POSTCROP: COMPOSE BEFORE CROP FOR CONTEXT-AWARE FINE-GRAINED VISUAL PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) have demonstrated strong fine-grained perception of relations among multiple entities by controlling *visual context*—expanding or reducing the surrounding context of multiple entities (e.g., zoom-out or locate-then-zoom-in). While existing training-free methods typically rely on either a coarse-to-fine search paradigm or a *Crop Before Compose* paradigm that crops entity regions before composing them with context, both face a context paradox: crops containing only the entity regions can obscure relations among entities, whereas crops keeping context introduce visual interference. To characterize this paradox, we fix the entity regions and vary the kept context. Across HR-Bench-4K and HR-Bench-8K, we find that the preferred amount of context varies across settings, motivating adaptive context selection. Further analysis shows that the early separation of entity regions in the *Crop Before Compose* paradigm makes this adaptation difficult. Leveraging these insights, we propose **PostCrop**, a training-free framework following the **Compose Before Crop** paradigm that considers relation-relevant context before the final crop is formed. The **Evidence Localization and Composition** module localizes the region that best matches each queried entity and composes the cue maps of multiple entities on the original visual grid to identify surrounding regions that support the queried relation. The **Evidence Cropping and Answering** module uses the composed map to determine how much surrounding context to retain and selects a compact crop containing the entity regions and the kept context, which is combined with the original image for answering. Extensive experiments on VSTAR, HR-Bench-4K, and HR-Bench-8K demonstrate that PostCrop consistently outperforms the compared baselines across all three evaluated LVLMs (e.g., 74.0% accuracy on the Fine-grained Cross-instance Perception (FCP) subtask of HR-Bench-4K with Qwen2.5-VL-7B), validating the effectiveness of the **Compose Before Crop** paradigm.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) have achieved remarkable progress in visual understanding and reasoning, with strong performance across tasks including visual question answering, visual grounding, and spatial reasoning (Bai et al., 2023; Wang et al., 2024a; Yue et al., 2024). Fine-grained perception of relations among multiple entities (e.g., spatial, containment, and logical relations) requires not only recognizing the key entities but also capturing visual context that conveys their relations (Ranasinghe et al., 2024; Wu et al., 2025; Li et al., 2026b; Cui et al., 2026; Chen et al., 2024a). Controlling visual context can therefore improve such perception, as a visible region with less surrounding information can make it easier to focus on key entities, while a region with more surrounding information can help establish more comprehensive relations among entities.

Despite its value, existing methods generally treat context as a whole, either removing it or keeping it without further filtering. Some studies adopt search-based or retrieval strategies that scan images in a coarse-to-fine manner (Wu & Xie, 2024; Shen et al., 2025; Li et al., 2025; 2026c), mainly determining where to look and zooming out to include more surrounding context, without explicitly considering the relations between entities. Alternatively, some methods follow a *Crop Before Compose* paradigm: they first localize individual entity crop regions and then combine these regions to construct the answering view. One line of work crops entities individually and removes surrounding context to reduce visual interference (Ma et al., 2026; Li et al., 2025; Wang et al., 2026; Liu et al., 2026a; Li et al.,

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

2026a; Li & Peng, 2026). For example, Layout-Preserving Decoupling (LPD) (Liu et al., 2026a) compresses gaps between entity regions while preserving their relative spatial layout. While this can suppress irrelevant information, discarding context may also remove relational cues and make the answer be more susceptible to localization errors (Yu et al., 2025; Jiang et al., 2026). Another line of work follows this order by constructing Union (i.e., the minimum bounding box enclosing all localized entity boxes), which keeps the entity regions and the full intervening context (Zhong et al., 2025; Yu et al., 2025; Wu & Xie, 2024; Shen et al., 2025; Li et al., 2026c). However, keeping more context can expose LVLMs to distracting visual evidence, including attention sinks and attention drift toward semantically similar objects (Xiao et al., 2024; Cheng et al., 2017). These approaches reveal a **context paradox** (Fig. 1(a)): *crops containing only the entity regions can obscure relations among entities, whereas crops keeping context introduce visual interference.*

In this paper, we investigate how perception of relations among entities changes with the kept context while keeping entity regions fixed. Our analysis reveals a trade-off between *evidence loss* and *contextual interference*, suggesting that the context kept after localization should depend on the queried relation. However, such adaptive context selection is difficult under the *Crop Before Compose* paradigm, where entity regions are separated before their relation is jointly considered. Context selected from an individual entity response map may strongly support that entity while providing little evidence about its relation to another (Li et al., 2019). Moreover, presenting entities and context in separate crops requires the model to recover objects' spatial correspondence (Guo et al., 2026).

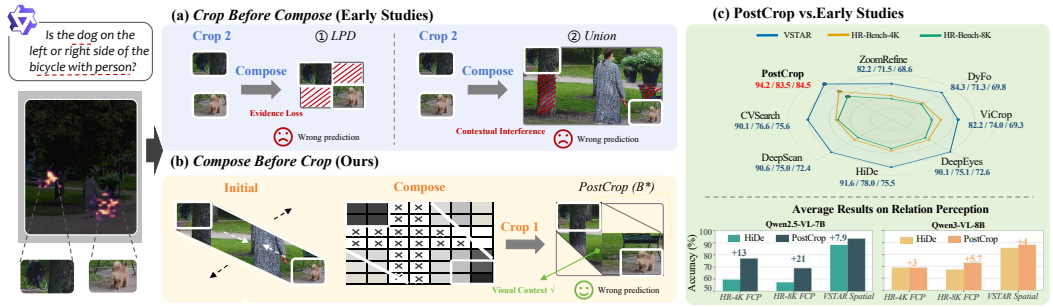


Figure 1: **Motivation and performance of PostCrop.** (a) Crop Before Compose either discards useful context (LPD) or keeps distracting context (Union). (b) PostCrop composes entity evidence on the original grid before selecting B^*. (c) PostCrop improves performance on VSTAR and HR-Bench.

Building on these insights, we propose **PostCrop**, a training-free framework that follows **Compose Before Crop** to select relation-relevant context (Fig. 1(b)). Specifically, **Evidence Localization and Composition** obtains separate maps for the object and attribute cues mentioned in the question and answer options. It then combines these maps and verifies spatial candidates to select one region for each entity, while keeping the full entity maps on the original visual-token grid. For multi-entity questions, it composes these maps to guide context selection for the queried relation. **Evidence Cropping and Answering** uses the verified entity view directly for single-entity questions and, for multi-entity questions, searches a two-dimensional space of context expansion and gap compression, selecting the view with the most relevant visual context (i.e., the highest context score). The model uses the selected view for focused evidence and the original image as a global reference.

Our contributions are:

- We analyze the competing risks of *evidence loss* and *contextual interference* in relational reasoning, and the limitations of separating entities from context before composition.
- We introduce **PostCrop**, a training-free **Compose Before Crop** framework that composes entity evidence on the original grid to adaptively select relation-relevant context.
- PostCrop achieves the highest overall accuracy among the compared methods on VSTAR and HR-Bench across both evaluated backbones (Fig. 1(c)).

2 Diagnostic Analysis

In this section, to characterize the context paradox, we address two questions: (1) when context is helpful and when additional content introduces interference, and (2) whether the *Crop Before*

Compose paradigm limits relation-aware context selection. During evaluation, accuracy is the primary metric. Reference-answer log-probability measures the model's support for the reference answer under teacher forcing. It is averaged over answer content tokens within each question and then equally across questions (Appendix C).

2.1 Context: Evidence Loss and Contextual Interference

To investigate evidence loss and contextual interference, we keep entity regions fixed and compare layouts, then vary context within and beyond Union.

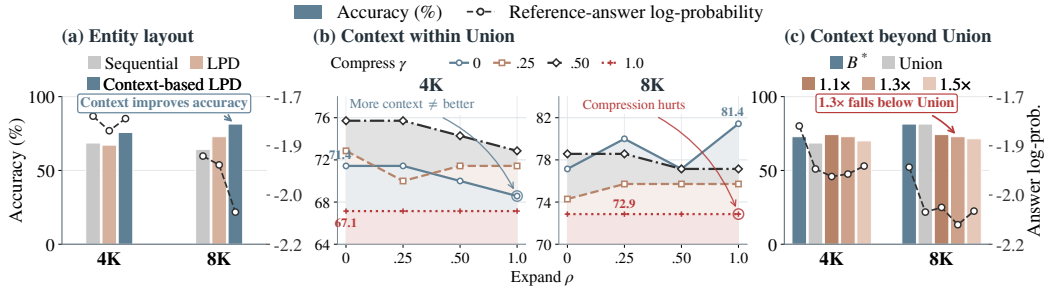


Figure 2: **Context can help, but keeping more does not consistently improve accuracy.** On 70 HR-Bench questions per resolution, (a) compares entity-only layouts with a contextual reference, (b) varies expansion ρ and compression γ within Union, and (c) compares Union, expanded Union, and adaptive B^*. In (a,c), bars show accuracy and lines show reference-answer log-probability.

Does context help? We first compare entity-only layouts with a contextual reference on 70 HR-Bench questions per resolution using Qwen2.5-VL-7B, with entity regions fixed. Sequential tiles entity crops in cue order, while LPD compresses gaps and preserves their relative arrangement. Context-based LPD uses the fixed state with the highest accuracy in Fig. 2(b) as a descriptive reference for this diagnostic sample. LPD improves over Sequential on 8K, while the contextual reference outperforms LPD at both resolutions (Fig. 2(a)), suggesting that evidence outside entity boxes also contributes to relation perception. ***How much context should be kept?*** We next vary context within Union through the two-dimensional candidate space $B(\rho, \gamma)$, where γ compresses non-entity rows and columns and ρ expands context from the remapped entity convex hull. We sample four rank fractions along each dimension (Fig. 2(b)). On 4K, moderate compression outperforms full Union, but further compression to LPD reduces accuracy. On 8K, full Union achieves the highest accuracy. This difference reflects the trade-off between removing interference and keeping evidence, with the preferred context extent varying across the two question sets. ***Is more context always beneficial?*** We further expand Union by factors of 1.1, 1.3, and 1.5 and directly use the enlarged crops, comparing them with Union and the adaptive B^* selected from the full candidate space within Union (Fig. 2(c)). Small expansion benefits 4K, but the gain diminishes as expansion increases, while all three expansions perform below Union on 8K. These results motivate context selection for each question, guided by evidence about the queried relation (Appendix B).

2.2 Context Selection under Crop Before Compose

To assess whether context selection under *Crop Before Compose* can capture useful relation cues, we compare six controls: *Random*, *Source*, *Partner*, *Sum*, *Zoom-out*, and *Union*. In these controls, context is selected after the entity regions have been cropped. *Random* uses a random window with a similar area, while *Source*, *Partner*, and *Sum* select windows from the response map of one entity, the other

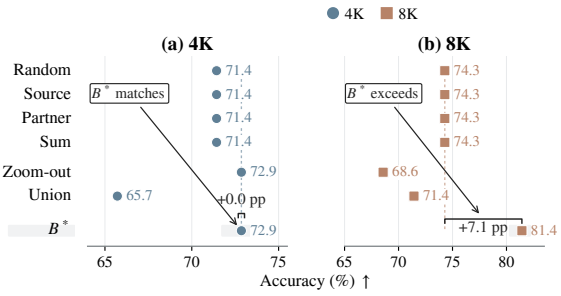


Figure 3: **Context selection under Crop Before Compose.** Accuracy on HR-Bench at 4K and 8K.

entity, or their sum. *Zoom-out* and *Union* keep broader regions. Each selected context is combined with the fixed entity regions in one unified view. We compare these views with B^*, whose context is selected after composing entity evidence on the original visual grid. Across both HR-Bench resolutions, B^* shows an overall advantage over the *Crop Before Compose* controls, supporting the benefit of composing evidence before cropping.

The results also explain why these controls behave differently. *Union* performs worst overall, suggesting that keeping the full union also keeps irrelevant regions and increases interference. *Zoom-out* can recover useful relational cues, but its broader view can also add distracting content, so its performance varies across resolutions. *Random*, *Source*, *Partner*, and *Sum* produce similar results. With the entity regions fixed and the context area kept similar, response maps from individual entities provide little information for selecting context that depends on their relation. Consequently, neither one response map nor their sum improves reliably over random selection. In contrast, B^* first composes the entity evidence, preserves its spatial correspondence, and then selects context from the composed map. This provides a more suitable context for questions about relations. Table 5 further tests how the composed evidence is constructed.

3 METHOD

3.1 OVERVIEW

Motivated by the above analysis, **PostCrop** follows **Compose Before Crop**, composing evidence from the queried entities before forming the final crop. This training-free design allows the crop to keep relation-relevant context while suppressing distracting content. Given an image I, a question q, and answer options O, PostCrop first performs *Cue-Level Extraction* to obtain separate maps for the object and attribute cues of each entity (Fig. 4). *Single-entity Localization* combines these cue maps, generates candidate regions, and uses verification to select one component for each entity. *Multi-entity Composition* then connects verified entity maps on the original visual-token grid, while a single-entity question directly uses its verified entity map. *Cropping & Answering* uses the verified entity view for single-entity questions and, for multi-entity questions, searches over compression and context expansion, selects a view B^* supported by the evidence, and answers using both the original image I and B^*. The model answers without option labels and verifies each option only when the answer is ambiguous.

3.2 EVIDENCE LOCALIZATION AND COMPOSITION

To obtain the evidence used for cropping, PostCrop first transforms cues relevant to the question into a composed map on the original visual-token grid (Fig. 4). This stage proceeds in three steps. (i) *Cue-level extraction*, inspired by HiDe's Token-wise Attention Decoupling (TAD) (Liu et al., 2026a), separately maps object and attribute cues and assembles them into entity maps. (ii) *Single-entity localization* uses verification to select one spatial component from each entity map. (iii) *Multi-entity composition* connects and combines the verified entity maps on the original visual-token grid.

Cue-level extraction. We first prompt the LVLM with "Which visual cues are needed to answer the question?" together with the answer options to extract the entity set $\mathcal{E} = \{e_1, \ldots, e_N\}$ (Fig. 4 and Appendix G). Decomposing an entity phrase into its constituent cues helps capture regions jointly supported by those cues (Zhong et al., 2025). Each entity e, such as *bicycle with person*, comprises an object cue o_e (e.g., *bicycle*) and attribute cues $\mathcal{M}_e$ (e.g., *with person*), yielding $e := \{o_e\} \cup \mathcal{M}_e$. Middle layers provide the strongest localization signal (Liu et al., 2026a), so we extract every cue map from the middle decoder layer ℓ_{mid}. For each cue token t, *attention contrast* subtracts mean non-cue attention $\bar{A}_{\neg c}$ from its visual attention map A_t and keeps the positive part (Han et al., 2026). The resulting maps are normalized and averaged to form the cue map:

$$E_c = \mathcal{N}\left(\frac{1}{|Q_c^+|} \sum_{t \in Q_c^+} \mathcal{N}\big(\mathcal{S}\big([A_t - \bar{A}_{\neg c}]_+\big)\big) \right), \tag{1}$$

Here, $\mathcal{S}$ rescales each contrast map using its minimum and maximum, whereas $\mathcal{N}$ normalizes a map so that its entries sum to one. Q_c^+ contains cue tokens whose contrast maps vary across visual tokens. If this set is empty, we use the uniform fallback described in Appendix B.

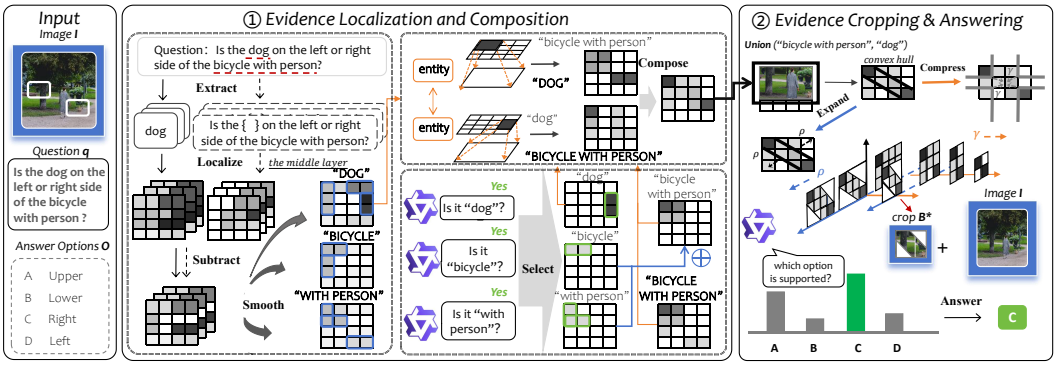


Figure 4: **The Framework of PostCrop. (1) Evidence Localization and Composition.** Object and attribute cues extracted from the question and answer options are localized using attention from a middle decoder layer. Subtracting attention from the remaining prompt tokens yields cue maps, which form entity maps through object and attribute composition. Smoothing provides candidate regions (blue outlines), whose local crops the model verifies with "Is it `{cue}`?" to select matching regions (green outlines) that guide bidirectional composition on the original visual grid. **(2) Evidence Cropping & Answering.** From Union, we form the entity convex hull and vary context expansion ρ and gap compression γ to generate candidate views. The context score selects B^*.

Single-entity localization. An entity is jointly supported by its object cue o_e and attribute cues $\mathcal{M}_e$. To emphasize locations supported by both, we define its entity map $E_e(z)$:

$$E_e(z) = \mathcal{N}\left(E_{o_e} + \mathcal{N}\left(\prod_{c \in e} E_c\right)\right)(z). \tag{2}$$

The product is normalized before addition, giving it and the object cue equal mass while the *object residual* preserves object support after cue conjunction. Subsequently, we smooth E_e, partition it into watershed basins, and rank them by summed evidence. To suppress responses from visually similar instances (Wang et al., 2026), for each candidate we ask the LVLM "Is it `{cue}`?" using only its local crop and select the first candidate accepted as e. If none is verified, we use the first basin in this ranking. Denoting the selected basin mask by M_e (see Appendix B for verification details), we obtain

$$E_e^*(z) = M_e(z)\, E_e(z). \tag{3}$$

Multi-entity composition. Section 2.2 motivates composition on the original visual-token grid to preserve layout and expose context supported by multiple entities. A single-entity question can therefore use E_e^* directly. For a question involving multiple entities, however, we also need responses of visual tokens jointly supported by the entities, in addition to their individual responses (Abnar & Zuidema, 2020). Taking two entities as an example, we perform bidirectional transfer on the original visual-token grid to construct their joint support as follows.

Proposition 1 (Bidirectional support and symmetry)**.** *For entity maps E_a, E_b with verified source maps E_a^*, E_b^*, denote by $T^{\ell_{\text{mid}}}(z, x \mid p_{a\to b})$ the attention from target token z to source token x at the middle layer, normalized over source visual tokens in each head and then averaged across heads. The transfer input $p_{a\to b}$ uses two copies of the original image with a relation instruction between them (see Appendix B for the input and normalization). The transfer from a to its partner b is*

$$\Gamma_{a\to b}(z) = \sum_{x \in \Omega} T^{\ell_{\text{mid}}}(z, x \mid p_{a\to b}) \mathcal{N}(E_a^*)(x). \tag{4}$$

Defining $\Gamma_{b\to a}$ analogously by exchanging a and b gives the bidirectional relation map

$$E_{\mathcal{E}}(z) = \mathcal{N}(\mathcal{N}(E_b \Gamma_{a\to b}) + \mathcal{N}(E_a \Gamma_{b\to a}))(z), \tag{5}$$

where each directional product is normalized before addition. For positive product mass, its support is the intersection of the full target response and the transferred source response. This construction is symmetric with respect to exchanging the two entities and transfer directions (see Appendix B).

For $N > 2$, incoming transfers are averaged for each target j before normalizing the product with its full entity map E_j. The target fields are then summed and normalized, with the exact aggregation given in Appendix B.

3.3 EVIDENCE CROPPING & ANSWERING

Guided by Sec. 2.1, this stage uses composed evidence to select a crop and answer the question.

Cropping based on composed evidence. To balance keeping evidence and contextual interference, we select context according to its support for the queried relation. A single-entity question uses its verified bounding box, while multi-entity candidates vary between LPD and Union to recover relational cues or remove unnecessary context. The verified entity boxes form R, and $U = \mathrm{BBox}(R)$ defines Union, a sufficient search boundary under complete cue coverage, complete localization, and relation locality (see Appendix B for the assumptions and proof). Their starting region is the *entity convex hull*, the smallest convex region containing the verified boxes. In Fig. 4, this is the diagonal region covering the object boxes and the space between them, from which ρ expands context while γ compresses gaps while preserving entity pixels and axis order. The resulting views are

$$C_{\rho,\gamma} = \{z \in V_\gamma \setminus R_\gamma : \mathrm{dist}(z, H_\gamma) \le \rho\}, \qquad B_{\rho,\gamma} = R_\gamma \cup C_{\rho,\gamma}, \tag{6}$$

where R_γ contains entity tokens at their original resolution, remapped onto the valid compressed grid V_γ, and H_γ is the convex hull of their boxes after compression (see Appendix B).

To favor context with both strong and broad relational support, we score candidates in $\mathcal{R}_q = \{B_{\mathrm{LPD}}\} \cup \{B_{\rho,\gamma}\}$ using $E_{\mathrm{rel}} = E_{\mathcal{E}}$, combining evidence above a shared background level with context coverage:

$$\mu_D = \mathrm{mean}(E_{\mathrm{rel}}[D \setminus R]), \qquad J(C) = \left[\sum_{i \in C}(E_i - \mu_D)\right] \frac{|C|}{|C_{\mathrm{Union}}|}, \tag{7}$$

where E_i is the response at token i's position on the original visual-token grid, $C_{\mathrm{Union}} = U \setminus R$ contains the non-entity tokens within Union, and D is the fixed observed domain, including observed zeros (see Appendix B for its boundary). Responses below μ_D reduce the evidence sum, while coverage rewards broader context when the sum is positive. For each view B, $C(B)$ denotes its kept context, with $C(B_{\mathrm{LPD}}) = \varnothing$. With $J(\varnothing) = 0$, LPD remains available when no added context offers positive excess, and we select

$$B^* = \arg\max_{B \in \mathcal{R}_q} J(C(B)). \tag{8}$$

Proposition 2 (Context coverage and excess evidence)**.** *Fix D, R, and $|C_{\mathrm{Union}}| > 0$. A nonempty context scores above LPD exactly when its mean evidence exceeds μ_D. At fixed positive excess mean, $J(C)$ increases quadratically with $|C|$. If every nonempty candidate has mean evidence at most μ_D, LPD is a maximizer (see Appendix B for the derivation and the Union boundary).*

Answering with selective option verification. For multi-entity questions, the original image provides a global reference, while B^* presents the selected evidence in a focused view. Generated text can be more robust to option reordering than first-token probability readout (Wang et al., 2024b), so the LVLM first produces a concise answer without option letters (Fig. 4), which is matched to the options by content. If the match is ambiguous, each unresolved option is verified with a Yes/No question (Liu et al., 2026b; Hu & Liu, 2026).

4 EXPERIMENTS

Benchmarks. We evaluate PostCrop on four challenging high-resolution benchmarks: VSTAR (Wu & Xie, 2024), HR-Bench-4K, HR-Bench-8K (Wang et al., 2025), and MME-RealWorld-Lite (Zhang et al., 2025b). VSTAR contains 191 questions spanning attribute recognition and spatial reasoning, with an average image resolution of 2246×1582. Each HR-Bench resolution contains 200 original questions, split evenly between fine-grained single-instance perception (FSP) and fine-grained cross-instance perception (FCP). HR-Bench-8K images average 7680 pixels in resolution, while HR-Bench-4K is derived by cropping the 8K images around the relevant objects. MME-RealWorld covers five real-world scenarios and 43 subtasks. Its Lite version contains 1,919 questions, divided into 1,169 perception and 750 reasoning questions.

Main Results. We compare PostCrop with the base models and two representative approaches: ViCrop (Zhang et al., 2025a) and HiDe (Liu et al., 2026a). ViCrop localizes key regions before

cropping and zooming, while HiDe reconstructs localized entity regions into a compact view through LPD. On VSTAR and HR-Bench, completed PostCrop evaluations use Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-8B (Bai et al., 2025a), alongside InternVL2.5-4B (see Appendix C.1 for the detailed experimental configurations).

Table 1: **Accuracy on VSTAR and HR-Bench.** Bold and underlined scores mark the best and second best results for each backbone and metric.

Method	Training free	VSTAR			HR-Bench-4K			HR-Bench-8K		
		Attr.	Spatial	Overall	FSP	FCP	Overall	FSP	FCP	Overall
Qwen2.5-VL-7B										
Base	–	80.9	76.3	79.1	88.8	54.8	71.8	84.2	51.5	67.9
ViCrop (ICLR'25)	✓	89.6	71.1	82.2	90.5	57.5	74.0	85.5	53.0	69.3
HiDe (ICML'26)	✓	**94.8**	86.8	91.6	**95.0**	61.0	78.0	**96.0**	55.0	75.5
PostCrop (ours)	✓	93.9	**94.7**	**94.2**	93.0	**74.0**	**83.5**	93.0	**76.0**	**84.5**
Qwen3-VL-8B										
Base	–	87.0	78.9	83.8	90.2	63.7	77.0	84.0	62.3	73.1
ViCrop (ICLR'25)	✓	89.6	75.0	83.8	91.8	64.3	78.0	88.0	63.3	75.6
HiDe (ICML'26)	✓	92.9	85.5	89.5	93.5	69.0	81.2	93.5	67.3	80.4
PostCrop (ours)	✓	**93.0**	**89.5**	**91.6**	**96.0**	**72.0**	**84.0**	**94.0**	**73.0**	**83.5**
InternVL2.5-4B										
Base	–	68.7	68.4	68.6	77.5	53.8	65.6	63.0	49.5	56.3
ViCrop (ICLR'25)	✓	44.0	58.3	51.1	61.3	49.5	55.4	60.8	44.5	52.6
HiDe (ICML'26)	✓	**89.6**	67.1	80.6	**89.5**	48.8	69.1	**87.3**	46.3	66.8
PostCrop (ours)	✓	86.1	**80.3**	**83.8**	81.0	**60.0**	**70.5**	84.0	**57.0**	**70.5**

Table 1 shows that PostCrop achieves the best overall performance on all three evaluated backbones. Its strongest performance appears on VSTAR Spatial and HR-Bench FCP, where answers depend on relations among entities. HiDe remains stronger on attribute and FSP questions for Qwen2.5-VL-7B and InternVL2.5-4B, while the Qwen3-VL-8B results also show PostCrop performing well on these single-entity subsets. This pattern follows the scope of context selection: single-entity questions use the verified entity crop directly, whereas multi-entity questions additionally use composed evidence to keep context and its correspondence with the entities. Locating each entity alone can leave the evidence needed to relate them incomplete, which gives context selection a more direct role in relation-focused tasks.

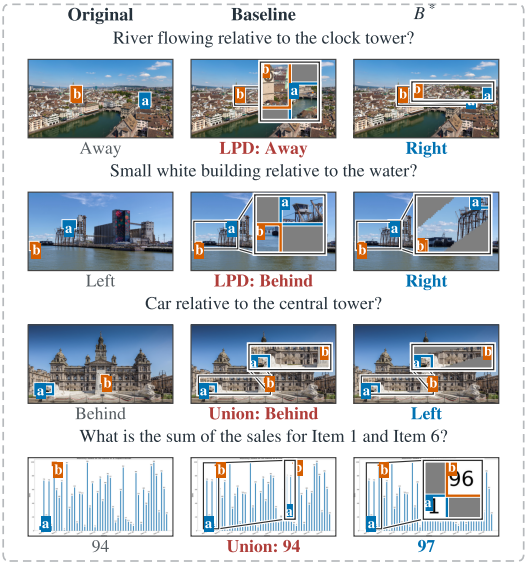


Figure 5: **Final crops for four HR-Bench questions.** Each row compares the Original, Baseline, and B^* views. Blue labels indicate the correct answers; a and b mark the queried entities.

Table 2: **Accuracy on MME-RealWorld-Lite.**

Method	Per.	Rea.	Overall
Base	51.41	39.60	46.80
HiDe	**58.43**	42.67	52.27
PostCrop	55.69	**47.47**	**52.48**

On MME-RealWorld-Lite, we use Qwen2.5-VL-7B to compare the methods, and PostCrop outperforms Base on both subsets (Table 2). PostCrop keeps context around composed entity evidence, whereas HiDe reconstructs a more compact entity view. This distinction matches the split: PostCrop's gains are concentrated in monitoring calculations and driving intention prediction, while relative losses occur in OCR and motion-attribute recognition. This split reflects their different visual demands: reasoning benefits from context around composed entity evidence, whereas OCR and recognition of motion attributes favor a compact entity view.

Qualitative Analysis. Fig. 5 shows LPD losing relational context, Union keeping distractors, and PostCrop answering all four cases correctly.

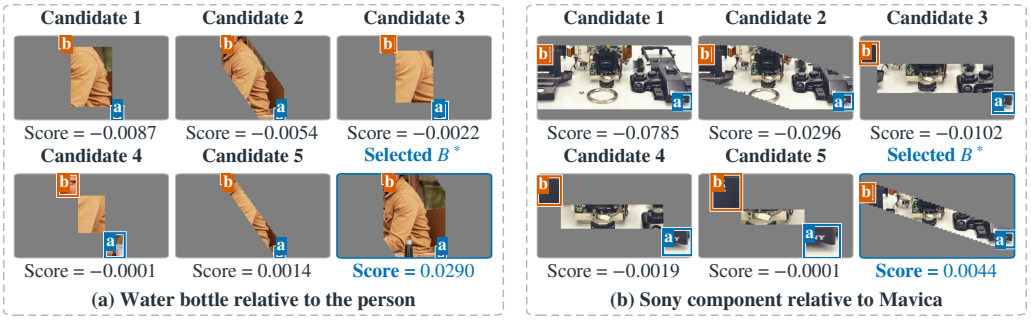


Figure 6: **Selected crops keep different amounts of context.** Candidates for two HR-Bench-4K questions are ordered by score (Eq. 7). Blue marks the selected B^*, and gray marks excluded regions.

Fig. 6 further shows how B^* is selected from the two-dimensional expansion–compression candidate space on HR-Bench-4K. For the bottle and person, B^* receives the highest score with stronger contextual support. In the camera example, the score instead favors a narrower view that connects the queried components over the other expanded and compressed views.

Ablation Study. We ablate PostCrop with Qwen2.5-VL-7B on VSTAR for localization and on HR-Bench for crop selection, relation composition, and context scoring. Each comparison changes one component, with composition and scoring evaluated on the complete FCP subset.

Table 3: **Localization ablation on VSTAR.**

Variant	VSTAR		
	Attr.	Spatial	Overall
Base	**93.9**	92.1	93.2
+ Sub-cue product	**93.9**	92.1	93.2
+ Object residual	**93.9**	**94.7**	**94.2**

Table 4: **Crop ablation on HR-Bench.**

Variant	4K		8K	
	FSP	FCP	FSP	FCP
LPD	**93.0**	67.0	**93.0**	72.0
Union	**93.0**	68.0	**93.0**	**78.0**
PostCrop (B^*)	**93.0**	**71.0**	**93.0**	**78.0**

Evidence localization. Table 3 shows that the object residual improves spatial accuracy while leaving attribute accuracy unchanged. The sub-cue product alone leaves accuracy unchanged: it emphasizes shared cue responses, but a weak cue can also suppress useful object responses. The residual preserves this support for localization, consistent with the spatial gain (Appendix D).

Adaptive crop selection. B^* improves 4K FCP over LPD and Union but ties Union on 8K FCP (Table 4). With other components fixed, keeping Union provides little benefit over LPD on 4K but accounts for the gain on 8K. These results support a candidate space containing both compact views and the full Union, allowing selection to keep broader context when it is useful.

Relation composition. Bidirectional composition matches the best accuracy at both resolutions (Table 5). Here a, b are full entity maps, a_R, b_R restrict them to verified boxes, and $E_{a\to b}, E_{b\to a}$ denote directional evidence. These controls vary both support and transfer (Appendix C). Either direction improves over local addition on 4K. On 8K, bidirectional composition exceeds both one-way variants.

Table 5: **Relation composition.**

Composition	4K	8K
$a_R + b_R$	68.0	**78.0**
$a + E_{a\to b}$	**71.0**	77.0
$b + E_{b\to a}$	**71.0**	77.0
$\boldsymbol{a \leftrightarrow b}$	**71.0**	**78.0**

Table 6: **Context scoring.**

Metric	4K	8K
Mean	69.0	76.0
Excess only	71.0	77.0
Coverage only	68.0	78.0
Context score	**71.0**	**78.0**

Context scoring. The context score matches the best ablated metric at both resolutions and exceeds mean scoring (Table 6). On 4K, Excess alone ties the combined score, whereas Coverage alone selects Union and performs worse, pointing to the importance of distinguishing useful evidence from distracting context. On 8K, Coverage ties the combined score while Excess falls slightly behind. The combined criterion accounts for both evidence above background and its extent, with their contributions differing across resolutions.

Context-source ablation. Since the original image already contains visual context similar to that in B^*, we test whether a focused view keeping this context provides an additional benefit. We fix B^* and independently block context access in the two answering views on HR-Bench-4K and HR-Bench-8K, keeping entity pixels, crop geometry, and image order fixed (Appendix C.2).

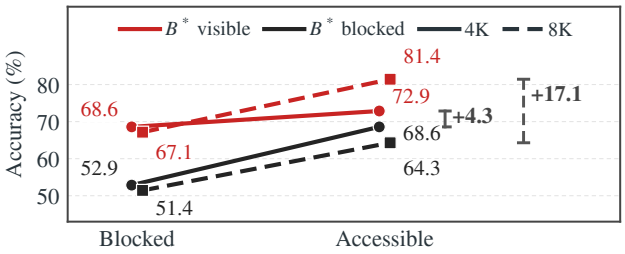


Figure 7: **Context-source ablation.** The horizontal axis shows original-image context access. Color denotes B^* context access; line style denotes resolution.

With the original image intact, keeping context in B^* improves accuracy by $+4.3$ and $+17.1$ percentage points on 4K and 8K, respectively (Fig. 7). The gain shows that B^* concentrates relation-relevant evidence while the original image preserves broader context. The two sources therefore provide complementary evidence: blocking either one reduces accuracy, whereas keeping both accessible performs best, especially on 8K.

Efficiency Analysis. PostCrop trades additional inference time for higher accuracy. In separate matched timing runs on VSTAR with Qwen2.5-VL-7B, it improves accuracy over HiDe while taking $2.22\times$ as long (Table 7). Peak allocated memory is similar, although PostCrop reserves more memory. Evidence extraction, identity verification, and bidirectional transfer take 75.8% of runtime, while candidate enumeration takes 0.42 seconds for the full VSTAR run. Evidence construction and verification dominate the runtime, while searching the candidate views adds little cost.

Table 7: **End-to-end inference cost on VSTAR.**

Method	Memory	Time
ViCrop	∼20 GB	∼28 min
HiDe	19.3 GiB	17.7 min
PostCrop	19.1 GiB	39.2 min

5 Related Work

5.1 High-Resolution LVLMs and Multi-Entity Perception

LVLMs improve visual understanding through multimodal alignment and visual instruction tuning (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023; Chen et al., 2024b; Bai et al., 2023). To preserve fine-grained visual information, high-resolution architectures use resolution-adaptive tokenization (Wang et al., 2024a; Bai et al., 2025b) and restore continuity across image patches (Huang et al., 2025). Such details are particularly important for relating entities whose evidence is distributed across an image, motivating studies of relation perception and visual grounding (Ranasinghe et al., 2024; Li et al., 2026b; Cui et al., 2026; Chen et al., 2024a). Related work further studies region-level and interactive spatial reasoning (Cheng et al., 2024; Mayer et al., 2025).

5.2 Visual Evidence Selection and Context Construction

Training-based approaches learn visual operations during reasoning (Zheng et al., 2026; Su et al., 2025; Zhang et al., 2026), whereas training-free methods use attention or gradients (Zhang et al., 2025a; Ma et al., 2026), search candidate regions (Shen et al., 2025; Li et al., 2025; 2026a; Li & Peng, 2026), or refine local context after localization (Yu et al., 2025; Li et al., 2026c). Beyond region selection, semantic cue separation distinguishes queried entities from distractors (Wang et al., 2026; Liu et al., 2026a). The resulting evidence can then be supplied through visual working memory (Wu & Xie, 2024), textual patch descriptions (Wang et al., 2025), or a constructed image view. For multiple entities, *Crop Before Compose* combines localized entity crops, with LPD compressing intervening gaps (Liu et al., 2026a) and Union keeping the enclosing rectangle (Zhong et al., 2025; Yu et al., 2025). In contrast, PostCrop follows *Compose Before Crop*, composing entity evidence on the original visual-token grid to select context through expansion and gap compression.

6 Conclusion

In this paper, we characterize a visual context paradox in fine-grained multi-entity perception: entity crops can obscure relations, whereas retained context can introduce visual interference. To address it, we propose PostCrop, a training-free framework based on a new *Compose Before Crop* paradigm that composes entity evidence before forming the final crop. More broadly, our findings suggest a shift in fine-grained visual perception from locating evidence to organizing it for reasoning.

AI USE STATEMENT

Generative AI tools were used to polish the writing, retrieve relevant literature, support research ideation and execution, including the refinement of experimental analyses, draft and revise portions of the paper, and assist with mathematical claims and proofs. The authors reviewed all AI-assisted work, verified the cited literature, mathematical arguments, and reported results, and take responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

The method and inference procedure are described in Sections 3.1–3.3, with the assumptions and proofs provided in Appendix B. Section 4 defines the benchmarks, question counts, and evaluation protocol. Appendix C specifies the models, baseline sources, and experimental settings; Appendix D reports supplementary experiments; and Appendix G provides the prompt templates. The implementation, experiment scripts, configurations, and run manifests will be included as supplementary material.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385/.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022. doi: 10.52202/068431-1723. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. URL https://arxiv.org/abs/2308.12966.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024a. doi: 10.1109/CVPR52733.2024.01370. URL https://openaccess.thecvf.com/

content/CVPR2024/html/Chen_SpatialVLM_Endowing_Vision-Language_Models_with_Spatial_Reasoning_Capabilities_CVPR_2024_paper.html.

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. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_InternVL_Scaling_up_Vision_Foundation_Models_and_Aligning_for_Generic_CVPR_2024_paper.html.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 135062–135093, 2024. doi: 10.52202/079017-4293. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f38cb4cf9a5eaa92b3cfa481832719c6-Abstract-Conference.html.

Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5076–5084, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Cheng_Focusing_Attention_Towards_ICCV_2017_paper.html.

Kelly Cui, Nikhil Prakash, Shoval Messica, Ayush Raina, David Bau, Antonio Torralba, and Tamar Rott Shaham. The dual mechanisms of spatial variable binding in vision-language models. *arXiv preprint arXiv:2603.22278*, 2026. URL https://arxiv.org/abs/2603.22278.

Tianchen Guo, Chen Liu, Ling Chen, and Xin Yu. SSMNBench: Diagnosing image-based cross-view human-object understanding via single-view sufficiency and multi-view necessity. *arXiv preprint arXiv:2606.25634*, 2026. URL https://arxiv.org/abs/2606.25634.

Su Ho Han, Jeongseok Hyun, Pilhyeon Lee, Minho Shim, Dongyoon Wee, and Seon Joo Kim. Decomposed attention fusion in MLLMs for training-free video reasoning segmentation. In *International Conference on Learning Representations*, pp. 28342–28370, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2fc6b8a3fc23108f184daa4759024c25-Abstract-Conference.html.

Rongbin Hu and Jeffrey Liu. Binary verification for zero-shot vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 11498–11506, 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/VAR/html/Hu_Binary_Verification_for_Zero-Shot_Vision_CVPRW_2026_paper.html.

Runhui Huang, Xinpeng Ding, Chunwei Wang, Jianhua Han, Yulong Liu, Hengshuang Zhao, Hang Xu, Lu Hou, Wei Zhang, and Xiaodan Liang. HiRes-LLaVA: Restoring fragmentation input in high-resolution large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29814–29824, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Huang_HiRes-LLaVA_Restoring_Fragmentation_Input_in_High-Resolution_Large_Vision-Language_Models_CVPR_2025_paper.html.

Yubo Jiang, Xin Yang, Abudukelimu Wuerkaixi, Zheming Yuan, Xuxin Cheng, Cao Liu, Ke Zeng, Fengying Xie, Zhiguo Jiang, and Haopeng Zhang. Global context or local detail? adaptive visual grounding for hallucination mitigation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15133–15152, 2026. doi: 10.18653/v1/2026.findings-acl.745. URL https://aclanthology.org/2026.findings-acl.745/.

Geng Li and Yuxin Peng. BVS: Bayesian visual search with multimodal large language model for fine-grained perception. *arXiv preprint arXiv:2607.03184*, 2026. URL https://arxiv.org/abs/2607.03184.

Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. DyFo: A training-free dynamic focus visual search for enhancing LMMs in fine-grained visual understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9098–9108, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_DyFo_A_Training-Free_Dynamic_Focus_Visual_Search_for_Enhancing_LMMs_CVPR_2025_paper.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10313–10322, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Li_Relation-Aware_Graph_Attention_Network_for_Visual_Question_Answering_ICCV_2019_paper.html.

Liupeng Li, Haoqian Kang, Zhenyu Lu, Jinpeng Wang, Bin Chen, Ke Chen, and Yaowei Wang. CVSearch: Empowering multimodal LLMs with cognitive visual search for high-resolution image perception. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026a. URL https://icml.cc/virtual/2026/poster/65958.

Rang Li, Lei Li, Shuhuai Ren, Hao Tian, Shuhao Gu, Shicheng Li, Zihao Yue, Yudong Wang, Wenhan Ma, Zhe Yang, Jingyuan Ma, Zhifang Sui, and Fuli Luo. GroundingME: Exposing the visual grounding gap in MLLMs through multi-dimensional evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2412–2422, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_GroundingME_Exposing_the_Visual_Grounding_Gap_in_MLLMs_through_Multi-Dimensional_CVPR_2026_paper.html.

Yangfu Li, Hongjian Zhan, Jiawei Chen, Yuning Gong, Qi Liu, and Yue Lu. DeepScan: A training-free framework for visually grounded reasoning in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19155–19164, 2026c. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_DeepScan_A_Training-Free_Framework_for_Visually_Grounded_Reasoning_in_Large_CVPR_2026_paper.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Xianjie Liu, Yiman Hu, Yixiong Zou, Liang Wu, Jian Xu, and Bo Zheng. HiDe: Rethinking the Zoom-IN method in high resolution MLLMs via hierarchical decoupling. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026a. URL https://icml.cc/virtual/2026/poster/60950.

Yesheng Liu, Hao Li, Haiyu Xu, Baoqi Pei, Jiahao Wang, Mingxuan Zhao, Jing-Shu Zheng, Zheqi He, JG Yao, Xi Yang, Bowen Qin, and Jiajun Zhang. Beyond multiple choice: Verifiable OpenQA for robust vision-language RFT. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18859–18869, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_Beyond_Multiple_Choice_Verifiable_OpenQA_for_Robust_Vision-Language_RFT_CVPR_2026_paper.html.

Mingjie Ma, Yichao Ma, Zhong Yang, and Guohui Li. Seeing what matters: A training-free self-guided framework for multimodal detail perception and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8727–8736, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ma_

Seeing_What_Matters_A_Training-Free_Self-Guided_Framework_for_Multimodal_Detail_CVPR_2026_paper.html.

Julius Mayer, Mohamad Ballout, Serwan Jassim, Farbod Nosrat Nezami, and Elia Bruni. iVISPAR—an interactive visual-spatial reasoning benchmark for VLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26757–26781, 2025. doi: 10.18653/v1/2025.emnlp-main.1359. URL https://aclanthology.org/2025.emnlp-main.1359/.

Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12977–12987, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ranasinghe_Learning_to_Localize_Objects_Improves_Spatial_Reasoning_in_Visual-LLMs_CVPR_2024_paper.html.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. ZoomEye: Enhancing multimodal LLMs with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6602–6618, 2025. doi: 10.18653/v1/2025.emnlp-main.335. URL https://aclanthology.org/2025.emnlp-main.335/.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel Reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0280. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0c38f54740062529aa4117a04b583f3c-Abstract-Conference.html.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a. URL https://arxiv.org/abs/2409.12191.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8):7907–7915, 2025. doi: 10.1609/aaai.v39i8.32852. URL https://ojs.aaai.org/index.php/AAAI/article/view/32852.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. In *First Conference on Language Modeling*, 2024b. URL https://openreview.net/forum?id=qHdSA85GyZ.

Yajing Wang, Chao Bi, Junshu Sun, Shufan Shen, Zhaobo Qi, Shuhui Wang, and Qingming Huang. ActiveScope: Actively seeking and correcting perception for MLLMs. *arXiv preprint arXiv:2606.24292*, 2026. URL https://arxiv.org/abs/2606.24292.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

Qiucheng Wu, Handong Zhao, Michael Saxon, Trung Bui, William Yang Wang, Yang Zhang, and Shiyu Chang. VSP: Diagnosing the dual challenges of perception and reasoning in spatial planning tasks for MLLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2270–2280, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wu_VSP_Diagnosing_the_Dual_Challenges_of_Perception_and_Reasoning_in_ICCV_2025_paper.html.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, pp. 21875–21895, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5e5fd18f863cbe6d8ae392a93fd271c9-Abstract-Conference.html.

Xuan Yu, Dayan Guan, Michael Ying Yang, and Yanfeng Gu. Zoom-Refine: Boosting high-resolution multimodal understanding via localized zoom and self-refinement. *arXiv preprint arXiv:2506.01663*, 2025. URL https://arxiv.org/abs/2506.01663.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *International Conference on Learning Representations*, pp. 68194–68213, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/aaa0ac4253da75faf9b0dc0dda062612-Abstract-Conference.html.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images. In *International Conference on Learning Representations*, pp. 79536–79572, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8080912b919ad0344c0d34e68983dfcb-Abstract-Conference.html.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. MME-RealWorld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, pp. 89655–89701, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/df29d63af05cb91d705cf06ba5945b9d-Abstract-Conference.html.

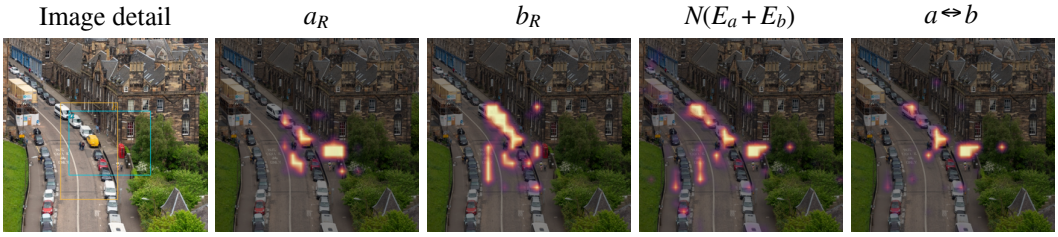
Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, pp. 126775–126798, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/cdb347d7516d52a9280ea9d6708911f6-Abstract-Conference.html.

Liangyu Zhong, Fabio Philipp Rosenthal, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, and Leo Schwinn. FOCUS: Internal MLLM representations for efficient fine-grained visual question answering. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3227. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8b02cdb04bc6351d36565b5c874bb804-Abstract-Conference.html.

A SUPPLEMENTARY VISUALIZATIONS

Figure 8 compares entity responses restricted to verified boxes, the sum of full entity maps, and bidirectional composition. The VSTAR examples cover left/right, above/below, and distance from the camera. Examples were chosen for visible response differences and scene readability, independently of answer correctness.

VSTAR q180. Is the telephone booth on the left or right side of the yellow van?



VSTAR q181. Is the red balloon above or below the white balloon?



VSTAR q182. Which one is closer to the camera, the black vehicle or the silver vehicle?



Figure 8: **Bidirectional evidence composition on VSTAR.** Fresh Qwen2.5-VL-7B captures at physical layer 15. From left to right: detail from the original image with automatically verified entity boxes, entity maps restricted to verified boxes a_R, b_R, the sum of full entity maps $\mathcal{N}(E_a + E_b)$, and bidirectional composition (Eq. 5). Maps are normalized so that their entries sum to one on the original visual-token grid. All panels in a row share the same display window.

Spatial coverage and response selectivity. The maps restricted to verified boxes, $a_R = \mathcal{N}(E_a \mathbf{1}_{R_a})$ and $b_R = \mathcal{N}(E_b \mathbf{1}_{R_b})$ concentrate responses within the selected entity regions by construction. Adding the full entity maps, $\mathcal{N}(E_a + E_b)$, extends coverage to the scene around both entities, while retaining scattered background responses. Bidirectional composition preserves the main responses around the queried spatial configuration and attenuates many of these scattered activations. This is particularly visible on the wall and road surrounding the balloons in q181 and in the background around the vehicles in q182. Compared with the sum of full entity maps, composition thus gives

a more selective response pattern in these examples. This is a qualitative observation of response selectivity, whose contribution to answering requires separate evaluation.

B Theoretical and Method Details

For entity regions $R = \bigcup_{e\in\mathcal{E}} R_e$, define $U = \mathrm{BBox}(R)$ and $C_{\mathrm{Union}} = U \setminus R$. Content X_V is observed at original coordinates, before resizing or compression. Probabilities condition on q and fixed entity boxes. The Union-boundary result uses three assumptions:

- *Complete cue coverage.* $\mathcal{E}$ contains all cues needed for q, including headers, legends, and scales when relevant.
- *Complete localization.* Each entity state Z_e (identity, attributes, and geometry) is recoverable from R_e: $Z_e = f_e(X_{R_e}, q)$.
- *Relation locality.* Context in $U \setminus R$ and entity states $Z_{\mathcal{E}} = \{Z_e\}_{e\in\mathcal{E}}$ determine the answer posterior for the full image through

$$Z_{\mathrm{rel}} = g(X_{U\setminus R}, Z_{\mathcal{E}}, q), \qquad P(Y = y \mid X_\Omega, q) = h_y(Z_{\mathcal{E}}, Z_{\mathrm{rel}}, q). \tag{9}$$

Proposition 3 (Union as a sufficient search boundary)**.** *Under these assumptions, every W with $U \subseteq W \subseteq \Omega$ has the same answer posterior and Bayes accuracy as U:*

$$P(Y \mid X_W, q) = P(Y \mid X_U, q). \tag{10}$$

Proof. All states in Eq. 9 are recoverable from X_U and hence from X_W for $U \subseteq W \subseteq \Omega$. The law of total expectation gives

$$P(Y = y \mid X_W, q) = \mathbb{E}[h_y(Z_{\mathcal{E}}, Z_{\mathrm{rel}}, q) \mid X_W, q] = h_y(Z_{\mathcal{E}}, Z_{\mathrm{rel}}, q).$$

The identity for $W = U$ gives equal posteriors and hence equal Bayes accuracy. □

Union may contain smaller sufficient crops, and expansion can recover cues omitted by estimated entity boxes. The result concerns Bayes accuracy before resizing or compression, which can affect a fixed LVLM's predictions.

The directional fields in Proposition 1 are nonnegative, and normalization preserves the support of each positive-mass product. Thus, for $F_{a\to b} = \mathcal{N}(E_b\Gamma_{a\to b})$ and $F_{b\to a} = \mathcal{N}(E_a\Gamma_{b\to a})$,

$$\mathrm{supp}(F_{a\to b}) = \mathrm{supp}(E_b) \cap \mathrm{supp}(\Gamma_{a\to b}).$$

Exchanging entities and transfer directions swaps the two fields for fixed maps and operators, leaving $E_{\mathcal{E}} = \mathcal{N}(F_{a\to b} + F_{b\to a})$ unchanged.

For the fixed map, D, and R in Proposition 2, write $K = |C_{\mathrm{Union}}| > 0$ and $r_i = E_i - \mu_D$. For nonempty $C \subseteq C_{\mathrm{Union}}$,

$$J(C) = \frac{|C|}{K}\sum_{i\in C} r_i = \frac{|C|^2}{K}(\bar{E}_C - \mu_D). \tag{11}$$

Thus $J(C) > 0$ exactly when $\bar{E}_C > \mu_D$, and the score grows quadratically with $|C|$ at fixed positive excess mean. Since $J(\varnothing) = 0$, LPD is a maximizer when every candidate mean is at most μ_D.

Conditional utility bound. Under the same fixed setup, assume token utilities u_i satisfy $|r_i - u_i| \le \epsilon$ for all $i \in C_{\mathrm{Union}}$, with $\epsilon \ge 0$. The corresponding proxy utility is $\mathcal{U}(C) = \frac{|C|}{K}\sum_{i\in C} u_i$. For each candidate $C \subseteq C_{\mathrm{Union}}$, the triangle inequality gives

$$|J(C) - \mathcal{U}(C)| \le \frac{|C|}{K}\sum_{i\in C}|r_i - u_i| \le \epsilon\frac{|C|^2}{K} \le \epsilon K. \tag{12}$$

For maximizers C_J of J and $C_{\mathcal{U}}$ of $\mathcal{U}$ over the same candidate family, Eq. 12 gives

$$\mathcal{U}(C_{\mathcal{U}}) - \mathcal{U}(C_J) \le J(C_{\mathcal{U}}) - J(C_J) + 2\epsilon K \le 2\epsilon K.$$

This bounds proxy utility, whose relation to answer accuracy requires empirical evaluation.

Beyond Union, for $C \subseteq C_{\mathrm{Union}}$ and nonempty $A \subseteq \Omega \setminus U$ with mean evidence μ_D, $\sum_{i\in A} r_i = 0$. With the map, background, and denominator fixed, $J(C \cup A) - J(C) = \frac{|A|}{K}\sum_{i\in C} r_i$. Positive total residual in C therefore yields a score gain from coverage alone.

Cue maps and verification. Eq. 1 uses the middle decoder layers listed in Appendix C.1, with normalization so that entries sum to one on the original visual-token grid Ω. Constant contrast maps are skipped, with a uniform cue field if all are constant. Directional products require positive mass and are normalized separately before addition. In Eq. 2, a zero conjunction is replaced by the normalized geometric mean of cue responses floored at the smallest positive normal float64 value. For Qwen localization, a 3×3 average filter smooths each map once, while InternVL uses the settings in Appendix C.1. Tokens follow the largest strictly higher value in their 8-neighborhood to form watershed basins. The ten basins with the highest summed smoothed evidence are verified through their bounding rectangles. We select the first basin with $\log P(\texttt{Yes}) - \log P(\texttt{No}) \geq 0$, or the first basin in the ranking if none passes.

Transfer and multiple entities. The input $p_{i \to j}$ contains two copies of the original image with the relation instruction between them and the partner cue last (Appendix G). Transfer attention uses second-copy queries and first-copy keys. For N participating roles, incoming transfers are averaged before normalizing each target product:

$$\widetilde{\Gamma}_{i \to j}(z) = |\Omega| \sum_{x \in \Omega} T^{\ell_{\mathrm{mid}}}(z, x \mid p_{i \to j}) \mathcal{N}(E_i^*)(x), \qquad \bar{\Gamma}_j = \frac{1}{N-1} \sum_{i \neq j} \widetilde{\Gamma}_{i \to j},$$

$$E_{\mathcal{E}} = \mathcal{N}\left(\sum_j \mathcal{N}(E_j \bar{\Gamma}_j) \right).$$

The common factor $|\Omega|$ cancels in normalization. HR-Bench map and MME-RealWorld-Lite relation questions aggregate all participating pairs. For HR-Bench tables, Item labels bind the requested values to columns, and cell overlap projects local value fields onto the original visual-token grid before bidirectional composition. A six-cue fallback groups label, bar, and value maps by Item, using their normalized sum as the target field and the verified mask union as the transfer root. MME structured-data questions bind values to label anchors, while counting uses verified component sets and their enclosing crop.

Candidate geometry and observed domain. Compression partitions each Union axis at entity-box boundaries and keeps every interval intersecting an entity box. For an empty interval $[a, a + L)$, the kept indices are

$$n_\gamma = \mathrm{round}((1 - \gamma)L), \qquad \left\{ a + \left\lfloor \frac{L - n_\gamma}{2} \right\rfloor + k : k = 0, \ldots, n_\gamma - 1 \right\},$$

with the set empty when $n_\gamma = 0$. Sorting kept rows as $s_1 < \cdots < s_m$ and columns as $t_1 < \cdots < t_n$ defines

$$\phi_\gamma(s_u, t_v) = (u, v), \qquad \phi_\gamma^{-1}(u, v) = (s_u, t_v), \qquad E_\gamma(u, v) = E_{\mathrm{rel}}(s_u, t_v).$$

Index selection preserves entity pixels and axis order. For $\gamma > 0$, masking non-entity cells whose row or column intersects an entity box gives the valid grid V_γ, whereas all Union cells are valid at $\gamma = 0$. The search pairs distinct row and column layouts for $\gamma \in [0, 1]$ with distinct distances ρ from valid context tokens to the entity convex hull. LPD supplies the candidate with no context, and boxes covering Union yield the entity view directly. Ties favor more context, then enumeration order. Rendering copies kept image rows and columns, resamples the token mask by nearest neighbors, and fills excluded pixels with RGB $(127, 127, 127)$.

For fields defined on the full image, the observed domain is $D = \Omega$. For local captures with rectangles P_t,

$$D = \bigcup_t \{ z \in \Omega : \mathrm{area}(\mathrm{cell}(z) \cap P_t) > 0 \},$$

where $\mathrm{cell}(z)$ is the token footprint in the original visual-token coordinates. Observed zeros enter the background mean, while cells outside D remain unobserved. If $D \setminus R$ is empty, $\mu_D = 0$. The domain, background, evidence map, and Union denominator remain fixed during search.

C Experimental Configuration

C.1 Models and Evaluation

PostCrop and reproduced HiDe runs use deterministic greedy decoding ("do_sample=False") on one NVIDIA GPU per process.

Qwen2.5-VL. The 3B and 7B models use between 256 and 16,384 visual tokens per image and capture attention at the middle decoder layers 19 and 15, respectively, counted from one. The cap of 16,384 visual tokens also applies to Base, HiDe, and PostCrop on MME-RealWorld-Lite. Base/ViCrop and published HiDe scale results come from Table 1 of HiDe (Liu et al., 2026a). Our Qwen2.5-VL-7B HiDe reproduction uses the same HR-Bench questions and option order as PostCrop, retaining the released pipeline and decoding with Gaussian smoothing $\sigma = 3$ and threshold 0.7. Predictions are matched to the recorded answer letters.

Qwen3-VL-8B. This model uses the middle decoder layer 19 and between 256 and 16,384 visual tokens per image for all question families, including structure questions. PostCrop structure branches omit softplus. Base/ViCrop results and HiDe results on VSTAR and HR-Bench-4K follow Table 1 of HiDe (Liu et al., 2026a), while we reproduce HiDe on HR-Bench-8K with the released greedy decoding, Gaussian smoothing $\sigma = 3$, threshold 0.5, and scoring based on the first generated option character.

InternVL2.5-4B. Base/ViCrop results follow Table 2 of ActiveScope (Wang et al., 2026), and we reproduce HiDe. Both HiDe and PostCrop capture attention at the middle decoder layer 15 (counted from one) using 448×448 visual tiles. HiDe uses up to 128 visual tiles per image, Gaussian smoothing $\sigma = 1$, threshold 0.4, and thumbnails throughout (settings taken from the HiDe paper and its released source code (Liu et al., 2026a)). PostCrop uses up to 36 visual tiles per image and verifies the ten basins with the highest response, smoothing attribute maps by a 3×3 average filter while leaving relation maps unsmoothed. Local tiles serve localization and B^* construction, while answering adds thumbnails to I and B^*, with one overview tile in addition to each image's tile budget when an input uses multiple tiles. HiDe places the original image before reconstructed views and applies the released rule based on the first generated character, whereas PostCrop uses $[I, B^*]$, generates an answer phrase without option labels, and uses selective Yes/No verification. Both evaluate all VSTAR questions, and HiDe averages the four cyclic option orders on HR-Bench.

C.2 Diagnostic Controls

Diagnostics use Qwen2.5-VL-7B at the middle decoder layer 15 on 70 questions per resolution: 48 relation, 12 map, and 10 table. These question types involve multiple independently identifiable visual entities whose evidence must be considered jointly, matching our focus on selecting and composing task-relevant context across entities. The within-Union grid in Fig. 2(b) samples rank fractions $0, 0.25, 0.5, 1$ on each candidate-space axis. Reference-answer log-probabilities serve only as diagnostics and are computed by teacher forcing, averaging answer content tokens (excluding headers and end tokens) and then averaging over questions.

Composition and context selection. Table 5 distinguishes full entity maps $a = E_a$, $b = E_b$ from $a_R = \mathcal{N}(E_a \mathbf{1}_{R_a})$ and $b_R = \mathcal{N}(E_b \mathbf{1}_{R_b})$, restricted to entire verified boxes, including cells outside selected basins. Direct addition uses $\mathcal{N}(a_R + b_R)$, while one-way variants normalize $a + E_{a\to b}$ or $b + E_{b\to a}$, with $E_{a\to b} = \mathcal{N}(E_b \Gamma_{a\to b})$ and an analogous reverse. For multiple roles, direct addition groups the other entities in b_R, and one-way variants sum normalized pairwise maps. The restricted sum has support only in R, giving $\mu_D = 0$ and zero context scores, so the area tie rule selects Union. This ablation varies support and transfer, whereas the Sum control uses full entity maps.

Context-selection controls. The rows of Fig. 3 mask entity tokens and match selectors by rectangle shape and valid-context count closest to B^*'s context area. Source, Partner, and Sum maximize mean response over translations, with Partner summing other roles' normalized maps. Count ties favor larger areas and response ties later windows. For cases with no context, we use LPD and the original image.

The context-selection controls fix entity maps, boxes, and masks while each normalized selector applies Eq. 7 with its own non-entity background. Random, Source, Partner, Sum, Zoom-out, and Union use the same input order, with the original image presented first, and one auxiliary view containing the verified entities together with the selected context. The B^* reference is the condition that presents the original image first and uses the composed evidence map. Auxiliary views use the 8,192-token cap and the original image uses the 16,384-token cap.

Context sources. Context-source interventions fix B^*, pixels, geometry, prompts, options, and image order $[I, B^*]$ while independently blocking context in the crop and the original image. A second control blocks only the copy of the original image containing selected C^*, mapped from crop source coordinates. When C^* is empty, this control reuses conditions with the original image intact. Masks block merged visual-token groups in every ViT and decoder self-attention layer while preserving positions. Readout uses answer phrases and selective Yes/No verification in the original option order.

C.3 Supplementary Controls and Timing

The layout control compares moderate and complete gap compression with identical entity pixels and no contextual content on 18 eligible 4K and 22 eligible 8K questions. Comparing native views with presentations on a shared gray canvas at $1\times$ and $2\times$ nearest-neighbor scale controls the presentation, with token grids and entity areas matched within scale.

We assess localization and search sensitivity separately: kernel size and the number of verified basins vary on fixed cue fields, whereas paired searches at the Union boundary and at $1.1\times$ that boundary fix boxes, evidence, background, denominator, and prompts. Wider limits reach $1.5\times$ the Union boundary, retaining the first enumerated candidate on ties. Under perturbed localization, one box shifts by 25% of its width or height, or each side shrinks to 75%, after which views containing only entities and Union views share the canvas, original image, and readout.

Full inference timing. For Table 7, both methods process all VSTAR questions on the same GPU, with fresh cue parsing for PostCrop and generated decomposition for HiDe. Timing covers planning, model calls, image processing, and synchronization, excluding model loading and dataset export, while peak allocated CUDA memory includes weights. HiDe's timing run supplies its main table VSTAR row, while PostCrop uses separate timing and main table runs.

D Supplementary Experiments

Layout and scale with fixed entity content. This control asks whether accuracy changes arise from presenting identical entity evidence differently rather than from differences in the selected visual content. Table 8 compares two layouts containing the same entity pixels under native presentation and presentation on a common canvas (Appendix C).

Table 8: **Layout and scale sensitivity with identical entity content.** Accuracy (%) for moderate and complete gap compression on matched entity pixels.

Presentation	4K		8K	
	Moderate	Complete	Moderate	Complete
Native	61.1	55.6	77.3	77.3
Common canvas, $1\times$	66.7	61.1	54.5	54.5
Common canvas, $2\times$	66.7	66.7	59.1	59.1

Moderate compression has no consistent accuracy advantage over complete compression (Table 8). Both layouts decline on 8K when placed on a shared canvas, despite keeping identical source pixels. The shared canvas changes the displayed scale and the spatial presentation of visual tokens without changing the evidence itself, so the decline points to sensitivity in visual presentation and readout rather than to a content difference. The control uses entity-only views and therefore does not rank contextual candidates.

Localization and search-boundary sensitivity. We use this ablation to determine how localization granularity and the candidate search boundary affect entity coverage, context area, and answering accuracy. Table 9 varies localization settings and the search boundary with the other answering conditions fixed (Appendix C).

Table 9: **Localization and search-boundary sensitivity.** Adaptive B^* selection uses the indicated search limits. Areas are mean percentages of the original image, accuracy is in percent, and shading marks the default localization setting.

Kernel and basin count	4K				8K			
	Area		B^* accuracy		Area		B^* accuracy	
	Entity	Union	Union limit	1.1× limit	Entity	Union	Union limit	1.1× limit
1×1, 10	0.386	12.842	74.29	75.71	0.295	6.939	64.29	64.29
3×3, 5	1.094	14.406	80.00	77.14	0.880	9.346	72.86	72.86
3×3, 10	1.101	14.522	80.00	78.57	0.857	9.746	71.43	71.43
3×3, 20	1.102	15.230	78.57	77.14	0.837	9.683	71.43	71.43
5×5, 10	1.752	17.777	78.57	78.57	1.372	10.513	74.29	72.86

Larger kernels produce larger average boxes without a consistent accuracy gain (Table 9). The modest gain for the smallest boxes on 4K indicates that a narrow localization can omit cues near an entity boundary. Once the box is large enough, further expansion adds context without improving the evidence used for answering, so box area alone does not predict accuracy.

Wider search ranges. The experiment with a wider search range tests whether enlarging the candidate space beyond Union produces an answering gain when the evidence and answering procedure remain fixed. The nested search limits use fixed evidence and answering inputs (Appendix C).

Table 10: **Adaptive search beyond Union with automatic boxes.** Accuracy (%) at each search limit.

Search limit	4K	8K
Union	77.14	72.86
1.1×	77.14	74.29
1.3×	75.71	68.57
1.5×	75.71	71.43

A larger candidate pool raises the maximum context score but does not consistently improve accuracy (Table 10). Because the score rewards coverage, extra area can increase the score even when the added regions do not help answer the question. The selector often moves beyond Union under expansion, so the weak accuracy trend cannot be attributed to an unchanged crop. The result separates a score gain from an answering gain and shows why the coverage term should be interpreted together with the evidence content (Appendix B).

Context under perturbed localization. To probe robustness to localization errors, we measure whether selected context compensates for those errors and whether the compensation depends on the perturbation direction. Under the perturbation protocol in Appendix C, keeping Union context helps some perturbed localizations and harms others (Fig. 9). Union helps when the displaced box misses a nearby cue that remains inside the surrounding region, but it can hurt when the same expansion admits a distractor. The effect therefore depends on both the error direction and the evidence around the displaced box.

Results across Qwen2.5-VL scales. Across model sizes, we examine whether the relation-focused behavior of PostCrop persists rather than being specific to one checkpoint. With both 3B and 7B models, PostCrop performs better on FCP than HiDe, while HiDe keeps stronger FSP results (Table 11). FCP requires combining evidence across entities, which matches PostCrop's composed map and context selection. FSP is more localized, so the compact entity views retained by HiDe can be sufficient and avoid additional context. The same split across model sizes supports a relation-focused advantage rather than a gain tied to one checkpoint.

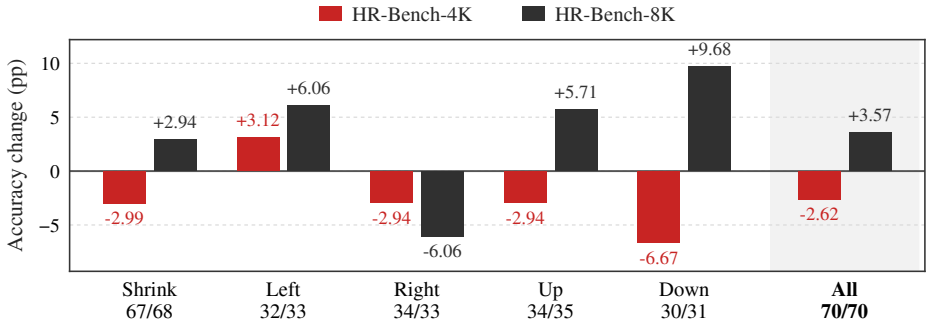


Figure 9: **Context effects vary with the localization perturbation.** Paired accuracy change (percentage points) after keeping Union context, with positive values favoring Union. Counts below the bars give eligible questions for 4K/8K. Averages first pool perturbations within each question and then across questions.

Table 11: **Accuracy across Qwen2.5-VL scales.** HiDe results follow its published protocol (Liu et al., 2026a).

Scale	Method	VSTAR			HR-Bench-4K			HR-Bench-8K		
		Attr.	Spatial	Overall	FSP	FCP	Overall	FSP	FCP	Overall
3B	HiDe	85.2	72.4	80.1	87.8	51.5	69.6	88.3	51.3	69.8
	PostCrop	87.8	80.3	84.8	85.0	62.0	73.5	84.0	61.0	72.5
7B	HiDe	94.8	88.2	92.1	95.5	59.5	77.5	93.5	57.3	75.4
	PostCrop	93.9	94.7	94.2	93.0	74.0	83.5	93.0	76.0	84.5

Supplementary localization ablation. The final localization ablation disentangles the contributions of candidate generation, verification, and proposal diversity to localization quality. Target coverage is $|M \cap G|/|G|$ for candidate mask M and tokens G overlapping the annotated target box, averaged over 267 VSTAR entity cues. With the same cues and layer, watershed provides higher candidate coverage than released HiDe TAD (88.7% versus 84.9%). Verification improves target coverage over choosing the first ranked basin (82.5% versus 73.0%) because the strongest response is not always the basin that covers the target most completely. A connected-component control supplies fewer candidate regions and performs worse on VSTAR, consistent with the need for diverse proposals before verification can select a useful crop.

E Scope and Applicability

PostCrop represents task-relevant objects, map landmarks, and table or chart labels and values as entities. Headers, legends, and scales are included when needed to interpret those entities. For counting questions, a cue may refer to multiple instances, whose verified regions jointly supply the evidence for the answer.

Context selection depends on the visual evidence covered by these cues and their localized regions. Questions involving metric distances additionally require recoverable scale and geometry in the supplied views, because gap compression changes the displayed distances. These requirements distinguish retaining the relevant entities from preserving the spatial measurements needed to answer a particular question.

G PROMPT TEMPLATES

VSTAR and HR-Bench use fixed question-specific cues. Object questions specify the object identity and requested attribute, while relation questions specify the source, partner, and relation operator. Compound nouns remain intact when separating object and attribute cues, which fill the templates below.

Cue-map capture and identity verification. We present the original image with the following prompt to capture the response to each cue:

> Is the following visual cue present in the image? Cue: `{cue}`. Answer Yes or No only.

After generating candidate regions, we verify each entity using only its candidate crop and the following prompt, abbreviated as "Is it `{cue}`?" in the main text:

> Does this local crop contain the visual entity described as '`{cue}`'? Answer Yes or No only.
> Answer:

Bidirectional transfer. For each direction, we supply two copies of the original image with the relation instruction between them and the partner cue after the second copy:

> `[Image 1]`
> Image 1 is the reference scene. Apply this relation operator before reading Image 2: `{operator}`.
> `[Image 2]`
> Image 2 is unchanged. Locate the partner evidence: `{partner}`. Do not answer the question and do not return a box.

Exchanging the source and partner roles gives the reverse-direction input.

Answer generation and option verification. Final readout supplies the original image I followed by B^* and asks for an answer phrase using the following template:

> Question type: `{type}`.
> Question: `{q}`
> Question-specific visual cues: `{visual cues}`.
> Task-specific reasoning: `{instruction}`
> Answer the question from the visible evidence. Give only the concise answer phrase.
> Answer:

The task-specific reasoning line is included only when an instruction is provided. Greedy generation produces at most 16 new tokens, which are matched to option content through normalized text, counts, map pairs, or spatial-relation terms. When the match is ambiguous, we retain the question, cues, and any reasoning instruction, replacing the answer request with:

> Proposed answer: `{o_i}`
> Is the proposed answer correct based on the visible evidence? Answer Yes or No only.
> Answer:

Verification considers options compatible with a partially resolved compound relation, or all options otherwise, and selects the largest next-token margin $s_i = \log P(\texttt{Yes} \mid o_i) - \log P(\texttt{No} \mid o_i)$, with ties resolved by option order. Each query uses option content without letters and the same image pair, keeping localization, composition, and crop selection fixed throughout answering.