Beyond the Panorama: Training-Free Hierarchical Perception-Reasoning for Fine-Grained Vision in MLLMs

Xiaoyang Yi[1,3,4], Jing Chen[2,3,4], Li Peng[1,3,4], Yuru Bao[1,3,4], Jian Zhang[1,2,3,4]*
[1]College of Cryptology and Cyber Science, Nankai University
[2]College of Computer Science, Nankai University
[3]Tianjin Key Laboratory of Network and Data Security Technology
[4]Key Laboratory of Data and Intelligent System Security, Ministry of Education
***Correspondence:** zhang.jian@nankai.edu.cn

Abstract

Multimodal large language models (MLLMs) enable cross-modal semantic understanding and generation by learning semantic alignment and fusion across modalities. However, existing MLLMs still face challenges in fine-grained visual tasks. Their uniform encoding for global understanding tends to blur or lose local details, while the lack of explicit modeling of intermediate visual evidence leads them to rely on semantic priors or the statistical patterns of language models rather than grounded visual information, resulting in potential hallucinations. To address these issues, we propose HiPerson, a training-free hierarchical perception-reasoning framework that enhances fine-grained visual understanding by simulating human perception mechanisms. Specifically, HiPerson fuses internal relative attention and gradient activation signals to generate a task-aware semantic heatmap, providing explicit perceptual anchors for precise localization. Then, it employs a dual-scale adaptive cropping strategy to extract visual cues for interactive reasoning, simulating the process of human visual focus shifting and detail attention. Finally, by combining local-global dual-image cooperative input with a multi-step reasoning prompting mechanism, HiPerson guides the model to complete a full perception loop from detail observation to contextual verification. Experiments show that HiPerson achieves competitive results on multiple datasets, demonstrating its generalizability and scalability.

1 Introduction

In recent years, with the continuous maturation of visual pre-training and large language models (LLMs), multimodal LLMs (MLLMs) have experienced a strong momentum for cross-modal semantic understanding and generation (Xu et al., 2025; Cai et al., 2025; Park et al., 2025). Typically, MLLMs employ a visual encoder to transform input images into a series of visual tokens, which are then passed through a lightweight connector and concatenated with text tokens. The combined tokens are subsequently processed by a frozen LLM for unified autoregressive modeling, achieving deep integration of image content and linguistic semantics (Zhang et al., 2025b; Li et al., 2025a). MLLMs leverage the powerful generalization and composition abilities of LLMs to convert visual perception problems into sequence understanding tasks, leading to breakthroughs in cross-modal semantic understanding and generation capabilities (Li et al., 2025b; Lu et al., 2025).

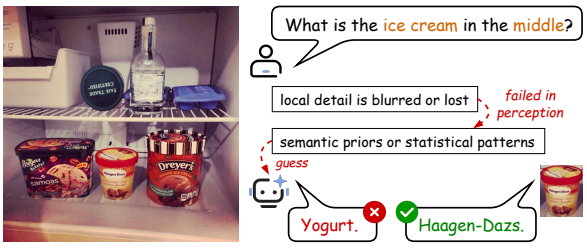


Figure 1: When faced with fine-grained visual tasks, MLLMs may struggle to perceive local details and rely on semantic priors or statistical patterns for guess.

Unfortunately, when confronted with fine-grained visual tasks, existing MLLMs still exhibit fundamental limitations (Zhang et al., 2023, 2024). As shown in Figure 1, fine-grained visual tasks require MLLMs to precisely localize the task-relevant region within the high-dimensional image and extract sufficient details for reliable reasoning (Yang et al., 2025; Feng et al., 2025). If the target region occupies only a minimal portion of the input image, traditional MLLMs struggle to provide correct answers since their uniform encoding scheme for global understanding compresses all image information into a fixed-length low-dimensional sequence (Wu and Xie, 2024), causing an amount of local details to be blurred or lost during encoding. While several methods enhance local perception via image region cropping (Jiang et al., 2025a; Su

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30534–30549

et al., 2025), they rely on additional training or explicit region annotations, making it difficult to achieve generalization improvements without incurring training costs.

Moreover, current MLLMs generally adopt an black-box inference mode, which simplifies complex visual understanding tasks into a perception-as-decision mapping process, lacking explicit modeling and traceability of intermediate visual evidence (Chen et al., 2024b). They are forced to respond despite insufficient information, tending to rely on semantic priors or the statistical patterns of the model to guess, rather than deriving answers based on genuine visual evidence (Wang et al., 2025a; Liu et al., 2025). This dependence on linguistic priors reduces the reliability of the model, especially in the presence of visual interference or when tasks involve atypical scenarios, potentially leading to hallucinations (Qi et al., 2025). Although some studies attempt to introduce chain-of-thought prompting mechanisms (Mondal et al., 2024; He et al., 2024), they treat visual perception and language-based reasoning as independent modules, neglecting substantial cross-modal interaction and communication.

In contrast, humans often exhibit a highly structured perception process when understanding complex visual scenes (Shao et al., 2024). They first rapidly locate the potential target region in the image based on linguistic cues, then focus on those regions to capture key details, and finally integrate visual evidence with contextual information through multi-step reasoning to reach reliable conclusions. This process combines top-down semantic guidance with context-driven perception, forming a tightly coupled cross-modal reasoning system that integrates perception and reasoning. In comparison, current MLLMs complete tasks from perception to decision in a single forward pass. Therefore, bridging their gap in fine-grained visual tasks requires reconstructing their reasoning paradigm to simulate the phased reasoning architecture of human cross-modal perception.

Armed with this insight, we propose HiPerson, a training-free **Hi**erarchical **Per**ception-rea**son**ing framework, aimed at guiding MLLMs to enhance their perception of fine-grained visual content. Specifically, HiPerson employs a dual-path region refinement mechanism to collaboratively extract the model's internal relative attention and gradient activation signals, constructing a comprehensive semantic heatmap that mimics human context-sensitive focusing to achieve task-aware localization. Subsequently, it introduces a dual-scale adaptive cropping strategy, which can adaptively amplify the key region from the heatmap via multi-scale sliding window search and saliency assessment, simulating the process of human visual focus shifting and detail attention. Finally, by constructing dual-view enhanced reasoning, HiPerson concurrently inputs both the global scene and local details into the model, and introduces a structured multi-step reasoning prompting mechanism, guiding the model to complete the full perception loop from detail perception to contextual verification within a unified reasoning process.

To summarize, our contributions are as follows:

- We propose HiPerson, a training-free hierarchical perception-reasoning framework for MLLMs, designed to enhance perception of fine-grained visual content.
- We achieve task-aware localization and extract reliable visual cues for interactive reasoning through a dual-path region refinement and a dual-scale adaptive cropping strategy.
- We combine a local-global cooperative input a multi-step reasoning mechanism, guiding the model to complete the perception loop from detail perception to contextual verification.
- Experimental results demonstrate that HiPerson can improve performance on fine-grained visual tasks while enhancing the interpretability of the reasoning process.

2 Related Work

Existing MLLMs primarily focus on efficiently aligning powerful LLMs with visual encoders to achieve understanding, reasoning, and conversational capabilities regarding visual content. Specifically, LLaVA (Liu et al., 2023) connects a pre-trained CLIP visual encoder with an LLM via a simple trainable projection layer. BLIP (Li et al., 2022) employs a unified vision-language pre-training (VLP) framework to jointly optimize image-text understanding and generation tasks. BLIP-2 (Li et al., 2023) proposes a lightweight query transformer to efficiently connect and bridge the frozen visual encoder and the LLM. Building upon BLIP-2, InstructBLIP (Dai et al., 2023) introduces instruction-aware visual feature extraction and fine-tuning

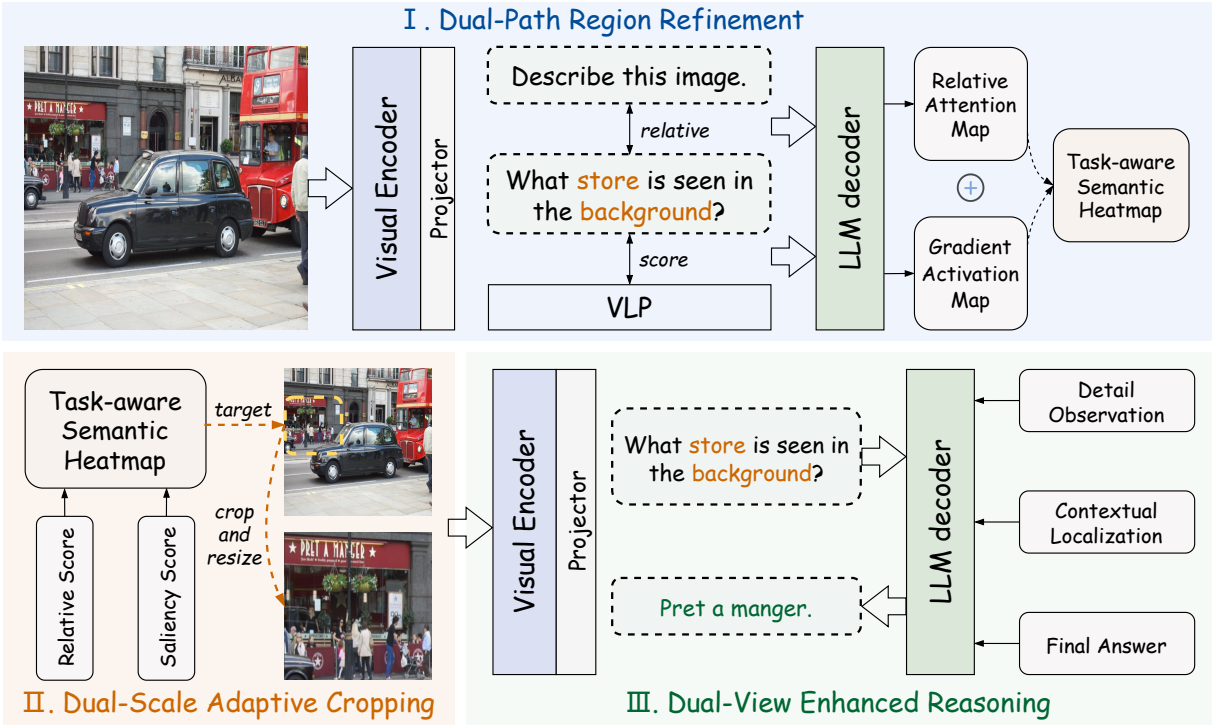


Figure 2: The overall framework of HiPerson. It first extracts relative attention map and gradient activation map in parallel, fusing them to generate a task-aware semantic heatmap, which is then used to precise focusing via relative search and saliency assessment. Finally, it coordinates the global scene and local details to complete a full perception closed-loop within a unified reasoning process.

based on large-scale instruction data. Qwen-VL (Bai et al., 2023) introduces a multi-resolution visual tokenizer and fine-grained language-vision alignment training to support detail-aware visual understanding. InternVL (Chen et al., 2024a) aims to bridge the capability gap between vision and language models by constructing a giant visual encoder that matches the scale of LLMs.

The difficulty of MLLMs in fine-grained tasks has been observed in some prior work (Zhang et al., 2023, 2024). Among these, SEAL (Wu and Xie, 2024) integrates information in visual working memory through collaboration with visual search models, leveraging the commonsense knowledge to generate target and contextual cues. ViCrop (Zhang et al., 2025a) uses relative attention, gradient-weighted attention, or pure gradient mapping to intelligently crop and amplify on parts of the image relevant to the question. VisCoT (Shao et al., 2024) predicts the image regions most relevant to the question and generates their bounding boxes, then uses a visual sampler to crop and extract features from these regions. R-GRPO (Jiang et al., 2025a) teaches the model through supervised fine-tuning when to invoke a cropping tool during reasoning to locate and amplify on specific areas of the image. Pixel Reasoner (Su et al., 2025) employs templated instruction fine-tuning to equip the model with the ability to invoke visual operations, and introduces a curiosity-driven reinforcement learning training mechanism.

3 Method

We propose HiPerson, a training-free hierarchical perception-reasoning framework, which is designed to guide MLLMs in simulating the human cross-modal perception mechanism to enhance the perception of fine-grained visual content. As illustrated in Figure 2, HiPerson consists of three core modules: (I) Dual-Path Region Refinement: Extracts relative attention and gradient activation signals in parallel, fusing them to generate a task-aware semantic heatmap. (II) Dual-Scale Adaptive Cropping: Achieves the transition from coarse localization to precise focusing via multi-scale sliding window search and saliency assessment. (III) Dual-View Enhanced Reasoning: Coordinates the global scene and local details, guiding the model to complete a full perception closed-loop within a unified reasoning process.

3.1 Dual-Path Region Refinement

In fine-grained visual tasks, target regions are often small and dependent on the query content. Directly feeding the entire image into MLLMs can cause their attention to be easily dispersed by large background areas or common objects, leading to insufficient detail perception. To address this, we propose a dual-path region refinement mechanism, which aims to generate a discriminative semantic heatmap that precisely locates the visual region most relevant to the current query.

Specifically, given an input image $I \in \mathbb{R}^{H\times W\times 3}$ and a natural language query q, we feed them into the target MLLM and extract the attention weight matrix $A \in \mathbb{R}^{L\times P}$ from its last layer, representing query token attention to image patches. Here, L is the number of query tokens, and $P = G \times G$ is the total number of image patches with grid size G. Since the attention of a single token can be influenced by local syntax and may not reflect the overall semantic intent, we aggregate a comprehensive cognitive focus vector by averaging the attention of all query tokens along the language dimension as follows:

$$M_p^{\text{focus}} = \frac{1}{L}\sum_{l=1}^{L} A_{l,p}, \quad p = 1, \ldots, P \tag{1}$$

The averaging operation mitigates noise from irrelevant tokens and highlights the visual region consistent with the overall query semantics.

Nevertheless, the attention of MLLMs often includes an inherent preference for generic backgrounds, which can dilute task-specific signals. To eliminate such interference, we introduce a neutral reference query q_{ref} (e.g., "Describe this image"), compute its corresponding cognitive focus M^{rel}, and construct a relative attention map:

$$M_p^{\text{rel}} = \frac{M_p^{\text{focus}}(I, q)}{M_p^{\text{focus}}(I, q_{\text{ref}})} \tag{2}$$

This division-based normalization suppresses generic regions that are highly responsive to any image, retaining only the parts specifically activated by the current query.

Meanwhile, we introduce a frozen VLP model as a semantic prior. Given the image-text matching score y_{itm}, we compute the gradient of this output with respect to the last cross-modal attention map A_{cross} as follows:

$$\nabla A = \frac{\partial y_{\text{itm}}}{\partial A_{\text{cross}}} \tag{3}$$

The magnitude of the gradient reflects the contribution of each attention region to the final semantic matching decision.

To further aggregate discriminative signals from multi-head attention, we multiply each attention head's response with its corresponding gradient, sum them, and suppress negative responses via ReLU (Nair and Hinton, 2010):

$$M^{\text{grad}} = \text{ReLU}\left(\sum_h \nabla A_h \odot A_h\right) \in \mathbb{R}^{G\times G} \tag{4}$$

where h is the attention head index and $\odot$ denotes element-wise multiplication.

Finally, this heatmap is fused with M^{rel}:

$$M^{\text{sem}} = \text{norm}(M^{\text{rel}}) + \text{norm}(M^{\text{grad}}) \tag{5}$$

where $\text{norm}(\cdot)$ represents min-max normalization to $[0, 1]$. This achieves complementarity between endogenous reasoning and external priors, where the former captures query-specific dynamics, and the latter provides generalization guarantees.

3.2 Dual-Scale Adaptive Cropping

After obtaining the comprehensive semantic heatmap M^{sem}, we search for the most semantically prominent local region on the heatmap via multi-scale sliding windows, cropping it to amplify details for model perception. First, we construct a set of windows covering typical target scales. Let $\mathcal{B} = \{\beta_1, \beta_2, \ldots, \beta_K\}$ be a set of scale coefficients (e.g., $\{1.0, 1.2, 1.4, 1.6, 1.8, 2.0\}$), then the window sizes are defined as:

$$\mathcal{W} = \{(\text{w}_j, w_j) \mid \text{w}_j = \beta_j W,\ \beta_j \in \mathcal{B}\} \tag{6}$$

where W is the width of the base window size.

For each candidate window $w \in \mathcal{W}$, we slide it over the heatmap with a fixed stride δ (e.g., 1) and compute the sum of semantic responses for all pixels within its coverage as follows:

$$S_{\text{sum}}(w, x, y) = \sum_{u=x}^{x+w-1}\sum_{v=y}^{y+h-1} M^{\text{sem}}(u, v) \tag{7}$$

where (x, y) is the coordinates of the upper left corner of the window. We treat the window as a semantic integrator, and a higher value indicates greater overall relevance of the region. Then we select the position that maximizes this value as the optimal location for that scale:

$$(x_w^*, y_w^*) = \arg\max_{(x,y)} S_{\text{sum}}(w, x, y) \tag{8}$$

However, regions with high absolute response values are not necessarily semantically salient targets. For instance, large areas of grass or sky may receive uniformly high scores due to VLP priors. To avoid such mis-selection, we introduce a local contrast mechanism, computing the difference between the window response and the average response of its neighborhood as a saliency score:

$$S_{\text{saliency}}(w) = S_{\text{sum}}(w, x_w^*, y_w^*) - \frac{1}{|\mathcal{N}|} \sum_{(x',y') \in \mathcal{N}(x_w^*, y_w^*)} S_{\text{sum}}(w, x', y') \tag{9}$$

where $\mathcal{N}(x, y)$ denotes the neighborhood centered at (x, y). By emphasizing local peaks rather than absolute high values, we can filter out truly prominent target regions, effectively suppressing interference from uniform backgrounds.

Finally, we select the candidate window with the highest saliency score as the cropping region:

$$(w^*, x^*, y^*) = \arg\max_{w \in \mathcal{W}} S_{\text{saliency}}(w) \tag{10}$$

The corresponding image region is cropped and resized to a standard dimension, yielding the local image I_{crop}.

3.3 Dual-View Enhanced Reasoning

While the cropped I_{crop} contains high-resolution details, being detached from the global context can easily lead to misjudgments (e.g., mistaking a "red car" for a "fire hydrant"). Conversely, using only the global image makes it difficult to discern subtle differences. Therefore, we propose a dual-view cooperative input strategy, enabling the model to simultaneously access the macroscopic scene and microscopic details during reasoning.

Specifically, the original image I and the cropped image I_{crop} are fed into the MLLM's visual encoder to obtain the global feature sequence $\mathbf{V}_{\text{global}} \in \mathbb{R}^{P \times d}$ with dimension d and the local feature sequence $\mathbf{V}_{\text{local}} \in \mathbb{R}^{P_{\text{crop}} \times d}$, respectively. They are concatenated with the query text embedding $\mathbf{T}_q \in \mathbb{R}^{L \times d}$ to form a cooperative input sequence:

$$\mathbf{F}_{\text{input}} = [\mathbf{V}_{\text{global}};\ \mathbf{V}_{\text{local}};\ \mathbf{T}_q] \tag{11}$$

The global view provides scene layout and object relationships, while the local view injects discriminative details such as texture, text, and color. Their parallel input allows the model to dynamically weigh information at different granularities within the same reasoning process.

To further structure the reasoning process, we design a multi-step prompt template that explicitly guides the model to perform three subtasks in sequence as follows:

Detail Observation. Describe the visual attributes most relevant to the query based on I_{crop}.

Contextual Localization. Map this detail back to the global image I and describe its surrounding environment.

Final Answer. Generate the above analysis to generate the answer.

The prompt is encoded as a textual prefix $\mathbf{T}_{\mathcal{P}}$ and combined with the cooperative input sequence to form the complete model input. This not only enhances the model's perception of fine-grained details but also improves the interpretability and robustness of the reasoning process through explicit verification steps.

4 Experiments

4.1 Experiment Setup

We evaluate the comparative performance of HiPerson against VisCoT (Shao et al., 2024), ViCrop (Zhang et al., 2025a), and FitPrune (Ye et al., 2025) on three fine-grained visual question answering (VQA) datasets TextVQA (Singh et al., 2019), V* (Wu and Xie, 2024), DocVQA (Mathew et al., 2021) that rely on local visual details in images, along with three general VQA datasets AOKVQA (Schwenk et al., 2022), VQAv2 (Goyal et al., 2017), InfographicVQA (Mathew et al., 2022), which cover challenges of knowledge reasoning, generalization, text and graphics integration, respectively. Our evaluation on TextVQA follows the same setting as ViCrop, where no optical character recognition (OCR)-extracted tokens are provided to MLLMs, thereby assessing the model's genuine perceptual ability on TextVQA without textual assistance. Accuracy across all datasets is computed using the VQA-score [1]. Meanwhile, to verify the general applicability of HiPerson as a training-free framework, we integrate it into multiple mainstream MLLMs, including LLaVA (Liu et al., 2023), InstructBLIP (Dai et al., 2023), Qwen-VL (Team, 2025), and InternVL (Wang et al., 2025b). All experiments are conducted on a single Tesla A100 40GB GPU. Additional experiments and details are provided in the appendix.

[1] https://visualqa.org/evaluation.html

Figure 3: Examples of the fine-grained visual question answering task.

	Fine-Grained			General		
	TextVQA	V*	DocVQA	AOKVQA	VQAv2	InfographicVQA
VisCoT	58.11	49.21	16.07	65.10	77.13	14.73
ViCrop `rel-att`	57.72	44.50	15.10	60.75	77.15	13.23
ViCrop `grad-att`	58.75	43.98	15.33	60.71	77.41	13.65
ViCrop `pure-grad`	53.62	42.41	13.60	59.20	76.74	13.23
FitPrune	51.80	43.46	13.37	75.11	74.07	17.24
HiPerson	**62.99**	**54.45**	**19.69**	**78.86**	**78.44**	**20.75**
LLaVA-7B	49.74	40.31	12.42	59.82	76.29	12.51
+HiPerson	62.99↑**13.25**	54.45↑**14.14**	19.69↑**7.27**	78.86↑**19.04**	78.44↑**2.15**	20.75↑**8.24**
InstructBLIP-7B	35.41	38.22	5.87	55.45	77.88	12.65
+HiPerson	43.81↑**8.40**	41.88↑**3.66**	5.94↑**0.07**	55.88↑**0.43**	78.29↑**0.41**	12.75↑**0.10**
Qwen3-VL-8B	78.34	51.83	60.23	70.99	82.84	41.21
+HiPerson	87.38↑**9.04**	74.87↑**23.04**	71.76↑**11.53**	91.83↑**20.84**	87.16↑**4.32**	58.71↑**17.50**
InternVL3.5-8B	64.08	47.12	33.47	65.19	75.38	24.41
+HiPerson	79.18↑**15.10**	65.45↑**18.33**	53.75↑**20.28**	89.50↑**24.31**	82.63↑**7.25**	46.81↑**22.40**

Table 1: Comparison results on six datasets, baselines are all using LLaVA-7B as the backbone model.

4.2 Main Results

Examples of Fine-Grained Tasks. Figure 3 presents a comparison of the original LLaVA-7B and the enhanced HiPerson through four concrete examples of fine-grained VQA. For tasks requiring reasoning such as "What store is seen in the background?" and "Who is the lady who sponsored the calendar?", LLaVA merely guesses based on semantic priors, whereas HiPerson accurately perceives and reasons, correctly identifying the text "Pret a manger" on the store sign and the name "Joanna kreuz". For tasks that demand extremely fine-grained observation, such as "How much is a polos crazy bike?" and "What is the year on the bottom left coin?", LLaVA incorrectly answers "9.80" and "1979", while HiPerson successfully focuses on the tiny regions and provides the correct answers. These examples collectively demonstrate that HiPerson, through its hierarchical localization-reasoning mechanism, effectively alleviating the hallucination problem caused by over-reliance on semantic priors.

Comparison with Baselines. The upper part of Table 1 presents the comparison between HiPerson and baselines. Among baselines, VisCoT enables the model to locate fine-grained objects through training, while FitPrune enhances the model's focus on target regions via token pruning. ViCrop includes three region localization and cropping methods, but lacks the ability to perform reasoning using the cropped images. The results show that HiPerson achieves the best performance on all tasks. On fine-grained tasks, HiPerson achieves improvements over the best-performing ViCrop variant, respectively. On general VQA tasks, HiPerson attains a notably higher accuracy on AOKVQA, which requires knowledge reasoning. This advan-

tage stems from its structured multi-step reasoning mechanism, which successfully integrates local details with the global context, thereby resolving semantic disconnection in complex scenarios. Overall, HiPerson achieves robust understanding in both fine-grained and general settings.

Experiments with Different Backbones. The lower part of Table 1 verifies the general applicability of HiPerson across different MLLMs. When applied to the relatively weaker LLaVA-7B, HiPerson brings substantial performance gains, indicating that HiPerson can effectively compensate for the model's inherent deficiency in fine-grained perception. On InstructBLIP-7B, HiPerson yields limited improvements due to the model's strict instruction-following architecture and strong modality alignment strategy. Furthermore, for already capable advanced models, HiPerson still delivers consistent performance gains, demonstrating that even strong backbone models lose some local details through global encoding. Notably, HiPerson works effectively across models of varying capabilities and architectures, showcasing the broad applicability of its hierarchical cognitive simulation mechanism.

		Acc.	↑
Fine-Grained	TextVQA	49.74	-
	+crop	58.46	8.72
	+reasoning	**62.99**	**13.25**
	V*	40.31	-
	+crop	42.93	2.62
	+reasoning	**54.45**	**14.14**
	DocVQA	12.42	-
	+crop	15.00	2.58
	+reasoning	**19.69**	**7.27**
General	AOKVQA	59.82	-
	+crop	60.37	0.55
	+reasoning	**78.86**	**19.04**
	VQAv2	76.29	-
	+crop	77.21	0.92
	+reasoning	**78.44**	**2.15**
	InfographicVQA	12.51	-
	+crop	13.05	0.54
	+reasoning	**20.75**	**8.24**

Table 2: Ablation results of components on six datasets using LLaVA-7B as the backbone model.

4.3 Ablation Study

Component Ablation. Table 2 presents the performance impact of different components in HiPerson across six datasets. Here, "+crop" includes the region refinement and cropping modules since region refinement serves the cropping process, while "+reasoning" incorporates all components. Specifically, on fine-grained tasks, cropping based solely on region refinement already brings significant improvements, indicating that local enhancement effectively mitigates the detail loss caused by global encoding in conventional MLLMs. When enhanced reasoning is further introduced, performance on all tasks sees another substantial leap. This suggests that although local cropping alone can boost detail perception, it may lead to bias due to detachment from the global context. In contrast, by combining structured multi-step reasoning, the model can jointly exploit local details and global scene information, forming a complete perception loop from detail observation to contextual verification, thereby greatly improving answer reliability.

	Accuracy	GPU Time	Memory
Original	49.74	0.47s	15G
Attention	53.59	0.91s	20G
CLIP	50.86	9.75s	21G
SAM	51.80	13.93s	25G
YOLO	52.72	8.68s	22G
HiPerson	62.99	1.12s	22G

Table 3: Ablation results of external tools on TextVQA using LLaVA-7B as the backbone model.

Region Refinement Ablation. We investigate the accuracy, average GPU computation time, and memory usage of the dual-path region refinement mechanism compared to using external tools to assist cropping, as shown in Table 3. Here, CLIP (Radford et al., 2021) refers to iteratively cropping the image at fixed crop ratios (e.g., 0.9) and selecting the cropping region with the highest CLIP similarity each time. SAM (Kirillov et al., 2023) refers to taking the bounding box corresponding to the segmentation mask computed by SAM as the cropping region. YOLO (Redmon et al., 2016) refers to selecting the bounding box among the predicted boxes whose confidence exceeds a threshold (e.g., 0.25) as the cropping region. The results show that although using these external tools can bring accuracy gains, the improvements are limited and come at a substantial cost, increasing inference time to 8.68s-13.93s. Meanwhile, compared to simple attention-based cropping, it adds only 0.21 seconds on GPU time while improving accuracy by 9.4%. In contrast, by solely leveraging the model's

internal attention and gradient signals, HiPerson captures query-related key regions and achieves the highest accuracy while maintains excellent computational efficiency.

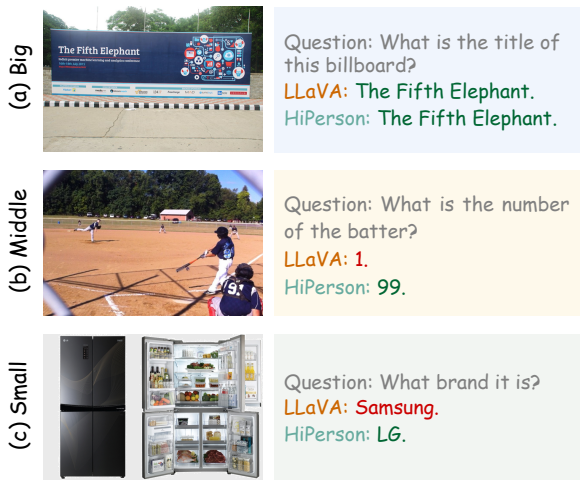


Figure 4: Comparison results on objects of different scale with LLaVA-7B.

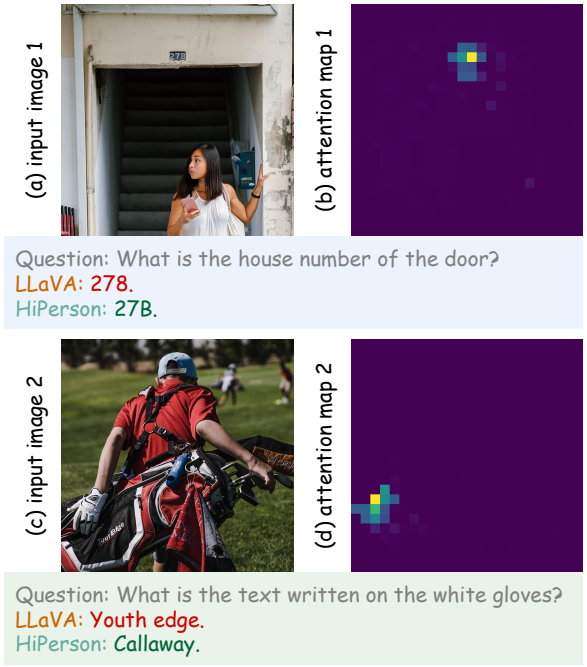


Figure 5: Comparison results and visualization of attention maps on high resolution images with LLaVA-7B.

4.4 Object Scale Study

Figure 4 demonstrates the performance differences between LLaVA and HiPerson in handling visual information at different scales through three recognition examples of varying granularity. In large-scale scenarios such as billboard title recognition, LLaVA can still generate a reasonably plausible answer, "The Fifth Elephant", based on its perception, indicating a certain level of macro-scene understanding capability. However, when the target region size decreases to medium granularity (e.g., athlete number recognition) and small granularity (e.g., brand logo recognition), LLaVA incorrectly answers "1" and "Samsung", respectively, showing that its attention is easily dispersed by background or similar objects, failing to accurately locate key details. In contrast, HiPerson provides correct answers in all three types of tasks through active focusing and local reasoning, showing that HiPerson's local enhancement mechanism effectively maintains cross-granularity robustness.

4.5 High Resolution Study

In the VQA task involving high-resolution images from HR-Bench (Jiang et al., 2025b) (about 4K resolution), as shown in Figure 5, LLaVA performs poorly due to its global uniform encoding strategy, often leading to perceptual biases and hallucinations. For instance, when asked, "What is the house number of the door?" LLaVA incorrectly identifies it as "278", whereas the correct answer is "27B". Similarly, when recognizing text on white gloves, LLaVA misinterprets it as "Youth edge", while the actual text is "Callaway". In contrast, HiPerson employs a task-aware semantic heatmap and the cropping strategy to magnify the tiny regions containing the target text, which enhances its perception of local details while preserving the global context, thereby accurately identifying the letter "B" in the house number and the fine characters of the brand name "Callaway" on the gloves. These results demonstrate HiPerson's ability to improve the accuracy of fine-grained visual understanding in high-resolution scenarios.

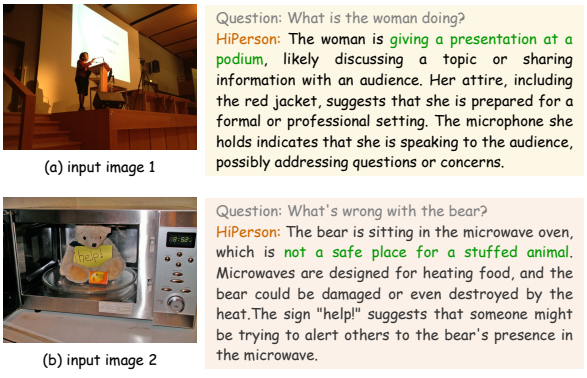


Figure 6: Examples of the fine-grained visual description task.

5 Description Task Study

Figure 6 demonstrates HiPerson's capability in the fine-grained visual description task. In the first example, when asked "What is the woman doing?", HiPerson goes beyond superficial action recognition and generates a multi-dimensional description that includes scene localization, behavioral inference, attire analysis, and tool support. This reflects its ability to integrate visual cues from global to local levels and perform semantic reasoning. In the second example, in response to "What's wrong with the bear?", HiPerson accurately identifies the unusual combination of "the bear is sitting in the microwave oven". It associates common knowledge about microwaves being used for heating food, points out the potential damage to the toy, and incorporates the warning sign into the explanation, suggesting it might serve to alert others. These two cases indicate that through dual-view input and a structured multi-step reasoning mechanism, HiPerson guides the model beyond surface-level feature capture, achieving a complete perception loop from detail perception and contextual localization to commonsense-based inference. This effectively enhances the model's reasoning reliability in fine-grained visual description tasks.

Figure 7: Examples of the fine-grained visual question answering task with multi-regions.

6 Multi-Region Study

Figure 7 demonstrates HiPerson's capability in understanding cross-regional scenes. In the example of "What are the colors of the hats of the two athletes?", both LLaVA and HiPerson provided the correct answer. Specifically, HiPerson first identifies the positions of the two target athletes, then extracts visual features from their respective head regions, sequentially determined the hat colors, and finally responded with a structured statement. This step-by-step reasoning enhances the credibility of the model's decisions. In contrast, in the example of "What are the numbers on the leftmost and rightmost people respectively?", LLaVA output the incorrect answer "41 and 5", confusing the number of the person in the middle region, which reflects its attentional bias when processing decentralized targets. By introducing a hierarchical perception mechanism, HiPerson first localizes all potential person regions at the global level and then performs refined feature extraction on the specified regions at the local stage, ultimately accurately identifying the numbers "10" and "3". This demonstrates that HiPerson can effectively overcome the model's perceptual limitations in complex spatial layouts.

7 Conclusion

This paper proposes HiPerson, a training-free hierarchical perception-reasoning framework that enhances MLLMs' fine-grained recognition capability. Specifically, HiPerson fuses the model's internal relative attention and gradient activation signals to generate a task-aware semantic heatmap, mimicking human context-sensitive focusing for task-aware localization. Subsequently, it amplifies key regions from the heatmap via multi-scale sliding window search and saliency assessment. Finally, by combining local-global dual-image cooperative input with a multi-step reasoning prompting mechanism, HiPerson guides the model to complete a full perception loop. Experiments demonstrate that HiPerson achieves expressive results on multiple datasets, showcasing its generalizability and scalability in fine-grained visual tasks.

Limitations

Despite the competitive performance and general applicability demonstrated by HiPerson, our work has several limitations that warrant consideration. For example, HiPerson relies on the internal signals of the underlying MLLMs. While this design ensures broad compatibility and avoids additional training costs, it may also constrain peak performance in specialized scenarios where task-specific adaptation could be beneficial. Moreover, the multi-step reasoning prompts are designed based on common granularity patterns, which may not generalize seamlessly to all types of fine-grained queries. Finally, our evaluation primarily focuses on benchmark datasets under controlled settings. Performance in open-world scenarios with extreme visual diversity, severe occlusions, or highly domain-specific content remains to be thoroughly validated.

Ethical Considerations

In conducting this research, we have considered its potential ethical implications. We introduces HiPerson, a training-free method to enhance fine-grained visual reasoning in MLLMs. Our experiments utilize established public datasets intended for research, and we cannot fully rule out that they may reflect societal biases or contain personal information, despite the curation efforts of their original creators. The improved perceptual capabilities of our method could support positive applications, such as assistive technologies. However, they also carry risks of misuse in surveillance or generating misleading content. We strongly discourage any deployment that violates privacy or promotes harm. Meanwhile, our framework inherits any biases present in the base models and datasets, and its evaluation is limited to common benchmarks.

Acknowledgments

This work was supported by the National Key R&D Program of China (2022YFB3103202) and the National Science and Technology Major Project of China (2025ZD1501602).

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Zhenyang Cai, Junying Chen, Rongsheng Wang, Weihong Wang, Yonglin Deng, Dingjie Song, Yize Chen, Zixu Zhang, and Benyou Wang. 2025. Exploring compositional generalization of multimodal llms for medical imaging. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 13057–13079. Association for Computational Linguistics.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024a. Intern vl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. 2024b. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 1254–1262. AAAI Press.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Dong Feng, Ping Guo, Encheng Peng, Mingmin Zhu, Wenhao Yu, and Peng Wang. 2025. Posellava: Pose centric multimodal LLM for fine-grained 3d pose manipulation. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 2951–2959. AAAI Press.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. 2024. Multi-modal latent space learning for chain-of-thought reasoning in language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18180–18187. AAAI Press.

Chaoya Jiang, Yongrui Heng, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. 2025a. VLM-r³: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yutao Jiang, Qiong Wu, Wenhao Lin, Wei Yu, and Yiyi Zhou. 2025b. What kind of visual tokens do we need? training-free visual token pruning for multi-modal large language models from the perspective of graph. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 4075–4083. AAAI Press.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3992–4003. IEEE.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. 2025a. Tokenpacker: Efficient visual projector for multimodal LLM. *Int. J. Comput. Vis.*, 133(10):6794–6812.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. 2025b. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(5):3424–3439.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yexin Liu, Zhengyang Liang, Yueze Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, and Bo Zhao. 2025. Unveiling the ignorance of mllms: Seeing clearly, answering incorrectly. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 9087–9097. Computer Vision Foundation / IEEE.

Zefeng Lu, Ronghao Lin, Yap-Peng Tan, and Haifeng Hu. 2025. Prompt-guided transformer and MLLM interactive learning for text-based pedestrian search. *IEEE Trans. Inf. Forensics Secur.*, 20:7181–7196.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. Infographicvqa. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2582–2591. IEEE.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18798–18806. AAAI Press.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 807–814. Omnipress.

Seokhyeon Park, Yumin Song, Soohyun Lee, Jaeyoung Kim, and Jinwook Seo. 2025. Leveraging multimodal LLM for inspirational user interface search. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, YokohamaJapan, 26 April 2025- 1 May 2025*, pages 579:1–579:22. ACM.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, and Jie Tang. 2025. Cogcom: A visual language model with chain-of-manipulations reasoning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788. IEEE Computer Society.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pages 146–162. Springer.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. 2025. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. 2025a. MLLM can see? dynamic correction decoding for hallucination mitigation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Penghao Wu and Saining Xie. 2024. V*: Guided visual search as a core mechanism in multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13084–13094. IEEE.

Yunzhe Xu, Yiyuan Pan, Zhe Liu, and Hesheng Wang. 2025. FLAME: learning to navigate with multimodal LLM in urban environments. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 9005–9013. AAAI Press.

Xuzheng Yang, Junzhuo Liu, Peng Wang, Guoqing Wang, Yang Yang, and Heng Tao Shen. 2025. New dataset and methods for fine-grained compositional referring expression comprehension via specialist-mllm collaboration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(10):8598–8612.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. 2025. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 22128–22136. AAAI Press.

Jiarui Zhang, Jinyi Hu, Mahyar Khayatkhoei, Filip Ilievski, and Maosong Sun. 2024. Exploring perceptual limitation of multimodal large language models. *Preprint*, arXiv:2402.07384.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2023. Visual cropping improves zero-shot question answering of multimodal large language models. *CoRR*, abs/2310.16033.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2025a. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Yuting Zhang, Hao Lu, Qingyong Hu, Yin Wang, Kaishen Yuan, Xin Liu, and Kaishun Wu. 2025b. Period-llm: Extending the periodic capability of multimodal large language model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 29237–29247. Computer Vision Foundation / IEEE.

A Dataset Statistics

Table 4 provides detailed information about the datasets used in our main experimental evaluation. Specifically, in terms of image resolution, the average dimensions of the datasets span a wide range. For example, InfographicVQA possesses a high aspect ratio, such images typically contain dense textual-graphical information and complex layouts. In contrast, AOKVQA and VQAv2 have relatively small average dimensions, focusing more on semantic understanding of everyday scenes. Moreover, regarding task scale, for VQAv2, we conduct experiments on a random 5K subset of the official validation set. For TextVQA, we experiment on the filtered subset of 4370 samples as selected by ViCrop. For all other datasets, we use the entire validation set for experimentation. Among these, TextVQA and DocVQA feature a large number of questions, emphasizing fine-grained question-answering on textual content within images. Although V* has a small sample size, its extremely wide images and complex visual search tasks pose unique challenges to the model's localization and reasoning capabilities. This disparity in scale requires the evaluation framework to not only demonstrate statistical effectiveness on large datasets but also exhibit stable performance improvements on small yet challenging benchmarks.

B Attention Study

B.1 Attention Visualization

Figure 8 presents the visualization results of attention maps at layers 0, 3, 7, 10, 15, 20, 26, 29, and 31

	TextVQA	V*	DocVQA	AOKVQA	VQAv2	InfographicVQA
Avg. Width (px)	949.3	2246.3	1782.8	581.6	571.8	1161.5
Avg. Height (px)	823.2	1583.0	2098.6	480.4	485.6	3002.4
Images	2897	191	1284	1122	883	500
Questions	4370	191	5349	1145	5001	2801

Table 4: Statistical results of the average width, average height, number of images, number of questions on datasets.

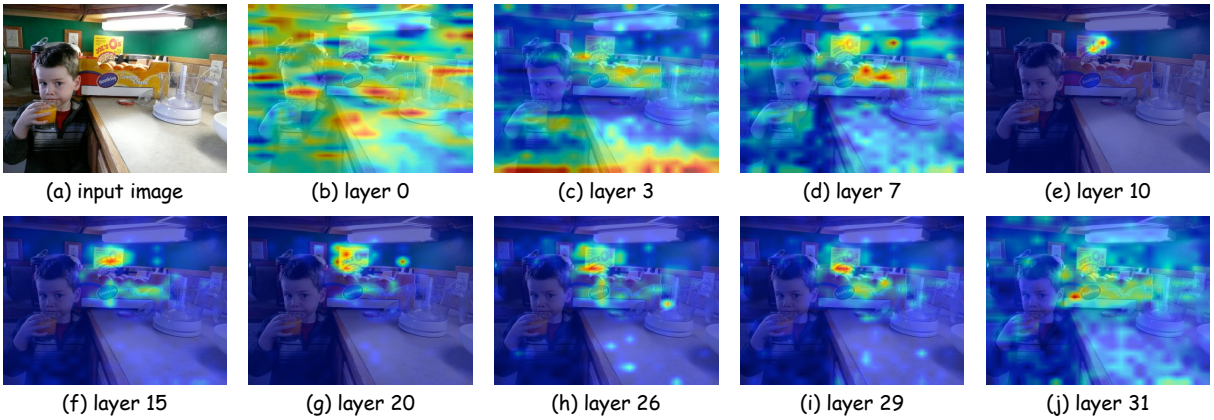
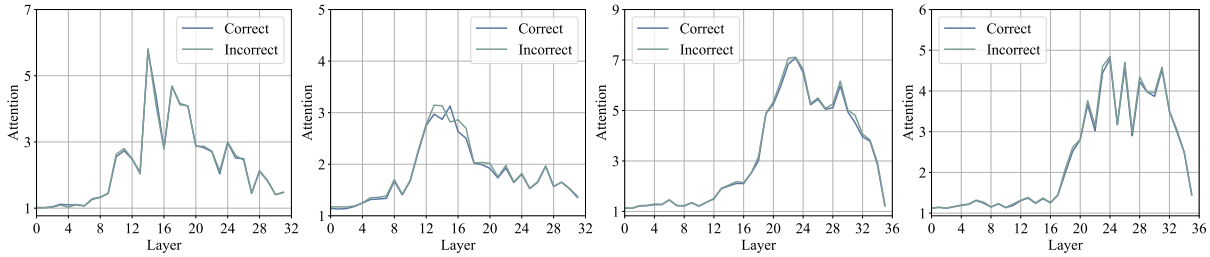


Figure 8: Visualization results of attention maps for different layers in LLaVA-7B.

for LLaVA when responding to the question "What kind of o's are shown on the box in the back?" regarding the image. In the initial stage, attention is highly scattered, indicating that the model has not yet established an effective association between the language query and the visual regions after receiving multimodal input. As network depth increases, attention gradually converges and directs toward potential areas semantically related to the question, specifically the location of the box in the image that may contain the "O"-shaped pattern. Upon entering the deeper network layers, attention further intensifies and stabilizes its concentration on the target region, showing that the model has successfully locked onto the image area most relevant to the question and is performing detailed feature extraction and integration. This stage represents the core phase for the model to accomplish fine-grained recognition, where high-level representations achieve a deeper understanding of visual cues. At deeper layers, attention exhibits a divergence trend, where the model shifts toward comprehensive reasoning involving the overall context, language priors, or cross-modal alignment during high-level semantic processing.

B.2 Attention Ratios

Figure 9 shows the average attention ratio for the top 30 correctly answered and top 30 incorrectly answered images in TextVQA across different MLLMs. This ratio is defined as the sum of relative attention within the ground-truth answer bounding box divided by the average of all bounding boxes of the same size in the image. The attention ratios of all MLLMs in the target region are significantly above 1, confirming that the multimodal alignment capability enables MLLMs to semantically relevant areas in the image based on linguistic queries. However, this inherent attentional localization mechanism is coarse and unstable. For LLaVA and InstructBLIP, the attention curves for correct and incorrect samples largely overlap across most network depths, with very close numerical values, indicating that models may still answer incorrectly due to insufficient detail resolution or reasoning biases. In contrast, Qwen-VL and InternVL exhibit stronger attentional discriminability, with showing higher attention ratios in deeper network layers. This suggests that their internal mechanism possesses a superior ability to sharpen localization and focus on details. Overall, it demonstrates that differences in model performance depend more on the stability of transitioning from coarse semantic localization to precise detail focus, rather than mere attentional intensity. This provides empirical support for HiPerson, which enhances fine-grained perceptual capability through task-aware heatmap generation and adaptive region cropping.



(a) LLaVA-Vicuna-7B-v1.5 (b) InstructBLIP-Vicuna-7B (c) Qwen3-VL-8B-Instruct (d) InternVL3.5-8B

Figure 9: Attention ratios across layers of different MLLMs when answering correctly or incorrectly.

Prompt1: Based on the cropped image and the full image above, answer the question:
{question}
What types of alcoholic beverages are sold in the store to the right?
Reasoning: Based on the image, it is not possible to definitively determine the specific types of alcoholic beverages sold in the store to the right. However, we can infer that the store offers a variety of alcoholic beverages, as it is described as a liquor store. The store could potentially sell different types of alcoholic beverages such as wine, beer, and spirits.
Answer: Wines.

Prompt2: 1) Observe the cropped image above and provide a brief description.
2) Now observe the original full image below:
<image>
Re-describe the content of the cropped region using the full image for context.
3) Based on the information from both images, answer the question:
{question}
What types of alcoholic beverages are sold in the store to the right?
Reasoning: 1)The cropped image shows a yellow taxi driving past a storefront with a sign that reads "Wines & Liquors." 2) The full image reveals that the store is located in a city, with a large building and a truck parked nearby. 3) Based on the information from both images, the store sells various types of alcoholic beverages, including wines and liquors. The storefront sign indicates that the store specializes in these types of beverages, and the presence of a truck parked nearby suggests that the store may also be involved in the transportation or distribution of these products.
Answer: Wines and liquors.

Figure 10: Comparison results of different textual prompts using LLaVA-7B as the backbone model.

C Textual Prompt Study

Figure 10 illustrates the impact of two structurally different textual prompts on the same VQA task. When faced with the same question "What types of alcoholic beverages are sold in the store to the right?", the simple single-step prompt (Prompt 1) merely leads the model to independently observe and guess based on the cropped image and the global image. Although it can infer the broad category of "alcoholic beverages", its final answer "Wines" fails to accurately reflect the complete information on the sign. In contrast, under the guidance of the multi-step structured prompt (Prompt 2), the model is explicitly instructed to perform three sequential steps: local observation, re-description within the global context, and integrated reasoning for the answer. This mechanism forces the model to first focus on the textual detail "Wines & Liquors" in the cropped region, then use the global image to confirm the surrounding environment of that region (such as the city street and vehicles), and finally explicitly combine the local textual evidence with the scene information during reasoning, thereby generating a complete and accurate answer "Wines and liquors". This comparison clearly demonstrates that simply providing a local image is insufficient to ensure that the model spontaneously accomplishes context fusion, whereas a hierarchical prompt design can effectively guide the model to establish a cognitive chain from detail perception to scene verification, improving both the precision of the answer and the reliability of the reasoning process.

Reasoning:
1) The cropped image shows a man is smiling and looking at the camera.
2) The full image shows a man wearing a blue shirt and holding a ladder.
3) The man is wearing a shirt with the year 2014 on it.

Question: What year is displayed on his shirt?
LLaVA: 2013.
HiPerson: 2014.

Figure 11: An example of the fine-grained visual question answering task with the reasoning process.

D Reasoning Study

Figure 11 illustrates the detailed reasoning process and results of HiPerson on the fine-grained VQA task "What year is displayed on his shirt?". In such task, LLaVA tends to rely on semantic priors or ambiguous local features for guessing, leading to the incorrect answer "2013". In contrast, HiPerson achieves a closed-loop perception process from local perception to global verification through a structured three-step reasoning pipeline. In the detail observation, the model performs preliminary perception based on the cropped local region, yet fails to directly derive a definitive answer from this stage alone, highlighting the limitations of relying solely on a local perspective for fine-grained recognition. After the contextual reasoning, the model is explicitly guided to recalibrate and redescribe the local observation within the global image, integrating contextual cues such as the person's posture and overall clothing to disambiguate the local clues. Finally, in the comprehensive reasoning, the model accurately outputs "2014". This process demonstrates that relying solely on observing local details is sometimes insufficient to arrive at the correct answer, underscoring the necessity of actively associating local details with and verifying them within the global context.

Question: What is this poster about?
HiPerson: The poster is about the movie "Leprechaun: Origins", which is a horror film featuring a character named Leprechaun. The gold necklace with a pendant is a key visual element in the movie's marketing campaign, and the movie poster is likely promoting the film's connection to mythical creatures.

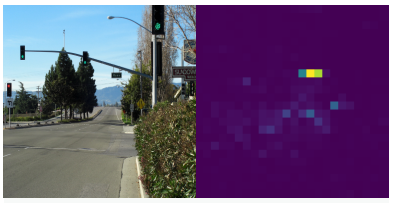
Figure 12: An example of the optical character recognition task with the reasoning process.

E OCR Reasoning Study

Figure 12 presents a fine-grained visual reasoning example based on text recognition, with the question being "What is this poster about?". HiPerson's response not only accurately identifies the poster's theme as the movie "Leprechaun: Origins", but also conducts semantic reasoning on this basis, demonstrating its ability to go beyond surface-level perception in complex visual scenes. First, HiPerson completes the extraction of basic visual information by locating and recognizing the key text on the poster. Then, the model further performs cross-modal semantic association and inference by combining the term "Leprechaun" from the movie title with the prominent visual element of the "gold necklace pendant" in the poster. It points out that the pendant is a "key visual element" in the movie's marketing campaign and infers that the poster aims to "promote the film's connection to mythical creatures". This indicates that HiPerson can integrate local textual information with the global visual design, and even external cultural common sense, to construct a coherent and in-depth interpretation of the scene. This example verifies the effectiveness of HiPerson in fine-grained text recognition tasks, and more importantly, demonstrates that the proposed reasoning framework can guide the model beyond pixel-level feature capture, advancing towards an understanding of the design logic and cultural context behind visual content, which is crucial for achieving fine-grained visual understanding.

F Extended Examples

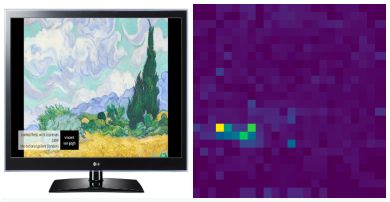
Figure 13 presents the performance and corresponding attention visualizations of LLaVA and HiPerson on additional fine-grained examples. Since HiPerson does not modify LLaVA's attention generation mechanism, the attention maps of both models are identical. HiPerson demonstrates significant advantages on several fine-grained tasks. For instance, for the question "What is the cross street?", although the attention of both models roughly covers the relevant area, LLaVA tends to directly output a guess based on this coarse perception (e.g., "Ridge"), which is often misled by semantic priors or ambiguous local features. In contrast, after obtaining the initial visual cues, HiPerson recalibrates and verifies the local observation within the global context through semantic association, thereby producing a more reliable answer (e.g., "Ross"). For some other fine-grained examples, both LLaVA and HiPerson perform well, as the underlying attention mechanism already provides sufficiently discriminative visual features. Despite this, HiPerson still has cases of failure. In the example "What is the building number?", neither HiPerson nor LLaVA provided the correct answer. HiPerson's response of "1931", while closer to the numerical form, is still incorrect. This case suggests that when the target details are highly ambiguous or blend with the background, relying solely on attention-based localization and cropping may still be insufficient for completely reliable recognition.



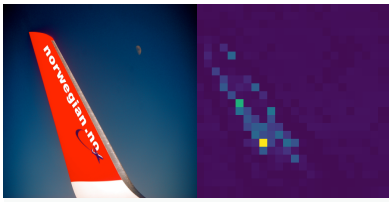
Question: What is the cross street?
LLaVA: Ridge.
HiPerson: Ross.

Question: What isle can be found to the right?
LLaVA: Couto mic.
HiPerson: Cotu mic.

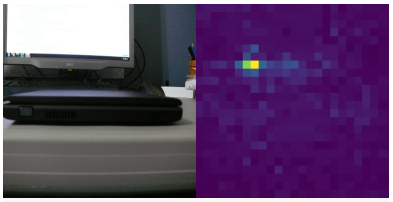
Question: What letter is on the sticker's hardhat?
LLaVA: None.
HiPerson: T.



Question: What year was this painting commissioned?
LLaVA: 1886.
HiPerson: 1889.



Question: Is the text white?
LLaVA: No.
HiPerson: Yes.

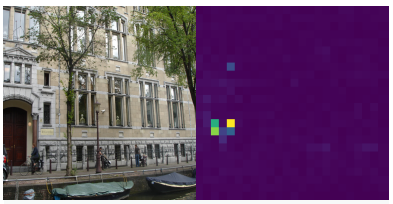


Question: What kind of computer monitor is shown?
LLaVA: Dell.
HiPerson: Acer.

Question: What team does this player play for?
LLaVA: Detroit.
HiPerson: Detroit.

Question: What is the title of the book at the bottom?
LLaVA: The idiot.
HiPerson: The idiot.

Question: How many milliliters does this hold?
LLaVA: 750.
HiPerson: 750.



Question: What is the building number?
LLaVA: 10.
HiPerson: 1931.
Correct Answer: 197.

Question: What is the beginning?
LLaVA: End.
HiPerson: End.
Correct Answer: Every day.

Question: What is the number on the last player in blue?
LLaVA: 27.
HiPerson: 3.
Correct Answer: 19.

Figure 13: Comparison results of the fine-grained visual question answering task with LLaVA-7B.