# REGION-LEVEL POLICY OPTIMIZATION FOR FINE-GRAINED MLLM PERCEPTION

**Yuheng Shi**[1], **Xiaohuan Pei**[1], **Minjing Dong**[2], **Chang Xu**[1]
[1]University of Sydney [2]City University of Hong Kong
{yuheng.shi, xiaohuan.pei, c.xu}@sydney.edu.au, minjdong@cityu.edu.hk



## ABSTRACT

Fine-grained visual perception in multimodal large language models (MLLMs) is commonly improved by raising the resolution, but the added visual tokens inflate vision-encoding and language-model prefilling costs. We show that the two operations underlying fine-grained perception, *localizing* the region of interest (RoI) and *recognizing* its content, have different resolution requirements. In a controlled diagnostic, localization tolerates roughly 3–4$\times$ stronger token compression than recognition, which motivates localizing from a coarse view and concentrating resolution on the selected evidence. Decoding coordinates with the MLLM can be trained end-to-end from answers, but costs a full model pass per query and depends on grounding ability. A lightweight proposal network distilled from the model's attention is fast, but inherits the noise of its attention targets. The RoI from the proposal network reaches the answer through a discrete region choice, so its faithfulness to the answer cannot supervise the network. We therefore optimize the proposal network with region-level reinforcement learning, which we call Vision-RL$^2$. It treats coherent regions as actions, and a frozen MLLM reader scores each one by how its removal changes the answer likelihood. Complementary subtractive and additive objectives suppress distracting proposals and recover missing evidence, updating only the predictor without region annotations, response sampling, or reasoning trajectories. The refined proposal further enables a sparse encoding that magnifies evidence and excludes background tokens. Across six fine-grained benchmarks and four MLLM backbones, Vision-RL$^2$ improves accuracy over the base model at every token budget and surpasses its largest-budget accuracy with about $4\times$ fewer visual tokens. Under the same token limit, Vision-RL$^2$ also outperforms previous state-of-the-art methods that fully finetune the base model, while training an order of magnitude fewer parameters. Code is available at `https://github.com/YuHengsss/VisionRL2`.

## 1 INTRODUCTION

Multimodal large language models (MLLMs) (Bai et al., 2025; Liu et al., 2023) still struggle with fine-grained details, as small text, distant objects, and cluttered high-resolution scenes reduce accuracy even when coarse scene understanding remains reliable (Tong et al., 2024; Wu & Xie, 2023; Zhang et al., 2025a). The direct remedy is to increase the input image resolution, which produces more visual tokens and increases both vision-encoding and language-model costs. As answer-relevant evidence often occupies a small fraction of the image, efficient perception requires allocating high-resolution processing selectively rather than uniformly across the image.

Answering a fine-grained visual question generally involves two perceptual operations, first *localizing* the relevant evidence and then *recognizing* its content. Standard MLLMs and recent privileged-view distillation methods (Wei et al., 2026; Yuan et al., 2026) optimize fine-grained perception within the full model, without exposing localization and recognition as separate stages. Prior work shows quantitatively that MLLMs can concentrate attention on the ground-truth region even when they answer incorrectly (Zhang et al., 2025a), suggesting that recognition may fail despite an intact localization signal. Following this insight, we build a controlled diagnostic on ZoomBench (Wei et al., 2026) with Qwen3.5-4B. The model predicts an RoI box from the scene, then answers from the resulting crop under a cap of at most 128 visual tokens. On filtered cases that this pipeline answers correctly at full resolution and that are not solvable from the question alone, the *localization* sweep compresses only the scene from which the box is re-predicted, while the *recognition* sweep

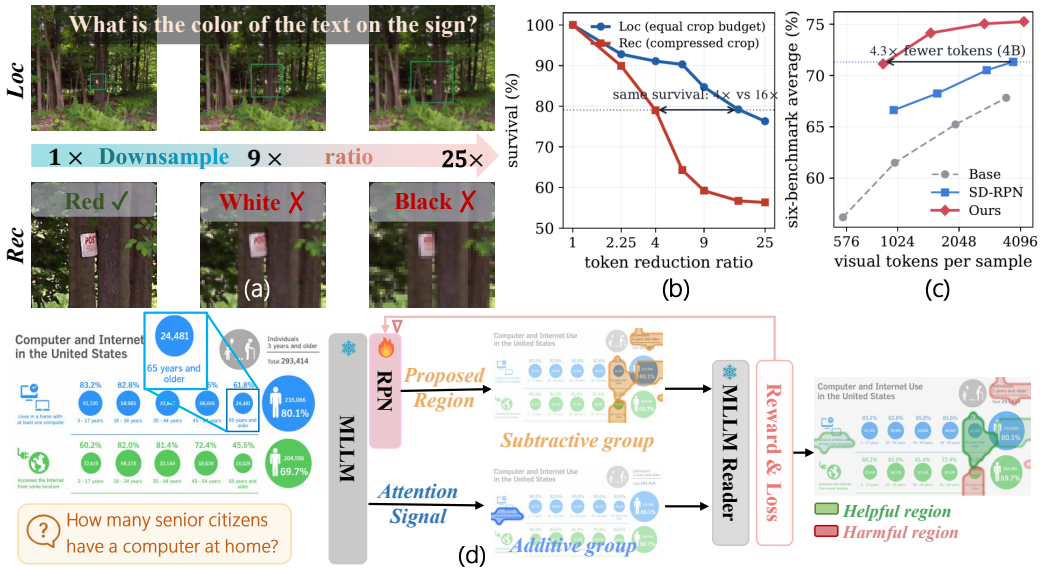


Figure 1: (a) The box is still re-predicted correctly at $25\times$ token reduction, while reading the crop fails at $9\times$. (b) Population survival ($n$=238) when compressing the locator's scene or the recognizer's crop in isolation. Recognition degrades more rapidly than localization. (c) Our method matches the SD-RPN's accuracy using $4.2\times$ fewer visual tokens on Qwen3.5-4B. (d) Overview of Vision-RL$^2$. The RPN proposes candidate regions and the MLLM supplies an attention-derived candidate. A reader scores every region's contribution to the answer and the reward updates the RPN.

freezes the box and compresses its crop. *Survival* is the fraction of these cases still answered correctly. At matched survival, localization tolerates roughly 3–4$\times$ stronger token compression than recognition (Fig. 1a,b). The full protocol is given in Appendix C.

This measured gap provides a quantitative basis for *resolution-decoupled inference*, which localizes from a coarse view and reserves higher resolution for the selected evidence. It places the burden on the RoI predictor, which must be reliable from a coarse view and add little overhead. Existing two-stage systems implement such routing through attention-guided or iterative visual search (Zhang et al., 2025a; Shen et al., 2025), or through autoregressive coordinate generation (Lee et al., 2026; Li et al., 2026). Although effective, these interfaces require costly full-model operations, multiple interaction rounds, or explicit localization sequences. Their routing overhead can therefore yield a less favorable accuracy–latency trade-off than privileged-view distillation methods. SD-RPN (Shi et al., 2026a) provides an efficient routing interface. A proposal network attached to intermediate MLLM layers predicts an answer-free RoI map in a single pass, trained on token-wise pseudo-labels distilled from response-to-image attention. This surrogate objective may retain spurious activations or omit weakly attended evidence, and neither outcome is checked against the answer. Fig. 1d shows such a case, where the proposal keeps a distracting region and misses the evidence the answer needs. Unlike a decoded box, which is part of the model's response and can be optimized from the answer directly, the RoI map reaches the answer only through a discrete region choice. The map is binarized, regions are extracted, and the selected crop is re-encoded, so no gradient connects the answer back to the map. Direct geometric supervision is also insufficient because the RoI in fine-grained VQA is not uniquely annotated and may be an arbitrary question-relevant region rather than a well-defined object (Man et al., 2025). What the predictor lacks is therefore an answer-level signal delivered with region-level credit, that is, a per-region measure of whether keeping it helped or harmed the answer.

We introduce Vision-RL$^2$, a **R**egion-**L**evel **R**einforcement **L**earning method that supplies exactly this signal (Fig. 1d). Built on the trained SD-RPN predictor to retain its single-pass efficiency, Vision-RL$^2$ treats coherent visual regions as actions, and a frozen MLLM reader scores each region by how its removal changes the teacher-forced likelihood of the gold answer. A subtractive objective prunes predictions whose contribution falls below a calibrated noise margin, and an additive objective recovers evidence the proposal missed. We further propose sparse visual encoding that processes the selected evidence at finer granularity at the token level.

Across multiple fine-grained benchmarks and backbones, Vision-RL$^2$ outperforms both the base model and SD-RPN at every shared token budget (Fig. 1c). Under the full-resolution setting, it is

further competitive with recent state-of-the-art methods that fully finetune the MLLM, while updating only the proposal network. In summary, we make three contributions. First, a controlled resolution intervention isolates localization and recognition and reveals their asymmetric compression tolerance. Second, our region-level policy optimization operates natively on a dense RoI map and trains a decode-free predictor from a frozen reader's functional answer signal without region annotations. Third, the learned predictor and sparse evidence encoding improve fine-grained accuracy and efficiency across models and benchmarks.

## 2 Related Work

MLLMs improve fine-grained perception primarily by admitting more visual tokens, through tiled high-resolution encoding (Liu et al., 2024; Chen et al., 2024; Li et al., 2024) or native dynamic-resolution encoders (Wang et al., 2024a; Lu et al., 2025; Dehghani et al., 2023; Zhu et al., 2025). Uniform scaling, however, spends encoding and prefilling compute regardless of where the evidence lies. Beyond scaling, thinking-with-images methods optimize the full model with reinforcement learning to interleave decoding with crop and zoom actions (OpenAI, 2025; Zheng et al., 2025; Lai et al., 2025; Yang et al., 2025b; Zhang et al., 2025b; Lee et al., 2026; Li et al., 2026), and latent-reasoning variants compress this search into the visual latent space (Wang et al., 2025; Li et al., 2025). These methods obtain answer-aware behavior, but full-model RL is memory-intensive and unstable, and inference pays for decoded trajectories or repeated generation rounds at every query. Recent analysis further suggests that their gains are dominated by improvements in the RL-optimized model itself rather than by the interleaved tool use (Ma et al., 2026; Wei et al., 2026). Privileged-view distillation instead trains the backbone offline with region-enhanced teachers (Wei et al., 2026; Yuan et al., 2026), transferring an answer-level signal at the cost of full-model finetuning and without exposing a reusable localizer. Another line separates localization from recognition at inference time. Training-free systems guide the zoom with handcrafted rules over internal attention or search trees (Zhang et al., 2025a; Zhong et al., 2025; Liu et al., 2025). They need no training but incur multiple prefilling or decoding passes, and their heuristics transfer poorly across models and scenes. Supervised alternatives predict RoIs from curated coordinate annotations (Jiang et al., 2025; Shi et al., 2025), at heavy data-curation and finetuning cost. SD-RPN (Shi et al., 2026a;b) removes both requirements by self-distilling response-to-image attention into a lightweight, answer-free RoI predictor over intermediate MLLM features. Yet its token-wise surrogate supervision never verifies that the proposed regions support the answer.

## 3 Method

### 3.1 Preliminaries

**Response-to-image attention.** Given an image $I$ and a question $q$, an MLLM maps the image to $N_v$ visual embeddings $\boldsymbol{H}_v^0 = \mathcal{P}(\mathcal{E}_v(I))$ through vision encoder $\mathcal{E}_v$ and projector $\mathcal{P}$, then generates a response $y = (y_1, \dots, y_T)$. At layer $l$, with response-token queries $\boldsymbol{Q}_y^l \in \mathbb{R}^{T\times d}$ and visual-token keys $\boldsymbol{K}_v^l \in \mathbb{R}^{N_v\times d}$, the response-to-image attention and its spatial map are:

$$\boldsymbol{A}_{y\to v}^l = \mathrm{softmax}\Big(\boldsymbol{Q}_y^l(\boldsymbol{K}_v^l)^\top/\sqrt{d}\Big), \qquad \boldsymbol{M}^l = \mathcal{G}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{A}_{y\to v}^l[t,:]\right). \tag{1}$$

where $\mathcal{G}: \mathbb{R}^{N_v} \to \mathbb{R}^{H_g\times W_g}$ reshapes the $N_v = H_g W_g$ tokens onto the visual-token grid. $\boldsymbol{M}^l$ often highlights answer-relevant evidence, but it typically becomes available only after decoding.

**SD-RPN.** We build on the Self-Distilled Region Proposal Network (SD-RPN) (Shi et al., 2026a), which distills this latent grounding signal into an efficient, answer-free predictor. It reuses the first $B$ frozen MLLM blocks as its backbone, attaches a small stack of trainable blocks, and predicts a dense RoI logit map from the last prefilling token, $\boldsymbol{Z}_\theta = \mathcal{G}\big(\boldsymbol{Q}_{\mathrm{RoI}}\boldsymbol{K}_{\mathrm{RoI}}^\top\big)$, where $\boldsymbol{Q}_{\mathrm{RoI}}$ and $\boldsymbol{K}_{\mathrm{RoI}}$ are normalized linear projections of the last prefilling token's hidden state and of the visual features refined by the trainable blocks. Their trainable parameters are denoted by $\theta$, and $\boldsymbol{Z}_\theta, \boldsymbol{P}_\theta \in \mathbb{R}^{H_g\times W_g}$ with $\boldsymbol{P}_\theta = \sigma(\boldsymbol{Z}_\theta)$, where $\sigma$ is the element-wise sigmoid. SD-RPN is trained by self-distilling Eq. 1, where denoised response-attention maps label only high-confidence foreground and background tokens and ambiguous tokens are ignored. This token-wise supervision remains a local surrogate. Spurious regions may survive, incomplete attention may omit necessary evidence, and ignored cells

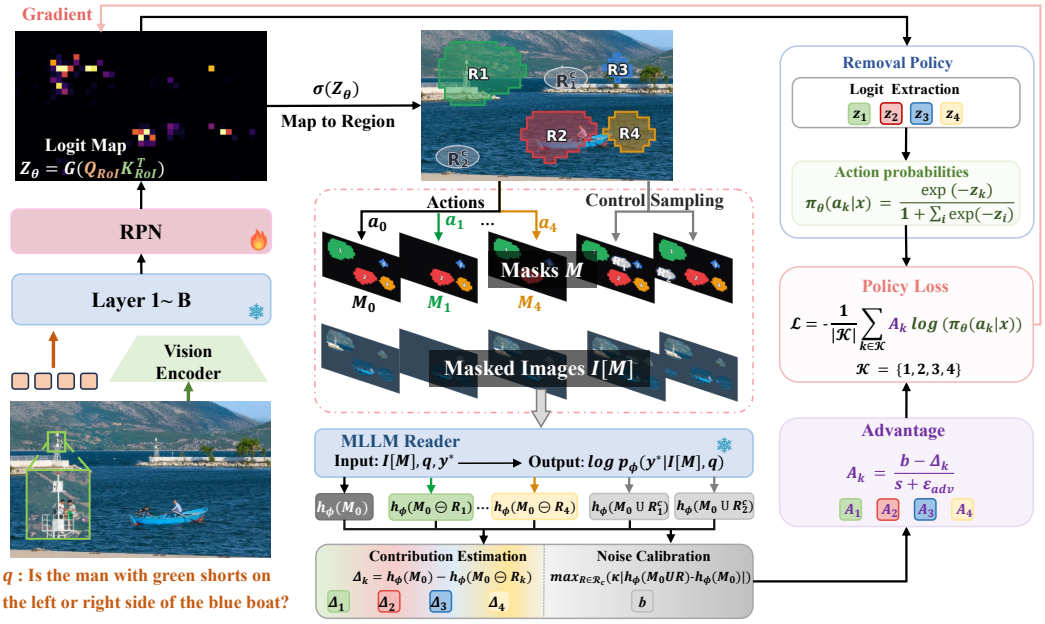


Figure 2: **Overview of the proposed region-level RL.** Predicted regions $R_1$–$R_4$ and control regions $R_1^c, R_2^c$ are extracted from the RoI map, the frozen reader scores each action's masked image, and the resulting contributions $\Delta_k$ and margin $b$ form the advantages and policy loss. Group markers are dropped for clarity, so $R_k, \Delta_k, A_k, s, \pi_\theta$ denote $R_k^{\mathrm{p}}, \Delta_k^{\mathrm{p}}, A_k^{\mathrm{p}}, s_{\mathrm{p}}, \pi_\theta^{\mathrm{sub}}$ of Sec. 3.3, with $M_0 = M_{\mathrm{p}}$ and $M_k = M_0 \ominus R_k$. The additive group is omitted here for brevity.

receive no task-level signal. Our region-level RL instead optimizes what this surrogate cannot see. The predictor should retain a *minimal sufficient support*, exactly the evidence the answer requires, keeping every functionally necessary region and recovering evidence the proposal missed.

### 3.2 Regions, Masks, and Functional Contributions

To realize this goal, we build three ingredients, namely regions as atomic units, masks that render retained regions into reader inputs, and each region's functional contribution to the answer. Unlike a language model with fixed vocabulary, an RPN produces a dense heatmap rather than a distribution over RoI candidates. Treating the $H_g W_g$ cells as independent binary actions yields $2^{H_g W_g}$ masks and credits isolated cells, whereas cropping requires spatially coherent evidence and an answer-based signal is informative only for sufficiently complete supports. We therefore take coherent visual regions as the primary objects of the formulation. A *region* $R$ is a connected set of visual-grid cells, produced from a spatial source map by one shared map-to-region operator. The map is smoothed, binarized at a peak-relative threshold, and split into connected components (detailed in Appendix A.1). Two sources instantiate this operator. The first is the current policy map $\boldsymbol{P}_\theta$, and the second, used during training only, is a frozen response-to-image map of the form of Eq. 1.

Each region carries the grid indicator mask $m_R(u) = \mathbf{1}[u \in R]$, and a set $\mathcal{A}$ of retained regions is rendered by the mask constructor, $M(\mathcal{A}) = \bigvee_{R\in\mathcal{A}} m_R$, the cell-wise union of its members' masks. A grid mask $M$ is upsampled to the image resolution, and $I[M]$ denotes the masked image, in which pixels outside the mask are replaced by the per-channel image mean (Fig. 2). The reader below consumes only $I[M]$.

Rather than scoring geometry, we score function. A mask $M$ is good exactly when the evidence it retains still supports the correct answer. Given the gold answer $y^\star = (y_1^\star, \ldots, y_T^\star)$, a frozen reader $p_\phi$, instantiated by the underlying MLLM itself, is conditioned on $I[M]$ and teacher-forced on the answer tokens. The geometric mean answer probability and its log-odds are:

$$P_\phi(M) = \exp\left(\frac{1}{T}\sum_{t=1}^{T} \log p_\phi(y_t^\star \mid I[M], q, y_{<t}^\star)\right), \qquad h_\phi(M) = \mathrm{logit}(P_\phi(M)). \tag{2}$$

Since $I$, $q$, and $y^\star$ are fixed per sample, $h_\phi(M)$ is a function of the mask alone, high when $M$ preserves the evidence required for $y^\star$ and dropping when that evidence is masked out. We use it as the functional score. Its additive, unbounded scale supports relative comparisons between masks, scoring needs neither response sampling nor a reasoning trajectory, and it is treated as a constant (stop-gradient) during optimization. The bridge from a single region to the functional score is its leave-one-out contribution against a reference mask $M_{\mathrm{ref}}$:

$$\Delta_\phi(R \mid M_{\mathrm{ref}}) = h_\phi(M_{\mathrm{ref}}) - h_\phi(M_{\mathrm{ref}} \ominus R), \tag{3}$$

where $M \ominus R = M \odot (1 - m_R)$ removes the cells of $R$ from a mask. A positive contribution indicates that removing $R$ reduces functional support. Contributions are clipped to $[-\delta, \delta]$ with $\delta = 5$. The estimator is agnostic to how regions are produced and needs only $n + 1$ reader evaluations for $n$ regions (one reference mask and $n$ singleton leave-one-out masks), rather than a search over subsets. The two action groups of Sec. 3.3 instantiate it with different region sources and masks.

### 3.3 Functional Region-Level Policy Optimization

Searching for this minimal sufficient support over candidate masks directly is impractical. Their number grows combinatorially with the number of regions, and a proposal that omits necessary evidence cannot be repaired by pruning alone. We therefore approximate it with two marginal action groups, both instantiating Eq. 3. The subtractive group removes predicted regions whose contribution does not exceed the noise level, and the additive group recovers unpredicted regions with positive contribution (Fig. 2 shows one step).

**Subtractive policy with noise calibration.** The first group operates on the policy's own predictions. Applying the shared map-to-region operator of Sec. 3.2 to the policy map $\boldsymbol{P}_\theta$ yields candidate components. We rank the candidates by their mean RoI logit $z_\theta(R) = \frac{1}{|R|} \sum_{u \in R} \boldsymbol{Z}_\theta(u)$ and keep the top $K \le K_{\max}$ as the **p**olicy-predicted regions $\mathcal{R}_{\mathrm{p}} = \{R_1^{\mathrm{p}}, \ldots, R_K^{\mathrm{p}}\}$, with mask $M_{\mathrm{p}} = M(\mathcal{R}_{\mathrm{p}})$. Region extraction is non-differentiable, but $z_\theta(R)$, the region's confidence, remains differentiable in the logits it averages. Against the intact prediction, each region's contribution is $\Delta_k^{\mathrm{p}} = \Delta_\phi(R_k^{\mathrm{p}} \mid M_{\mathrm{p}})$.

Small likelihood changes arise even when a removal does not alter answer-relevant evidence, and their scale varies across samples and reader sizes, so $\Delta_k^{\mathrm{p}}$ alone does not tell whether a change exceeds the reader's noise under masking. We estimate a sample-specific margin from a small set $\mathcal{R}_{\mathrm{c}}$ of **c**ontrol regions (gray in Fig. 2), grown outside the dilated prediction in areas with $\boldsymbol{P}_\theta < 0.02$ and matched in area to the median predicted region:

$$b = \min\Big(\kappa \max_{R \in \mathcal{R}_{\mathrm{c}}} \big|h_\phi(M(\mathcal{R}_{\mathrm{p}} \cup \{R\})) - h_\phi(M_{\mathrm{p}})\big|, b_{\max}\Big), \tag{4}$$

where $\kappa$ controls the calibration strength and $b_{\max} = 1$ caps the margin. We set $|\mathcal{R}_{\mathrm{c}}| = 2$ here.

The region confidence $z_\theta$ routes the resulting credit into the dense map. Denote the intact action by $a_0$, the singleton removal of $R_k^{\mathrm{p}}$ by $a_k$, write $z_k = z_\theta(R_k^{\mathrm{p}})$ and $x = (I, q)$, and let $\mathcal{K}$ index the removals that keep the foreground non-empty (emptying the mask would measure the loss of all visual support rather than a region's marginal effect). Since $z_k$ is the mean logit of the region's per-cell Bernoulli inclusion, the removal $a_k$ scores $-z_k$ while the intact action scores 0, and a softmax over these scores gives the removal policy:

$$\pi_\theta^{\mathrm{sub}}(a_k \mid x) = \frac{\exp(-z_k)}{1 + \sum_{j \in \mathcal{K}} \exp(-z_j)}, \qquad k \in \mathcal{K}, \tag{5}$$

where the unit term is the intact action. All removals in $\mathcal{K}$ are evaluated, so $\pi_\theta^{\mathrm{sub}}$ is never sampled. It is the differentiable channel that assigns each action's credit to its region. The margin-calibrated advantage and policy loss are:

$$A_k^{\mathrm{p}} = \frac{b - \Delta_k^{\mathrm{p}}}{s_{\mathrm{p}} + \varepsilon_{\mathrm{adv}}}, \qquad \mathcal{L}_{\mathrm{sub}} = -\frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} A_k^{\mathrm{p}} \log \pi_\theta^{\mathrm{sub}}(a_k \mid x), \tag{6}$$

where $s_{\mathrm{p}}$ is the standard deviation of $\{\Delta_k^{\mathrm{p}}\}_{k \in \mathcal{K}}$ and $\varepsilon_{\mathrm{adv}} = 1$. If $\Delta_k^{\mathrm{p}} > b$, removing $R_k^{\mathrm{p}}$ receives negative credit and the region is preserved by increasing $z_k$. Otherwise its confidence is decreased. With $K = 1$ no relative removal exists, so $\mathcal{L}_{\mathrm{sub}} = 0$ and the anchor below handles the region.

**Additive policy.** Removal alone cannot recover evidence the policy misses, so the second group tests supplementary regions, marked by s, proposed by the training-only source. Frozen response-to-image maps of the form of Eq. 1, taken from six MLLM layers spread across depth (Appendix A.1), are each passed through the shared region operator. Residual regions outside $M_{\mathrm{p}}$ are then merged across layers, and the top merged candidates form $\mathcal{R}_{\mathrm{s}} = \{R_1^{\mathrm{s}}, \ldots, R_J^{\mathrm{s}}\}$ with $J \leq 4$. With the augmented reference mask $M_{\mathrm{aug}} = M(\mathcal{R}_{\mathrm{p}} \cup \mathcal{R}_{\mathrm{s}})$, each supplementary region's contribution is $\Delta_j^{\mathrm{s}} = \Delta_\phi(R_j^{\mathrm{s}} \mid M_{\mathrm{aug}})$. No control-region margin is applied here. For a candidate with no true contribution, exclusion is already the zero-gradient default of the loss below, so a margin would only bias against weak genuine recoveries. Credit reaches supplementary regions through the mean inclusion log-likelihood $\ell_\theta^+(R) = \frac{1}{|R|}\sum_{u\in R} \log \boldsymbol{P}_\theta(u)$:

$$A_j^{\mathrm{s}} = \frac{\Delta_j^{\mathrm{s}}}{s_{\mathrm{s}} + \varepsilon_{\mathrm{adv}}}, \qquad \mathcal{L}_{\mathrm{add}} = -\sum_{j=1}^{J} A_j^{\mathrm{s}}\, \ell_\theta^+(R_j^{\mathrm{s}}), \tag{7}$$

where $s_{\mathrm{s}}$ is the standard deviation of $\{\Delta_j^{\mathrm{s}}\}_{j=1}^{J}$. Positive contribution $\Delta_j^{\mathrm{s}}$ raises a missing region, while negative contribution suppresses it. The two credit channels match their decision structures. The subtractive group resolves a categorical choice among mutually exclusive removals, so its removals compete under the normalizer of Eq. 5, whereas each supplementary candidate is an independent inclusion decision judged on its own contribution, which a shared normalizer would distort.

**Stabilization and overall objective.** Within-group normalization can amplify noise when the reader assigns low likelihood to the gold answer under every tested mask, so both policy losses are scaled by a detached attainability weight $w_i$, an EMA-normalized, clipped ratio of the sample's best attained gold likelihood (Appendix A.3). A frozen copy of the initial SD-RPN provides a KL anchor $\mathcal{L}_{\mathrm{KL}}$, and when $K_i = 1$ a binary cross-entropy term $\mathcal{L}_{\mathrm{K1}}$ preserves the single-component proposal, both defined in Appendix A.2. The per-sample objective is:

$$\mathcal{L}_i = \mathcal{L}_{\mathrm{KL},i} + w_i\,(\mathcal{L}_{\mathrm{sub},i} + \mathcal{L}_{\mathrm{add},i}) + \mathbf{1}[K_i = 1]\,\mathcal{L}_{\mathrm{K1},i}. \tag{8}$$

Training updates only the RPN parameters $\theta$ with the MLLM frozen. At inference, a single answer-free RoI prediction from $(I, q)$ is derived as in SD-RPN.

### 3.4 Sparse Visual Encoding

At inference, the trained RoI predictor enables a simple principle: *spend the visual-token budget on evidence, not on area* (Fig. 3). Dense pipelines allocate tokens uniformly over the image or a crop, so background inside the crop dilutes the effective resolution of the evidence. Given the RoI prediction, we take the bounding box of the predicted foreground, measure its foreground occupancy $\rho_{\mathrm{fg}} \in (0, 1]$, and re-encode the crop at spatial zoom $\eta = \min\left(\sqrt{1/\rho_{\mathrm{fg}}}, \eta_{\max}\right)$, retaining only the visual tokens inside the foreground while background tokens never enter the vision encoder or the language model. Position embeddings in both the vision encoder and the LLM are assigned on the full bbox-crop grid before background tokens are dropped, so the sparse token set is positionally indistinguishable from a dense encoding of the same crop. Since the crop area grows by $\eta^2$, the retained-token count $\rho_{\mathrm{fg}}\eta^2 N_{\mathrm{crop}}$ approximately matches the dense crop budget $N_{\mathrm{crop}}$, while the evidence is viewed at $\eta\times$ finer spatial resolution. When the prediction contains multiple regions, we crop their single enclosing bounding box rather than each region independently. Independent crops would be equally token-efficient, but each carries its own coordinate frame, weakening the spatial relations between regions that the shared crop grid encodes through position embeddings. The source-image tokens are not re-encoded in the second stage, and their KV cache from the proposal pass is reused in part of the LLM layers. Since RoI prediction needs neither the gold answer nor a decoded response, inference adds only one lightweight proposal pass before decoding.

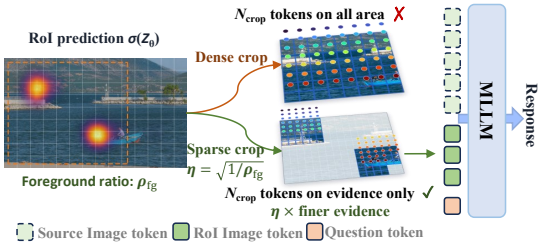


Figure 3: Sparse visual encoding. A dense crop spends the budget $N_{\mathrm{crop}}$ uniformly over the bbox, whereas the sparse crop encodes only foreground tokens at $\eta\times$ finer resolution. Source-image tokens are not re-encoded, and their KV cache is reused in part of the LLM layers.

Table 1: Comparison with SoTA MLLMs under a shared $16{,}384$ source-image token limit. Rows are grouped by base-model family, with large-scale open/closed-source models as reference.

| Model | Size | V* Bench | ZoomBench | HR-Bench 4K | HR-Bench 8K | MME-RW EN | MME-RW CN | Average |
|---|---|---|---|---|---|---|---|---|
| **Large-Scale Open/Closed-Source Models** | | | | | | | | |
| GPT-5.4 | — | 77.0 | 52.7 | 84.0 | 77.9 | 74.2 | 70.9 | 72.8 |
| Gemini-3.1-Pro | — | 88.0 | 61.2 | 89.6 | 86.9 | 76.5 | 73.3 | 79.3 |
| Qwen3-VL-Instruct | 235B | 91.1 | 56.1 | 86.1 | 80.4 | 71.7 | 69.0 | 75.8 |
| Qwen3.5 | 397B | 88.0 | 57.2 | 89.4 | 85.5 | 74.8 | 69.8 | 77.4 |
| Kimi-K2.6 | 1T | 88.5 | 53.1 | 81.9 | 78.0 | 69.2 | 66.1 | 72.8 |
| **Qwen2.5-VL Based** | | | | | | | | |
| DeepEyes | 7B | 85.9 | 46.5 | 75.1 | 72.6 | 64.1 | 64.1 | 68.1 |
| Thyme | 7B | 82.2 | 45.1 | 77.0 | 72.0 | 64.8 | 64.6 | 67.6 |
| DeepEyesV2 | 7B | 81.7 | 45.0 | 77.9 | 73.8 | 64.9 | 65.1 | 68.0 |
| ZwZ | 7B | 86.9 | 55.6 | 75.9 | 72.4 | 65.0 | 63.5 | 69.9 |
| Vision-RL$^2$ (Ours) | 7B | 91.6 | 59.8 | 78.8 | 75.0 | 62.2 | 58.7 | 71.0 |
| **Qwen3-VL Based** | | | | | | | | |
| Qwen3-VL-Instruct | 8B | 84.8 | 43.0 | 79.6 | 75.3 | 63.2 | 64.6 | 68.4 |
| ZwZ | 8B | 90.6 | 58.0 | 84.4 | 81.6 | 69.9 | 69.2 | 75.6 |
| P2R | 4B | 93.2 | – | 81.9 | 80.5 | – | – | – |
| P2R | 8B | 93.7 | – | 81.5 | 82.6 | – | – | – |
| **Qwen3.5 Based** | | | | | | | | |
| Qwen3.5 | 4B | 85.9 | 51.5 | 83.6 | 80.1 | 59.1 | 60.6 | 70.1 |
| Qwen3.5 | 9B | 83.8 | 54.9 | 84.9 | 83.5 | 72.5 | 67.9 | 74.6 |
| Vision-OPD | 4B | 90.6 | 59.5 | 82.0 | 79.1 | 74.2 | 70.6 | 76.0 |
| Vision-OPD | 9B | 90.6 | 65.1 | 87.1 | 85.6 | 73.2 | 70.3 | 78.7 |
| Vision-RL$^2$ (Ours) | 4B | 91.1 | 65.1 | 84.3 | 80.3 | 65.8 | 65.3 | 75.3 |
| Vision-RL$^2$ (Ours) | 9B | 95.3 | 68.4 | 86.8 | 86.1 | 73.4 | 70.6 | 80.1 |
| **Gemma-4 Based** | | | | | | | | |
| Gemma-4 | 12B | 72.8 | 46.5 | 75.5 | 67.5 | 65.2 | 52.4 | 63.3 |
| SD-RPN | 12B | 78.0 | 57.0 | 82.4 | 77.5 | 65.7 | 53.1 | 69.0 |
| Vision-RL$^2$ (Ours) | 12B | 82.2 | 60.7 | 85.0 | 79.4 | 67.6 | 61.6 | 72.8 |

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

**Models and training.** We validate our method on Qwen3.5 4B and 9B (Qwen Team, 2026), Qwen2.5-VL-7B (Bai et al., 2025), and the encoder-free Gemma-4-12B (Team et al., 2026) (Appendix B.4). For each model, the RoI predictor follows the SD-RPN configuration, RL starts from the supervised SD-RPN checkpoint, and the full frozen MLLM serves as the reader $p_\phi$. The RL pool contains 7K QA pairs from the VisualCoT training corpus (Shao et al., 2024), with 5K from its InfographicVQA (Mathew et al., 2022) split and 1K each from its TextVQA (Singh et al., 2019) and DocVQA (Mathew et al., 2020) splits. Within each split of 10K candidates, samples are ranked by the standard deviation of their region-removal rewards under the initial SD-RPN and drawn from the top half. The source image is encoded under a $576$ visual-token limit during training, so all region actions and rewards operate in this regime. All models are trained for one epoch with a single recipe (batch size 32 and learning rate $1.5 \times 10^{-5}$, full configuration in Appendix B.1).

**Benchmarks and evaluation protocols.** We evaluate on the high-resolution fine-grained perception benchmarks V* Bench (Wu & Xie, 2023), ZoomBench (Wei et al., 2026), HR-Bench 4K and 8K (Wang et al., 2024b), and MME-RealWorld (Zhang et al., 2025c). The *main protocol* (Table 1) follows the evaluation setting of recent fine-grained perception works (Wei et al., 2026; Yuan et al., 2026), where all methods share a $16{,}384$ source-image token limit, produce free-form responses, and are scored by rule-based parsing with an LLM judge for the remaining cases. The *training-aligned protocol*, used for the ablations and token-budget analyses, instead evaluates under the $576$-token training limit, extended to $\{576, 1024, 2048, 4096\}$ for the token-budget curves, with short-answer prompting, direct responses, and purely rule-based scoring, keeping the ablated components in the regime the objective optimizes (per-benchmark prompts in Appendix B.2).

## 4.2 MAIN RESULTS

Table 1 compares our models with SoTA fine-grained perception methods. Base-model and Vision-OPD/ZwZ rows are re-evaluated by us on released weights, the large-scale reference models (Team et al., 2025; Qwen Team, 2026) are quoted from Vision-OPD (Yuan et al., 2026), and the rest from their publications (Zhang et al., 2025b; Hong et al., 2026). A key distinction is the trainable pa-

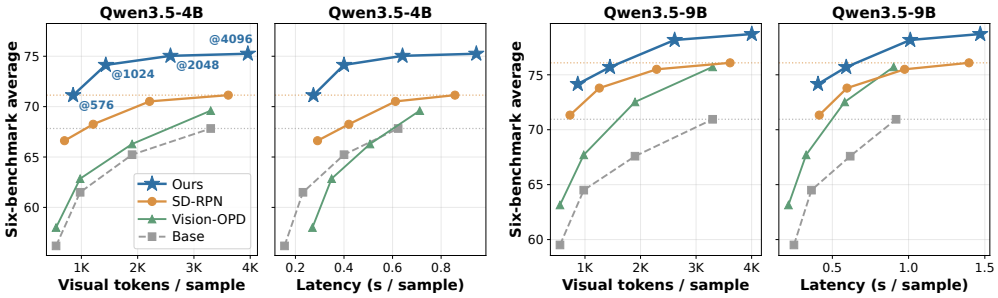


Figure 4: Six-benchmark average accuracy versus inference cost under the training-aligned short-answer protocol, for source-image token limits $\{576, 1024, 2048, 4096\}$. For each model scale, the left panel measures cost in visual tokens per sample (source plus crop) and the right panel as end-to-end latency. Per-benchmark token curves are given in Appendix B.3.

rameter budget. DeepEyes (Zheng et al., 2025), ZwZ (Wei et al., 2026), P2R (Li et al., 2026), and Vision-OPD (Yuan et al., 2026) fine-tune the full MLLM, whereas our method updates only the small attached predictor. Despite this, the 9B model attains the highest average in the table, ahead of Gemini-3.1-Pro and Vision-OPD-9B, with the best V* Bench and ZoomBench scores overall and a lead over Vision-OPD-9B on every remaining benchmark except HR-Bench 4K. At the 4B scale our model surpasses Vision-OPD-4B on V* Bench, ZoomBench, and both HR-Bench splits, while trailing on the MME-RealWorld splits, where full fine-tuning can adapt the reader itself, a gap our frozen-reader design deliberately forgoes. On the Qwen2.5-VL backbone, our 7B model attains the highest average of its block, ahead of ZwZ-7B's 69.9, and leads it on every benchmark except the MME-RealWorld splits, the same trade as the smaller Qwen3.5 model. Gemma-4-12B has no native-resolution mode, so its rows use the largest visual-token tier (1,120 tokens), where our predictor adds 3.8 points over SD-RPN and 9.5 over the base model on average. Qualitative comparisons are given in Appendix B.6.

### 4.3 Efficiency Analysis

All values here are produced under one shared protocol with lmms-eval (Zhang et al., 2024) on the same RTX A6000 GPU. To keep the analysis tractable, the large MME-RealWorld EN/CN splits are replaced by MME-RealWorld-Lite and the InfoVQA validation split, and models answer directly rather than free-form, following the two-stage inference convention of SD-RPN (Shi et al., 2026a). Values are therefore not comparable with the main table.

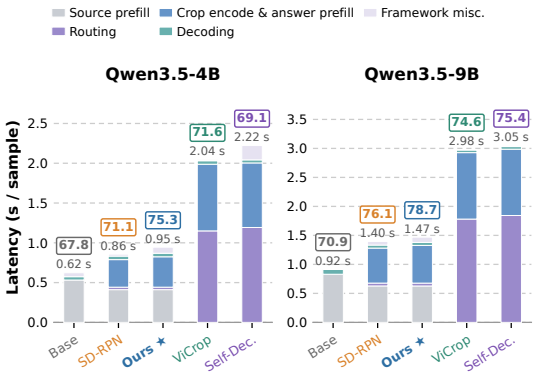


Figure 5: Latency decomposition of two-stage routing mechanisms at the 4,096-token source limit, with the six-benchmark average accuracy above each bar. ViCrop (Zhang et al., 2025a) and Self-Dec. incur substantial cost for RoI routing. All methods share the protocol and GPU.

Fig. 4 compares the base model, SD-RPN, Vision-OPD, and ours over source-image token limits. Our model lies above all baselines at every budget on all six benchmarks (per-benchmark curves in Appendix B.3). At both scales, our model under the 576 limit exceeds the base model at the 4,096 limit by more than three points while consuming about one quarter of its tokens. Our model also matches the SD-RPN accuracy at the 4,096 limit with $4.2\times$ fewer visual tokens on the 4B model and reaches within half a point with $2.5\times$ fewer on the 9B model. In wall-clock latency (right panels), our curve dominates the accuracy–latency plane at both scales. The 4B model at the 1,024 limit exceeds SD-RPN at the 4,096 limit with $2.1\times$ lower latency, the 9B model reaches within one point of that level at $2.4\times$ lower latency, and both models at the 576 limit surpass the base model at the 4,096 limit at $2.3\times$ lower latency. Against the released Vision-OPD weights, our 4B model at the 576 limit exceeds Vision-OPD-4B at its largest budget while responding $2.6\times$ faster.

Table 2: Ablations on Qwen3.5-4B under the training-aligned protocol (576-token source limit). Top block, mechanism ablation, where R.L. denotes region-level RL and S.E. the sparse visual encoding of Sec. 3.4. Bottom block, RL design ablation grouped by design axis with the sparse encoding disabled. $\Delta$ is the change in six-benchmark average vs. the full recipe.

| Variant | V* Bench | ZoomBench | HR-Bench 4K | HR-Bench 8K | MME-RW Lite | InfoVQA | Average | $\Delta$ |
|---|---|---|---|---|---|---|---|---|
| Qwen3.5-4B (base) | 66.0 | 40.5 | 63.5 | 56.4 | 41.0 | 69.8 | 56.2 | — |
| *Mechanism:* | | | | | | | | |
| SD-RPN | 82.7 | 55.6 | 71.1 | 63.3 | 48.9 | 78.1 | 66.6 | $-3.0$ |
| SD-RPN + R.L. (full recipe) | 82.2 | 58.7 | 77.3 | 69.1 | 50.4 | 80.1 | 69.6 | — |
| SD-RPN + R.L. + S.E. | **85.3** | **61.8** | **77.4** | **70.9** | **51.0** | **80.5** | **71.1** | $+1.5$ |
| *Reward: functional score* $\rightarrow$ | | | | | | | | |
| Generation accuracy | 80.6 | 53.3 | 73.9 | 65.1 | 50.7 | 80.5 | 67.3 | $-2.3$ |
| Raw log-probability | 81.7 | 53.1 | 75.8 | 67.6 | 50.0 | 80.4 | 68.1 | $-1.5$ |
| *Actions: singleton region removals* $\rightarrow$ | | | | | | | | |
| Singleton + pair removals | 83.3 | 56.7 | 76.8 | 69.8 | 51.4 | 80.5 | 69.7 | $+0.1$ |
| Cell-level keep-sets | 84.3 | 57.0 | 72.9 | 65.6 | 50.6 | 80.2 | 68.4 | $-1.2$ |
| *Credit assignment:* | | | | | | | | |
| No control-region margin ($\kappa = 0$) | 82.2 | 54.2 | 76.0 | 69.1 | 49.7 | 79.8 | 68.5 | $-1.1$ |
| No additive group | 82.2 | 55.6 | 75.5 | 68.4 | 50.1 | 80.3 | 68.7 | $-0.9$ |

Figure 5 compares the three routing interfaces of Sec. 1 under one setting. Attention routing (Zhang et al., 2025a) and coordinate decoding (the MLLM itself decodes the box) spend a full source pass, plus box tokens for the latter, purely on localization, taking 1.1–1.8 s per query, or 2.1–2.2$\times$ our end-to-end latency. The RPN head adds three blocks on the answer call's own prefill and routes in 31–49 ms, roughly 30$\times$ cheaper, while reaching the best accuracy at both scales.

### 4.4 Ablation Studies

All ablations use the training-aligned protocol of Sec. 4.1 on Qwen3.5-4B with all remaining recipe constants fixed, and the sparse visual encoding is used only where marked. Tables report the six-benchmark average of the efficiency analysis, which adds InfoVQA (Mathew et al., 2022) and differs from the main-table average. The top block of Table 2 isolates the two mechanisms added on top of the supervised predictor. Region-level RL alone contributes $+3.0$ average over SD-RPN, and the sparse visual encoding adds a further $+1.5$ at the same source-token limit. The gains are complementary, with RL improving which evidence is proposed and the sparse encoding how the fixed crop budget is spent on it. Relative to the frozen base model, the full system gains $+14.9$.

The lower block ablates the RL objective itself, grouped by design axis, with the sparse encoding disabled throughout. On the reward axis, replacing the functional score with generation accuracy costs $-2.3$, since binary correctness gives no relative credit when all actions in a group share an outcome, and replacing the clipped log-odds score with the raw mean log-probability costs $-1.5$, since saturated samples then dominate the gradient scale. On the action axis, adding pair removals (with a group-mean baseline, as the margin is calibrated for singletons) performs at parity with the full recipe at clear extra cost, so we keep the leave-one-out design. A budget-matched cell-level control (not single-token pruning), which samples keep-sets of the predicted-foreground size cell-wise from the map logits and rewards them as complete masks, costs $-1.2$, confirming that spatially coherent regions are the right action unit. On the credit axis, removing the control-region margin costs $-1.1$ and removing the additive recovery group costs $-0.9$, so an explicit noise floor is needed even with the intact-prediction reference, and subtractive-only training misses recoverable evidence.

## 5 Conclusion

We start from a measurement that localizing evidence tolerates several times stronger token compression than recognizing its content. Vision-RL$^2$ turns this asymmetry into a training principle. A lightweight RoI predictor is made answer-accountable through region-level reinforcement learning, where a frozen MLLM reader scores each coherent region by its functional contribution to the gold answer, without region annotations or response sampling. Across fine-grained benchmarks and backbones, it outperforms both the base model and its supervised predecessor at every token budget while routing RoIs efficiently. The trained predictor is a reusable, decode-free localizer, and we expect its recipe of scoring proposals by their measured effect on a frozen reader to extend to other interfaces that decide where to spend computation.

## REFERENCES


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*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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, 2024.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *NeurIPS*, 2023.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. In *ICLR*, 2026.

Yitong Jiang, Jinwei Gu, Tianfan Xue, Ka Chun Cheung, Pavlo Molchanov, Hongxu Yin, and Sifei Liu. Token-efficient vlm: High-resolution image understanding via dynamic region proposal. In *ICCV*, 2025.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. *arXiv preprint arXiv:2509.07969*, 2025.

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 *International Conference on Learning Representations*, 2026.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Hongxing Li, Xiufeng Huang, Dingming Li, Wenjing Jiang, Zixuan Wang, Haolei Xu, Hanrong Zhang, Haiwen Hong, Longtao Huang, Hui Xue, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Perceive-to-reason: Decoupling perception and reasoning for fine-grained visual reasoning. *arXiv preprint arXiv:2607.01191*, 2026.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 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. *arXiv preprint arXiv:2510.00054*, 2025.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.

Yan Ma, Weiyu Zhang, Tianle Li, Linge Du, Xuyang Shen, and Pengfei Liu. What does vision tool-use reinforcement learning really learn? disentangling tool-induced and intrinsic effects for crop-and-zoom. *arXiv preprint arXiv:2602.01334*, 2026.

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Argus: Vision-centric reasoning with grounded chain-of-thought. In *CVPR*. IEEE, 2025.

Minesh Mathew, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. Docvqa: A dataset for vqa on document images. corr abs/2007.00398 (2020). *arXiv preprint arXiv:2007.00398*, 2020.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, 2022.

OpenAI. Thinking with images. https://openai.com/index/thinking-with-images/, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *NeurIPS*, 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 *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Baifeng Shi, Boyi Li, Han Cai, Yao Lu, Sifei Liu, Marco Pavone, Jan Kautz, Song Han, Trevor Darrell, Pavlo Molchanov, et al. Scaling vision pre-training to 4k resolution. In *CVPR*, 2025.

Yuheng Shi, Xiaohuan Pei, Minjing Dong, and Chang Xu. Catching the details: Self-distilled roi predictors for fine-grained mllm perception. In *ICLR*, 2026a.

Yuheng Shi, Xiaohuan Pei, Linfeng Wen, Minjing Dong, and Chang Xu. Q-zoom: Query-aware adaptive perception for efficient multimodal large language models. *arXiv preprint arXiv:2604.06912*, 2026b.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

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

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.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond images and language. *arXiv preprint arXiv:2511.21395*, 2025.

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

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, Zhuosheng Zhang, and Weiran Huang. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for "important tokens" in multimodal language models: Duplication matters more. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *CVPR*, 2025a.

Senqiao Yang, Junyi Li, Xin Lai, Bei Yu, Hengshuang Zhao, and Jiaya Jia. Visionthink: Smart and efficient vision language model via reinforcement learning. In *NeurIPS*, 2025b.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

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 *ICLR*, 2025a.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. URL https://arxiv.org/abs/2407.12772.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025b.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *ICLR*, 2025c.

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. *arXiv preprint arXiv:2505.14362*, 2025.

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. In *NeurIPS*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

# A METHOD DETAILS

## A.1 REGION CONSTRUCTION

Both candidate sources of the main text share one map-to-region operator. For a spatial source map $\boldsymbol{M}$:

$$\widetilde{\boldsymbol{M}} = G_{\sigma_g} * \boldsymbol{M}, \qquad \mathfrak{R}(\boldsymbol{M}) = \mathrm{CC}\Big(\mathbf{1}\Big[\widetilde{\boldsymbol{M}} > \rho \max_u \widetilde{\boldsymbol{M}}(u)\Big]\Big), \tag{9}$$

where $G_{\sigma_g}$ is a Gaussian kernel, $\rho$ is a peak-relative threshold, and CC extracts connected regions. Source maps whose peak-to-mean ratio indicates diffuse, non-discriminative activation are discarded.

For the subtractive source, $\mathfrak{R}$ is applied to the current policy map $\boldsymbol{P}_\theta$. Region extraction and top-$K_{\max}$ ranking are non-differentiable, while the region confidences $z_\theta(R)$ remain differentiable with respect to the logits inside each retained region. For the additive source, frozen response-to-image maps from layers 7, 11, 15, 19, 23, and 27 are each passed through $\mathfrak{R}$. The layers are spaced evenly across the depth of the MLLM because answer-relevant attention emerges at different depths for different samples, and no single layer covers all of them. The predicted mask $M_{\mathrm{p}}$ is subtracted, overlapping residual regions are merged across layers, and the top merged candidates form $\mathcal{R}_{\mathrm{s}}$.

## A.2 ANCHOR LOSSES

Let $\boldsymbol{P}_{\theta_0}$ denote the RoI probability map of the frozen initial SD-RPN, computed at the same grid resolution as $\boldsymbol{P}_\theta$. The KL anchor is the mean per-cell Bernoulli divergence from the current map to this reference:

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{H_g W_g} \sum_u D_{\mathrm{KL}}\big(\mathrm{Bern}(\boldsymbol{P}_\theta(u)) \,\big\|\, \mathrm{Bern}(\boldsymbol{P}_{\theta_0}(u))\big), \tag{10}$$

applied to all samples so the map does not drift on prompts the policy gradient cannot reach. When the component set collapses to a single region ($K = 1$), no valid relative removal exists and the policy loss is zero. The sample is instead regularized by a binary cross-entropy toward the binarized reference foreground:

$$\mathcal{L}_{\mathrm{K1}} = \mathrm{BCE}(\boldsymbol{P}_\theta,\ \mathbf{1}[\boldsymbol{P}_{\theta_0} > 0.5]), \tag{11}$$

which preserves the single-component proposal in that regime.

## A.3 STABILIZATION DETAILS

**Attainability weight.** Both policy losses are scaled by a detached attainability weight that suppresses samples on which the frozen reader cannot attain the gold answer under any tested intervention:

$$P^{(i)}_{\max} = \max_{M \in \mathcal{M}^{(i)}_{\mathrm{sub}}} P_\phi(M), \qquad w_i = \mathrm{clip}\left(\frac{P^{(i)}_{\max}}{\max(\mathrm{EMA}_{0.99}[P_{\max}], 0.05)}, 0, 3\right), \tag{12}$$

where $\mathcal{M}^{(i)}_{\mathrm{sub}}$ collects the subtractive-group masks of sample $i$, namely the intact prediction $M_{\mathrm{p}}$ and all retained singleton-removal masks.

# B EXPERIMENTAL DETAILS

## B.1 HYPERPARAMETERS

Table 3 lists the training configuration and the constants of the method section, all of which are shared across the four backbones. Two settings are model-specific and stated here instead. The RoI predictor attaches three trainable blocks after the first $B$ frozen MLLM blocks, with $B = 21$ for the two Qwen3.5 models, $B = 18$ for Qwen2.5-VL-7B, and $B = 27$ for Gemma-4-12B, following each backbone's SD-RPN configuration (Shi et al., 2026a). The control-margin scale is $\kappa = 1.25$ for Qwen3.5-4B and $\kappa = 1.0$ for Qwen3.5-9B, Qwen2.5-VL-7B, and Gemma-4-12B, whose source images are encoded at the 560-token tier, the closest tier to the 576-token limit.

Table 3: Training hyperparameters, shared across the four backbones.

| Config | Setting |
|---|---|
| Optimizer | AdamW |
| Weight decay | 0.0 |
| Optimizer momentum | $(\beta_1, \beta_2) = (0.9, 0.999)$ |
| Global batch size | 32 |
| Learning rate schedule | cosine decay |
| Peak learning rate | $1.5 \times 10^{-5}$ |
| Warm-up strategy / ratio | linear / 0.2 |
| Source-image token limit | 576 |
| Max predicted regions $K_{\max}$ | 6 |
| Max supplementary regions $J$ | 4 |
| Margin cap $b_{\max}$ | 1.0 |
| Contribution clip $\delta$ | 5.0 |
| Advantage regularizer $\varepsilon_{\text{adv}}$ | 1.0 |
| Region-operator smoothing $\sigma_g$ / threshold $\rho$ | 1.0 / 0.3 |
| Diffuse-map discard (peak-to-mean) | $< 3.0$ |

### B.2 Evaluation Prompts

**Prompts of benchmarks.** For reproducibility, we list the complete prompts used under both evaluation protocols. MME-RealWorld-CN uses the Chinese translation of the EN template, and MME-RealWorld-Lite reuses the same template under the training-aligned protocol with the trailing short-answer line.

**V* Bench / HR-Bench (main protocol)**

```
<image>
{question} Select from the following choices.
{options}
```

**ZoomBench (main protocol)**

```
<image>
{question}
{options}
```

**MME-RealWorld EN (main protocol)**

```
<image>
{question} The choices are listed below:
{options}
Select the best answer to the above multiple-choice question based
on the image.  Respond with only the letter (A, B, C, D, or E) of the
correct option.
The best answer is:
```

**V* Bench (training-aligned protocol)**

```
<image>
{question}
{options}
Answer with the option's letter from the given choices directly.
```

**HR-Bench (training-aligned protocol)**

```
<image>
{question}
{options}
Answer the option letter directly.
```

**ZoomBench (training-aligned protocol)**

```
<image>
{question}
{options}
Answer with the option's letter from the given choices.
```

**InfoVQA (training-aligned protocol)**

```
<image>
{question}
Answer the question using a single word or phrase.
```

### B.3 Per-Benchmark Token Curves

Figure 6 breaks the six-benchmark average of Figure 4 down into per-benchmark accuracy-versus-token curves. Our model lies above both the base model and the SD-RPN pipeline at every source-image token limit on all six benchmarks and at both scales.

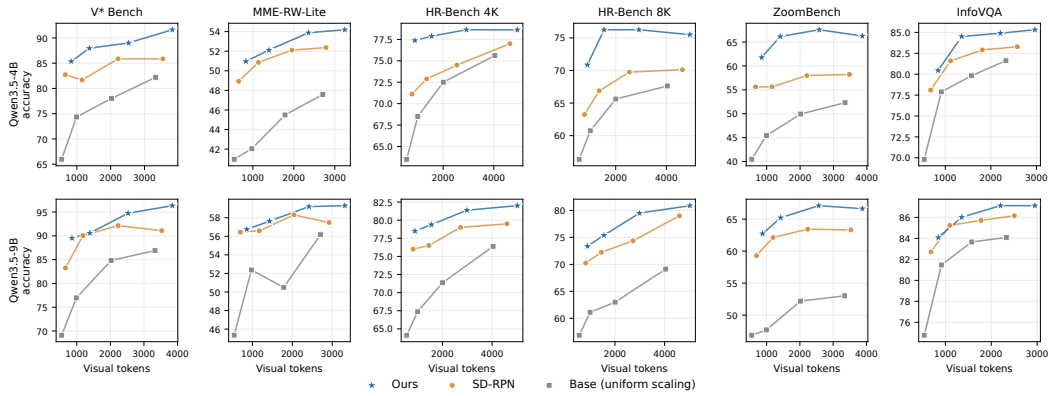


Figure 6: Per-benchmark accuracy versus measured visual tokens per sample (source plus crop) for source-image token limits $\{576, 1024, 2048, 4096\}$ under the training-aligned short-answer protocol (top row Qwen3.5-4B, bottom row Qwen3.5-9B).

### B.4 Generalization to an Encoder-Free Backbone

Gemma-4-12B (Team et al., 2026) has no vision transformer: images are cut into $48 \times 48$ patches that a linear embedder maps directly into the language model at one of five visual-token tiers (70 to 1,120 tokens), and the processor always fills the chosen tier. We follow the same pipeline as for the Qwen backbones with the tier as the only budget knob. The SD-RPN predictor is trained with the Qwen3.5 recipe on the same corpus regenerated by Gemma-4 at the 560-token tier, RL uses this predictor's own 7K pool at the same tier, and the source image is encoded at the largest tier (1,120 tokens) at test time. The RoI crop is encoded at the tier that quantizes the crop budget of Sec. 3.4, and sparse visual encoding drops background crop tokens before the embedder.

### B.5 Comparison with Visual Token Pruning

Visual token pruning reduces inference cost by selecting a subset of the encoded tokens, and is the main alternative route to visual-token efficiency. We compare against VisionZip (Yang et al., 2025a) and DART (Wen et al., 2025), the two methods with released Qwen2.5-VL implementations, on Qwen2.5-VL-7B. All methods run in the same evaluation harness with identical prompts, using each method's released code. Every method and benchmark uses one uniform pixel range, from $4$ to $4{,}096$ visual tokens at native aspect ratio.

**Token accounting.** The reference point is the base model under a $4{,}096$-token source limit ($100\%$). The pruning methods encode this full-budget source and then retain a fixed fraction of each sample's tokens. VisionZip keeps dominant-plus-contextual tokens after the vision encoder and DART prunes to the same nominal fraction at the second LLM layer, so their $50\%$ and $25\%$ budgets are exact by construction. Our method instead re-allocates resolution before encoding. On the high-resolution suite we report the $1{,}024$- and $576$-token rungs of the source ladder, whose measured totals (source plus RoI crop) average $41.2\%$ and $24.7\%$ of the reference. On the document suite, where native sizes vary widely, we downsample the source to $\mathrm{src}/a$ ($a{=}3, 6$) and cap the crop at half the downsampled source, so the per-sample total is close to $1.5\,\mathrm{src}' \approx 50\%$ and $25\%$ of native, with measured means of $46.6\%$ and $27.2\%$.

Table 4 reports accuracy with per-benchmark retention relative to the full-budget reference, and Figure 7 plots the suite averages against the retained-token budget. At the $50\%$ tier, VisionZip is nearly lossless on the high-resolution suite, consistent with its published claims. Our model does not merely retain the reference accuracy but exceeds it at both tiers. On text-dense benchmarks pruning collapses. At the $25\%$ tier VisionZip retains $73.3\%$ on InfoVQA and $56.5\%$ on OCRBench, and DART falls further, while our method holds $94$–$99\%$ on both. The one cell where pruning wins is ChartQA at the $25\%$ tier, where chart understanding depends on the global layout, so at one-sixth source resolution uniform token selection preserves more structure than region re-allocation. Finally, the two approaches spend compute differently. Pruning still pays the full vision encoding and saves only LLM-side tokens, whereas our routing reduces both the encoder and LLM budgets.

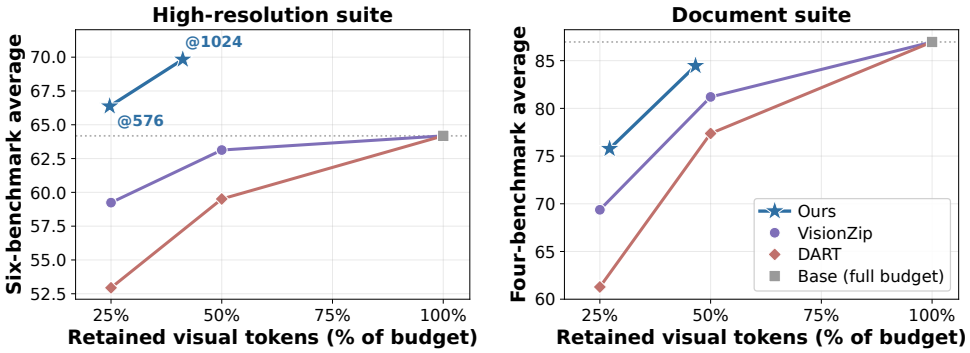


Figure 7: Suite-average accuracy versus retained visual-token budget for the pruning comparison of Table 4. Pruning points sit at their exact keep fractions and converge to the base model at $100\%$. Our points sit at their measured totals. On the high-resolution suite our curve lies above the full-budget reference at every budget.

### B.6 Qualitative Comparison

Figure 8 illustrates how the two predictors differ on cases where their answers diverge. The supervised SD-RPN exhibits the two failure modes of its pseudo-label surrogate. It scatters probability over question-irrelevant regions and covers the evidence only partially, so the reader is fed the wrong tokens and answers incorrectly. After region-level RL refinement, the same predictor suppresses the spurious regions and retains the required evidence, from the sundial plaque and the gated archways to all four bridge pillars, so the reader answers correctly at the same token budget. This is the intended effect of the objective. Rewards are assigned to regions by their functional contribution to

Table 4: Comparison with visual token pruning on Qwen2.5-VL-7B under a shared 4,096-token source limit. Gray subrows give accuracy retention relative to the full-budget base model. Pruning budgets are exact per-sample keep fractions. Ours are measured source-plus-crop totals, quoted per tier in the group headers.

| Method | V* Bench | ZoomBench | HR-Bench 4K | HR-Bench 8K | MME-RW Lite | InfoVQA | Avg. ret. |
|---|---|---|---|---|---|---|---|
| *High-resolution suite* | | | | | | | |
| Base | 77.49 | 46.86 | 72.50 | 64.00 | 42.31 | 81.89 | 100% |
| *Retain 50% (ours measured at 41.2%)* | | | | | | | |
| VisionZip | 79.06 | 44.62 | 72.00 | 63.75 | 43.04 | 76.28 | 98.4% |
| | 102.0% | 95.2% | 99.3% | 99.6% | 101.7% | 93.1% | |
| DART | 78.53 | 45.21 | 68.75 | 60.50 | 39.19 | 64.88 | 92.7% |
| | 101.3% | 96.5% | 94.8% | 94.5% | 92.6% | 79.2% | |
| **Ours** | 85.86 | 54.79 | 76.50 | 70.87 | 47.79 | 83.16 | **108.8%** |
| | 110.8% | 116.9% | 105.5% | 110.7% | 113.0% | 101.6% | |
| *Retain 25% (ours measured at 24.7%)* | | | | | | | |
| VisionZip | 81.68 | 43.43 | 69.25 | 61.38 | 39.66 | 60.02 | 92.3% |
| | 105.4% | 92.7% | 95.5% | 95.9% | 93.7% | 73.3% | |
| DART | 73.82 | 42.37 | 65.00 | 58.00 | 33.82 | 44.61 | 82.5% |
| | 95.3% | 90.4% | 89.7% | 90.6% | 79.9% | 54.5% | |
| **Ours** | 80.10 | 51.95 | 72.00 | 66.38 | 46.59 | 81.18 | **103.4%** |
| | 103.4% | 110.9% | 99.3% | 103.7% | 110.1% | 99.1% | |

| Method | OCRBench | ChartQA | DocVQA | TextVQA | Avg. ret. |
|---|---|---|---|---|---|
| *Document suite* | | | | | |
| Base | 85.50 | 84.48 | 94.86 | 82.98 | 100% |
| *Retain 50% (ours measured at 46.6%)* | | | | | |
| VisionZip | 70.20 | 78.96 | 94.03 | 81.62 | 93.4% |
| | 82.1% | 93.5% | 99.1% | 98.4% | |
| DART | 70.10 | 70.84 | 87.96 | 80.57 | 89.0% |
| | 82.0% | 83.9% | 92.7% | 97.1% | |
| **Ours** | 80.60 | 79.92 | 94.92 | 82.38 | **97.1%** |
| | 94.3% | 94.6% | 100.1% | 99.3% | |
| *Retain 25% (ours measured at 27.2%)* | | | | | |
| VisionZip | 48.30 | 66.80 | 86.91 | 75.46 | 79.8% |
| | 56.5% | 79.1% | 91.6% | 90.9% | |
| DART | 51.60 | 53.56 | 68.70 | 71.27 | 70.5% |
| | 60.4% | 63.4% | 72.4% | 85.9% | |
| **Ours** | 70.20 | 60.60 | 94.03 | 78.25 | **87.1%** |
| | 82.1% | 71.7% | 99.1% | 94.3% | |

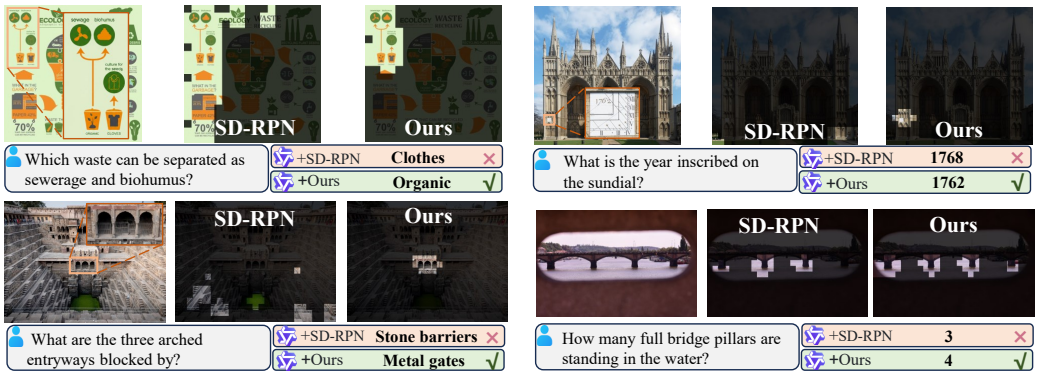


Figure 8: RoI predictions of SD-RPN and ours. Dimmed regions are the discarded tokens.

the answer, so the predictor learns to keep exactly the evidence the answer depends on rather than whatever correlates with the teacher's attention.

## C LOCALIZATION-VS-RECOGNITION PROBE: PROTOCOL AND CONTROLS

**Setup.** The probe uses the Qwen3.5-4B base model with no RoI predictor attached, on ZoomBench (Wei et al., 2026), whose items are four-way multiple-choice and blank-answer questions. The mixed chance level of this pool is the shaded band in Fig. 1b. The model itself serves as the locator. Given the question and an instruction to output the evidence box, it decodes a bounding box greedily, with the assistant reply prefilled by the JSON prefix `{"bbox_2d": [` so that every response is parseable. We restrict the pool to samples that are answerable from the ground-truth-box crop, and remain so under a $15\%$ area-preserving jitter of that box, but are not confidently solvable from the question text alone under a first-token margin filter. This leaves $|E|{=}276$ samples and removes the text-prior contamination that would otherwise grow as image information vanishes. The anchor run at $r{=}1$ presents the native-resolution scene. We record the predicted box and keep the $n{=}238$ eligible samples that emit a parseable box and are answered correctly from the model's own predicted RoI. Survival at compression $r$ is the probability of a correct answer at $r$ given a correct answer at the anchor, measured on this fixed set.

**Ladders.** The localization ladder downsamples the scene by $r$, so that its visual tokens scale as $1/r^2$ with no processor floor, re-predicts the box from the degraded scene, and reads the answer from the corresponding crop at native resolution. The recognition ladder freezes the anchor box, excludes the background, and downsamples the crop itself by $r$ before encoding. The processor floors the crop at about six tokens from $r{\geq}3$ onward, so recognition survival at the deep rungs is, if anything, optimistic.

**Fixed-crop-budget control.** Along the localization ladder the re-predicted box grows with $r$, reaching a median area ratio of $7.4\times$ at $r{=}6$, and a larger box buys a larger native-resolution crop. To rule out this budget effect, a control re-runs the ladder with the crop hard-capped at $128$ tokens at every rung (Table 5). Even at a fixed crop budget, localization retains an $18.2$-point advantage over recognition at $r{=}6$.

Table 5: Survival (%) on the $n{=}238$ anchor-correct set under scene compression (localization, with and without the crop cap) and crop compression (recognition).

| $r$ | 1.5 | 2 | 2.5 | 3 | 4 | 5 | 6 | 10 |
|---|---|---|---|---|---|---|---|---|
| Localization (uncapped) | 93.7 | 91.2 | 91.2 | 84.0 | 84.0 | 82.8 | 82.4 | 76.9 |
| Localization (crop $\leq$128 tok) | 92.8 | 91.1 | 90.3 | 84.7 | 79.2 | 76.3 | 69.5 | 55.0 |
| Recognition | 89.9 | 79.0 | 64.3 | 59.2 | 56.7 | 56.3 | 51.3 | 44.5 |

**Box behavior.** Along the localization ladder the re-predicted box loses precision but keeps containing the evidence. Across the rungs from $r{=}1.5$ to $r{=}6$, median IoU against the anchor box falls as $0.73/0.62/0.54/0.42/0.23/0.16/0.11$, while median ground-truth coverage rises from $0.44$ to $1.00$ and the hit rate, defined as coverage of at least $0.5$, rises from $47\%$ to $75\%$. Localization survival therefore reflects a graceful soft failure, an imprecision that downstream cropping recovers, rather than precision robustness.

## D LIMITATIONS

Our design deliberately keeps the reader frozen, so it cannot repair recognition failures that finetuning the full model can, which is visible on MME-RealWorld at the smallest scale. The compression-asymmetry probe is a diagnostic on one backbone and one benchmark rather than a general law, and training requires question–answer pairs to score contributions, though no region annotations.