding in vlms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9661–9675, 2025.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, et al. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision function layer in multimodal llms. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268 (5217):1632–1634, 1995.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Tu, and Jing Zhang. More thought, less accuracy? on the dual nature of reasoning in vision-language models. *arXiv preprint arXiv:2509.25848*, 2025.

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

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.

Hao Yin, Guangzong Si, and Zilei Wang. Lifting the veil on visual information flow in mllms: unlocking pathways to faster inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9382–9391, 2025.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *13th International Conference on Learning Representations, ICLR 2025*, pp. 28669–28688. International Conference on Learning Representations, ICLR, 2025a.

Zefeng Zhang, Hengzhu Tang, Jiawei Sheng, Zhenyu Zhang, Yiming Ren, Zhenyang Li, Dawei Yin, Duohe Ma, and Tingwen Liu. Debiasing multimodal large language models via noise-aware preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9423–9433, 2025b.

Chenyue Zhou, Mingxuan Wang, Yanbiao Ma, Chenxu Wu, Wanyi Chen, Zhe Qian, Xinyu Liu, Yiwei Zhang, Junhao Wang, Hengbo Xu, et al. From perception to cognition: A survey of vision-language interactive reasoning in multimodal large language models. *arXiv preprint arXiv:2509.25373*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. In *International Conference on Machine Learning*, pp. 80873–80899. PMLR, 2025.

A Appendix

A.1 Detailed Discussion of Dataset Preparation

We analyze the internal attention dynamics of MLLMs on a curated diagnostic dataset. As illustrated in Figure 2, we decouple semantic grounding from visual reasoning by identifying semantic tokens highlighted in blue text within the query and annotating both semantic objects via blue bounding boxes and essential visual evidence via red bounding boxes in the image.

Annotation Protocol and Agreement. To construct reliable probing benchmarks across diverse multimodal domains, we establish a standardized three-stage annotation pipeline. First, candidate instances are randomly sampled from three representative datasets: GQA, TextVQA, and DocVQA, ensuring that each sampled instance corresponds to a unique image to maintain high visual diversity. Second, annotators identify the query semantic tokens and delineate the corresponding semantic and reasoning bounding boxes for each candidate. Third, human experts verify all annotated candidates, retaining 73.3% of the pool to ensure high labeling fidelity. This curation yields a clean diagnostic benchmark comprising 315 rigorously validated samples. To ensure boundary consistency and annotation rigor, an experienced researcher developed standardized guidelines alongside 10 calibrated reference examples. All bounding boxes were independently marked by three annotators, achieving an inter-annotator mean IoU of 0.72 and a Cohen's κ of 0.79, which confirms substantial agreement across the decoupled targets.

A.2 Detailed Discussion of Attention Dynamics in MLLMs

What are the Specific Manifestations of MLLM's Structured Perception? As established in Section 3.3, semantic perception is significant at the semantic token, whereas reasoning perception intensifies at query-boundary tokens. Notably, we find that semantic perception is not restricted to object nouns but extends to diverse lexical categories, including attributes, materials, and spatial relations. For instance, the token *"wearing"* in Figure A1(a) triggers attention on apparel; *"left"* in Figure A1(c) grounds the model's focus to the left region of the image; and *"aluminum"* in Figure A1(f) directs attention to the metal can based on the context *"the can made of aluminum."* These examples underscore the comprehensive scope of vision-language alignment within MLLMs, suggesting a pathway for more granular investigations into their semantic perceptual mechanisms.

Which sequence terminals catalyze reasoning perception? To investigate how different terminal tokens influence reasoning perception, we perform an intervention by substituting the original question mark (*"?"*) with a period (*"."*), a comma (*","*) and a hash symbol (*"#"*). As illustrated in Figure A2, standard punctuation marks (specifically *"?"*, *"."* and *","*) consistently elicit reasoning perception, with the *"?"* yielding the highest intensity. This suggests that conventional punctuation serves as a contextual aggregation signal that triggers the reasoning process. In contrast, the *"#"* symbol, which typically does not function as a syntactic delimiter, fails to activate this perceptual mechanism, indicating the importance of terminal markers in internal visual evidence retrieval.

We further disentangle the identity of the punctuation symbol from its position in the query. Table A1 shows that conventional query-boundary markers, including *"?"*, *"."*, and *","*, consistently produce high terminal-token reasoning BalGR, whereas removing the terminal punctuation or replacing it with the non-syntactic symbol *"#"* substantially weakens this effect. The paraphrase control preserves the same pattern, suggesting that the effect is not tied to a single surface wording. Table A2 further shows that punctuation-like probes are most effective when they serve as genuine query-boundary aggregation points. Randomly inserting a question mark inside the query produces only weak reasoning BalGR, while boundary question marks in multi-question variants remain strong. Overall, reasoning perception is triggered by the interaction between syntactic boundary symbols, query-terminating position, and the accumulated question context.

Cross-Modality Robustness. We evaluate our findings across multiple models spanning different architectures and scales, including LLaVA-1.5 (Liu et al., 2023), Qwen2-VL-7B (Wang et al., 2024), and Qwen2-VL-2B. As shown in the perceptual intensity distributions in Figure A3 (a), both semantic and reasoning perception are primarily concentrated within the middle and later layers across the evaluated models. Notably, the perceptual patterns in LLaVA-1.5 are localized more toward the middle of the network. In addition, the marked decline in intensity within the final layers suggests that the model's deeper architecture may be dedicated to non-perceptual functions. Furthermore, the analysis

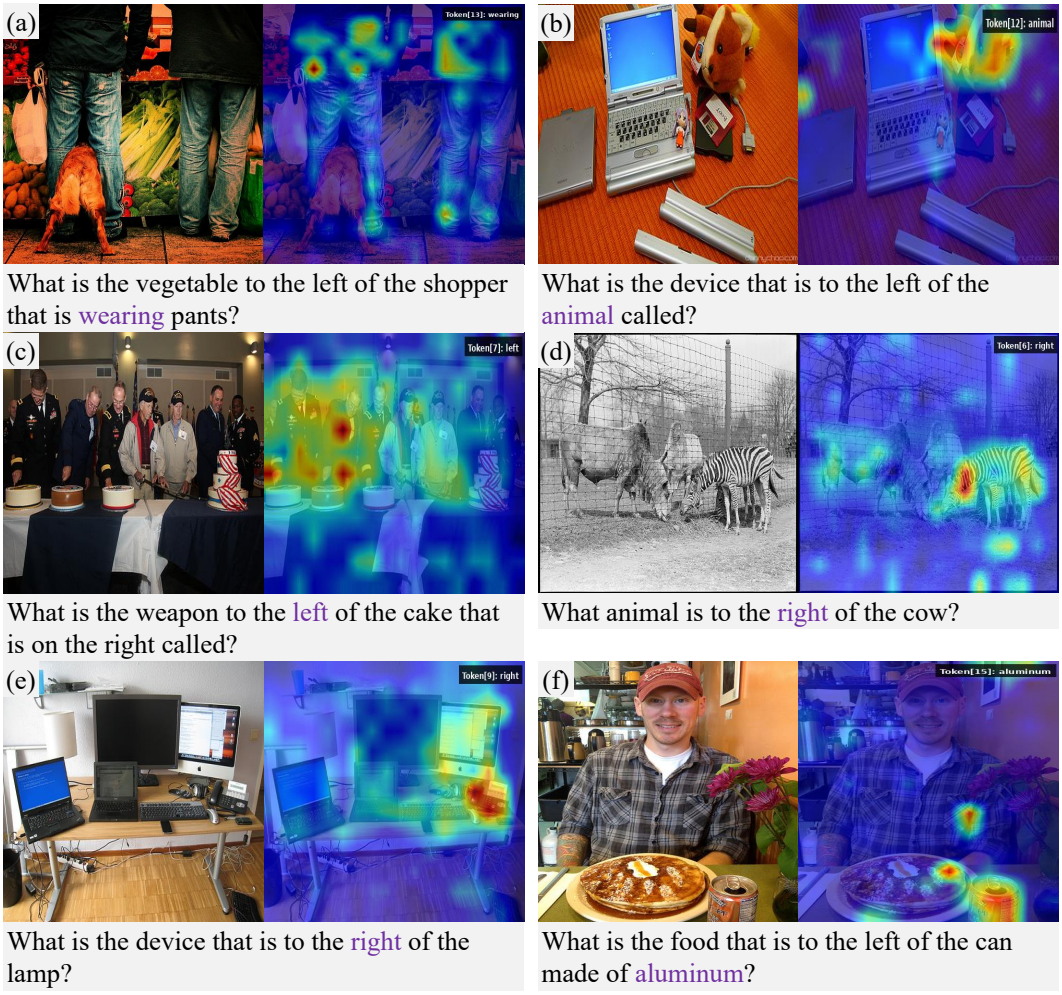


Figure A1: Qualitative visualization of token-level attention maps. Purple highlights indicate query tokens corresponding to each heatmap. Note that semantic perception is not restricted to object nouns but extends to diverse lexical categories, including attributes, spatial relations and materials.

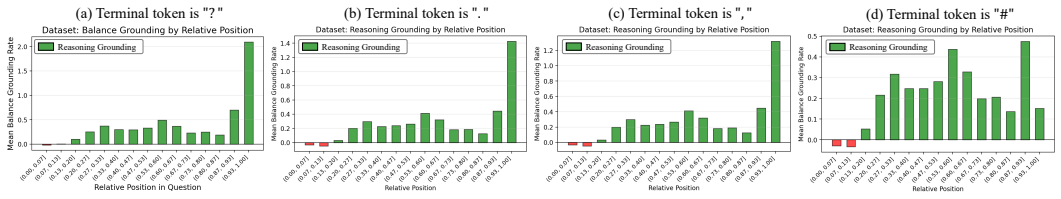


Figure A2: Influence of Terminal Sequence Markers to Reasoning Perception. (a) Terminal token is "?". (b) Terminal token is ".". (c) Terminal token is ",". (d) Terminal token is "#". The results demonstrate that standard syntactic delimiters (specifically "?", ".", and ",") consistently catalyze a localized spike in reasoning perception at the sequence terminus, with the question mark yielding the most pronounced intensity. In contrast, the non-syntactic symbol "#" fails to elicit such a perceptual mechanism, validating the role of conventional punctuation as a contextual aggregation signal essential for internal visual evidence retrieval.

of token distributions and relative positions (Figure A3 (b) and (c)) reveals a consistent trend: semantic perception is highly correlated with semantic query tokens, while reasoning perception is highly correlated with the sequence terminus. This observation validates our core findings that **"semantic**

Table A1: Terminal-marker variants for reasoning perception. "?" is the original question terminator; "no_punctuation" removes the terminal marker; ".", ",", and "#" replace the terminal marker; and "Paraphrase" rewrites the query while preserving its semantics. "TermTok" denotes the special terminal token, "OtherTok" denotes non-terminal query tokens, and "Sem"/"Reas" denote semantic and reasoning target boxes. Conventional query-boundary punctuation maintains strong "TermTok"→"Reas" BalGR, whereas no punctuation or a non-syntactic marker weakens reasoning perception.

Variant	TermTok → Sem	TermTok → Reas	OtherTok → Sem	OtherTok → Reas
"?"	0.162	3.068	0.146	0.418
no_punctuation	1.949	0.869	0.122	0.422
"."	0.073	2.612	0.150	0.413
","	0.170	2.205	0.143	0.433
"#"	0.773	0.326	0.137	0.458
Paraphrase	0.138	3.080	0.103	0.310

Table A2: Probe-position variants for reasoning perception. "Probe position" specifies the token whose BalGR is measured: an appended suffix "Please answer now.", an appended aggregation marker "Answer:", a randomly inserted question mark, the first question mark in a multi-question query, or the third question mark after reordering multiple questions. High "ProbeTok"→"Reas" BalGR appears when the probe functions as a query-boundary aggregation point rather than merely being a punctuation token.

Variant	Probe position	ProbeTok → Sem	ProbeTok → Reas
dummy_suffix	Please answer now.	-0.218	1.168
moved_aggregation	Answer:	-0.163	1.485
random_question_mark	inserted "?"	-0.027	0.523
multi_question	"?" #1	0.181	3.114
multi_question_reordered	"?" #3	0.040	2.402

perception exhibits highly visual-language alignment" and that **"reasoning perception retrieves task-relevant visual evidence for internal visual understanding"**. These results demonstrate that both semantic and reasoning perception mechanisms are universally present across different architectures and scales, exhibiting consistent layer-wise distribution, positional relationships within sequences, and behavioral patterns.

Cross-Modality Robustness. To verify the robustness of our findings, we conduct manual annotations on GQA, TextVQA and DocVQA, focusing on semantic and reasoning perception across different visual modalities (e.g., general objects vs. characters), as shown in Figure A9. We categorize the perceptual targets based on the dataset's domain: (i) in GQA, both semantic and reasoning targets are general objects; (ii) in TextVQA, semantic targets are generic objects while reasoning targets are characters; and (iii) in DocVQA, both targets consist of characters. Qualitative evidence in Figure A4 demonstrates that our observations remain consistent regardless of visual content. Specifically, both perceptual types are predominantly localized within the middle and later layers across all modalities. A subtle exception occurs in DocVQA, where semantic percepting for text emerges earlier in the shallow layers. Notably, the semantic intensity of the "question mark token" significantly exceeds that of other tokens on TextVQA, particularly at the sequence terminus. This stems from the reason that the reasoning bounding box is situated within the semantic bounding box in the majority of annotated samples. As can be seen from Figure A4 (c), the relative positional distribution of semantic perception intensity largely tracks that of the semantic tokens, whereas the reasoning perception intensity consistently peaks at the end of the query sequence. These results confirm that the identified attention dynamics are not dataset-specific artifacts but represent a foundational mechanistic trait of MLLM visual understanding.

Comparison of Samples from Different Modalities. Figures A8, A10, A11, A12 present sample examples from the GQA, TextVQA, and DocVQA datasets. As can be observed, both semantic perception and reasoning perception clearly focus on relevant visual regions and exhibit a structured perceptual process. A special observation emerges during the perception of characters with an image: MLLMs tend to simultaneously attend to the surrounding context and exhibit a heightened intensity toward subsequent characters. This phenomenon may stem from the "next-token prediction" mechanism employed during language model training. Just as linguistic tokens aggregate prefix

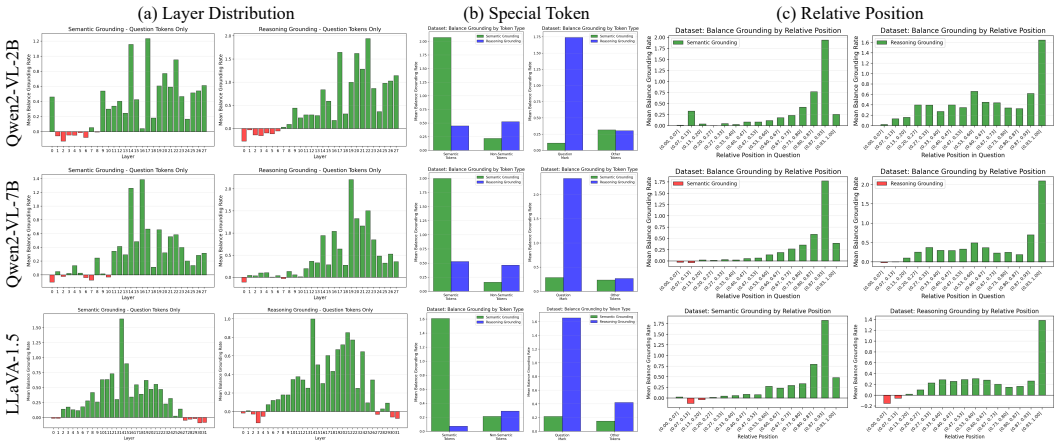


Figure A3: Statistical analysis of attention dynamics on various models. (a) Perception intensity trends across model layers. (b) Comparison of perceptual intensity across token categories. *left*: Semantic *vs.* non-semantic tokens; *right*: Terminal token *vs.* other tokens. (c) Statistical distribution of perceptual intensity across relative positions. *left*: Frequency of semantic tokens; *right*: Reasoning perception intensity. Rows from top to bottom correspond to Qwen2-VL-2B, Qwen2-VL-7B and LLaVA-1.5-7B respectively. The results demonstrate that semantic and reasoning perception mechanisms are universally present across different architectures and scales, exhibiting consistent layer-wise distribution, positional relationships within sequences, and behavioral patterns.

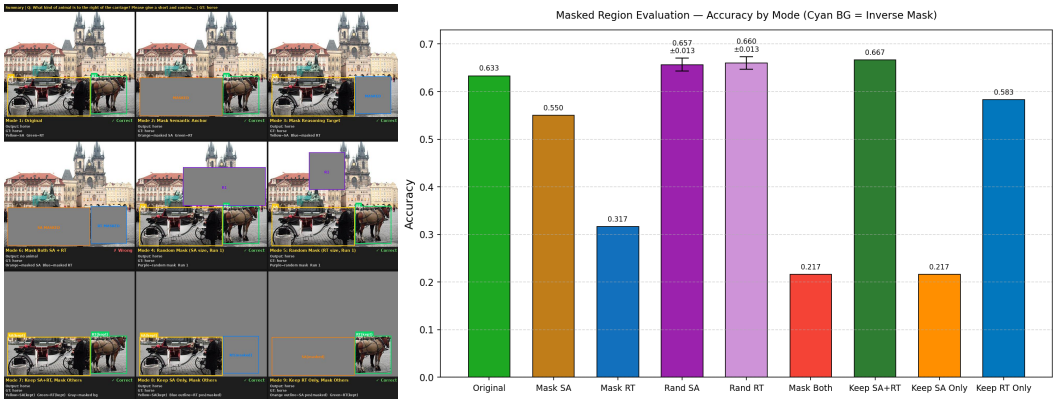


Figure A4: Statistical analysis of attention dynamics on various datasets. (a) Perception intensity trends across model layers. (b) Comparison of perceptual intensity across token categories. *left*: Semantic *vs.* non-semantic tokens; *right*: Terminal token *vs.* other tokens. (c) Statistical distribution of perceptual intensity across relative positions. *left*: Frequency of semantic tokens; *middle*: Semantic perception intensity; *right*: Reasoning perception intensity. Rows from top to bottom correspond to TextVQA, GQA and DocVQA respectively. The results demonstrate that the observed attention dynamics are a fundamental characteristic of MLLM visual understanding, independent of the specific visual modality.

context to anticipate future tokens in a sequence, the model gives rise to a similar "next-character" effect within the process of semantic perception.

A.3 Detailed Discussion of Structured Attention Focus

Implementation Details. We perform experiments on the LLaVA-1.5-7B (Liu et al., 2023), Gemma3-4B (Team et al., 2024), Qwen2-VL-7B(Wang et al., 2024), InternVL3-8B (Zhu et al., 2025) and ministral-3-8B (Liu et al., 2026) models to validate the effectiveness of the proposed method. For main experiments, we fix the hyperparameters to $k = 2$ and $m = 2$. In addition, based on the layer-wise perceptual intensities measured on the annotated dataset, we select half of the layers with



(a) Sample Example (b) Dataset Accuracy

Figure A5: Qualitative and quantitative evaluation of perceptual masking strategies. (a) Visualization of diverse masking configurations, with semantic (yellow box) and reasoning (green box) anchors annotated within the image in the top-left corner. (b) Task performance across different masking strategies. "Original": No Masking. "Mask SA": Masking semantic target. "Mask RT": Masking reasoning target. "Rand SA": Random masking using the semantic bounding box, avoiding overlap with semantic and reasoning targets. "Rand RT": Random masking using the reasoning bounding box, avoiding overlap with semantic and reasoning targets. "Mask Both": Masking both semantic and reasoning targets. "Keep SA+RT": Retaining semantic and reasoning targets. "Keep SA Only": Retaining only semantic targets. "Keep RT Only": Retaining only reasoning targets. For two random masking strategies, we report the mean accuracy and variance across 10 independent trials.

the highest intensity as the candidate layers $\mathcal{L}$. All experiments are implemented on a single NVIDIA A100 (80GB) GPU, with input images resized to a maximum side length of 1024 pixels. For objective and standardized evaluation, we report accuracy across all benchmarks during inference. We employ a prompt to encourage the model to think before generating an answer (e.g., *"Please reason step by step, and then provide the final answer."*). For open-ended tasks, we verify the model's output against the ground-truth short-form answers.

Are Semantic and Reasoning Perceptions Beneficial? To evaluate the individual contributions of semantic and reasoning perception to multimodal question answering, we conduct an ablation study using various masking strategies. As illustrated in Figure A5, masking either semantic or reasoning targets leads to a decline in performance. Specifically, masking semantic targets impairs the model's contextual understanding, while masking reasoning targets prevents the model from capturing essential visual cues and leads to a significant drop in performance. Notably, comparing to masking reasoning targets, masking both targets simultaneously results in a further decline in performance. This may be attributed to the model relying on the semantic context to generate "hallucinated" outputs (e.g., predicting a *"horse"* based on a *"carriage"* even when the horse is masked as shown in Figure A5). Furthermore, random background masking slightly boosts performance by mitigating retrieval complexity. Crucially, retaining reasoning targets without their semantic counterparts leads to below-baseline performance, proving that visual cues alone are insufficient without semantic context. These results demonstrate that semantic and reasoning perception are not redundant but synergistic, both being essential for robust multimodal visual understanding.

A.4 Detailed Discussion of Causal Validation

Causal validation via attention-path ablation. To determine whether MLLMs functionally exploit this structured perception, we perform causal intervention experiments on our annotated probing set. Specifically, keeping the input image intact, we selectively ablate targeted token-to-image attention paths. As the result of Table A3, we find that suppressing either semantic token $\rightarrow$ semantic region or boundary token $\rightarrow$ reasoning region paths induces a substantially steeper accuracy degradation than ablating randomly matched token-to-region paths. These ablation results substantiate the necessity of the identified perception pathways for accurate visual inference.

Table A3: Attention-path ablation on annotated dataset. "SemTok," "TermTok," and "RandomTok" denote semantic tokens, question mark tokens, and randomly selected query tokens, respectively. "semantic bbox" and "reasoning bbox" denote the annotated semantic target region and answer-evidence region. The results show that ablating the structured semantic/reasoning paths causes larger drops than the random token-path control.

Masked attention path	Accuracy (change)
Baseline	58.3 (0.0)
SemTok → semantic bbox	51.7 (-6.6)
TermTok → reasoning bbox	53.3 (-5.0)
RandomTok → semantic/reasoning bbox	57.2±0.78 (-1.1)

Table A4: blation study on functional heads and failure-mode distribution. Panel A reports accuracy under different head-ablation settings, and Panel B reports the error-type distribution of the resulting failure cases. The results indicate that semantic-head ablation causes a larger accuracy drop, while semantic and reasoning heads induce different dominant error types.

A: Study on functional head ablation. "Semantic," "Reasoning," and "Random" denote the ablation of the top 5% of semantic heads, the top 5% of reasoning heads, and a random 5% of heads. Random is reported as mean±std.

Head intervention	**A-OKVQA**
Base	61.9
Semantic	24.5
Reasoning	52.7
Random	59.4±3.0

B: Study on perception-head ablation error analysis. We leverage GPT-5.5 to classify failure cases into two modes: (1) Perception error, where the model introduces ungrounded visual content or factual inaccuracies; (2) Reasoning error, where visual perception is faithful but the deductive chain fails to derive the correct conclusion. "Semantic," "Reasoning," and "Random" follow the same ablation settings as Panel A.

Head intervention	**Perception error**	**Reasoning error**
Semantic	**42.2**	41.3
Reasoning	22.2	**55.3**
Random	30.3±2.8	46.3±1.3

Functional Evidence from Head Ablation. To verify whether semantic and reasoning perception exhibit functional divergence, we ablate the top 5% of each head type identified on our diagnostic set. As reported in Table A4 (A), ablating either group precipitates a sharp degradation in accuracy. To pinpoint the underlying failure mechanisms, we employ GPT-5.5 as an automated evaluator to classify failure cases under a structured rubric (prompt detailed in Table A5). Specifically, incorrect instances are categorized into two primary failure modes: (1) *perception errors*, where the ablated response misses, misidentifies, or hallucinates visual entities and attributes; and (2) *reasoning errors*, where visual perception remains largely grounded, but the model falters in evidence synthesis, relational deduction, or final answer mapping. As the failure-mode distribution reveals in Table A4 (B), semantic-head ablation predominantly provokes perceptual hallucinations and grounding failures, whereas reasoning-head ablation selectively impairs deductive synthesis and final inference. Table A6 illustrates representative examples of this functional dissociation.

A.5 Detailed analysis on robustness and efficiency.

Robustness of BalGR. To ensure our findings reflect intrinsic model mechanics rather than metric artifacts, we benchmark BalGR across computational alternatives, bounding-box jittering, and spatial-corruption controls as shown in Table A7. First, while alternative numerical formulations inevitably alter absolute metric scales, the core qualitative pattern persists: semantic tokens consistently exhibit stronger semantic $BalGR$, while boundary tokens dominate reasoning $BalGR$. Second, adding jitter to ground-truth bounding boxes leaves this separation intact, underscoring the metric's robustness against annotation noise. Finally, destroying the visual-spatial alignment via randomized or center-biased boxes causes this perceptual divergence to collapse entirely. This double dissociation verifies that semantic tokens genuinely ground semantic entities, while boundary tokens selectively retrieve critical reasoning evidence.

Comparison with generic masking and saliency baselines. To verify whether our gains stem from genuine structured perception rather than visual priors, we benchmark SAF against various generic masking and saliency baselines under identical evaluation settings. As shown in Table A9,

Table A5: Instruction for failure-case annotation. The table lists the GPT-5.5 prompt used to annotate failure cases, including input fields, label definitions, annotation rules, and the required JSON output format.

You are an annotation judge for failure cases.
You should compare the Ground Truth Lable and model output. Your task is to classify why the ablated output failed.
Input fields:
- question: {question}
- choices: {choices}
- gold_answer_letter: {gold_answer_letter}
- gold_answer_text: {gold_answer_text}
- direct_answers: {direct_answers}
- rationales: {rationales}
- base_prediction: {base_prediction}
- base_response: {base_response}
- ablated_prediction: {ablated_prediction}
- ablated_response: {ablated_response}

Definitions:
1. perception_hallucination_error:
The ablated response appears to miss, misidentify, or hallucinate visual evidence. This includes wrong object/attribute/count/color/OCR/location, or unsupported visual claims that contradict the gold answer, direct answers, or rationales.
2. reasoning_final_inference_error:
The ablated response contains relevant visual evidence or is not clearly hallucinating, but fails the reasoning, commonsense inference, relation inference, answer selection, or final option mapping.
3. other_unclear:
Use this only when the failure cannot be determined from the provided text and annotations.

Annotation rules:
- Multi-label annotation is allowed.
- Do not treat the copied multiple-choice option list itself as a visual claim.
- If the response states the correct evidence but selects the wrong option, mark reasoning_final_inference_error.
- If the response selects a wrong option because it describes a wrong visual object/attribute/count/text, mark perception_hallucination_error.
- If hallucination/perception and reasoning errors both appear, mark both.
- Judge only the ablated output; use the base output only as a reference for what a successful answer looked like.

Return only valid JSON:

```
{
  "perception_hallucination_error": true/false,
  "reasoning_final_inference_error": true/false,
  "other_unclear": true/false,
  "primary_label": "perception_hallucination_error" | "reasoning_final_inference_error" | "other_unclear",
  "rationale": "One concise sentence explaining the label."
}
```

these baselines systematically test alternative hypotheses across two dimensions: first, spatial priors, including random and center-biased preservation; second, unstructured attention, such as averaging across all query tokens or isolating the terminal token. Simple spatial heuristics like Center same-area at 54.6% average and heuristic attention maps like Last-token at 55.0% average offer only modest improvements over the 53.7% Base baseline. In contrast, SAF achieves superior performance across all benchmarks with a 57.2% average, notably outperforming its isolated components, Semantic only at 54.3% and Reasoning only at 56.7%. This confirmatory advantage demonstrates that the performance gain cannot be explained by generic background suppression or center-bias effects, but fundamentally arises from the decoupled semantic-reasoning synergy.

Comparison with training-free test-time methods. We benchmark SAF against two closely related test-time enhancement paradigms under identical backbone and evaluation protocols (Table A8):

Table A6: Examples of failure cases after functional-head ablation. Each case reports the question, choices, ground-truth answer, base output, ablated output, and takeaway. Case 1 corresponds to semantic-head ablation and illustrates a perception failure; Case 2 corresponds to reasoning-head ablation and illustrates a final-inference failure.

Case 1 (Semantic-head ablation -> perceptual failure).

Q: What sport are they mimicking? (A: baseball; B: lacrosse; C: cricket; D: basketball; GT: A)

Base output: "...The person in the foreground is holding a bat and appears to be swinging it... the person in the background is holding a ball and appears to be throwing it... the sport they are mimicking is: A: baseball" (correct)

After ablation: "...The image shows a room with shelves, cabinets, and various items that resemble a kitchen or a storage area... the most likely answer is C: basketball." (incorrect)

Takeaway: Ablating semantic heads shifts attention to irrelevant regions, causing a perceptual error.

Case 2 (Reasoning-head ablation -> reasoning failure).

Q: What vehicle is shown in the photo? (A: airplane; B: bus; C: train; D: car; GT: A)

Base output: "...a large, stationary aircraft with a cockpit and wings... markings that read 'USAF'... mounted on a concrete pedestal... the vehicle shown in the photo is: A: airplane" (correct)

After ablation: "The vehicle shown in the photo is a car." (incorrect, no reasoning given)

Takeaway: Ablating reasoning heads disrupts the required visual reasoning, yielding a direct unsupported answer.

Table A7: BalGR robustness under annotation and computation variants. "SemTok," "TermTok," and "OtherTok" denote semantic, question mark, and other query tokens; "Sem" and "Reas" denote semantic and reasoning target boxes. "no minmax" removes min-max normalization, "head average" averages heads before normalization, "jitter" perturbs annotated boxes, and center/random boxes are spatial controls. The semantic-token/semantic-box and terminal-token/reasoning-box preferences remain stable under valid variants.

Ablation	SemTok → Sem	SemTok → Reas	TermTok → Sem	TermTok → Reas	OtherTok → Sem	OtherTok → Reas
origin	2.215	0.718	0.162	3.068	0.146	0.418
no minmax	4.369	0.587	0.199	4.841	0.090	0.453
head average	2.348	0.785	0.189	3.421	0.163	0.446
jitter 5%	2.213	0.715	0.164	3.051	0.149	0.417
jitter 10%	2.188	0.724	0.158	3.013	0.150	0.412
jitter 20%	2.154	0.719	0.193	2.869	0.162	0.403
center box	0.701	0.693	0.467	0.429	0.264	0.229
random box, area	-0.052	0.129	0.177	-0.155	0.106	0.063
random box, +center-dist	0.015	0.222	0.166	-0.145	0.086	0.092

Visual Attention Redistribution (VAR) (Kang et al.), which reallocates internal visual attention weights, and Automatic Visual Cropping (ViCrop) (Zhang et al., 2025a), which relies on spatial zoom-in crops for fine-grained details. As shown in Table A8, SAF consistently outperforms both alternatives across all benchmarks, achieving an average accuracy of 57.2%—surpassing VAR and ViCrop by +2.4% and +2.6%, respectively. Notably, on detail- and layout-intensive tasks like InfoVQA and V*, SAF widens the performance gap (up to +3.7% over ViCrop). These consistent margins demonstrate that rather than relying solely on internal weight rebalancing or heuristic visual cropping, jointly harnessing complementary semantic and reasoning guidance yields a substantially more effective and generalizable test-time intervention.

Robustness of SAF across evaluation regimes. To verify that the observed improvements are not artifacts of a specific generation or scoring setup, we evaluate SAF under varied random seeds, prompt templates, decoding configurations, and the official A-OKVQA answer-matching protocol (Table A10). As shown, SAF delivers remarkably consistent performance gains across all conditions, yielding an improvement of $+3.0\%$ to $+3.3\%$ over the Base model. Notably, SAF exhibits lower variance across different seeds and decoding settings. These results substantiate that SAF's performance advantages are genuine, robust, and independent of specific prompt engineering or evaluation hyperparameters.

Table A8: Comparison with related training-free test-time methods under the same backbone and evaluation protocol. "VAR" denotes implementation of Visual Attention Redistribution, while "ViCrop" denotes Automatic Visual Cropping. "Avg." is the macro-average across the three benchmarks.

Method	A-OKVQA	InfoVQA	V*	Avg.
Base	61.9	42.1	57.1	53.7
VAR (Kang et al.)	63.9	42.9	57.6	54.8
ViCrop (Zhang et al., 2025a)	64.4	41.9	57.6	54.6
SAF	**65.4**	**45.6**	**60.7**	**57.2**

Table A9: Generic masking and saliency baselines. "Random same-area" and "Center same-area" preserve random or center regions with the same area as the SAF mask. "Random layer" uses attention maps from randomly selected layers; "All-query attention" aggregates attention over all query tokens; "Last-token attention" uses only the terminal-token attention map. "Semantic only" and "Reasoning only" use one SAF branch, while "SAF" fuses both branches and obtains the best average accuracy.

Method	A-OKVQA	InfoVQA	V*	Avg.
Base	61.9	42.1	57.1	53.7
Random same-area	62.7	42.2	55.3	53.4
Random layer	64.1	42.3	55.5	54.0
Center same-area	64.5	42.6	56.7	54.6
All-query attention	63.0	41.8	58.6	54.5
Last-token attention	63.4	43.4	58.1	55.0
Reasoning only	64.5	44.2	60.3	56.7
Semantic only	64.8	43.6	54.5	54.3
SAF	**65.4**	**45.6**	**60.7**	**57.2**

Hyperparameter sensitivity of mask construction. We evaluate the sensitivity of SAF to key hyperparameter choices in the mask-generation pipeline on A-OKVQA (Table A11). Our default configuration employs a Gaussian smoothing kernel of size 5, a median filter size of 3, a threshold coefficient $\alpha = 0.10$, and a minimum connected-region area of 0.001. We systematically perturb each hyperparameter across a standard range while holding the others fixed. As shown in Table A11, SAF maintains consistently performance across all variations (ranging from 64.7 ± 0.74 to 65.8 ± 0.39), exhibiting minimal variance (≤ 0.74). In particular, the threshold coefficient demonstrates zero performance fluctuation (± 0.0). These results verify that SAF is not brittle to mask-filtering heuristics and functions reliably without tedious hyperparameter tuning.

Extraction of Semantic Perceptual Attention Maps. We identify semantic tokens by quantifying the spatial concentration of their attention maps via entropy. As illustrated in Figure A6 (a), the semantic token *"lamp"* exhibits the lowest entropy value. This phenomenon arises because semantic tokens facilitate a highly localized focus on specific visual targets, resulting in lower entropy. This hypothesis is further substantiated by our dataset-wide analysis in Figure A6(b). We represent semantic tokens with "$\times$" and non-semantic tokens with "$\cdot$", where the horizontal axis of the scatter plot denotes entropy values and the vertical axis represents semantic perception intensity. In this scatter plot, semantic tokens predominantly cluster in the upper-left quadrant, which demonstrates that semantic tokens usually exhibit high semantic perception intensity and low entropy. This confirms that attention entropy can serve as a robust and effective metric for isolating semantic anchors within the query sequence.

SAF Improves VQA Accuracy. Figure A7 presents several qualitative examples where SAF facilitates self-correction within MLLMs. For the question *"What is the metal basket near the net used to hold?"*, the enhanced image retains the semantically relevant region *"net"* and highlights the critical visual evidence required to address the query. This refinement enables the correction of the response of the model from *"bats"* to *"tennis balls."* Similarly, for the question *"What type of city district is this?"*, background information within the image is instrumental in determining the urban category, whereas the figure occupying the foreground provides no assistance. Consequently, by masking this central figure, the enhanced image redirects visual attention toward background cues, thereby assisting the model in reaching the correct determination. Comprehensive quantitative evaluations across multiple VQA benchmarks, as summarized in Table 2, demonstrate that the proposed training-free approach consistently yields performance gains. Specifically, when applied to the Qwen-VL model, SAF achieves a 5.5% improvement on MME and a 3.6% increase on

Table A10: Robustness across evaluation settings on A-OKVQA. "Different seeds," "Different prompts," and "Different decoding settings" evaluate variations in random seed, instruction template, and decoding strategy, respectively. "Official answer matching" uses the official A-OKVQA evaluation protocol. Values with ± report dispersion across variants, and SAF consistently improves over the matched Base setting.

Evaluation setting	Base acc.	SAF acc.	Gain
Different seeds	62.5±0.46	65.6±0.18	+3.1
Different prompts	62.6±1.25	65.7±1.76	+3.1
Different decoding settings	62.3±0.49	65.3±0.12	+3.0
Official answer matching	76.4	79.7	+3.3

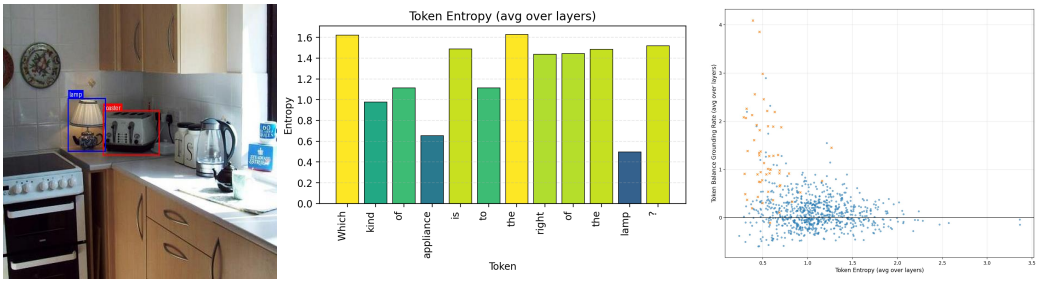
Table A11: Sensitivity of mask-construction components on A-OKVQA. "Gaussian smoothing," "Median filter size," "Threshold coefficient," and "Connected-region min. area" denote the four mask-generation hyperparameters varied one at a time around the default setting. Each entry reports mean±std accuracy over the tested values, showing that SAF is stable to these mask-construction choices.

Mask-construction component	A-OKVQA
Gaussian smoothing	64.7±0.74
Median filter size	65.4±0.39
Threshold coefficient	65.4±0.10
Connected-region min. area	65.8±0.39

V*, which prioritize visual hallucination and fine-grained perception, respectively. These results substantiate the efficacy of SAF in mitigating multimodal hallucinations and sharpening visual perception. Furthermore, SAF demonstrates robust cross-architecture generalizability and remains effective across various models.

Ablation Study on the Quantity of Perceptual Tokens. We investigate the impact of the number of perception tokens on model performance as summarized in Table A12. The results demonstrate that performance begins to degrade once the token count exceeds an optimal threshold, suggesting potential interference from redundant information. Notably, this indicates that the emergence of structured perception is primarily driven by a localized, sparse subset of tokens within the sequence rather than a uniform global distribution. This observation is consistent with the sequence-wise attention dynamics characterized in Section 3.3, further validating the necessity of precise token selection.

Ablation Study on the Perception Layer Selection. We examine the sensitivity of model performance to different layer selection strategies as shown in Table A13. The results indicate that utilizing early layers leads to the lowest performance, whereas middle, late, or all-layer configurations consistently yield more competitive results. This aligns with our findings in Section 3.4 that semantic and reasoning perceptions concentrate in the middle and later layers. Ultimately, SAF employs a filtered subset of perception layers to achieve the optimal results, thereby empirically validating the effectiveness of the layer filtering mechanism.



(a) Sample Example (b) Dataset Analysis

Figure A6: Entropy analysis of token-wise attention maps. (a) An illustrative case where the semantic token *"lamp"* yields the lowest entropy. (b) Dataset-level scatter plot, with the x-axis representing entropy values and the y-axis representing semantic perception intensity. "×" and "·" denote semantic and non-semantic tokens respectively. Semantic tokens are predominantly clustered in the upper-left quadrant, indicating that they are characterized by high semantic perception intensity and low entropy.

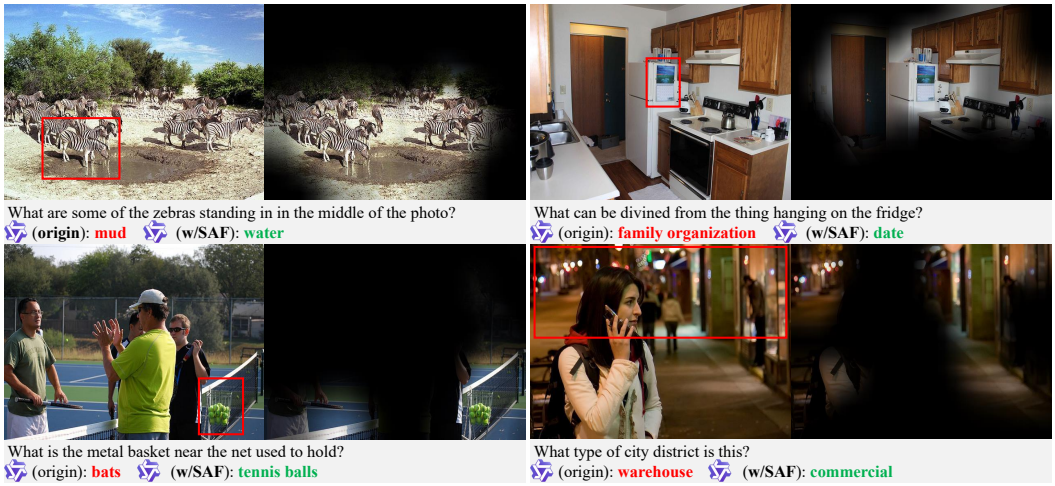


Figure A7: Qualitative comparison of raw and enhanced images. "origin" denotes the original image and response, while "w/SAF" denotes the SAF-processed image and response. Task-critical visual evidence is highlighted within red bounding boxes. By suppressing visually irrelevant background context, SAF effectively grounds the model's focus on essential cues, thereby rectifying erroneous origin responses.

Table A12: Ablation Study on the Quantity of Perceptual Tokens. k indicates the number of semantic tokens. m indicates the number of reasoning tokens. Configurations A–D set $k = m$ to 1, 2, 3, and 5, respectively. Configuration "E" indicates selecting all query token as semantic and reasoning tokens. The results show that selecting a small number of perceptual tokens is more effective than using too many or all query tokens.

Paradigm	k	m	AOKVQA	TextVQA	DocVQA	InfoVQA	MME	POPE	HallusionBench	V*
Base	/	/	61.9	77.8	78.6	42.1	65.5	84.0	32.0	57.1
A	1	1	64.3	81.6	80.1	43.6	71.1	87.0	36.0	57.1
B	2	2	65.4	81.8	79.5	45.6	71.0	87.5	34.2	60.7
C	3	3	64.6	81.4	79.7	43.8	70.1	87.5	33.1	57.1
D	5	5	63.0	82.0	79.9	44.0	70.4	87.7	33.2	58.1
E	all	all	63.3	81.8	80.0	43.7	70.3	87.6	33.0	57.4

Table A13: Ablation Study on the Perception Layer Selection. A, B, and C correspond to early, middle, and late layer subsets, respectively; D uses all layers, and E uses perception-intensity-selected layers. The results show that middle/late or selected perception layers are more effective than early layers

Paradigm	layers	AOKVQA	TextVQA	DocVQA	InfoVQA	MME	POPE	HallusionBench	V*
Base	/	61.9	77.8	78.6	42.1	65.5	84.0	32.0	57.1
A	Early Layers	63.8	81.6	79.7	42.6	70.9	87.2	33.9	56.5
B	Middle Layers	66.2	81.2	79.5	44.3	71.2	87.3	32.7	59.2
C	Late Layers	64.5	81.6	79.8	45.4	71.2	87.6	34.2	59.2
D	Full Layers	64.5	81.6	80.0	45.7	70.2	87.4	34.3	58.6
E	Selected Layers	65.4	81.8	79.5	45.6	71.0	87.5	34.2	60.7

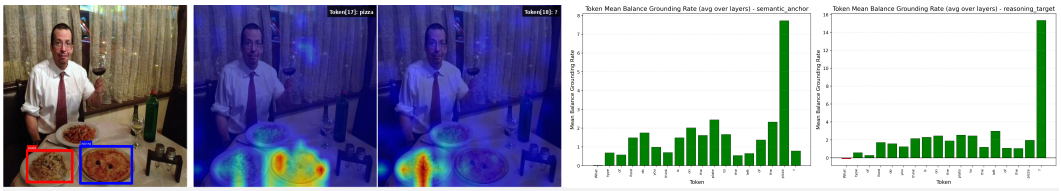


Figure A8: Visual comparison of semantic and reasoning perception on GQA samples. Semantic targets (blue box) and visual evidence (red box) are annotated within the image. Heatmaps illustrate the attention distributions for semantic tokens and the final token of the query sequence respectively. Bar charts present the semantic and reasoning perception intensities for each token within the query. Note that semantic perception peaks at the semantic tokens, whereas reasoning perception surges significantly at the sequence end.

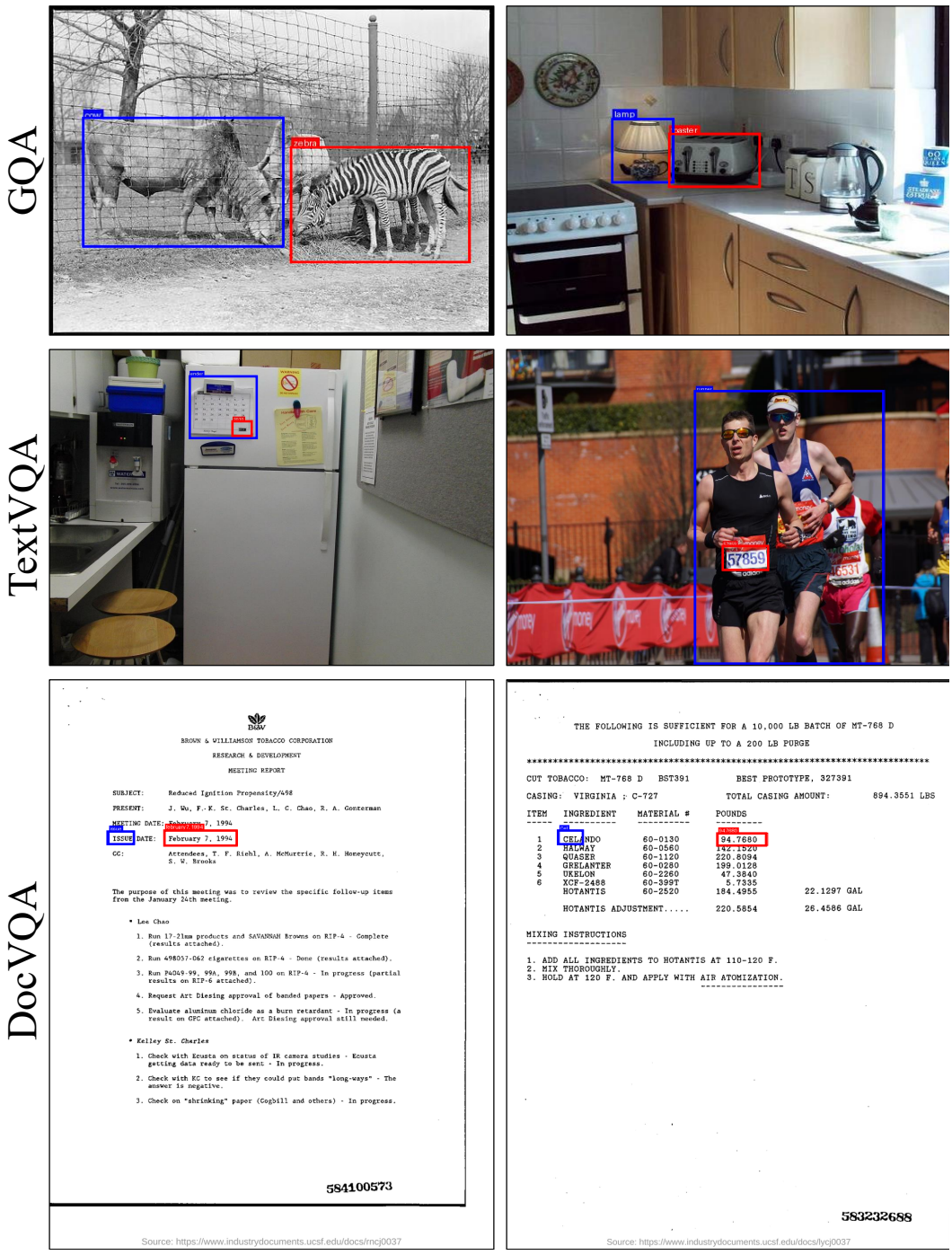
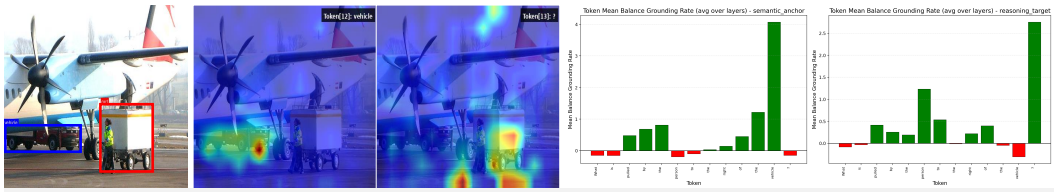
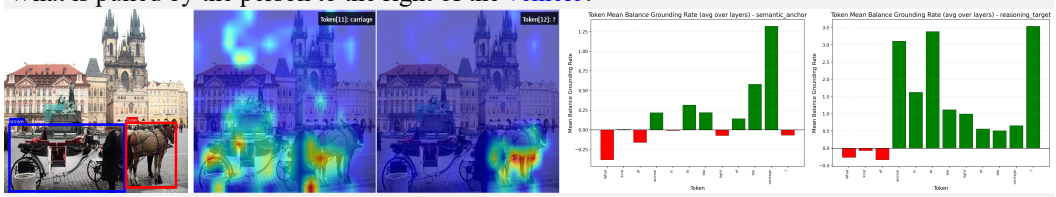
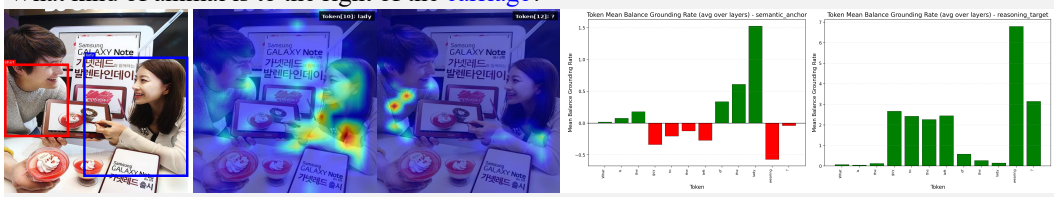
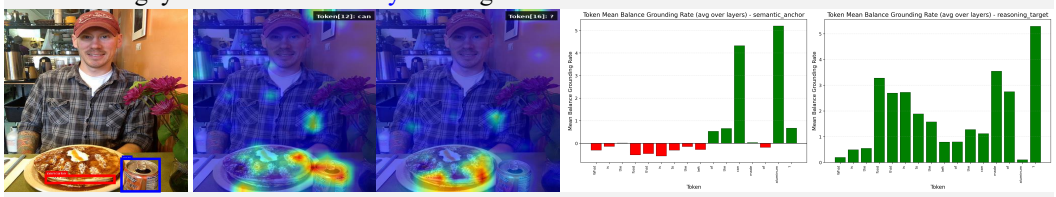


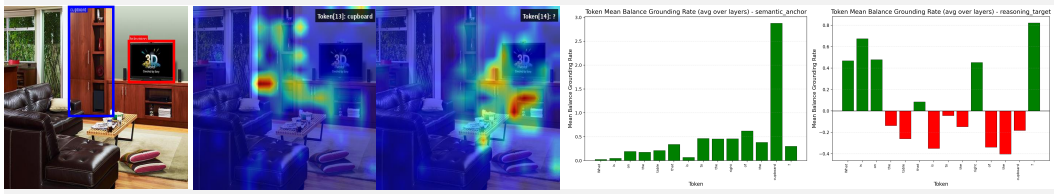
Figure A9: Examples of annotated samples from different datasets. We categorize the perceptual targets based on the dataset's domain: (i) in GQA, semantic targets (blue bounding box) and reasoning targets (red bounding box) are both general objects; (ii) in TextVQA, semantic targets are generic objects while reasoning targets are characters; and (iii) in DocVQA, both targets consist of characters.

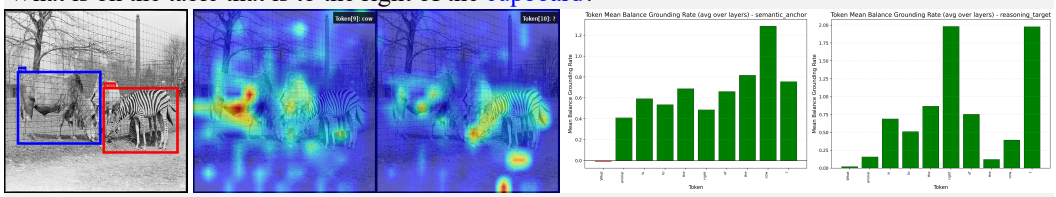












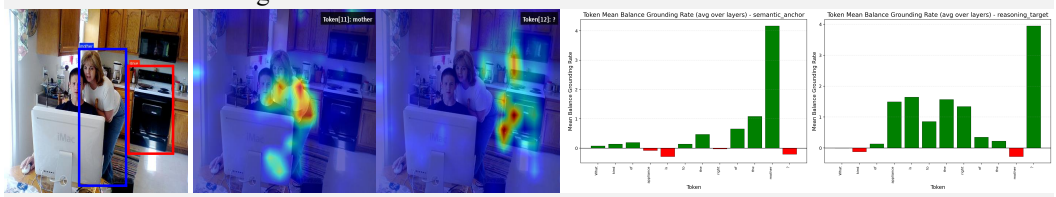
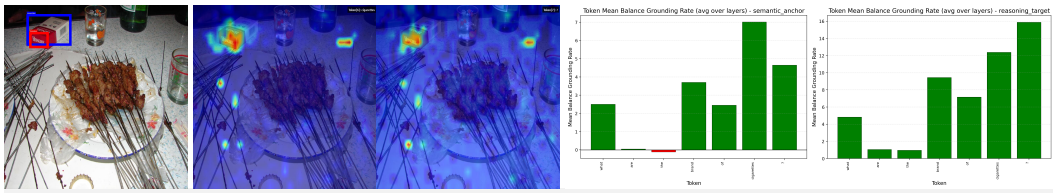
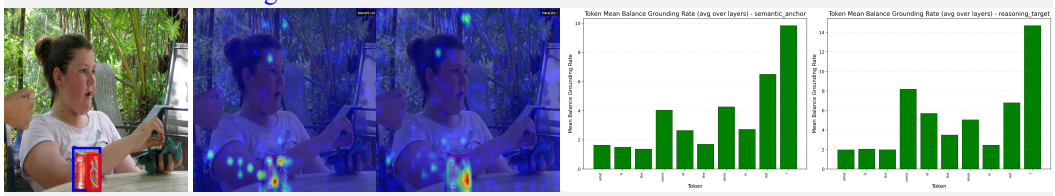
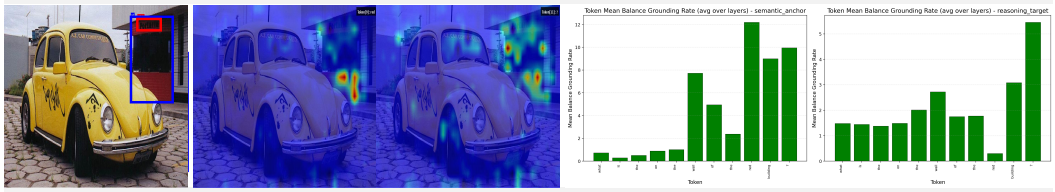
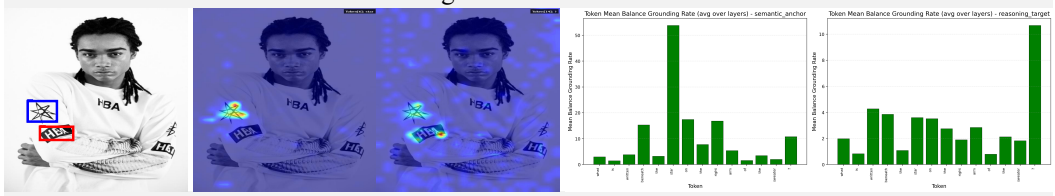


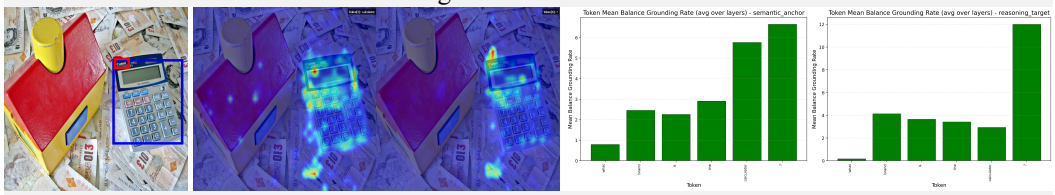
Figure A10: Visual comparison of semantic and reasoning perception on GQA samples. Semantic targets (blue box) and visual evidence (red box) are annotated within the image. Heatmaps illustrate the attention distributions for semantic tokens and the final token of the query sequence respectively. Bar charts present the semantic and reasoning perception intensities for each token within the query. Note that semantic perception peaks at the semantic tokens, whereas reasoning perception surges significantly at the sequence end.

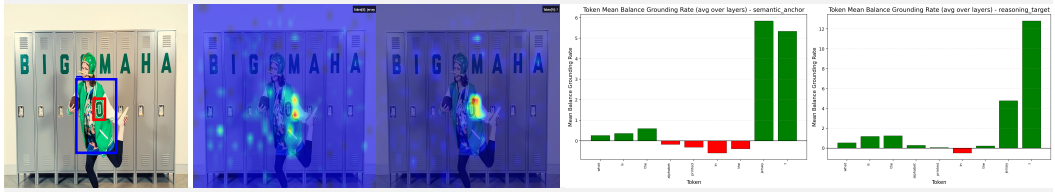












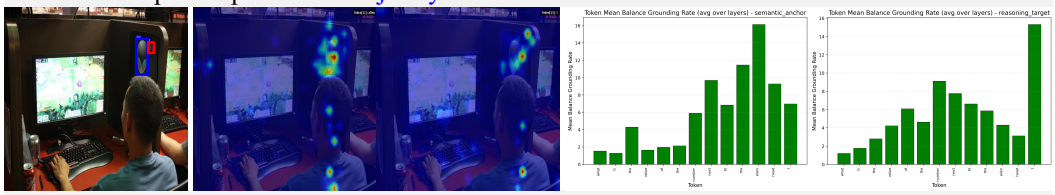
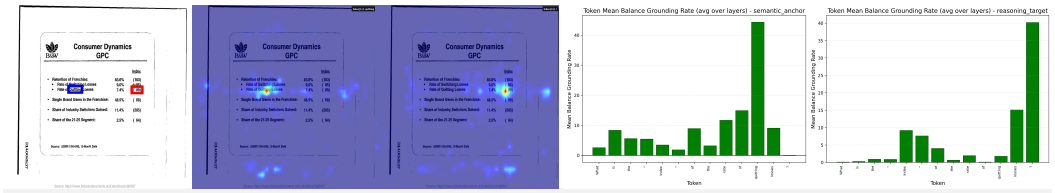
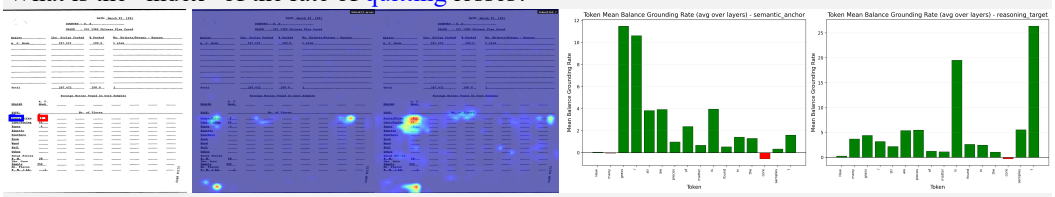


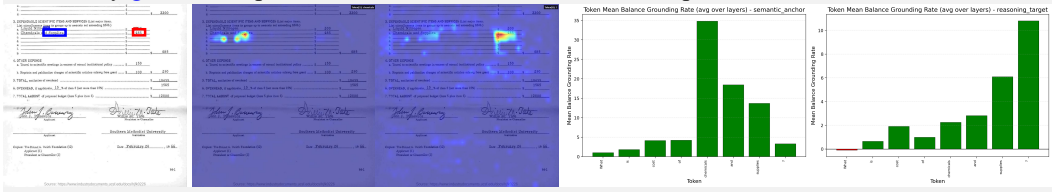
Figure A11: Visual comparison of semantic and reasoning perception on TextVQA samples. Semantic targets (blue box) and visual evidence (red box) are annotated within the image. Heatmaps illustrate the attention distributions for semantic tokens and the final token of the query sequence respectively. Bar charts present the semantic and reasoning perception intensities for each token within the query. Note that semantic perception peaks at the semantic tokens, whereas reasoning perception surges significantly at the sequence end.



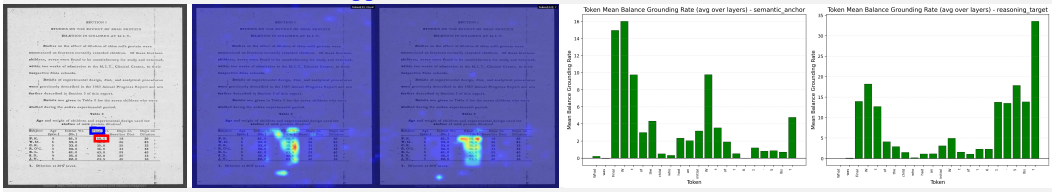
What is the “index” of the rate of quitting losses?



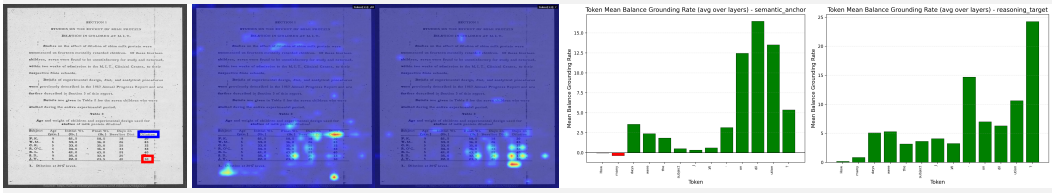
How many grass/straw pieces of matter is found in the core samples?



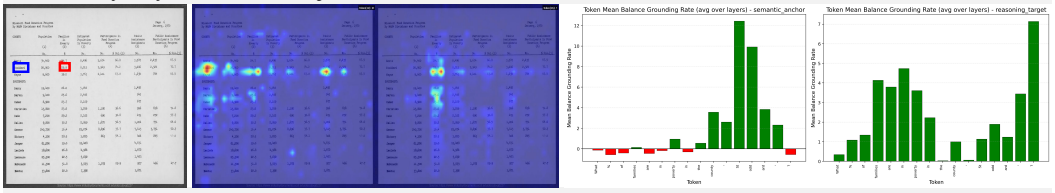
What is cost of chemicals and supplies?



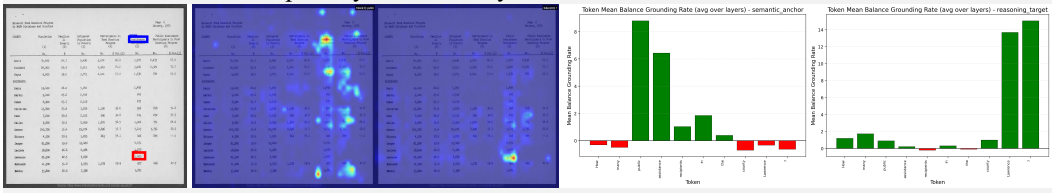
What was Final Wt of the child who had an initial Wt of 61.5 lbs?



How many days were the subject J.W. on dilution?



What % of families are in poverty in the county ‘Stoddard’?



How many public assistance recipients in the county Lawrence?

Figure A12: Visual comparison of semantic and reasoning perception on DocVQA samples. Semantic targets (blue box) and visual evidence (red box) are annotated within the image. Heatmaps illustrate the attention distributions for semantic tokens and the final token of the query sequence respectively. Bar charts present the semantic and reasoning perception intensities for each token within the query. Note that semantic perception peaks at the semantic tokens, whereas reasoning perception surges significantly at the sequence end.