RE-ATTENDING TO THE MISSING PIECES: TEST-TIME VISUAL EVIDENCE REFINEMENT FOR HIGH-RESOLUTION REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) can process increasingly high-resolution images, yet often miss small, question-relevant evidence in cluttered scenes. In our controlled study across three MLLMs, accuracy generally improves as the surrounding background is removed, while the queried object remains visible at every ratio. This exposes a visual evidence allocation bottleneck: more pixels do not ensure that finite encoding and attention budgets are directed to relevant evidence. Existing training-free methods may select or refine visual regions, but typically do not jointly diagnose missing evidence and verify the resulting crops. We introduce RECAP (**Re**-attending and **C**omposing **A**ttention-guided **P**roposals), a training-free closed-loop framework for adaptive test-time visual evidence refinement. RECAP identifies localizable entities and derives source-resolution proposals from their attention maps. Tentative answers direct refinement toward unresolved evidence, while comparative verification ranks proposals by target visibility. Selected regions form compact, order-aware canvases. Confidence-guided routing returns the answer once the model is sufficiently confident. Across V*Bench, HRBench-4K, and HRBench-8K, RECAP improves all five evaluated backbones. It reaches 95.8% with Qwen2.5-VL-7B and 92.7% with InternVL3-8B on V*Bench, achieving the best results among the compared methods for both backbones.

1 INTRODUCTION

Multimodal large language models (MLLMs) provide strong general-purpose visual reasoning (Alayrac et al., 2022; Dai et al., 2023; Liu et al., 2023). Their performance nevertheless remains brittle when an answer depends on a small object embedded in a high-resolution scene (Tong et al., 2024; Wu & Xie, 2024; Wang et al., 2025). Increasing input resolution appears to be a natural remedy because it preserves more source detail. In practice, however, larger inputs also introduce more irrelevant pixels, visual tokens, and competing objects. Under finite encoding and attention budgets, target visibility need not translate into effective evidence use.

Figure 1 isolates the difference between missing pixels and failing to use visible evidence. For each V*Bench question, we construct a sequence of views that always retains the annotated target while progressively reducing its surrounding background, and ask each backbone to answer from every view. Despite local fluctuations, all three MLLMs improve overall as the view becomes tighter. Because the queried object remains visible throughout, this trend shows that target visibility alone is insufficient and is consistent with competition from non-target context for limited visual processing.

We refer to this failure mode as the visual evidence allocation bottleneck. As the scene grows, a small target occupies a diminishing share of the encoded sequence; global resizing or token thinning can further weaken its detail, while background objects compete for attention. The required pixels may therefore be present, yet their contribution is diluted by a longer visual sequence and competing scene content.

This leads to the central question: how can inference concentrate finite visual processing on compact, question-relevant evidence while retaining the context needed for attribute and relational reasoning?

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

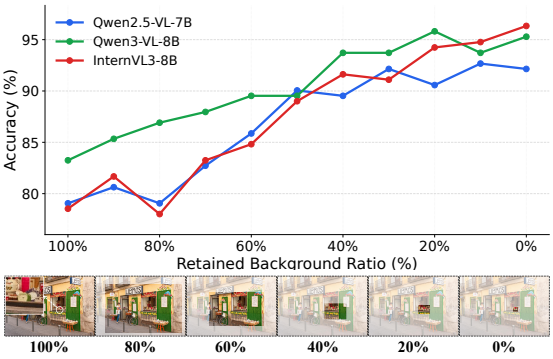

Figure 1: Target-preserving background study on V*Bench. Top: accuracy generally improves as background is removed. Bottom: corresponding views, with the queried object visible throughout.

Existing solutions address only parts of this problem. Dynamic-resolution architectures and tiling strategies (Bai et al., 2025b; Liu et al., 2024; Guo et al., 2024; Zhu et al., 2025) preserve more pixels but distribute computation broadly, without determining which evidence the current question still lacks. Training-based active-perception methods (Shao et al., 2024; Zhao et al., 2025; Zheng et al., 2026; Lee et al., 2026) learn where to inspect, but require additional supervision or policy optimization; task-specific fine-tuning can also degrade previously acquired capabilities (Zhai et al., 2024). Training-free methods (Wang et al., 2025; Liu et al., 2025; Shen et al., 2025; Li et al., 2025b; Zhang et al., 2025; Zhong et al., 2025) exploit heuristic or model-internal cues, such as sliding windows, search trees, attention maps, and internal cache signals, to identify question-relevant regions; however, they typically do not jointly diagnose an unresolved entity from a tentative answer and comparatively rank target-specific proposals with an explicit entity-existence score. Moreover, independent crops may obscure cross-region ordering cues in multi-object reasoning. An attribute question may require a clearer view of one object, whereas a relational question may require several objects to be represented together. The evidence needed can also change as reasoning proceeds: once one entity becomes clear, the remaining uncertainty may concern another. Effective refinement therefore needs to track unresolved entities and use their status to guide subsequent inspection. RECAP addresses these different evidence needs through entity-specific refinement followed by a shared composition step.

To address the visual evidence allocation bottleneck, we propose RECAP, a training-free framework that refines visual evidence at test time by re-attending to missing information and composing attention-guided proposals. As illustrated in Figure 3, RECAP turns evidence allocation into a closed loop between answer generation and visual evidence acquisition. Entity Proposal extracts localizable entities, caches their attention maps, and constructs candidate regions. A tentative answer then identifies an unresolved entity, allowing Target Re-attention to revisit the corresponding cached map. Verified Assembly uses comparative verification to select target-specific evidence and consolidates the accumulated regions into compact, order-aware canvases. Confidence Routing terminates the loop once the current answer is sufficiently confident. The entire procedure operates at test time without updating model parameters.

RECAP uses an entity-level feedback loop to turn tentative answers into targeted visual evidence updates. It couples answer-conditioned entity diagnosis with cached-map re-attention and comparative proposal verification, then composes the selected regions into an order-aware canvas. This coupling directs each refinement toward answer-critical evidence while retaining cross-region ordering cues. Entity diagnosis specifies what evidence to seek, while proposal verification tests whether a candidate makes that entity recognizable. Their interaction ties region selection to both the question and the currently available evidence.

Our contributions are summarized as follows:

- We expose and empirically characterize a visual evidence allocation bottleneck in high-resolution VQA. Our controlled target-preserving study across three MLLMs shows that preserving the queried object is insufficient when surrounding content dilutes its visible evidence.
- We formulate high-resolution reasoning as a bounded, answer-conditioned evidence-refinement loop and instantiate it in RECAP, a training-free framework. RECAP connects tentative-answer diagnosis with cached entity-conditioned re-attention, comparative proposal verification, and order-aware evidence composition without updating model parameters.
- Across V*Bench and HRBench-4K/8K, RECAP improves all five evaluated backbones. On V*Bench, it reaches 95.8% with Qwen2.5-VL-7B and 92.7% with InternVL3-8B, the highest results among the compared methods for these two backbones.

2 Related Work

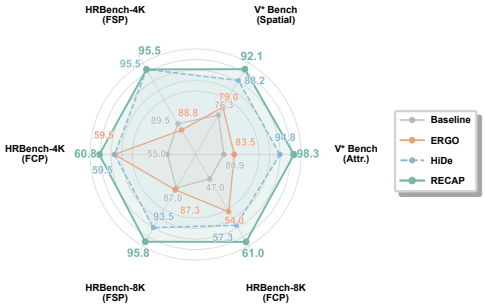

Figure 2: Visual summary of Qwen2.5-VL-7B performance across V*Bench and HRBench. Detailed results are reported in Table 1.

High-Resolution Multimodal Models. Early MLLMs commonly encode a fixed-size or globally resized view, which can suppress small objects and fine text (Alayrac et al., 2022; Liu et al., 2023; Dai et al., 2023; Zhu et al., 2024; Bai et al., 2023). Recent systems preserve more source detail through dynamic tiling or native-resolution encoding. LLaVA-NeXT and LLaVA-UHD use adaptive image partitions, InternVL3 dynamically tiles the input, and Qwen2.5-VL and LLaVA-OneVision support variable-resolution visual inputs (Liu et al., 2024; Guo et al., 2024; Zhu et al., 2025; Bai et al., 2025b; Li et al., 2025a). These designs reduce information loss from global resizing, but also lengthen the visual sequence. Lightweight models and token-reduction methods, including MiniCPM-V, PyramidDrop, HiRED, and general visual-token pruning, improve efficiency by retaining compact visual representations (Yao et al., 2024; Xing et al., 2025; Arif et al., 2025; Endo et al., 2025). Contrastive vision–language pretraining strengthens cross-modal alignment (Radford et al., 2021; Zhai et al., 2023), while ViT representations can contain register-like global tokens and visual attention sinks whose dominant weights need not reflect semantic relevance (Darcet et al., 2024; Kang et al., 2025). These observations connect high-resolution perception with both detail preservation and visual-token allocation. RECAP refines the question-relevant visual evidence supplied to these models during inference, with their parameters held fixed.

Test-Time Visual Evidence Selection. High-resolution VQA methods differ in how they acquire local evidence. V* performs guided visual search, while DeepEyes and ERGO learn zooming or region-selection behavior through additional optimization (Wu & Xie, 2024; Zheng et al., 2026; Lee et al., 2026). Training-free methods instead divide the image or exploit model-internal signals: DC^2 recursively processes image patches, ZoomEye and DyFo explore candidate crops, and ViCrop, FOCUS, and HiDe localize fine details from attention or internal representations (Wang et al., 2025; Shen et al., 2025; Li et al., 2025b; Zhang et al., 2025; Zhong et al., 2025; Liu et al., 2025). Together, these approaches establish compact local views as a means of concentrating computation on question-relevant content. Their selection mechanisms range from explicit spatial search to model-derived saliency, providing different ways to prioritize regions within a large image. RECAP guides visual evidence acquisition by diagnosing which question entity remains unclear in the current evidence state. At each refinement round, the diagnosed entity determines the cached attention map to revisit and the target to verify in candidate crops. The selected crop updates the evidence canvas for subsequent answering and diagnosis, forming an entity-level feedback loop between unresolved evidence and targeted visual inspection.

Answer-Conditioned Refinement. Re-Attention uses a predicted answer to reconstruct visual attention in VQA (Guo et al., 2020). Recent MLLM approaches revisit encoded evidence through token-coreset reuse or learned attention control (Ghosal et al., 2026; Ebouky et al., 2026). SIEVE retrieves and injects salient region embeddings into the reasoning chain, while TTSP combines entropy-gated exploration with evidence-guided refinement (Shi et al., 2026; Jiang et al., 2026). EGG uses next-token entropy gradients for multi-region cropping and spatial entropy for stopping (Gröpl et al., 2026). UniT and related studies further investigate multimodal test-time reasoning and its task-dependent gains (Chen et al., 2026; Ahmadpour et al., 2025). These lines of work motivate coupling intermediate reasoning with renewed visual inspection. RECAP uses answer-conditioned refinement to identify which question entity remains unclear and guide the next visual inspection. The resulting evidence updates a shared spatial representation, enabling subsequent reasoning to incorporate newly resolved details while retaining the relative order of relevant regions.

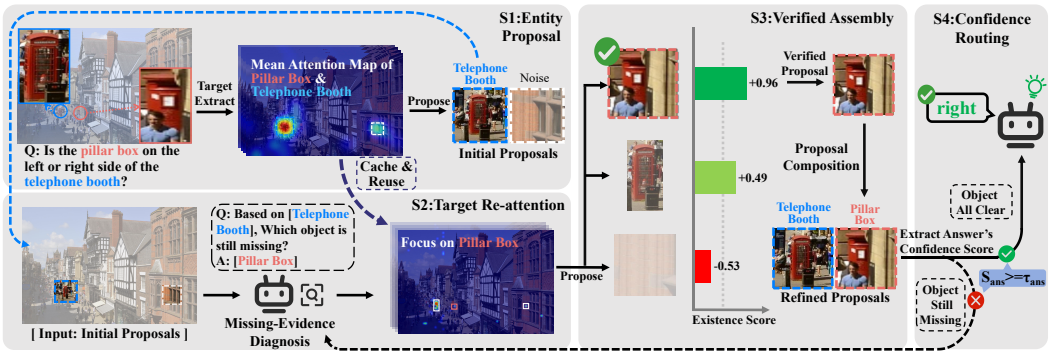

Figure 3: Overview of RECAP. Entity Proposal derives candidate regions and cached attention maps; Target Re-attention uses a tentative answer to revisit unresolved evidence; Verified Assembly verifies and composes selected regions; and Confidence Routing controls termination of the bounded refinement loop.

3 METHODOLOGY

As illustrated in Figure 3, RECAP constructs a bounded sequence of visual evidence states for a high-resolution image I and question Q. **Entity Proposal** constructs the initial evidence state $\mathcal{Z}^{(0)}$ from attention-derived proposals and caches the corresponding entity-conditioned maps. The frozen MLLM evaluates $\mathcal{Z}^{(0)}$, producing a tentative answer $a^{(0)}$ and an answer confidence $S^{(0)}_{\text{ans}}$. At refinement round $t \geq 1$, **Target Re-attention** diagnoses one unresolved entity from the preceding answer $a^{(t-1)}$ and generates target-specific proposals from its cached map. **Verified Assembly** ranks these proposals and incorporates the selected evidence into $\mathcal{Z}^{(t)}$, which the model evaluates to obtain $a^{(t)}$ and $S^{(t)}_{\text{ans}}$. Finally, **Confidence Routing** terminates the loop once the current answer reaches the confidence threshold, with at most $T_{\max}$ refinement rounds. RECAP runs at test time without updating the backbone.

This design separates two objectives that are difficult to satisfy with a single crop operation. Initialization gathers evidence across multiple candidate entities, whereas refinement concentrates each round on an entity whose visual evidence remains unclear. Every state supplements the visual input with compact canvases that increase the relative visual budget assigned to small targets. Localization attention is extracted once and cached; later rounds reuse these maps while performing only the answer, diagnosis, verification, and composition steps needed to update the evidence.

Entity Proposal. RECAP decomposes the question into a set of candidate object phrases:

$$\mathcal{E} = \{e_1, e_2, \ldots, e_n\} = f^{\text{dec}}_\theta(Q, C), \tag{1}$$

where C is a fixed instruction that requests physical objects that can serve as localization anchors. The same frozen MLLM performs this decomposition. The extracted phrases define the candidate targets available to later diagnosis.

To localize these phrases, RECAP aggregates attention from each retained post-image text token to the processed visual tokens across all decoder layers and heads (Abnar & Zuidema, 2020; Neo et al., 2025). The heads are first averaged within each layer, and the resulting layer maps are averaged before projection to the backbone's two-dimensional visual grid, yielding $\bar{A}_q$. Let $\mathcal{D}$ denote a fixed attention-denoising operator comprising spatial smoothing, prompt-noise subtraction, and normalization. The index q denotes a post-image text-token position, and $u = (r, c)$ denotes a spatial cell on the processed visual-token grid. For Qwen, $\mathcal{T}(e_k)$ contains positions whose lowercased, whitespace-trimmed decoded token fragments have more than one character and occur as substrings of the normalized entity phrase e_k. The token and entity maps are:

$$\begin{aligned} \widetilde{M}_q &= \mathcal{D}(\bar{A}_q), \\ M_k(u) &= \max_{q \in \mathcal{T}(e_k)} \widetilde{M}_q(u). \end{aligned} \tag{2}$$

The prompt-noise prior is averaged over the first eight post-image prompt-token positions, with adapter-specific smoothing detailed in the supplement. Matched token maps are combined by pixel-wise maximum pooling to form the cached entity map. Together, aggregation and prompt-noise subtraction reduce dependence on a single decoder view and suppress prompt-wide visual sinks. Each successfully matched entity map is cached in $\mathcal{M} = \{e_k \mapsto M_k\}$.

Initialization prioritizes broad coverage. RECAP extracts connected high-response regions from each retained denoised token map, converts them into boxes, and removes redundant boxes with non-maximum suppression. The remaining boxes are projected to source-image coordinates and form the initial component list $\mathcal{C}_{\text{init}}$ without entity-existence scoring. At this point no tentative answer has identified a missing entity, so applying target-specific verification would prematurely narrow the initial evidence. Instead, these components are composed directly to obtain the first compact canvas; target-specific ranking begins only during refinement. Entity Proposal thus outputs both the initial evidence and the cached map bank used by later rounds.

Target Re-attention. Because the initial state covers multiple entities, a small, occluded, or ambiguous entity may still be represented too weakly to support the answer. At refinement round $t \geq 1$, given $a^{(t-1)}$ from $\mathcal{Z}^{(t-1)}$, RECAP diagnoses which entity in the remaining candidate set is still unclear:

$$e^* = f_\theta^{\text{diag}}(Q, a^{(t-1)}, \mathcal{E}_{\text{remain}}^{(t-1)}). \tag{3}$$

The list begins with $\mathcal{E}_{\text{remain}}^{(0)} = \mathcal{E}$. A text-only diagnostic prompt supplies the question, tentative answer, and remaining entity phrases and requests the phrase corresponding to an entity that remains unclear. The returned identifier is recorded as attempted before map retrieval; subsequent rounds exclude entity phrases with the same normalized identifier.

RECAP retrieves the cached attention map for the diagnosed entity e^* through normalized entity-name matching. Cached entity names e_k and the diagnostic output are converted to lowercase and trimmed of surrounding whitespace; angle brackets, quotation marks, and periods are additionally removed from the diagnostic output. Retrieval prioritizes exact equality between the normalized strings. If no exact match exists, bidirectional substring matching selects the first cached entity for which either normalized string is a substring of the other, with candidates ordered by their cache position. If neither matching criterion is satisfied, the current evidence state is retained.

Given the retrieved map, RECAP computes $X_{e^*} = \text{Norm}(M_{e^*}^2)$, where Norm divides by the map maximum when it is positive. High-response anchors seed boxes that expand along each axis while the adjacent slice's mean response is at least a fixed fraction of the anchor response. Non-maximum suppression removes redundant boxes, and the survivors are projected to source-image coordinates to form $\mathcal{P}_{e^*}^{(t)}$. The supplement specifies the anchor, expansion, and suppression settings.

Target Re-attention uses the diagnosed entity's cached attention map to guide region refinement.

Verified Assembly. Attention-derived proposals provide broad candidate coverage but may still contain distractors. RECAP therefore assigns each candidate $b \in \mathcal{P}_{e^*}^{(t)}$ a visibility score $S_{\text{exist}}(e^*, b)$ using the frozen MLLM and selects the highest-scoring candidate:

$$b^* = \arg\max_{b \in \mathcal{P}_{e^*}^{(t)}} S_{\text{exist}}(e^*, b). \tag{4}$$

The frozen model assesses each candidate crop using the prompt "Is the [entity] clearly visible in this image? Answer Yes or No." Let ℓ_{yes} and ℓ_{no} denote the maximum next-token logits over the affirmative and negative token sets, respectively. The visibility score is:

$$S_{\text{exist}}(e^*, b) = 2\frac{\exp(\ell_{\text{yes}})}{\exp(\ell_{\text{yes}}) + \exp(\ell_{\text{no}})} - 1. \tag{5}$$

The score lies in $[-1, 1]$: a larger value indicates a stronger Yes preference relative to No. Candidates for the same entity are ranked by this score, and the Qwen adapter appends the highest-scoring usable crop. An empty candidate set leaves the evidence unchanged. Initial proposals enter the first

canvas directly; scoring is applied to the target-specific refinement proposals. The supplement gives response-token handling and adapter-specific acceptance rules.

Fold-down composition assembles selected crops on a shared canvas while preserving their source sizes and axis-aligned ordering. Along each axis, crop boundaries divide the image into intervals. Intervals covered by at least one crop retain their original lengths. Each uncovered gap of length d is compressed to $\hat{d} = \min(d, \delta)$, where δ controls the maximum retained gap. Cumulative interval lengths determine the new crop positions through a shared coordinate mapping. Each crop is placed at its mapped upper-left coordinate, with its width and height unchanged.

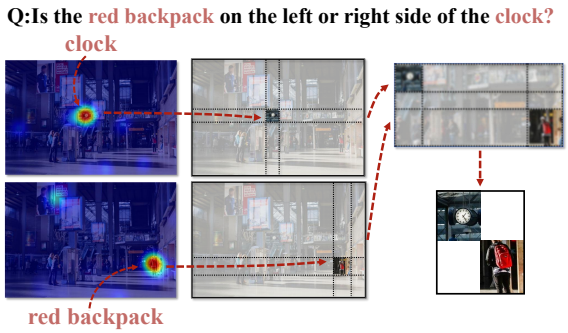

Figure 4: Fold-down composition preserves the clock and backpack crops' source sizes and left–right order before model preprocessing.

Applying the mapping along both axes preserves crop sizes and relative ordering, giving the model a shared spatial frame for cross-region reasoning. The supplement details the coordinate construction, gap setting, numerical example, and overlap handling.

Confidence Routing. Let $p^{(t)}(v)$ denote the model's next-token probability at the first answer-generation step, with $\mathcal{V}$ denoting the vocabulary. The routing score is:

$$S_{\text{ans}}^{(t)} = \max_{v \in \mathcal{V}} p^{(t)}(v). \tag{6}$$

This first-token score controls stopping: refinement ends when $S_{\text{ans}}^{(t)} \geq \tau_{\text{ans}}$, no candidate entities remain, or the iteration cap is reached. Each evidence update yields a new answer and score; continued refinement diagnoses an unclear entity from the remaining list.

4 Experiments

Benchmarks and backbones. We evaluate Qwen2.5-VL-3B/7B/32B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), and InternVL3-8B (Zhu et al., 2025) on V*Bench (Wu & Xie, 2024) and HRBench (Wang et al., 2025). V*Bench measures fine-grained attribute recognition and spatial reasoning. HRBench provides paired views of the same questions: the 8K version retains the full image, whereas the 4K version crops the question-relevant region. Both report fine-grained single-instance perception (FSP) and cross-instance perception (FCP). Both benchmarks use multiple-choice questions; local runs default to greedy option-letter decoding and normalize responses to valid options. We evaluate all samples and report split accuracies.

Protocol. We use $\sigma = 3$, $\tau_{\text{ans}} = 0.95$, and $T_{\max} = 4$, selected on V*Bench and frozen for HRBench. Default RECAP and all sensitivity experiments aggregate attention across all decoder layers and heads; restricted variants are used only in the attention ablation. Local comparisons share samples, source images, and image-size policy and run on RTX A6000 GPUs. The ERGO reproduction follows its released checkpoint and two-stage protocol under the shared image cap. Complete prompts, adapter settings, hardware details, and reproduction protocols are provided in the supplement.

Quantitative results. Table 1 shows that RECAP improves average accuracy over the original-image baseline across all five evaluated backbones on V*Bench and HRBench-4K/8K. On V*Bench, RECAP achieves 95.8% with Qwen2.5-VL-7B and 92.7% with InternVL3-8B. With Qwen2.5-VL-32B, it achieves 92.7%, 77.3%, and 77.1% on V*Bench, HRBench-4K, and HRBench-8K, respectively, exceeding the reported original-image baselines by 4.8, 3.4, and 6.7 percentage points.

Table 1: Results on V*Bench and HRBench, reported as percentage accuracy under each method's evaluation settings. T-Free marks training-free methods. V*Bench covers attribute recognition (Attr) and spatial reasoning (Spatial), while HRBench covers fine-grained single-instance perception (FSP) and cross-instance perception (FCP). † identifies a locally reproduced prior result; unmarked prior results are source-reported. Bold indicates the highest listed value in each block.

Method	Base Model	T-Free	V*Bench			HRBench-4K			HRBench-8K		
			Attr	Spatial	Avg	FSP	FCP	Avg	FSP	FCP	Avg
–	GPT-4o	–	–	–	66.0	70.0	48.0	59.0	62.0	49.0	55.5
–	InternVL3 8B	–	83.5	71.1	78.5	81.8	61.0	71.4	68.0	58.0	63.0
ViCrop	InternVL3 8B	√	88.7	75.0	83.3	88.0	57.0	72.5	82.8	54.8	68.8
HiDe	InternVL3 8B	√	92.2	90.8	91.6	90.0	63.5	76.8	92.2	**62.0**	77.1
RECAP	InternVL3 8B	√	**93.0**	**92.1**	**92.7**	**92.8**	**65.5**	**79.1**	**94.3**	60.0	**77.1**
–	Qwen2.5-VL 3B	–	81.7	61.8	73.8	81.8	50.0	65.9	79.5	45.2	62.4
ViCrop	Qwen2.5-VL 3B	√	81.7	65.8	75.4	86.3	49.8	68.0	80.3	45.5	62.9
HiDe	Qwen2.5-VL 3B	√	85.2	72.4	80.1	87.8	51.5	69.6	88.3	51.3	69.8
RECAP	Qwen2.5-VL 3B	√	**88.7**	**73.7**	**82.7**	**90.5**	**54.0**	**72.3**	**91.3**	**53.0**	**72.1**
–	Qwen2.5-VL 7B	–	80.9	76.3	79.1	89.5	55.0	72.2	87.0	47.0	67.0
ViCrop	Qwen2.5-VL 7B	√	89.6	71.1	82.2	90.5	57.5	74.0	85.5	53.0	69.3
DeepEyes	Qwen2.5-VL 7B	–	91.3	88.2	90.1	91.3	59.0	75.1	86.8	58.5	72.6
ERGO†	Qwen2.5-VL 7B	–	83.5	79.0	81.7	88.8	59.5	74.1	87.3	54.3	70.8
HiDe	Qwen2.5-VL 7B	√	94.8	88.2	92.1	**95.5**	59.5	77.5	93.5	57.3	75.4
EGG	Qwen2.5-VL 7B	√	–	–	86.9	–	–	–	–	–	–
RECAP	Qwen2.5-VL 7B	√	**98.3**	**92.1**	**95.8**	**95.5**	**60.8**	**78.1**	**95.8**	**61.0**	**78.4**
–	Qwen2.5-VL 32B	–	87.8	88.1	87.9	89.8	58.0	73.9	84.5	56.3	70.4
ViCrop	Qwen2.5-VL 32B	√	90.4	85.5	88.5	90.5	60.0	74.6	88.8	54.3	71.5
HiDe	Qwen2.5-VL 32B	√	91.3	89.5	90.6	92.0	60.0	76.2	90.5	57.5	74.0
RECAP	Qwen2.5-VL 32B	√	**93.9**	**90.8**	**92.7**	**92.8**	**61.8**	**77.3**	**92.0**	**62.3**	**77.1**
–	Qwen3-VL 8B	–	86.1	78.9	83.2	91.5	62.3	76.9	86.8	56.0	71.4
ViCrop	Qwen3-VL 8B	√	89.6	75.0	83.8	91.8	64.3	78.0	88.0	63.3	75.6
SIEVE	Qwen3-VL 8B	–	88.7	**86.8**	88.0	90.0	**73.0**	**81.5**	83.0	**73.5**	78.3
HiDe	Qwen3-VL 8B	√	**92.9**	85.5	89.5	93.5	69.0	81.2	**96.5**	71.2	**83.9**
RECAP	Qwen3-VL 8B	√	91.3	**86.8**	89.5	**94.3**	68.0	81.1	92.5	65.8	79.1

Table 2: Qwen2.5-VL-7B ablations on V*Bench. Strategy variants use the default all-layer/all-head configuration; Δ is measured from full RECAP.

(a) Inference strategy **Settings**	**Attr.**	**Spatial**	**Avg.**	**Δ**
Full RECAP	**98.3**	**92.1**	**95.8**	–
Single-pass diagnose	93.0	84.2	89.5	−6.3
Propose w/o diagnose	90.4	79.0	85.9	−9.9
Baseline	80.9	76.3	79.1	−16.7
(b) Attention configurations **Variant**	**Attr.**	**Spatial**	**Avg.**	**Δ**
All layers, all heads	**98.3**	**92.1**	**95.8**	–
Last layer, all heads	89.6	86.8	88.5	−7.3
Random-head subset	94.8	**92.1**	93.7	−2.1

Component ablation. Table 2 uses the same backbone, entity decomposition, attention maps, and canvas construction. Propose w/o diagnose answers using only the initial attention-derived crops, without entity diagnosis or subsequent verification and composition updates. Single-pass diagnose adds one diagnosis, verification, and composition update; Full RECAP repeats this process under confidence routing for up to $T_{\max}$ rounds. Accuracy rises from 79.1% to 85.9%, 89.5%, and 95.8%, respectively, showing successive gains from acquiring and integrating additional evidence. Full

RECAP improves over the proposal-only state by 9.9 points. Last-layer attention and random-head subsets score 88.5% and 93.7%, respectively, favoring localization evidence aggregated across all layers and heads.

Composition ablation. Using Qwen2.5-VL-7B, we hold the selected crop sets fixed and compare independent crop images, a grid sorted by vertical then horizontal crop centers, a shuffled grid, and fold-down composition (Figure 4). Each condition retains the original image, omits crop labels, and generates one answer under a total cap of 16,384 visual tokens.

Table 3: Composition ablation: accuracy (%) on the full V*Bench benchmark and the FCP subsets of HRBench-4K and HRBench-8K. Shuffled-grid results are averaged over three seeds.

Representation	V*Bench	HRBench-4K	HRBench-8K
Independent crops	88.0	55.5	50.3
Ordered grid	82.7	53.3	55.0
Shuffled grid	83.6	55.7	53.7
Fold-down composition	**89.0**	**60.8**	**64.3**

Table 3 shows that fold-down composition improves FCP accuracy over independent crops by 5.3 points at 4K and 14.0 points at 8K, and over the ordered grid by 7.5 and 9.3 points. With the same selected evidence, the compact canvas improves multi-instance reasoning, with the largest gain over independent crops at 8K.

Efficiency. Table 4(a) reports mean end-to-end latency under a shared image-size cap. RECAP takes 10.8 s per sample, versus 1.8 s for the base model and 6.6 s for HiDe. It improves over HiDe by 5.2 accuracy points at an additional 4.2 s. Compared with ERGO, RECAP gains 14.1 accuracy points with an additional 0.3 s per sample. Cached entity attention is reused across rounds; subsequent costs arise from entity diagnosis, crop verification, and answer generation. Confidence routing and $T_{\max}$ bound this evidence-acquisition process.

Matched model-request budget. We compare RECAP with repeated answer sampling followed by majority voting and sequential self-refinement on V*Bench. For each question, their request count matches RECAP's complete trajectory, including initialization, entity diagnosis, crop verification, and answer generation. This control executes up to four refinement rounds without confidence-based early stopping and returns the answer generated in RECAP's final round; the mean budget is 9.3 requests per question.

Table 4: Qwen2.5-VL-7B on V*Bench: (a) locally reproduced pipelines under a shared image-size cap; (b) a separate matched-request control. Time denotes mean end-to-end seconds per question.

Method	Acc. (%)	Time (s)
(a) Accuracy–latency comparison		
Qwen2.5-VL 7B	79.1	1.8
HiDe	90.6	6.6
ViCrop	83.8	4.5
ERGO	81.7	10.5
RECAP	**95.8**	10.8
(b) Matched model-request budget		
Repeated sampling + majority vote	79.1	14.2
Sequential self-refinement	78.5	14.2
RECAP	**95.8**	**8.9**

Table 4(b) reports 95.8% accuracy for RECAP, versus 79.1% for majority voting and 78.5% for self-refinement. These results support allocating requests to acquiring and integrating targeted visual evidence. Requests have different visual-token and forward-pass costs; the measured times report the resulting latency of each strategy.

Verification and sensitivity. Our verifier audit evaluates proposal selection against the V*Bench ground-truth boxes, comparing raw-attention ranking with comparative verification. The analysis examines the complementary roles of attention and verification: attention supplies spatial candidates, while verification ranks them for target-specific evidence selection. It complements the end-to-end ablation by examining the contribution of proposal ranking within the refinement loop. Proposal recall measures whether a target-containing candidate is available, whereas top-1 target-hit rate measures whether the ranking selects it. Considering both separates candidate coverage from selection quality and helps locate errors within the evidence-acquisition process. Quantitative results, split-level statistics, and evaluation details are provided in the supplement.

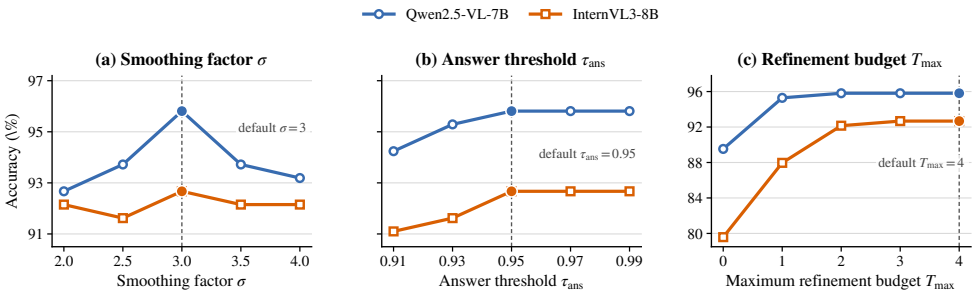

Figure 5: Hyperparameter sensitivity on V*Bench for Qwen2.5-VL-7B and InternVL3-8B. Each panel varies one parameter while holding the others fixed; dashed lines mark the adopted settings.

Figure 5 shows that $\sigma = 3$ is favored by both evaluated backbones and that $\tau_{\text{ans}} = 0.95$ is the smallest tested threshold attaining their best routing accuracy. Performance then plateaus within a short trajectory: Qwen2.5-VL-7B reaches its reported result by the second retained refinement state and InternVL3-8B by the third. We nevertheless keep $T_{\max} = 4$ as a common safety bound and freeze these choices for HRBench.

5 Limitations and Future Work

Coarse proposal maps. The limited spatial resolution of attention maps can make small objects difficult to localize and nearby instances difficult to distinguish. For spatially extended or dispersed targets, a crop may make the target recognizable while covering only part of it. Such partial coverage can omit evidence needed for reasoning about the target's spatial relations. Multi-scale proposals and coverage-aware selection could improve localization and preserve more complete target evidence.

Approximate geometry. Fold-down composition preserves axis-aligned order while compressing metric distances and potentially altering topology. Spatial reasoning therefore depends on the geometry retained in the compact canvas and the frozen backbone's capabilities. One direction is to associate crops with normalized source coordinates or relative offsets, allowing the model to access additional geometric information alongside the compact visual evidence. Evaluating these representations on distance-sensitive questions would help identify which geometric cues should be retained during compression.

6 Conclusion

RECAP connects tentative-answer entity selection, cached attention, visibility-ranked proposals, and order-aware composition for training-free high-resolution reasoning. It improves five backbones on V*Bench and HRBench. Composition controls show gains from compact evidence assembly, while matched-request controls support directing inference toward targeted visual evidence acquisition. Together, these results demonstrate the value of linking answer generation to entity-specific visual updates. By directing each inspection toward entities that remain unclear, RECAP progressively acquires the visual evidence needed for subsequent reasoning, enabling adaptive allocation of inference effort to question-relevant regions.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, the design of supplementary experimental controls, supplementary experiment code, and the analysis and interpretation of experimental results. We have not used generative AI tools to develop the proposed method framework or research hypotheses, formulate mathematical claims, assist with mathematical proofs, or perform qualitative or thematic data analysis. Synthetic dataset generation and dataset cleaning or reformatting are not applicable to this work, which uses existing public benchmark datasets. Additionally, we used generative AI tools for manuscript polishing and organization, LaTeX formatting, and consistency checks. We have reviewed the AI-assisted text and code and manually checked reported values against the original experimental records. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The main paper describes the RECAP pipeline, evaluation settings, and controlled comparisons. Supplementary Appendix A specifies the evaluation protocol, hyperparameter selection, and hardware and software environment. Appendices B and C provide the prompt templates and adapter-specific implementation details, while Appendix D presents the inference procedure and fold-down composition algorithm. These materials document the settings and operations used to reproduce the reported experiments.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385/.

Mohammadjavad Ahmadpour, Amirmahdi Meighani, Payam Taebi, Omid Ghahroodi, Amirmohammad Izadi, and Mahdieh Soleymani Baghshah. Limits and gains of test-time scaling in vision-language reasoning, 2025. URL https://arxiv.org/abs/2512.11109.

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

Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1773–1781, 2025.

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, 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.

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, 2025b. URL https://arxiv.org/abs/2502.13923.

Leon Liangyu Chen, Haoyu Ma, Zhipeng Fan, Ziqi Huang, Animesh Sinha, Xiaoliang Dai, Jialiang Wang, Zecheng He, Jianwei Yang, Chunyuan Li, Junzhe Sun, Chu Wang, Serena Yeung-Levy, and Felix Juefei-Xu. Unit: Unified multimodal chain-of-thought test-time scaling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30257–30267, June 2026.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 49250–49267. Curran Associates, Inc., 2023. doi: 10.52202/075280-2142. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9a6a435e75419a836fe47ab6793623e6-Paper-Conference.pdf.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Brown Ebouky, Gabriele Carrino, Niccolo Avogaro, Christoph Studer, Andrea Bartezzaghi, and Mattia Rigotti. Gazevlm: Active vision via internal attention control for multimodal reasoning, 2026. URL https://arxiv.org/abs/2605.07817.

Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22826–22835, October 2025.

Soumya Suvra Ghosal, Youngeun Kim, Zhuowei Li, Ritwick Chaudhry, Linghan Xu, Hongjing Zhang, Jakub Zablocki, Yifan Xing, and Qin Zhang. Visref: Visual refocusing while thinking improves test-time scaling in multi-modal large reasoning models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33404–33414, June 2026.

Marcel Gröpl, Jaewoo Jung, Seungryong Kim, Marc Pollefeys, and Sunghwan Hong. Entropy-gradient grounding: Training-free evidence retrieval in vision-language models, 2026. URL https://arxiv.org/abs/2604.08456.

Wenya Guo, Ying Zhang, Xiaoping Wu, Jufeng Yang, Xiangrui Cai, and Xiaojie Yuan. Re-attention for visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 91–98, 2020. doi: 10.1609/aaai.v34i01.5338.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: An lmm perceiving any aspect ratio and high-resolution images. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 390–406, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-73010-8.

Zheng Jiang, Yiming Chen, Nan He, Jiahui Chen, Chaoyang Li, Houde Qian, and Lifeng Sun. Test-time scaling over perception: Resolving the grounding paradox in thinking with images. *arXiv preprint arXiv:2604.11025*, 2026.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7uDI7w5RQA.

Jewon Lee, Wooksu Shin, Seungmin Yang, Ki-Ung Song, DongUk Lim, Jaeyeon Kim, Tae-Ho Kim, and Bo-Kyeong Kim. ERGO: Efficient high-resolution visual understanding for vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ehKzPoOReW.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

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 (CVPR)*, pp. 9098–9108, June 2025b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, January 2024.

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, 2025. URL https://arxiv.org/abs/2510.00054.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chanJGoa7f.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models, 2024.

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 Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6602–6618, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.335. URL https://aclanthology.org/2025.emnlp-main.335/.

Zeru Shi, Kai Mei, Yihao Quan, Dimitris N. Metaxas, and Ruixiang Tang. Improving visual reasoning with iterative evidence refinement, 2026. URL https://arxiv.org/abs/2603.14117.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, 2024. doi: 10.1109/CVPR52733.2024.00914.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025.

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 (CVPR)*, pp. 13084–13094, June 2024.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. Conical visual concentration for efficient large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14593–14603, June 2025.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone, 2024. URL https://arxiv.org/abs/2408.01800.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952, 2023. doi: 10.1109/ICCV51070.2023.01100.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In Yuejie Chi, Gintare Karolina Dziugaite, Qing Qu, Atlas Wang Wang, and Zhihui Zhu (eds.), *Conference on Parsimony and Learning*, volume 234 of *Proceedings of Machine Learning Research*, pp. 202–227. PMLR, 03–06 Jan 2024. URL https://proceedings.mlr.press/v234/zhai24a.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 *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DgaY5mDdmT.

Kesen Zhao, Beier Zhu, Qianru Sun, and Hanwang Zhang. Unsupervised visual chain-of-thought reasoning via preference optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2303–2312, October 2025.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xUyMXkI958.

Liangyu Zhong, Fabio Rosenthal, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, and Leo Schwinn. FOCUS: internal MLLM representations for efficient fine-grained visual question answering. *arXiv*, abs/2506.21710, 2025. doi: 10.48550/ARXIV.2506.21710. URL https://doi.org/10.48550/arXiv.2506.21710.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/50623630a2372839c078474efa6c0cb8-Paper-Conference.pdf.

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