FOCUS: Fine-grained Optimal Compression Under Semantic-guidance in High Resolution Image Perception

Anonymous ACL submission

Abstract

High-resolution (HR) visual question answering challenges Multimodal Large Language Models (MLLMs) with a tension between detail preservation and visual context overload: globally resizing an HR image can erase fine-grained evidence, whereas exposing the full image can bury question-relevant details among weakly related content. We introduce ***Fine-grained Optimal Compression Under Semantic-guidance (FOCUS)***, a training-free framework that treats HR compression as question-conditioned evidence construction. FOCUS first distills an answer-oriented question into visually grounded localization cues, then integrates local patch-text alignments from overlapping HR tiles into a dense full-image evidence map. This map guides semantic compression, preserving high-evidence visual content while removing redundant spatial extent; a confidence-guided adapter further adjusts the retention budget to the context required by each question. Across V* Bench and HR-Bench with Qwen and LLaVA backbones, FOCUS consistently improves the corresponding backbone inputs, yielding average gains of 6.5% on V* Bench and 9.4% on HR-Bench 8K.

1 Introduction

Multimodal Large Language Models (MLLMs) have made rapid progress in visual understanding and instruction following (OpenAI, 2024; Team and Google, 2024; Team, 2025a). Yet high-resolution (HR) visual question answering still exposes a persistent tension between detail preservation and visual context overload. The visual evidence needed to answer a question is often sparse and fine-grained, ranging from small objects and subtle attributes to short text spans. Global resizing may remove these cues, whereas processing the entire HR image can dilute them within weakly related context and consume visual capacity on question-irrelevant regions (Li et al., 2024a; Liu et al., 2024; Luo et al., 2024). The central challenge is therefore not simply to preserve more pixels, but to construct a visual input with higher density of answer-relevant evidence.

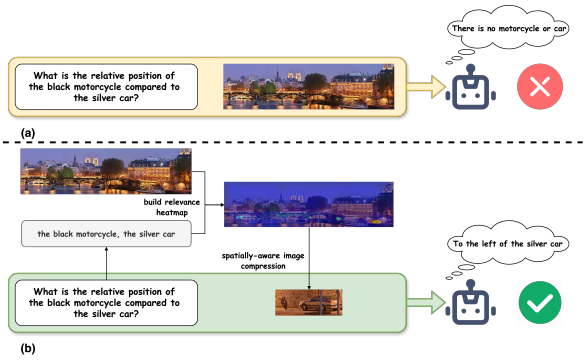

Figure 1: Motivating example of question-conditioned HR visual compression. (a) Direct HR input exposes the full scene and may dilute small answer-critical evidence. (b) FOCUS selects question-relevant visual support and compresses the input around it before final answering.

Existing approaches improve HR perception along different axes, but none directly models the fine-grained visual evidence required by a given question. HR encoders, dynamic tiling, and mixed-resolution adaptation increase access to image details, yet relevant and irrelevant regions are still preserved jointly (Li et al., 2024a; Liu et al., 2024; Luo et al., 2024). Visual token reduction lowers computational cost, but the reduction process is typically constrained by token budgets rather than question-specific evidence demands (Li et al., 2024b; Zhao et al., 2025). Region- or crop-selection methods introduce a stronger notion of relevance (Shen et al., 2024; Wang et al., 2024, 2025), but their evidence representation remains tied to discrete spatial units. This creates a granularity mismatch: answer-critical evidence can be tiny, fragmented, or distributed across an image, while a selected crop may include substantial irrelevant context or omit complementary visual support.

An effective HR VQA input should therefore increase the density of answer-relevant evidence without discarding the fine details on which the answer may depend. As illustrated in fig. 1, this objective requires more than preserving a fixed resolution or allocating a fixed visual-token budget. Parsing the question can expose the visual localization needs while reducing interference from reasoning intent and answer-format constraints. The relevant evidence also benefits from being identified below the level of whole images or coarse crops, as decisive cues may be small, fragmented, or distributed across the scene. Finally, spatially coherent support helps compression concentrate the downstream model on useful visual content rather than scattered or noisy fragments.

To this end, we introduce ***Fine-grained Optimal Compression Under Semantic-guidance (FOCUS)***, a training-free framework that constructs a question-conditioned visual input before MLLM reasoning. Rather than asking the downstream model to process a uniformly resized image, a full HR scene, or a set of coarse candidate regions, FOCUS first estimates where the question-relevant evidence lies and then compresses the image around that evidence. It separates evidence selection from answer generation by reconstructing the raw question into visually grounded semantic phrases for localization, while preserving the original question for final reasoning. These phrases are aligned with overlapping HR tiles to produce local dense responses, which are restored and refined into a coherent full-image semantic support map. Compression is then applied on this support: high-evidence pixels are retained, redundant spatial extent is trimmed, and a confidence-guided budget adapter chooses the amount of visual context to keep. By organizing the input around dense question-conditioned evidence support, FOCUS preserves local resolution where it is needed and suppresses weakly related visual content before final reasoning.

In summary, our contributions are three-fold:

- We formulate HR VQA as question-conditioned evidence construction, emphasizing answer-relevant evidence density beyond resolution preservation or token reduction.
- We propose ***FOCUS***, a training-free framework that turns localization-oriented question semantics into dense full-image evidence maps and uses the resulting support to guide HR image compression.
- We evaluate FOCUS across HR benchmarks and MLLM backbones, demonstrating consistent gains over backbone inputs and competitive performance against training-free HR perception frameworks.

2 Related Work

2.1 High-Resolution MLLMs

Modern MLLMs aim to connect visual perception with language reasoning through vision-language pre-training and instruction tuning (Radford et al., 2021; Li et al., 2022; Alayrac et al., 2022; Li et al., 2023; OpenAI, 2024; Team, 2025a). For HR scenarios, subsequent models and adaptations seek to make fine visual details accessible to the backbone by enlarging the input resolution, splitting images, or combining multiple resolutions, as in Monkey (Li et al., 2024a), LLaVA-NeXT (Liu et al., 2024), and mixture-of-resolution adaptation (Luo et al., 2024). These approaches improve the visibility of HR details, but they generally preserve relevant and irrelevant content together.

2.2 Training-Free HR Input Selection

Another line of work aims to improve HR perception without task-specific training, often by selecting or reorganizing the visual input before inference. Representative methods divide and recombine HR sub-images (DC^2) (Wang et al., 2024), explore informative regions through tree-based zooming (ZoomEye) (Shen et al., 2024), or use confidence feedback to choose among candidate HR inputs (RAP) (Wang et al., 2025). Related token-reduction methods reduce computation by pruning or merging visual tokens (Li et al., 2024b; Zhao et al., 2025). These methods share the goal of avoiding unnecessary visual processing, but their selection units are typically crops, regions, search states, or tokens. FOCUS instead estimates a dense question-conditioned support map over the original image and compresses around that support.

2.3 Dense Semantic Evidence Localization

Dense vision-language localization provides tools for identifying visual evidence from language. CLIP enables transferable text-image alignment (Radford et al., 2021), while NACLIP and PH-Reg improve dense open-vocabulary responses and reduce noisy localization artifacts (Hajimiri et al., 2025; Chen et al., 2025). Phrase-grounding work further shows the importance of discrimina-

tive referring expressions and attributes for localization (Plummer et al., 2015; Pan et al., 2023). FOCUS adapts dense semantic localization to HR VQA by deriving localization-oriented semantics from the question, aligning them over overlapping HR tiles, and converting the restored heatmap into compression-ready support.

3 Method

We propose ***FOCUS*** (Fine-grained Optimal Compression Under Semantic-guidance), a training-free framework for question-conditioned high-resolution visual compression, as illustrated in fig. 2. After formalizing the task in section 3.1, FOCUS reconstructs the query to expose visual semantic targets (section 3.2), maps them to a fine-grained relevance heatmap (section 3.3), and compresses the original image around the resulting support while preserving local details (section 3.4).

3.1 Preliminary

We consider high-resolution visual question answering. Given a high-resolution image I of size $h \times w$ and a visual question q, a downstream MLLM F predicts an answer a as

$$a = F(I, q). \tag{1}$$

For high-resolution inputs, the visual evidence required by q may occupy only a small part of I. Uniformly reducing the image can erase decisive details, while exposing the entire scene can devote substantial visual capacity to weakly related content. We therefore introduce a question-conditioned visual input constructor Ψ before the final MLLM call:

$$\hat{I}_q = \Psi(I, q), \qquad a = F(\hat{I}_q, q). \tag{2}$$

Here $\hat{I}_q$ is not a uniformly resized image, but a compact input that preserves high-resolution evidence relevant to q and suppresses visually redundant content. FOCUS instantiates Ψ by extracting a visual semantic specification from the question, mapping it to dense evidence over the original image, and compressing I around the resulting support. This construction changes the visual input presented to F while keeping the original question for final reasoning. The following subsections detail the semantic reconstruction, heatmap construction, and semantic compression stages that realize Ψ.

3.2 Semantic Reconstruction

Raw VQA queries are noisy localization signals: they mix visual targets with reasoning intent, relations, and answer-format constraints. Directly using the full query can diffuse dense alignment responses, while entity-only parsing may remove attributes needed for grounding. FOCUS therefore converts the question into attribute-preserving referring expressions for evidence localization.

We implement this conversion with a structured semantic parser based on few-shot MLLM prompting (Brown et al., 2020; Team and Google, 2024). The parser combines Chain-of-Thought-style query deconstruction (Wei et al., 2023) with Referring Expression Comprehension (REC)-style phrase extraction (Plummer et al., 2015; Pan et al., 2023), separating visual intent from answer-only instructions while preserving discriminative cues such as color, text, spatial role, and fine-grained attributes.

Formally, with prompt constructor $\mathcal{F}_{\text{prompt}}$, few-shot demonstrations P_{fs}, and MLLM $\mathcal{M}_{\text{MLLM}}$, the semantic phrases are extracted as

$$\begin{aligned} Z_q &= \mathcal{M}_{\text{MLLM}}(\mathcal{F}_{\text{prompt}}(q, P_{\text{fs}}))\,, \\ E &= \mathcal{P}_{\text{json}}(Z_q). \end{aligned} \tag{3}$$

Here $E = \{e_1, \ldots, e_n\}$ denotes the visual semantic phrases used to form the heatmap. The downstream answer model still receives the original question q, so reconstruction guides evidence selection without changing the final reasoning objective.

3.3 Heatmap Construction

Given E, FOCUS estimates image relevance without repeatedly scoring candidate regions with an additional MLLM. Region-search pipelines can localize useful evidence, but their crop-level scoring may introduce scorer noise and nontrivial overhead under resource constraints (Shen et al., 2024; Wang et al., 2024, 2025). Since section 3.2 removes many non-visual query terms while retaining grounding attributes, a frozen dense image encoder can provide reliable patch-level alignment at lower cost. FOCUS therefore constructs the heatmap through two stages: local dense responses on high-resolution tiles, followed by overlap-based fusion into a full-image evidence field.

Local dense responses. FOCUS decomposes I into overlapping high-resolution tiles $\{T_m\}_{m=1}^{M}$, avoiding global downsampling before localization while preserving the valid coordinates of each tile

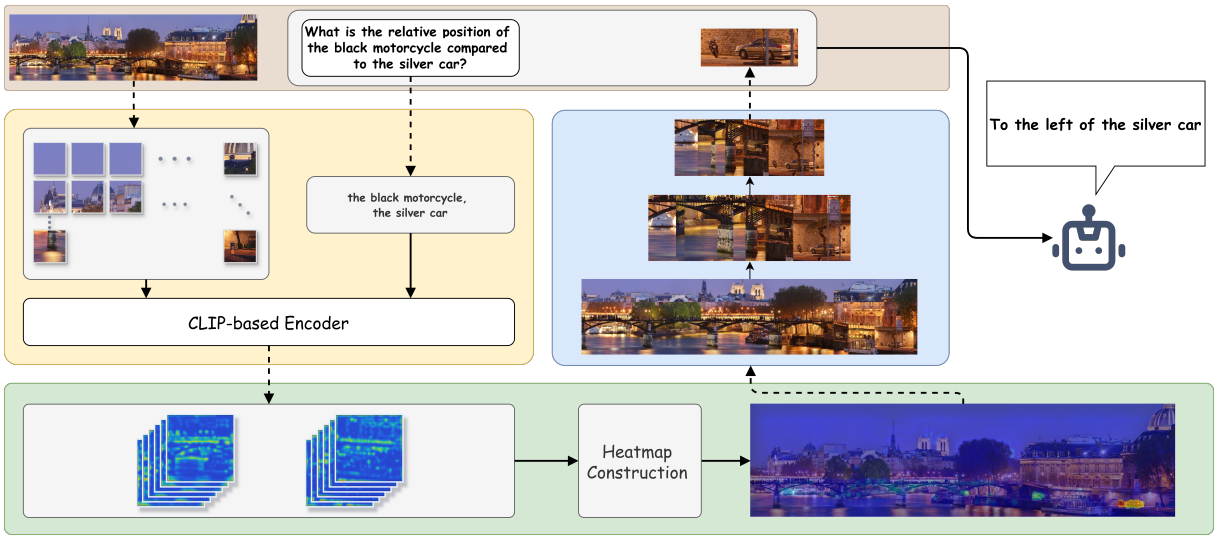

Figure 2: Overview of ***FOCUS*** for question-conditioned HR visual compression. FOCUS reconstructs the question into localization-oriented semantic phrases, aligns them with overlapping HR tiles to build a full-image evidence heatmap, and converts the heatmap into semantic support for compression. The compressed input retains high-support visual content, trims redundant spatial extent, and is passed to the downstream MLLM for final answering.

for later restoration. For each tile, CLIP-based encoders project the semantic specification E and local image patches into a shared feature space. We instantiate the text side with NACLIP (Hajimiri et al., 2025) and the visual side with a frozen PH-Reg-distilled StudentCLIP encoder. PH-Reg introduces additional register tokens to suppress noisy dense CLIP activations, making the resulting heatmap cleaner for local evidence estimation (Chen et al., 2025). Let $\mathbf{t}_E$ be the normalized text feature and $\mathbf{v}_m(i,j)$ the normalized visual feature at patch (i,j) in tile T_m. The local semantic response is:

$$S_m(i,j) = \mathrm{sim}(\mathbf{v}_m(i,j), \mathbf{t}_E)\,, \tag{4}$$

where sim is cosine similarity. Upsampling S_m to the tile resolution yields a local heatmap H_m, so each tile provides a dense relevance field rather than a single crop-level score.

Coherent heatmap restoration. The local maps $\{H_m\}_{m=1}^{M}$ are computed under separate tile contexts and therefore do not immediately form a globally calibrated evidence field. Direct stitching would assume comparable response scales across neighboring tiles, an assumption that is often violated in tiled inference, where local context and window boundaries can induce inconsistent predictions near restored seams (Reina et al., 2020; An et al., 2020). We therefore formulate restoration as estimating a full-image relevance field from all valid local observations.

Let $\Omega(p)$ denote the set of tile heatmaps covering pixel p after each H_m is projected back to the original image coordinates. FOCUS first obtains a raw relevance field by local consensus:

$$H_{\mathrm{raw}}(p) = \frac{1}{|\Omega(p)|} \sum_{m \in \Omega(p)} H_m(p). \tag{5}$$

This estimator uses overlapping observations to reduce dependence on any single tile and preserves the dense spatial support needed for compression. However, H_{raw} remains a raw relevance field: it may still contain low-frequency window bias and ambiguous background activations. FOCUS therefore applies a fixed adapter:

$$H = \mathcal{A}(H_{\mathrm{raw}}), \tag{6}$$

which converts restored relevance into semantic support for compression. The adapter first estimates a coarse support prior through stride-scale pooling, local smoothing, and upsampling:

$$B = \mathrm{Up}(\mathrm{AvgPool}_3(\mathrm{AvgPool}_s(H_{\mathrm{raw}})))\,. \tag{7}$$

The prior is then blended with the raw field,

$$H_{\mathrm{blend}} = (1-\alpha)H_{\mathrm{raw}} + \alpha B, \tag{8}$$

moderating window-level variation while retaining dense local responses. Finally, local expansion, normalization, thresholding, and power activation

suppress diffuse background and sharpen the retained support:

$$\begin{aligned} U &= \mathrm{Norm}(\mathrm{MaxPool}(H_{\mathrm{blend}}))\,, \\ \bar{H} &= \mathrm{Norm}(\mathrm{ReLU}(U-\tau))\,, \\ H &= \bar{H}^{\gamma}. \end{aligned} \tag{9}$$

where Norm denotes min–max normalization. The adapter is deterministic and fixed during evaluation; we use $\alpha = 0.4$, max-pooling kernel size 55, threshold $\tau = 0.1$, and power $\gamma = 2.0$.

3.4 Semantic Compression

Once H defines a coherent support over the original image, compression is carried out over semantic evidence. For a retention budget k, FOCUS computes the heatmap quantile threshold:

$$\theta_k = \mathcal{Q}_H(1-k) \tag{10}$$

and keeps pixels whose evidence exceeds that threshold:

$$\begin{aligned} M_k(p) &= \mathbf{1}[H(p) \geq \theta_k]\,, \\ \widetilde{I}_k(p) &= I(p) \odot M_k(p). \end{aligned} \tag{11}$$

Here M_k is a semantic mask and $\widetilde{I}_k$ is the masked image in the original coordinate system.

Trimming the selected support. Masking alone suppresses irrelevant content but leaves empty spatial extent. FOCUS removes all-black rows and columns from $\widetilde{I}_k$ to obtain the compressed input I_k. This operation differs from uniform downsampling: selected pixels retain their local resolution, weakly related regions are removed from the visual context, and trimming reduces invalid area introduced by the mask. The result is a compact input that concentrates the downstream MLLM on evidence selected by the question-conditioned map.

Confidence-guided retention. The retained visual extent should vary with the question: compact support may suffice for local attributes, whereas relational or OCR-heavy queries may require broader context. We follow the task-adaptive principle in RAP (Wang et al., 2025), which selects the number of retrieved crops using retrieval relevance and the downstream model's confidence that the current visual input is sufficient for answering. Based on the ordered evidence through H, we evaluate candidate compressed inputs $\{I_k\}$ and select:

$$k^* = \arg\max_k \mathrm{Conf}(F; I_k, q). \tag{12}$$

Here Conf can be instantiated by a sufficiency prompt or token-level confidence when available. This selector is an adopted task adapter; the core role of FOCUS is to construct dense semantic evidence so that each candidate I_k is already question conditioned before the final budget is chosen.

4 Experiments

4.1 Experimental Setup

Benchmarks and Metrics. We evaluate ***FOCUS*** on V* Bench (Wu and Xie, 2024), HR-Bench 4K/8K (Wang et al., 2024), and MME-RealWorld-lite (Zhang et al., 2024). All task results are reported as accuracy under the corresponding official evaluation protocol. V* Bench contains two splits: Attribute emphasizes fine-grained visual attributes, while Spatial focuses on spatial relations. HR-Bench also separates two categories: fine-grained single-instance perception (FSP), which stresses local evidence recognition, and fine-grained cross-instance perception (FCP), which requires comparing or relating multiple visual instances. For efficiency, we report the average number of downstream MLLM calls.

Models. FOCUS is evaluated as a model-agnostic visual input constructor on open-source Qwen2.5-VL (Team, 2025a), Qwen3-VL (Team, 2025b), and LLaVA-OneVision (Zheng et al., 2024) backbones. We pair each backbone with and without FOCUS to isolate the effect of question-conditioned evidence compression from the intrinsic capacity of the downstream MLLM. To contextualize these paired comparisons, we also include standalone results from LLaVA-v1.5 (Liu et al., 2023), VILA-HD (Shi et al., 2025), GPT-4o (OpenAI, 2024), and Qwen3-VL-Plus (Team, 2025b).

Baselines. Beyond direct backbone comparisons, we compare with representative training-free HR perception frameworks, including DC^2 (Wang et al., 2024), ZoomEye (Shen et al., 2024), and RAP (Wang et al., 2025). These systems provide the closest comparison to FOCUS because they improve HR perception without task-specific training of the downstream VQA backbone.

Implementation Details. We implement FOCUS as a training-free visual input constructor and keep all encoders and downstream MLLM backbones frozen. For heatmap construction, each image is divided into 448×448 tiles with overlap

Method	*V* Bench*			*HR-Bench 4K*			*HR-Bench 8K*		
	Attribute	Spatial	Overall	FSP	FCP	Overall	FSP	FCP	Overall
Open-source MLLMs									
LLaVA-v1.5-13B	43.5	50.0	46.1	42.0	33.0	37.5	44.0	35.0	39.5
Qwen3-VL-32B*	73.0	65.8	70.2	76.0	64.0	70.0	63.0	58.0	60.5
VILA-HD-8B-PS3-4K#	76.5	71.1	74.4	-	-	67.8	-	-	65.3
Closed-source MLLMs									
Qwen3-VL-Plus*	71.3	68.4	70.2	79.0	68.0	73.5	73.0	64.0	68.5
GPT 4o*	56.5	46.1	52.4	56.0	57.0	56.5	53.0	55.0	54.0
Baseline and Ours									
LLaVA-OneVision-4B	72.2	72.4	72.3	75.0	55.0	65.0	68.0	57.0	62.5
- w/ FOCUS	85.2	71.1	79.6	**93.0**	59.0	76.0	87.0	60.0	73.5
LLaVA-OneVision-8B	81.7	73.7	78.5	85.0	58.0	71.5	83.0	54.0	65.0
- w/ FOCUS	87.0	72.4	81.2	**93.0**	58.0	75.5	89.0	55.0	72.0
Qwen2.5-VL-3B	80.9	63.2	73.8	83.0	53.0	68.0	80.3	45.8	63.0
- w/ FOCUS	87.8	67.1	79.6	88.0	58.0	73.0	93.0	59.0	76.0
Qwen2.5-VL-7B	80.9	76.3	79.1	84.0	60.0	72.0	80.3	53.0	66.6
- w/ FOCUS	**96.5**	75.0	88.0	91.0	56.0	73.5	**94.5**	60.0	77.2
Qwen3-VL-4B	85.2	77.6	82.2	85.0	66.0	75.5	83.5	62.5	73.0
- w/ FOCUS	**96.5**	82.9	91.1	90.0	64.0	77.0	93.0	66.0	79.5
Qwen3-VL-8B	87.0	84.2	85.9	86.0	57.0	71.5	83.8	63.8	73.8
- w/ FOCUS	93.9	**88.2**	**91.6**	91.0	**69.0**	**80.0**	93.0	**71.0**	**82.0**

Table 1: Cross-backbone performance on HR benchmarks. Each shaded row pairs FOCUS with the immediately preceding open-source backbone, and FOCUS consistently improves overall accuracy on V* Bench and HR-Bench 4K/8K. Best results are in **bold** and second-best results are underlined for each metric; * denotes API-based results and # denotes results reported in the original paper.

ratio 0.5 and patch size 16; semantic phrases are encoded with NACLIP ViT-B/16, and local dense visual features are extracted with a frozen PH-Reg-distilled CLIP ViT-B/16 encoder using 16 register tokens. For confidence-guided budget selection, we use $\{0.25, 0.5, 0.75\}$ for both initial budgets and candidate expansion multipliers, stop at confidence 0.7. All experiments are conducted on 4 NVIDIA RTX 4090 GPUs, detailed settings in appendix C.

4.2 Main Experimental Results

Cross-backbone Generality. As shown in table 1, ***FOCUS*** improves all paired open-source backbones, suggesting that question-conditioned evidence construction provides benefits beyond the intrinsic capacity of a particular MLLM. Averaged over the six paired comparisons, FOCUS improves overall accuracy by 6.5% on V* Bench, 5.3% on HR-Bench 4K, and 9.4% on HR-Bench 8K. The improvements are not uniform across task types. Larger gains are observed on categories dominated by sparse local evidence, with average improvements of 11.8% on V* Attribute and 9.8% on HR-Bench 8K FSP, while Spatial and FCP show more moderate gains. This distribution is consistent with the intended role of FOCUS: the method primarily increases the visibility and density of localized evidence before final reasoning. In contrast, relational and cross-instance categories also depend on the downstream backbone's ability to compare multiple entities and infer spatial relations, which may limit the effect of input construction alone.

Training-free Competitiveness. Under the Qwen2.5-VL-7B setting, table 3 compares FOCUS with representative training-free HR perception frameworks. FOCUS achieves the highest overall accuracy on all three benchmarks, exceeding the strongest baseline by 2.66% on V* Bench, 1.50% on HR-Bench 4K, and 3.20% on HR-Bench 8K. The larger margin on HR-Bench 8K is informative:

Method	*MO*			*AD*			*OCR*		
	Property	*Orientation*	*Color*	*Intention*	*Attention*	*Visual*	*License*	*Text*	*Address*
Qwen2.5-VL-7B	54.0	24.0	32.0	28.0	48.0	59.0	88.0	88.0	92.0
- w/ FOCUS	63.0	24.0	36.0	33.0	50.0	71.0	94.0	90.0	94.0
Δ	+9.0	+0.0	+4.0	+5.0	+2.0	+12.0	+6.0	+2.0	+2.0

Table 2: Transfer results on MME-RealWorld-lite under Qwen2.5-VL-7B. The shaded rows report FOCUS and its absolute gain (Δ) over the base backbone, with green Δ values denoting non-negative gains. FOCUS improves eight of nine reported real-world subtasks, with larger gains on localized visual and OCR-centric cues.

Method	***V* Bench***	***HR-Bench 4K***	***HR-Bench 8K***
DC^2	71.2	59.5	48.5
ZoomEye	71.5	65.0	65.0
RAP	85.3	72.0	74.0
FOCUS	**87.96**	**73.50**	**77.2**

Table 3: Comparison with training-free HR perception frameworks under Qwen2.5-VL-7B. All entries report overall accuracy, and the best results are in **bold**.

as resolution increases, the answer-relevant evidence tends to occupy a smaller fraction of the input, making both fine-detail loss and irrelevant-context exposure more consequential. The comparison therefore highlights a difference in evidence representation. Region-search methods operate over discrete candidate crops or search states, which can still include irrelevant context or fail to isolate small cues within a selected region. FOCUS represents relevance as a dense semantic support map over the original image, allowing compression to preserve local detail while reducing the visual capacity assigned to weakly related content.

Real-world Transferability. We further evaluate FOCUS on MME-RealWorld-lite (Zhang et al., 2024) to test whether evidence-preserving compression transfers beyond HR-specific benchmarks. As shown in table 2, FOCUS improves eight of the nine reported subtasks under Qwen2.5-VL-7B, with larger gains on MO/Property (+9.0%), AD/Visual (+12.0%), and OCR/License (+6.0%). These categories rely heavily on localized properties, fine visual cues, or small textual regions, so their improvements provide evidence that FOCUS is most effective when the bottleneck is access to compact visual evidence. The unchanged result on MO/Orientation also clarifies the scope of the method: improving the visual input does not necessarily resolve errors that depend more strongly on the backbone's orientation or relational reasoning. The transfer results therefore suggest that FOCUS captures a general input-construction principle rather than a benchmark-specific heuristic; complete results are provided in table 7.

Inference-time Efficiency. As shown in table 4, FOCUS requires 15.1 calls on average, substantially fewer than DC^2 (173.4) and ZoomEye (156.4), which invoke the MLLM over many segmented or searched regions. Its call count is close to RAP (16.8), since both methods use confidence feedback to select the retained visual extent. However, downstream calls capture only part of the cost. RAP-style retrieval pipelines also rely on a separate visual relevance estimator, often instantiated with a model of nontrivial scale such as VisRAG (Wang et al., 2025; Yu et al., 2025). In FOCUS, semantic question reconstruction turns the answer-oriented query into localization-oriented cues, allowing relevance to be estimated with a lightweight CLIP-based dense alignment model. This design reduces the cost of relevance estimation while maintaining competitive accuracy, as reflected by the stronger results in table 3.

4.3 Ablation Study

We conduct method ablations on HR-Bench 8K and Qwen2.5-VL-3B to examine the three stages of FOCUS: semantic reconstruction, heatmap construction, and semantic compression. Unless otherwise specified, all variants use the same downstream backbone and confidence-guided budget selection, changing only the component specified in table 5.

Qwen2.5-VL-7B	***FSP***	***FCP***	***Avg.***
DC^2	173.1	173.6	173.4
ZoomEye	120.9	191.8	156.4
RAP	15.3	18.3	16.8
FOCUS	**14.5**	**15.7**	**15.1**

Table 4: Inference cost on HR-Bench 8K under Qwen2.5-VL-7B. We report average downstream MLLM calls, where lower values indicate more efficient HR evidence selection. Best results are in **bold**.

Variant	*FSP*	*FCP*	*Avg.*
Full model			
FOCUS	93.0	59.0	76.0
Semantic reconstruction			
w/o reconstruction	86.0	53.0	69.5
Heatmap construction			
w/o overlap	90.0	61.0	75.5
w/o fusion	89.0	55.0	72.0
w/o fusion + overlap	88.0	45.0	66.5
Semantic compression			
mask only	75.0	55.0	65.0
bbox crop	90.0	51.0	70.5
random mask	77.0	53.0	65.0

Table 5: Method ablation on HR-Bench 8K. The variants isolate the effects of semantic reconstruction, heatmap construction, and semantic compression.

Semantic Reconstruction. Removing semantic reconstruction reduces the average score from 76.0 to 69.5, with drops on both FSP and FCP. This indicates that the raw VQA question is not always an effective localization prompt: reasoning intent and answer-format constraints can interfere with dense evidence construction. Parsing the question into localization-oriented semantic phrases therefore provides a cleaner signal for the subsequent heatmap and compression stages.

Heatmap Construction. The heatmap variants evaluate how local tile responses are restored into full-image semantic support. Removing overlap alone causes only a slight drop from 76.0 to 75.5, while disabling fusion leads to a larger decline to 72.0. Removing both further reduces the score to 66.5, with the most pronounced degradation on FCP. These results indicate that overlap and fusion play complementary roles: overlap preserves cross-tile continuity, while fusion consolidates redundant local responses into coherent global evidence. Their joint removal causes a disproportionate drop, highlighting their synergy in maintaining semantic consistency for cross-instance reasoning.

Semantic Compression. The compression variants clarify why FOCUS applies both semantic masking and trimming. Keeping the semantic mask without trimming lowers the average score to 65.0, indicating that simply turning unselected pixels black does not remove their effect from the input; the remaining canvas can still introduce substantial visual noise and dilute the model's attention. Bounding-box cropping performs better at 70.5, showing that a coarse region around the semantic support is useful, but it still trails FOCUS because distant selected pixels or noisy activations can enlarge the box and retain irrelevant context. Random masking also falls to 65.0, confirming that reducing the number of input pixels alone is not sufficient. The benefit comes from preserving question-critical visual elements while removing weakly related spatial extent.

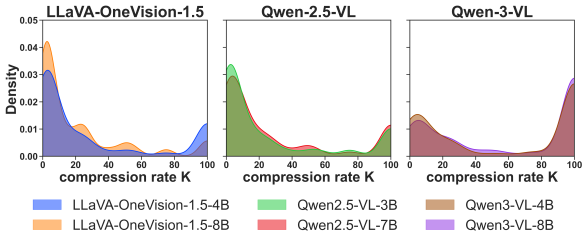

Figure 3: Distribution of the selected retention budget k across MLLMs and HR-Bench 8K task categories.

Retention Budget Analysis. The selected retention budgets in fig. 3 vary across backbones and task categories, indicating that a fixed compression ratio is not suitable for all VQA instances. This supports the use of confidence-guided budget selection after semantic evidence construction: compact support may suffice for localized attributes, whereas cross-instance or text-heavy questions can require broader visual context.

5 Conclusion

We present ***Fine-grained Optimal Compression Under Semantic-guidance (FOCUS)***, a training-free framework for high-resolution visual question answering. FOCUS reframes HR visual input construction as question-conditioned evidence selection: it reconstructs the question into a localization-oriented semantic specification, builds a dense full-image evidence map from local HR tile responses, and compresses the image around the resulting support before downstream MLLM answering. Across HR benchmarks and multiple MLLM backbones, this evidence-centered design consistently improves the corresponding backbone inputs and remains competitive with training-free HR perception frameworks. These results suggest that preserving all high-resolution pixels is less important than exposing a compact visual input with higher answer-relevant evidence density. Future work can further improve efficiency and better preserve distributed relational cues for spatial and cross-instance reasoning.

Limitations

FOCUS is designed as a training-free visual input constructor for MLLMs, and several limitations remain. First, our evaluation is conducted on open-source backbones for which intermediate inference settings and confidence-based selection can be controlled. Applying the same pipeline to closed-source systems may require alternative confidence estimates or API-specific adaptations. Second, due to computational constraints, our controlled experiments focus on 2B–8B open-source models; although the results suggest consistent cross-backbone gains within this range, further evaluation on larger open-source MLLMs would better characterize scaling behavior. Third, FOCUS reduces the cost of relevance estimation by replacing heavier retrieval-style scoring with CLIP-based dense alignment, but the end-to-end runtime is still affected by confidence-guided retention search, which may require multiple downstream MLLM calls to select an appropriate compression budget. Improving this budget selection procedure is an important direction for future work.

Acknowledgments

This document has been adapted by Steven Bethard, Ryan Cotterell and Rui Yan from the instructions for earlier ACL and NAACL proceedings, including those for ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTEX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Yu An, Qing Ye, Jiulin Guo, and Ruihai Dong. 2020. Overlap training to mitigate inconsistencies caused by image tiling in cnns. In *AI 2020: Advances in Artificial Intelligence*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yuxuan Chen, Bailey Miller, Trung Le, Siddhant Chopra, Ce Zhang, Zhi Yang, Ziheng Zhang, Devavrat Tomar, Juan Tang, Irwin King, and 1 others. 2025. Vision transformers don't need trained registers. *arXiv preprint arXiv:2505.21501*.

Sina Hajimiri, Ismail Ben Ayed Li, and Jose Dolz. 2025. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Bochen Li, Yuxiao Wang, Yuxuan Wang, Sen Wang, Ziqin Wang, Chen-Lin Zhang, Jiahong Li, Yiren Jian, Xvguoqing.L wu, Wenhao Wu, Yue Wang, Xin Wang, Yu Sun, Yifei Shen, Wei Liu, Yang yu, Yizhou Zhao, Yebin Liu, Weipeng Yan, and 5 others. 2024a. Monkey: An mllm that understands ultra-high-resolution images. *Preprint*, arXiv:2405.04547.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Shaohan Li, Zigang Geng, Zhaofeng He, Hong-Bin Wei, Jun-Jie Huang, Han-Lin Chen, Xiaoguang Zhu, and Zhaofeng He. 2024b. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *Preprint*, arXiv:2403.15388.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng-Yu Wang, and Yong Jae Lee.

2024. Llava-next: Improved llava-1.6 with high-resolution support and sota-level reasoning. *Preprint*, arXiv:2405.15088.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. 2024. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *Preprint*, arXiv:2403.03003.

OpenAI. 2024. Gpt-4o. *Preprint*, arXiv:2405.07127.

Tengchao Pan, Zhengyuan Yang, Fanyi Wei, Shaohan Huang, Wenhui Wang, Yichong Xu, Junjian Zhang, Zewen Chi, Li Dong, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *Preprint*, arXiv:2306.14824.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

G. Anthony Reina, Ravi Panchumarthy, Siddhesh Pravin Thakur, Alexei Bastidas, and Spyridon Bakas. 2020. Systematic evaluation of image tiling adverse effects on deep learning semantic segmentation. *Frontiers in Neuroscience*, 14:65.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. 2024. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. *arXiv preprint arXiv:2411.16044*.

Baifeng Shi, Boyi Li, Han Cai, Yao Lu, Sifei Liu, Marco Pavone, Jan Kautz, Song Han, Trevor Darrell, Pavlo Molchanov, and 1 others. 2025. Scaling vision pre-training to 4k resolution. *arXiv preprint arXiv:2503.19903*.

Gemini Team and Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2402.10170.

Qwen Team. 2025a. Qwen2.5-vl.

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

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, and Dacheng Tao. 2024. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. *Preprint*, arXiv:2408.15556.

Wenbin Wang, Yongcheng Jing, Liang Ding, Yingjie Wang, Li Shen, Yong Luo, Bo Du, and Dacheng Tao. 2025. Retrieval-augmented perception: High-resolution image perception meets visual rag. *Preprint*, arXiv:2503.01222.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

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

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *Preprint*, arXiv:2410.10594.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, and 1 others. 2024. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*.

Yang Zhao, Zhen Yang, Zhaofeng Wu, Xin Wang, and Zhaofeng He. 2025. Accelerating multimodal large language models by searching optimal vision token reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zilong Zheng, Yifei Huang, Xiaoxiang Yu, Yikun Wei, Haotian Liu, and Yong Jae Lee. 2024. Llava-onevision: A hallucination-free vision-language model by curing vision deficiency. *Preprint*, arXiv:2407.03433.

A Outline

This appendix provides implementation details and supplementary experimental materials for ***Fine-grained Optimal Compression Under Semantic-guidance (FOCUS)***. The structure of the appendix is summarized as follows:

- appendix B describes the usage of large language models during manuscript preparation.
- appendix C describes the implementation of semantic reconstruction, heatmap construction, semantic compression, confidence-guided budget selection, and the hyperparameter settings used in our experiments.
- appendix D provides detailed benchmark results, complete MME-RealWorld-lite results, and the supplementary protocol for method ablations.

B LLM Usage

The large language model (LLM) was utilized as a writing assistant during the preparation of this manuscript. Its application was strictly limited to improving the clarity and grammatical accuracy of the text. Specific uses included rephrasing sentences for better flow and translating initial concepts and drafts from Chinese to English. All core scientific contributions, the design of the methodology and experiments, and the analysis and interpretation of the results, are solely the work of the authors. The authors take full responsibility for all claims and the final content of this paper.

C Implementation Details

C.1 Details of Semantic Evidence Construction

Given an HR image I and question q, FOCUS first reconstructs the question into attribute-preserving semantic phrases for evidence localization. The reconstructed phrases form the semantic specification E used for heatmap generation, while the downstream MLLM still receives the original question for final answering.

FOCUS localizes the semantic specification on an overlapping HR tile grid rather than scoring discrete crop candidates. For each tile, CLIP-based encoders project the semantic phrases and local image patches into a shared feature space, and dense patch-text similarity produces a local heatmap. We use NACLIP text features together with the frozen PH-Reg-distilled visual encoder described in section 3.3. Tile coordinates and valid regions are retained when local heatmaps are projected back to the original image.

The restored raw heatmap averages all valid local responses covering each pixel. A fixed adapter then converts the restored relevance field into semantic support for compression: it blends stride-scale smoothed support with the dense response, expands sparse high-response regions with max pooling, and sharpens the support map through normalization, thresholding, and power activation. This procedure yields the full-image evidence map H used by semantic compression.

C.2 Details of Semantic Compression

Given the adapted full-image evidence map H and retention budget k, FOCUS selects the top-k support through the quantile mask in section 3.4. The masked image preserves original local pixels only where the evidence exceeds the selected threshold. FOCUS then trims all-black rows and columns introduced by masking, producing a compact input I_k without uniformly downsampling the retained visual detail.

C.3 Details of Confidence-Guided Budget Selection

We use the confidence-guided selector in algorithm 1 as a task adapter for VQA instances that require different visual extents. It evaluates an initial set of retention budgets, expands candidates around the current best budget when confidence remains insufficient, and stops once the confidence threshold is reached or the maximum number of evaluations is exhausted.

C.4 Hyperparameter Settings

FOCUS uses 448×448 overlapping tiles for heatmap construction with overlap ratio 0.5, stride 224, and patch size 16. Semantic phrases are encoded with NACLIP ViT-B/16, and local dense visual features are extracted with a frozen PH-Reg-distilled CLIP ViT-B/16 encoder with 16 register tokens. The fixed heatmap adapter uses blend weight $\alpha = 0.4$, max-pooling kernel size 55, threshold $\tau = 0.1$, and power $\gamma = 2.0$. For confidence-guided budget selection, the initial budgets and candidate expansion multipliers are both $\{0.25, 0.5, 0.75\}$, the confidence threshold is 0.7,

Algorithm 1 Confidence-Guided Budget Selection

procedure SELECTBUDGET($\mathcal{K}_0, R, T_{\max}, \tau$)
  $\mathcal{S} \leftarrow \emptyset$
  for each k in $\mathcal{K}_0$ **do**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{(k, \text{SCORE}(k))\}$
  end for
  while $|\mathcal{S}| < T_{\max}$ **and** $\max_{(k,c)\in\mathcal{S}} c < \tau$ **do**
    $(k_{best}, c_{best}) \leftarrow \arg\max_{(k,c)\in\mathcal{S}} c$
    for each r in R **do**
      $k' \leftarrow k_{best} \times r$
      $\mathcal{S} \leftarrow \mathcal{S} \cup \{(k', \text{SCORE}(k'))\}$
    end for
  end while
  return $\arg\max_{(k,c)\in\mathcal{S}} c$
end procedure

function SCORE(k)
  Require: I, q, H, downstream MLLM F
  $I_k \leftarrow$ SEMANTICCOMPRESS(I, H, k)
  return CONFIDENCE(F, I_k, q)
end function

and the maximum number of confidence evaluations is 20. All experiments are conducted on 4 NVIDIA RTX 4090 GPUs.

D Additional Experimental Results

D.1 Detailed Main Experiment Result

Main Results. We additionally compare FOCUS with RAP on Qwen backbones for V* Bench and HR-Bench. FOCUS improves the corresponding backbone inputs across the reported benchmark settings. Relative to RAP, the two training-free systems differ by benchmark and backbone, with FOCUS obtaining the strongest overall results for the Qwen2.5-VL-7B and Qwen3-VL-8B comparisons reported in table 6.

General Benchmark Results. The detailed results on MME-RealWorld-lite (Zhang et al., 2024) are shown in table 7. FOCUS improves all reported task groups. Gains are smaller for some OCR subtasks, where evidence can be distributed across text-bearing regions and the downstream model's intrinsic recognition capability remains important.

D.2 Supplementary Method Ablation Protocol

The main paper groups method ablations into semantic reconstruction, heatmap construction, and semantic compression. We use the following protocol to instantiate these variants. The full model uses semantic reconstruction, overlapping tiled heatmaps with fusion and the fixed adapter, semantic mask-and-trim compression, and confidence-guided budget selection. Each ablated variant changes only the component specified in table 8 while keeping the remaining pipeline unchanged.

Table 6: Detailed results on high-resolution benchmarks and comparison between ***FOCUS*** and ***RAP***. Best results are in **bold** and second-best results are underlined. * denotes API-based results, and # denotes results reported in the original paper.

Method	*V* Bench*			*HR-Bench 4K*			*HR-Bench 8K*		
	Attribute	Spatial	Overall	FSP	FCP	Overall	FSP	FCP	Overall
Open-source MLLMs									
LLaVA-v1.5-13B(Liu et al., 2023)	43.5	50.0	46.1	42.0	33.0	37.5	44.0	35.0	39.5
Qwen3-VL-32B*(Team, 2025b)	73.0	65.8	70.2	76.0	64.0	70.0	63.0	58.0	60.5
VILA-HD-8B-PS3-4K#(Shi et al., 2025)	76.5	71.1	74.4	-	-	67.8	-	-	65.3
Closed-source MLLMs									
Qwen3-VL-Plus*(Team, 2025b)	71.3	68.4	70.2	79.0	68.0	73.5	73.0	64.0	68.5
GPT 4o*(OpenAI, 2024)	56.5	46.1	52.36	56.0	57.0	56.5	53.0	55.0	54.0
Baselines and FOCUS									
InternVL-2-llama3-76B(DC^2)#	-	-	-	-	-	70.4	-	-	63.3
Qwen2.5-VL-3B(Team, 2025a)	80.9	63.2	73.8	83.0	53.0	68.0	80.3	45.8	63.0
- w/ RAP(Wang et al., 2025)	83.5	75.0	80.1	89.0	52.0	70.5	86.0	40.0	63.0
- w/ FOCUS	87.8	67.1	79.6	88.0	58.0	73.0	82.0	50.0	66.0
Qwen2.5-VL-7B(Team, 2025a)	80.9	76.3	79.1	84.0	60.0	72.0	80.3	53.0	66.6
- w/ RAP	91.3	76.3	85.3	84.0	60.0	72.0	93.0	55.0	74.0
- w/ FOCUS	**96.5**	75.0	88.0	91.0	56.0	73.5	**94.5**	60.0	77.2
Qwen3-VL-2B(Team, 2025b)	74.8	75.0	74.9	-	-	74.5	-	-	66.0
- w/ RAP	85.2	84.2	84.9	91.0	61.0	76.0	87.0	59.0	73.0
- w/ FOCUS	91.30	79.0	86.4	89.0	67.0	78.0	93.0	57.0	75.0
Qwen3-VL-4B (Team, 2025b)	85.2	77.6	82.2	85.0	66.0	75.5	83.5	62.5	73.0
- w/ RAP	93.0	81.6	88.5	90.0	64.0	77.5	92.0	66.0	79.0
- w/ FOCUS	**96.5**	82.9	91.1	90.0	64.0	77.0	93.0	66.0	79.5
Qwen3-VL-8B (Team, 2025b)	87.0	84.2	85.9	86.0	57.0	71.5	83.8	63.8	73.8
- w/ RAP	89.6	84.2	87.4	95.0	65.0	80.0	91.0	66.0	78.5
- w/ FOCUS	93.9	**88.2**	**91.6**	91.0	**69.0**	**80.0**	93.0	**71.0**	**82.0**

Table 7: Complete results on MME-RealWorld-lite under Qwen2.5-VL-7B. MO: Monitoring; AD: Autonomous Driving; DT: Diagram and Table; RS: Remote Sensing.

Task	Subtask	Qwen2.5-VL-7B	+FOCUS
MO	Property	54.00	63.00
	Orientation	24.00	24.00
	Color	32.00	36.00
AD	Intention	28.00	33.00
	Attention	48.00	50.00
	Visual	59.00	71.00
OCR	License	88.00	94.00
	Text	88.00	90.00
	Address	92.00	94.00
DT	All	71.50	73.50
RS	All	39.90	47.30

Table 8: Ablation protocol corresponding to table 5.

Variant	Controlled change
w/o reconstruction	Use the raw VQA question directly for heatmap construction.
w/o overlap	Construct local heatmaps on non-overlapping tiles before restoration.
w/o fusion	Keep tiled localization but remove overlap-based fusion when restoring the heatmap.
mask only	Apply the semantic mask but keep the empty canvas without trimming.
bbox crop	Crop the bounding box around the selected semantic support.
random mask	Retain the same amount of randomly selected pixels and apply trimming.