FOCUS WHEN NECESSARY: ADAPTIVE ROUTING AND COLLABORATIVE GROUNDING FOR TRAINING-FREE VISUAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

While Multimodal Large Language Models (MLLMs) excel in cross-modal reasoning, they often struggle to perceive fine-grained details in complex high-resolution images. Recent training-free methods address this through image scaling and localized cropping. However, applying these manipulations indiscriminately introduces computational redundancy for simple queries and can degrade accuracy by truncating essential global context or introducing irrelevant background noise. To this end, we propose **LazyMCoT**, a dynamic and training-free framework that adaptively allocates visual grounding efforts based on sample difficulty. The framework features an ***Adaptive Routing*** mechanism that evaluates predictive uncertainty using first-token statistics from a single forward pass. This efficiently bypasses confident cases while ensuring the recall of difficult samples via conformal calibration. For these challenging cases, a ***Collaborative Grounding*** module integrates the inherent cross-modal attention of the model with an external visual expert through a two-stage refinement process. This refinement process generates a precise localized display to recover small or occluded targets. Extensive experiments across diverse benchmarks demonstrate that LazyMCoT rivals training-based approaches by simultaneously improving reasoning accuracy and reducing average inference latency.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have recently achieved unprecedented success across vision-language tasks Li et al. (2023b); Dai et al. (2023); Liu et al. (2023); Bai et al. (2023). While early models with fixed resolution encoders struggle to capture fine details in complex images Li et al. (2023b); Zhang et al. (2023); Liu et al. (2024a), recent advancements resolve this limitation through dynamic resolution mechanisms Zhu et al. (2025); Bai et al. (2025b); Wang et al. (2025a); Bai et al. (2025a); Liu et al. (2024b). Building on these advanced base models, training-free visual grounding methods have emerged as a promising paradigm Wang et al. (2025c); Shen et al. (2025); Li et al. (2025). By utilizing visual experts, search algorithms, or attention decoupling, these methods extract key patches and perform localized cropping to enhance the model's perception of subtle visual evidence without the need for expensive retraining Li et al. (2026); Liu et al. (2025); Khayatkhoei et al. (2025); Morini et al. (2026).

Although effective on challenging instances, current visual grounding methods without additional training exhibit a critical flaw by indiscriminately applying complex visual manipulations to all samples Shen et al. (2025); Li et al. (2025); Gröpl et al. (2026). Empirical observations reveal that this uniform strategy is highly suboptimal. As shown in Fig. 2 and Fig. 3, it wastes significant computational resources since standard VLMs can accurately answer a large fraction of queries in a single forward pass. More importantly, forcibly applying localized cropping to straightforward samples can severely degrade performance. By truncating essential global context and introducing irrelevant background noise, redundant visual processing misguides the model and causes failures on otherwise solvable cases. This issue is particularly detrimental in tasks demanding complex reasoning, where blind grounding reduces the overall accuracy of VLMs.

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.

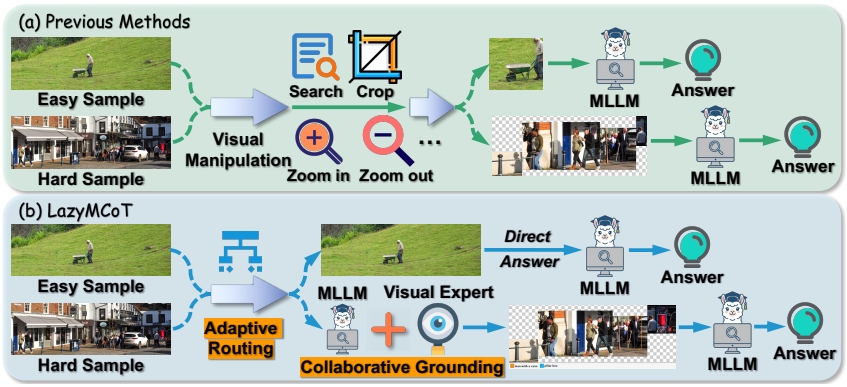


Figure 1: **LazyMCoT allocates visual grounding effort by sample difficulty.** (a) Previous training-free methods indiscriminately apply heavy visual manipulation to every sample. (b) LazyMCoT instead routes easy samples to a direct answer and dispatches only hard samples through Collaborative Grounding before re-querying the MLLM.

Motivated by these observations, we argue that visual grounding should be applied selectively. We hypothesize that the base VLM's initial predictive uncertainty is a strong indicator of whether further visual exploration is required. Through statistical analysis, we discover that zero-cost features extracted from the first answer token's logits, namely the option top probability and the option-versus-non-option logit gap, exhibit a strong monotonic correlation with the model's predictive entropy. These statistics cleanly separate the samples that the base VLM can solve directly from those that genuinely require dense visual grounding.

To translate this insight into an actionable solution, we propose **LazyMCoT**, a novel training-free framework that dynamically allocates visual grounding effort based on sample difficulty. LazyMCoT consists of two core components, **Adaptive Routing** and **Collaborative Grounding**. The Adaptive Routing uses a lightweight classifier to evaluate first-token statistics. It is fitted once on an external corpus disjoint from all evaluation benchmarks, with its threshold calibrated via conformal prediction Shafer & Vovk (2008) to guarantee a controlled recall on difficult samples. When confident, the router directly outputs the answer, bypassing further computation. Otherwise, the query is routed to the Collaborative Grounding module. As shown in Fig. 1, unlike previous methods that rely on inflexible search strategies Shen et al. (2025); Li et al. (2025), our grounding module couples an attention-driven branch derived from the VLM's cross-modal attention with an external visual expert Carion et al. (2025); Liu et al. (2024c) branch. Through a two-stage parallel detection and refinement process, it precisely extracts fine-grained evidence containing key targets, and renders a Localized Panel Display, which is then fed back to the VLM for reasoning.

We conduct extensive experiments on multiple challenging benchmarks using open-source VLM backbones. LazyMCoT consistently outperforms existing training-free methods and even matches or exceeds recent training-based grounding approaches. It not only achieves significant accuracy gains on fine-grained localization tasks, but also prevents the performance degradation commonly observed in other grounding methods on reasoning-heavy tasks. Furthermore, by short-circuiting easy samples, LazyMCoT significantly reduces the average end-to-end inference latency.

In summary, our main contributions are threefold:

- We identify and empirically validate the inherent limitations of indiscriminate visual grounding in current training-free methods, demonstrating that redundant visual manipulation may hurt performance on easy samples and wastes computational resources.
- We propose **LazyMCoT**, a dynamic framework featuring an **Adaptive Routing** that leverages zero-cost first-token statistics with conformal-calibrated decision rules to selectively trigger visual grounding, and a **Collaborative Grounding** module that synergizes VLM attention with visual experts for precise evidence extraction.
- Extensive experiments demonstrate that LazyMCoT achieves competitive performance among training-free methods across multiple benchmarks and VLMs, improving accuracy and reducing inference latency.

2 Related Work

Multimodal Large Language Models. Multimodal Large Language Models (MLLMs) have demonstrated remarkable potential in cross-modal tasks. Among these, Vision-Language Models (VLMs) have advanced with particular rapidity. However, early models Li et al. (2023b); Dai et al. (2023); Liu et al. (2023); Bai et al. (2023); Liu et al. (2024a) relying on Q-Former Li et al. (2023b) or trainable visual adapters Zhang et al. (2023) often lose fine-grained details due to their fixed-resolution visual encoders. To address complex visual reasoning, dynamic resolution MLLMs Zhu et al. (2025); Bai et al. (2025b); Gao et al. (2024); Chen et al. (2024); Liu et al. (2024b) have emerged to preserve spatial precision for high-resolution inputs. For instance, InternVL-3.5 Wang et al. (2025a) utilizes dynamic slicing for high resolution images, while Qwen3-VL Bai et al. (2025a) adopts DeepStack multi-layer visual injection for enhanced alignment. Building upon these advancements, we propose a training-free method that leverages the inherent capabilities of VLMs to enhance their perception and expand their performance boundaries.

Training-Free Visual Grounding. Globally uniform feature extraction often wastes computational resources and obscures critical details within background noise. Consequently, recent studies emphasize training-free visual grounding and dynamic focusing mechanisms. For example, RAP Wang et al. (2025c) pioneers spatially aware layouts, while ZoomEye Shen et al. (2025) and DyFo Li et al. (2025) utilize search algorithms for efficient visual navigation. Other approaches leverage visual experts for bottom-up evidence extraction Li et al. (2026), entropy gradients for region guidance Gröpl et al. (2026), or attention decoupling for key patch extraction Liu et al. (2025); Khayatkhoei et al. (2025); Morini et al. (2026). Unlike previous methods relying on inflexible search strategies, our proposed approach integrates attention guidance with visual experts to extract fine-grained evidence much more precisely. Additionally, we introduce a dynamic routing mechanism to adaptively schedule inference paths based on varying sample difficulty.

3 Motivation

3.1 Limitations of Existing Training-Free Visual Grounding Methods

Recent advancements Li et al. (2026); Liu et al. (2025) in training-free visual grounding heavily rely on image scaling and localized cropping to enhance the perception of fine-grained details. However, observations reveal that these operations are not universally beneficial across all instances. A substantial portion of straightforward samples can be accurately resolved by original VLMs without additional visual manipulation. We evaluated on a unified benchmark comprising V* Bench Wu & Xie (2024), HR-Bench 4K/8K Wang et al. (2025b), and TreeBench Wang et al. (2026). As shown in Fig. 2, an average of **67.17%** of the samples can be correctly answered relying solely on the vanilla VLM inference.

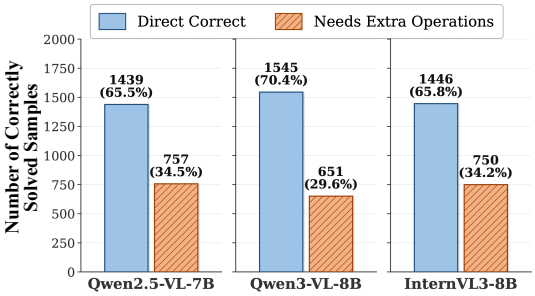


Figure 2: **Vanilla VLM inference successfully solves most samples.** Breakdown of samples solved without extra visual operations (**Direct Correct**, blue) vs. those requiring additional grounding (**Needs Extra Operations**, orange) for each VLM across four benchmarks.

Furthermore, forcibly applying these complex operations to such simple cases can introduce unnecessary background noise or truncate essential global context. As shown in Fig. 3, this redundant processing occasionally misguides the model and leads to incorrect predictions for samples that the base VLM would have otherwise answered correctly.

3.2 Statistical Features for Sample Routing

Motivated by these observations, we propose selective visual grounding based on sample difficulty. We hypothesize that the base VLM's initial predictive uncertainty indicates the need for further visual exploration. To verify this, we extract two zero-cost statistics from a single forward pass.

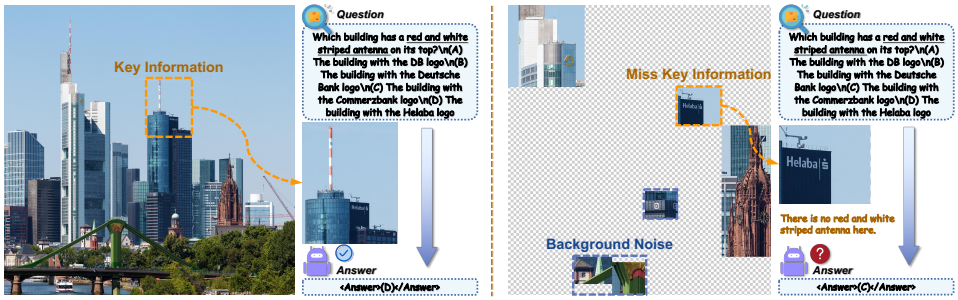


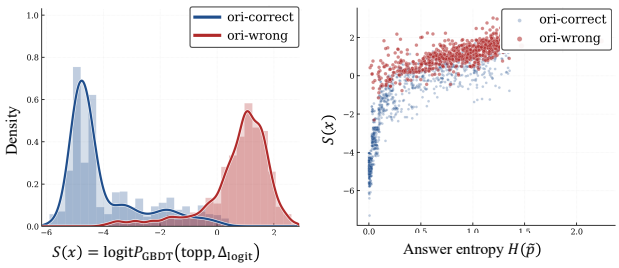
Figure 3: **Indiscriminately applied visual grounding can hurt easy samples.** (a) On the original image, the base VLM directly localizes the key region and answers correctly. (b) In existing methods, after image scaling and localized cropping, the truncated context omits the key information while introducing background noise indicated by blue box.

Let $\mathbf{z} \in \mathbb{R}^V$ denote the first answer token's logits, and $\mathcal{O}$ be the candidate option indices. Given the full vocabulary distribution $p_i = \mathrm{softmax}(\mathbf{z})_i$ and the renormalized option-restricted distribution $\tilde{p}$, we define the option top probability (topp) and option-versus-non-option logit gap (Δ_{logit}) as:

$$\text{topp} = \max_{i \in \mathcal{O}} \tilde{p}_i, \quad \Delta_{\text{logit}} = \max_{i \in \mathcal{O}} z_i - \max_{j \notin \mathcal{O}} z_j. \tag{1}$$

Intuitively, topp measures probability concentration on the top option, while Δ_{logit} reflects the model's confidence in choosing a valid option over other tokens. Unlike uncertainty metrics such as answer entropy, these statistics directly capture the option/non-option decision boundary, offering superior routing discrimination. To obtain a comparable scalar, we train a Gradient Boosting Decision Tree Friedman (2001) g_θ on a held-out set to classify $\mathbf{x} = (\text{topp}, \Delta_{\text{logit}})$ as ori-correct ($y = 0$) or ori-wrong ($y = 1$). The routing score is the logit of the predicted ori-wrong probability $\hat{p}(\mathbf{x}) = g_\theta(\mathbf{x})$:

$$s(x) = \log \frac{\hat{p}(\mathbf{x})}{1 - \hat{p}(\mathbf{x})}. \tag{2}$$



(a) Routing score distribution (b) Score versus answer entropy

Figure 4: **Statistical separability of ori-correct & ori-wrong samples on Qwen2.5-VL-7B.** (a) First-token routing score $s(x)$ shows distinct modes per class. (b) $s(x)$ correlates monotonically with answer entropy $H(\tilde{p})$, with ori-wrong samples clustering in the high-entropy region.

We evaluate this router on the pooled benchmarks from Sec. 3.1 alongside the answer entropy $H(\tilde{p}) = -\sum_{i \in \mathcal{O}} \tilde{p}_i \log \tilde{p}_i$, ensuring the subsequent results reflect true generalization instead of training fit. As shown in Fig. 4, routing scores separate ori-correct and ori-wrong samples (Fig. 4a), validating topp and Δ_{logit} as reliability indicators. Predictive entropy correlates monotonically with routing scores (Fig. 4b), peaking for ori-wrong samples. Thus, these statistics are ideal for a lightweight router to selectively trigger visual grounding. Please see Appendix Sec. C for these statistics on additional VLMs.

4 Method

4.1 Overview

Building on the observations in Sec. 3, we propose **LazyMCoT**, a training-free framework that dynamically allocates visual grounding effort based on sample difficulty. As illustrated in Fig. 5, given an image I and a multiple-choice question Q, first, a single-token forward pass yields a direct answer and statistics $\mathbf{x} = (\text{topp}, \Delta_{\text{logit}})$. An *Adaptive Routing* module evaluates $\mathbf{x}$ and returns the direct

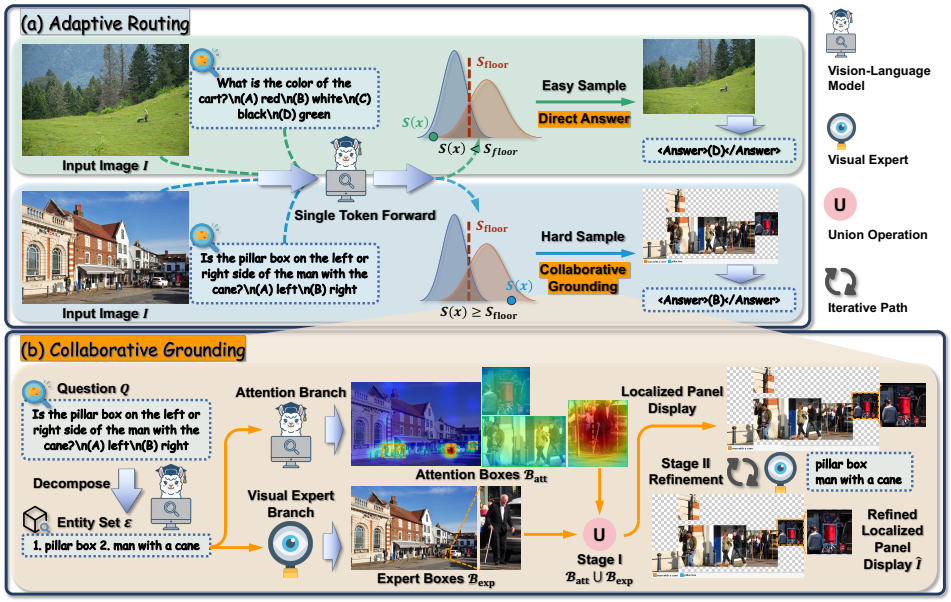


Figure 5: **Overview of the proposed LazyMCoT framework.** (a) The Adaptive Routing utilizes first-token statistics from a single forward pass to dynamically bypass simple cases or route hard samples. (b) Collaborative Grounding integrates an attention branch and a visual expert to construct a localized panel display for precise VLM re-querying.

answer immediately if the routing score $s(x) < s_{\text{floor}}$. Otherwise, a *Collaborative Grounding* module coupling attention and visual expert branches generates a Localized Panel Display (LPD) Liu et al. (2025) via a two-stage detection. This LPD is then used to re-query the VLM to obtain the final answer. This design selectively applies dense grounding to hard samples while preserving zero-shot efficiency on easy ones.

4.2 Adaptive Routing

The motivation of adaptive routing is to convert the empirical observations in Sec. 3.2 into an automatic decision rule that triggers Collaborative Grounding only when the base VLM is uncertain. Given the first-token feature vector $\mathbf{x}$, we obtain the routing score $s(x)$ in Eqn. 2 via a Gradient Boosting Decision Tree Friedman (2001) that is trained once on a held-out routing set $\mathcal{D}_{\text{cal}}$ with labels $y \in \{0, 1\}$ for ori-correct and ori-wrong respectively.

External routing set. $\mathcal{D}_{\text{cal}}$ comprises high-resolution multiple-choice questions from MME-RealWorld-Lite Zhang et al. (2024) and MagnifierBench Li et al. (2023a). We further enforce perceptual-hash deduplication against all evaluation images. The GBDT remains frozen during inference, introducing zero additional parameters at deployment. Please see Sec. A.3 for more details.

Conformal threshold calibration. A naive choice of decision threshold may either skip too many ori-wrong samples or run grounding on too many ori-correct ones. To make the routing behavior controllable, we adopt a conformal calibration on the must-recall subset $\mathcal{D}_{\text{mr}} \subseteq \mathcal{D}_{\text{cal}}$, which is defined as the set of ori-wrong samples within $\mathcal{D}_{\text{cal}}$ that are actually corrected by the grounding module. Let $\{s_i\}_{i \in \mathcal{D}_{\text{mr}}}$ be the out-of-fold scores produced by the GBDT. For a target miscoverage rate $\alpha \in [0, 1)$, the routing threshold is set to:

$$s_{\text{floor}} = Q_\alpha\big(\{s_i\}_{i \in \mathcal{D}_{\text{mr}}}\big), \tag{3}$$

where $Q_\alpha(\cdot)$ denotes the empirical α-quantile. By construction, at most an α fraction of must-recall samples falls below s_{floor}, guaranteeing a controlled lower bound on the recall of difficult samples. Smaller α yields a lower threshold and a more conservative router that triggers grounding more often, while larger α allows aggressive skipping of easy samples.

Routing rule. Let $\text{Direct}(I, Q)$ denote the direct answer obtained from the single-token forward pass and let $\text{CG}(I, Q)$ denote the answer produced by feeding the LPD $\hat{I}$ back to the VLM. The

Algorithm 1 Collaborative Grounding

Input: image I, question Q, base VLM f, visual expert $\mathcal{S}$
Parameter: kernel bandwidth σ, threshold τ
Output: localized panel display $\hat{I}$

Decompose Q into entity set $\mathcal{E} = \{e_1, \ldots, e_M\}$ via f.
$\mathcal{B}_{\text{exp}} \leftarrow \mathcal{S}(I, \mathcal{E})$ *(visual expert branch)*
Run f with prompt "Search $\mathcal{E}$ in I" and record cross-modal attention A.
Compute aggregated saliency $\mathcal{A}(I)$ by Eqn. 5.
Threshold $\mathcal{A}(I)$ at τ to obtain attention boxes $\mathcal{B}_{\text{att}}$.
$\mathcal{B}^{(1)} \leftarrow \mathcal{B}_{\text{att}} \cup \mathcal{B}_{\text{exp}}$.
$\Delta\mathcal{B} \leftarrow \emptyset$
for each $b \in \mathcal{B}^{(1)}$ not covered by $\mathcal{B}_{\text{exp}}$ **do**
    $I_b \leftarrow$ crop of I on the enlarged region of b.
    $\Delta\mathcal{B} \leftarrow \Delta\mathcal{B} \cup \mathcal{S}(I_b, \mathcal{E})$.
end for
$\mathcal{B}^{(2)} \leftarrow \mathcal{B}^{(1)} \cup \Delta\mathcal{B}$
Render $\mathcal{B}^{(2)}$ on I with color borders and legends to obtain $\hat{I}$.
return $\hat{I}$

final prediction of LazyMCoT is:

$$\hat{y} \;=\; \begin{cases} \text{Direct}(I, Q), & s(x) < s_{\text{floor}}, \\ \text{CG}(I, Q), & s(x) \geq s_{\text{floor}}. \end{cases} \tag{4}$$

Because $s(x)$ is computed from the same forward pass as the direct answer and $\mathcal{D}_{\text{cal}}$ is routing-data only, no additional inference cost is introduced for skipped samples and no training of the base VLM is required. This makes adaptive routing a lightweight plug-in that can be paired with any VLM.

4.3 Collaborative Grounding

Entity decomposition and parallel detection. For routed difficult samples, we first prompt the VLM with a rule-based template Liu et al. (2025) to decompose Q into a list of canonical entities $\mathcal{E} = \{e_1, \ldots, e_M\}$. With $\mathcal{E}$ as queries, we run two complementary detectors in parallel. The *visual expert branch* feeds $\mathcal{E}$ into SAM3 Carion et al. (2025) $\mathcal{S}$ to obtain a set of expert boxes $\mathcal{B}_{\text{exp}}$. The *attention branch* appends the prompt "Search the following entities in the images: $\mathcal{E}$" to the VLM input and records the cross-modal attention $A \in \mathbb{R}^{T \times N}$, where T is the number of entity tokens and N is the number of visual tokens. Per entity token, the attention map is reshaped to the spatial grid, smoothed by a Gaussian kernel of bandwidth σ, normalized, and aggregated across entity tokens into a single saliency map:

$$\mathcal{A}(I) \;=\; \frac{1}{T} \sum_{t=1}^{T} \mathcal{N}\big(g_\sigma * A_t\big), \tag{5}$$

where $\mathcal{N}(\cdot)$ rescales the map to $[0, 1]$. Connected components of the saliency map above a relative threshold τ on $\mathcal{A}(I)$ are converted into the attention boxes $\mathcal{B}_{\text{att}}$.

Two-stage refinement. We observe that $\mathcal{B}_{\text{exp}}$ tends to recall the most salient instances but may miss small or occluded ones, whereas $\mathcal{B}_{\text{att}}$ covers question-relevant regions but often yields coarse and imprecise boundaries. To exploit their complementarity, we couple the two sources by a two-stage detection procedure. In the first stage, we take the union $\mathcal{B}^{(1)} = \mathcal{B}_{\text{att}} \cup \mathcal{B}_{\text{exp}}$ as a coarse evidence pool. In the second stage, for each box $b \in \mathcal{B}^{(1)}$ that is not already covered by $\mathcal{B}_{\text{exp}}$, we crop I to the slightly enlarged region of b and re-query SAM3 with the same entities. The newly discovered boxes $\Delta\mathcal{B}$ inside the crop are mapped back to image coordinates and appended to the evidence pool. The refined box set $\mathcal{B}^{(2)} = \mathcal{B}^{(1)} \cup \Delta\mathcal{B}$ is finally rendered as a localized panel display $\hat{I}$, in which each region is assigned a color and a textual legend so that the VLM can reason over multiple evidence patches in one forward pass. The complete procedure is summarized in Alg. 1.

5 EXPERIMENTS

5.1 EXPERIMENT SETTINGS

Datasets & Metrics. We systematically evaluate our proposed method on three challenging benchmarks. (1) **V* Bench** Wu & Xie (2024), which contains 191 images with an average resolution of 2246×1582 and focuses on *Direct Attribute* recognition (Att.) and *Spatial Relationship* reasoning (Spa.). (2) **HR-Bench** Wang et al. (2025b), which provides 4K and 8K resolution images that are evenly split into *Single-Instance* (Sin.) and *Cross-Instance* (Cro.) perception tasks. (3) **TreeBench** Wang et al. (2026), comprising 405 images with an average resolution of 2152×1615, which covers fine-grained perception (e.g., Material Recognition, OCR) and multi-step reasoning (e.g., Occlusion, Comparative Analysis). Multiple-choice accuracy is adopted as the primary evaluation metric across all benchmarks.

Implementation Details. We employ SAM3 Carion et al. (2025) as the visual expert, setting the maximum number of detected objects to $k = 10$ to balance performance and latency. LazyMCoT is evaluated across three different VLMs: Qwen2.5-VL-7B Bai et al. (2025b), Qwen3-VL-8B Bai et al. (2025a), and InternVL3-8B Zhu et al. (2025). To eliminate manual tuning, we train a Gradient Boosting Decision Tree (GBDT) Friedman (2001) on the external routing set and calibrate its decision threshold via conformal prediction on out-of-fold scores, guaranteeing high recall on mispredicted samples. Please refer to Appendix Sec. A for further implementation details.

5.2 MAIN RESULTS

Results on HR-Bench and V*. As reported in Tab. 1, LazyMCoT delivers consistent gains over the base VLMs across all three backbones, raising the average accuracy by 3.3, 7.9, and 8.4 points on InternVL3-8B, Qwen3-VL-8B, and Qwen2.5-VL-7B, respectively. Among training-free methods, LazyMCoT ranks first on every aggregate column and surpasses the HiDe baseline on average. Notably, on the Qwen2.5-VL-7B, our training-free framework matches or exceeds recent training-based grounding methods, achieving the best V* average of 90.6% without any parameter update. The improvement is most pronounced on tasks that demand fine-grained spatial localization (V*-Spa. +17.1 points on Qwen2.5-VL-7B), confirming that our collaborative grounding effectively recovers small targets.

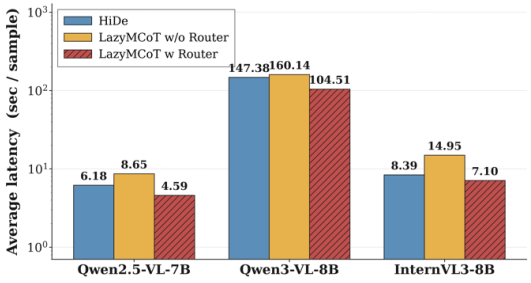


Figure 6: **Adaptive routing yields the lower end-to-end latency.** Average per-sample inference time on V* for three VLM backbones under three configurations: HiDe, w/o Adaptive Routing, and the full LazyMCoT. The first-token fast-path lets LazyMCoT skip confident samples.

Results on TreeBench. TreeBench probes perception and reasoning beyond pure localization, where indiscriminate grounding can disrupt models with already accurate reasoning chains. As shown in Tab. 2, on Qwen3-VL-8B HiDe degrades the average score from 43.0% to 41.7% because every sample is forced into the grounding pipeline, whereas LazyMCoT lifts it to 43.5% by routing only difficult cases through the visual expert. On the Qwen2.5-VL-7B backbone, LazyMCoT obtains 41.7% average accuracy with balanced gains over both Perception (+3.4 points vs. base) and Reasoning categories, outperforming all training-free competitors. These results confirm that adaptive routing is essential for benchmarks where blind grounding is detrimental.

Latency comparison. Fig. 6 reports the average per-sample wall-clock time on V* across three VLMs, comparing HiDe, w/o Adaptive Routing, and our full LazyMCoT. To ensure fair comparison, all experiments are conducted on a single NVIDIA H20 GPU with a batch size of 1. Adding collaborative grounding alone is slightly slower than HiDe due to the SAM3 verification stage, but enabling the router lets the framework short-circuit on samples whose first-token statistics already indicate confident answers, reducing average inference latency. Hence, LazyMCoT offers a balance of reasoning accuracy and efficiency for visual grounding.

Method	Training Free	HR-Bench-4K			HR-Bench-8K			V*		
		Sin.	Cro.	Avg.	Sin.	Cro.	Avg.	Att.	Spa.	Avg.
GPT-4o OpenAI (2024)	–	70.00	48.00	59.00	62.00	49.00	55.50	–	–	66.00
InternVL3-8B-Instruct										
Base Zhu et al. (2025)	–	82.80	**58.80**	70.80	80.00	**59.80**	69.90	81.70	78.90	80.60
ViCrop Khayatkhoei et al. (2025)	✓	88.00	57.00	72.50	82.80	54.80	68.80	88.70	75.00	83.30
HiDe† Liu et al. (2025)	✓	88.50	57.50	73.00	86.00	54.20	70.10	89.60	81.60	86.40
LazyMCoT (Ours)	✓	**89.50**	**58.80**	**74.10**	**89.80**	53.00	**71.40**	**89.70**	**83.90**	**87.30**
Δ *vs.* InternVL3-8B-Instruct	–	↑6.70	–	↑3.30	↑9.80	↓6.80	↑1.50	↑8.00	↑5.00	↑6.70
Qwen3-VL-8B-Instruct										
Base Bai et al. (2025a)	–	90.50	66.30	78.40	83.80	**65.00**	74.40	85.20	80.30	83.20
LazyMCoT (Ours)	✓	**94.30**	**66.50**	**80.40**	**90.80**	64.00	**77.40**	**92.20**	**89.50**	**91.10**
Δ *vs.* Qwen3-VL-8B-Instruct	–	↑3.80	↑0.20	↑2.00	↑7.00	↓1.00	↑3.00	↑7.00	↑9.20	↑7.90
Qwen2.5-VL-7B-Instruct										
Base Bai et al. (2025b)	–	88.80	54.80	71.80	84.20	51.50	67.90	80.90	76.30	79.10
DeepEyes Zheng et al. (2025)	✗	91.30	59.00	75.10	86.80	58.50	72.60	92.10	86.80	90.00
Thyme-VL Zhang et al. (2025)	✗	91.00	63.00	77.00	86.50	57.50	72.00	83.50	80.30	82.20
TreeVGR Wang et al. (2026)	✗	89.50	61.50	72.70	84.40	57.20	69.80	86.10	85.50	85.90
ViCrop Khayatkhoei et al. (2025)	✓	90.50	57.50	74.00	85.50	53.00	69.30	89.60	71.10	82.20
Dyfo Li et al. (2025)	✓	89.20	53.50	71.30	86.50	53.20	69.80	82.60	86.80	84.30
ZoomRefine Yu et al. (2025)	✓	88.50	55.30	71.50	83.90	54.00	68.60	85.30	77.60	82.20
DeepScan Li et al. (2026)	✓	90.10	**59.70**	75.00	87.20	**57.60**	72.40	93.00	86.80	**90.60**
HiDe† Liu et al. (2025)	✓	95.50	59.30	77.40	92.50	57.30	74.90	**94.80**	82.90	90.00
LazyMCoT (Ours)	✓	**96.20**	59.00	**77.60**	**93.80**	57.00	**75.40**	92.20	**88.20**	**90.60**
Δ *vs.* Qwen2.5-VL-7B-Instruct	–	↑5.70	↑1.50	↑3.60	↑8.30	↑4.00	↑6.10	↑2.60	↑17.1	↑8.40

Table 1: Quantitative results on HR-Bench and V*. For each backbone, the **best** result is in **bold** and the second-best is underlined. † denotes our reimplementation. LazyMCoT achieves competitive performance across all settings.

Method	Avg.	Perception						Reasoning				
		Attributes	Material	Phy. State	Obj. Retr.	OCR	Per. Trans.	Ordering	Con. & Oc.	Spa. Cont.	Comparison
GPT-4o OpenAI (2024)	46.90	51.70	61.50	65.20	43.80	69.10	18.80	38.60	48.80	72.40	43.20
Qwen2.5-VL-7B-Instruct											
Base Bai et al. (2025b)	37.00	55.20	53.80	**56.50**	62.50	27.90	20.00	**35.10**	39.00	44.80	43.20
Dyfo Li et al. (2025)	39.30	**58.60**	**69.20**	**56.50**	62.50	35.30	**21.20**	**35.10**	41.50	44.80	40.90
ZoomRefine Yu et al. (2025)	38.00	48.30	61.50	**56.50**	62.50	**39.70**	18.80	29.80	**46.30**	44.80	38.60
HiDe† Liu et al. (2025)	40.00	55.20	61.50	52.20	56.20	**39.70**	**21.20**	31.60	**46.30**	**48.30**	47.70
LazyMCoT (Ours)	**41.70**	**58.60**	61.50	**56.50**	**68.80**	**39.70**	**21.20**	31.60	**46.30**	**48.30**	**54.50**
Qwen3-VL-8B-Instruct											
Base Bai et al. (2025a)	43.00	**55.20**	**61.50**	**69.60**	**75.00**	**45.60**	**18.80**	28.10	41.50	69.00	50.00
HiDe† Liu et al. (2025)	41.70	48.30	46.20	65.20	**75.00**	**45.60**	15.30	24.60	**53.70**	65.50	**52.30**
LazyMCoT (Ours)	**43.50**	**55.20**	**61.50**	65.20	**75.00**	**45.60**	16.50	**29.80**	48.80	**72.40**	50.00

Table 2: Quantitative results on the TreeBench. The **best** results for each VLM are highlighted in **bold**, and the second-best is underline. † denotes our reimplementation. Our method effectively improves the perception capability.

5.3 Ablation Study

Effect of the two main components. Tab. 3 dissects Adaptive Routing (AR) and Collaborative Grounding (CG). Note that disabling AR leaves the SAM3 expert and Stage 2 re-query active on every sample, whereas HiDe relies solely on attention cropping without external models. Forcing CG on every sample lifts V* from 79.1% to 90.0% and HR-Bench-4K from 71.8% to 77.4%, but on TreeBench the gain is limited because indiscriminate grounding hurts easy reasoning samples. Plugging in AR further pushes V* to 90.6% and TreeBench to 41.7%, indicating that the two components are complementary. CG provides the precision needed for hard cases, while AR shields the base VLM from unnecessary grounding on easy ones.

Two-stage visual expert refinement. Tab. 4 ablates CG's inner structure. The attention branch ($\mathcal{B}_{\text{att}}$) and the visual expert branch ($\mathcal{B}_{\text{exp}}$) alone reach 80.3% and 85.9% on V* respectively, since each has its own failure mode, like noisy attention regions or missed small instances. Their Stage 1 union $\mathcal{B}^{(1)}$ lifts V* by 2.5 points over the stronger single source. Adding the Stage 2 refinement,



Figure 7: **Qualitative comparison on hard samples.** HiDe and LazyMCoT results are shown in the top and bottom rows. By recovering small or co-occurring targets missed by HiDe, our method provides more complete evidence for VLM re-querying.

AR	CG	V*-Avg.	HR-4K-Avg.	HR-8K-Avg.	TreeBench-Avg.
✗	✗	79.10	71.80	67.90	37.00
✗	✓	90.00	77.40	74.90	40.00
✓	✓	**90.60**	**77.60**	**75.40**	**41.70**

Table 3: Ablation study on Adaptive Routing (AR) and Collaborative Grounding (CG).

$\mathcal{B}_{\text{att}}$	$\mathcal{B}_{\text{exp}}$	Stage 2	V*-Att.	V*-Spa.	V*-Avg.
✓	✗	✗	84.30	76.30	80.30
✗	✓	✗	88.70	82.90	85.90
✓	✓	✗	91.30	85.50	88.40
✓	✓	✓	**92.20**	**88.20**	**90.60**

Table 4: Ablation study on the two-stage Collaborative Grounding pipeline on V* Bench.

α	s_{floor}	Skip Rate (%)	V*-Att.	V*-Spa.	V*-Avg.
0.00	-4.61	24.7	92.20	**88.20**	**90.60**
0.05	-0.74	51.8	93.30	86.20	90.20
0.10	-0.25	58.1	**94.10**	82.40	89.25
0.15	0.01	62.3	91.40	87.30	89.85
0.20	0.20	64.9	89.10	83.20	86.65
0.30	0.47	71.2	86.20	79.60	83.45
0.50	0.94	84.8	82.80	77.10	80.45
0.70	1.29	**94.2**	81.10	76.50	79.25

Table 5: Ablation study on the conformal miscoverage rate α. Results on V* with Qwen2.5-VL-7B.

which re-queries SAM3 inside enlarged attention crops to recover small or occluded targets, brings the final V* to 90.6%, validating that the two stages are necessary for accurate localization.

Effect of conformal miscoverage rate α. Tab. 5 sweeps the conformal miscoverage rate α that controls s_{floor} via Eqn. 3. A smaller α yields a lower threshold and a more conservative router that triggers grounding on most samples, while a larger α aggressively skips and falls back to the base VLM. The sweet spot at $\alpha = 0$ achieves the best V* of 90.6% while sparing 24.7% samples from the grounding pipeline, showing that conformal calibration offers a principled and tunable trade-off between accuracy and efficiency.

5.4 Case Study

Fig. 7 presents qualitative comparisons of the LPD produced by HiDe and our LazyMCoT on four representative hard samples from V* and HR-Bench. HiDe relies on attention-driven cropping and frequently misses the queried small targets (e.g., scooter, baby carriage), or recalls only a subset of the relevant instances when multiple objects co-occur. In contrast, LazyMCoT couples cross-modal attention with a visual expert and further refines the evidence pool through Stage 2 re-querying inside enlarged attention crops. Please refer to Appendix Sec. G and F for more qualitative results.

6 Conclusion

We presented **LazyMCoT**, a training-free framework allocating visual grounding effort based on sample difficulty. Its **Adaptive Routing** uses first-token statistics and conformal prediction to bypass easy samples. For routed samples, **Collaborative Grounding** combines cross-modal attention and a visual expert to generate precise localized panel displays robust to small or occluded targets. Experiments demonstrate LazyMCoT achieves competitive accuracy, surpasses recent training-based methods, and reduces inference latency. This selective grounding paradigm offers a practical recipe for efficient visual reasoning with frozen VLMs.

AI USE STATEMENT

We transparently disclose that Large Language Models (LLMs) were utilized as assistive tools during the preparation of this manuscript. Our use of these models was confined to refining the language of the text. The core scientific contributions, including the formulation of our hypothesis, experimental design, analysis of results, and the overall scholarly narrative, were exclusively the work of the human authors. We take full responsibility for the accuracy, originality, and all claims presented in this paper.

REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our research, we provide the complete source code of our LazyMCoT framework, including the Adaptive Routing module, the two-stage Collaborative Grounding pipeline, the router training scripts, and all evaluation protocols, at `https://anonymous.4open.science/r/LazyMCoT-16B5/`. Our implementation is built upon standard deep learning libraries, and our approach is training-free: it requires no fine-tuning or parameter modification of the base Vision-Language Models. All experiments were conducted using the base VLMs detailed in our experiment section. The only learned component, a lightweight GBDT router, is trained purely on first-token probability statistics derived from a single forward pass; we release its pre-trained parameters together with the exact commands to retrain it from scratch. To facilitate verification, we specify all key hyperparameters, including the conformal miscoverage rate α, in the experiment section, and include the exact prompts, inference configurations, and per-benchmark evaluation settings in the Appendix. We are confident that these resources provide a comprehensive basis for replicating our results and enabling further research.

REFERENCES

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, et al. 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`.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

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. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks Lakshmanan, and Ahmed H Awadallah. Hybrid llm: Cost-efficient and quality-aware query routing. In *International Conference on Learning Representations*, volume 2024, pp. 41348–41366, 2024.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2(1):32, 2024.

Marcel Gröpl, Jaewoo Jung, Seungryong Kim, Marc Pollefeys, and Sunghwan Hong. Entropy-gradient grounding: Training-free evidence retrieval in vision-language models. *arXiv preprint arXiv:2604.08456*, 2026.

Mahyar Khayatkhoei, Prateek Chhikara, Filip Ilievski, et al. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pp. 68194–68213, 2025.

Bo Li, Peiyuan Zhang, Jingkang Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. Otterhd: A high-resolution multi-modality model, 2023a. URL https://arxiv.org/abs/2311.04219.

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 Computer Vision and Pattern Recognition Conference*, pp. 9098–9108, 2025.

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

Yangfu Li, Hongjian Zhan, Jiawei Chen, Yuning Gong, Qi Liu, and Yue Lu. Deepscan: A training-free framework for visually grounded reasoning in large vision-language models. *arXiv preprint arXiv:2603.03857*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

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

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024c.

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.

Marco Morini, Sara Sarto, Marcella Cornia, and Lorenzo Baraldi. Look twice: Training-free evidence highlighting in multimodal large language models. *arXiv preprint arXiv:2604.01280*, 2026.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*, 2024.

OpenAI. Openai-gpt-4o, 2024. URL https://openai.com/zh-Hans-CN/index/gpt-4o-system-card/.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6613–6629, 2025.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Sule Bai, Zijian Kang, Jiashi Feng, Wang Zhuochen, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *The Fourteenth International Conference on Learning Representations*, 2026.

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

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, 2025b.

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

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.

Xuan Yu, Dayan Guan, and Yanfeng Gu. Zoom-refine: Boosting high-resolution multimodal understanding via localized zoom and self-refinement. *arXiv preprint arXiv:2506.01663*, 2025.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

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? *arXiv preprint arXiv:2408.13257*, 2024.

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*, 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. *arXiv preprint arXiv:2505.14362*, 2025.

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.

APPENDIX

This Appendix is organized as follows:

- **Implementation Details.** Hardware and reproducibility settings, VLM backbones and the SAM3 visual expert, the full Adaptive Routing training pipeline, the composition of the external routing set together with its leakage-control checks, and the Collaborative Grounding hyperparameters.

- **Two-Stage Refinement Details.** The coverage test selecting attention boxes missed by the visual expert, box enlargement and cropping for the second-stage re-query, and coordinate mapping rules for the evidence pool.
- **Dataset Information.** Detailed descriptions of the four high-resolution multiple-choice benchmarks used in our experiments, including image source, resolution, task taxonomy, and evaluation metric.
- **More Statistical Features.** Routing score and entropy diagnostic plots for all three VLMs, demonstrating that the proposed first-token statistics are VLM-agnostic.
- **Additional Related Work.** Unlike passive external routing, we utilize first token uncertainty as an intrinsic probe for zero cost active perception.
- **More Qualitative Results.** Additional side-by-side comparisons between HiDe and LazyMCoT, illustrating how Collaborative Grounding produces cleaner Localized Panel Displays for the VLM re-query.
- **Some Inference Cases.** End-to-end inference traces that compare the base VLM, HiDe, and LazyMCoT on some samples, showing the complementary roles of Adaptive Routing and Collaborative Grounding.

A Implementation Details

For each VLM backbone, we conduct experiments on two NVIDIA H20 GPUs. To ensure reproducibility, we fix the random seeds for all libraries (Python, CUDA, PyTorch, and NumPy) to 2077 during the training process.

A.1 VLM Backbones and Visual Expert

LazyMCoT is evaluated on three open-source VLM backbones: Qwen2.5-VL-7B-Instruct Bai et al. (2025b), Qwen3-VL-8B-Instruct Bai et al. (2025a), and InternVL3-8B-Instruct Zhu et al. (2025). All backbones remain frozen during inference and we use greedy decoding (`do_sample=False`). Input images are resized so that their longest edge is at most `maxp` $= 16{,}384$ pixels to fit the dynamic-resolution encoders. We adopt SAM3 Carion et al. (2025) as the unified visual expert and serve it as a separate FastAPI service that accepts entity texts as prompts and returns boxes, masks, and confidence scores.

A.2 Adaptive Routing

Feature extraction. For each test sample, we run a single forward pass with `max_new_tokens=1` and record the first answer token logits $\mathbf{z} \in \mathbb{R}^V$. We then parse the candidate option letters $\mathcal{O}$ from the question text via regular expression so that benchmarks with $K \geq 4$ options are handled uniformly. The same forward pass also produces the direct answer letter, which is reused when the router decides to skip Collaborative Grounding.

Router training. We split the routing set $\mathcal{D}_{\text{cal}}$ by the base VLM into ori-correct ($y = 0$) and ori-wrong ($y = 1$) samples and train a Gradient Boosting Decision Tree Friedman (2001) (g_θ) with 300 estimators, max depth 3, and learning rate 0.05 under 5-fold cross-validation with seed 2077. The out-of-fold (OOF) predicted probability $\hat{p}(\mathbf{x})$ is mapped to the routing score $s(x) = \log \hat{p}/(1-\hat{p})$. The must-recall subset $\mathcal{D}_{\text{mr}}$ is obtained by additionally running the full Collaborative Grounding pipeline offline on $\mathcal{D}_{\text{cal}}$ and keeping the ori-wrong samples that grounding actually corrects, and s_{floor} is the α-quantile of their OOF scores. Unless otherwise stated we adopt $\alpha = 0$ for the strictest must-recall target. The trained router is serialized once into a JSON report together with the pickled GBDT.

A.3 Routing Set Composition

Composition. $\mathcal{D}_{\text{cal}}$ is assembled exclusively from corpora that are external to our evaluation suite, as listed in Tab. 6. We draw high-resolution multiple-choice questions from MME-RealWorld-Lite Zhang et al. (2024), whose images cover sports broadcasts, video surveillance, remote sensing,

autonomous driving, and document scenes, and we take all questions of MagnifierBench Li et al. (2023a), in which the queried object occupies roughly one percent of a high-resolution frame. Both corpora match our test distribution in the properties the router actually needs, namely high input resolution, small query targets, and a letter-based answer format, while sharing no image source with V* Bench Wu & Xie (2024), HR-Bench Wang et al. (2025b), or TreeBench Wang et al. (2026). In particular, neither corpus is derived from the SA-1B pool from which V* Bench and TreeBench sample their images.

Source	#Samples	Image domain	Overlap with Tab. 1–5
MME-RealWorld-Lite Zhang et al. (2024)	1,500	sports, surveillance, remote sensing, driving, documents	none
MagnifierBench Li et al. (2023a)	283	object-dense indoor and outdoor scenes, target area $\approx 1\%$	none
$\mathcal{D}_{\mathrm{cal}}$ (total)	1,783	–	none

Table 6: Composition of the external routing set $\mathcal{D}_{\mathrm{cal}}$. No image, question, or sample identifier of V* Bench, HR-Bench-4K/8K, or TreeBench takes part in router fitting or threshold calibration.

Leakage control. Beyond the source-level disjointness, we run three explicit checks. First, we compare sample identifiers and question strings between $\mathcal{D}_{\mathrm{cal}}$ and every evaluation benchmark and find no exact match. Second, we compute a perceptual hash for all routing images and all evaluation images and discard any routing image whose Hamming distance to an evaluation image falls below a conservative threshold. Third, the conformal threshold s_{floor} is derived only from the out-of-fold scores of $\mathcal{D}_{\mathrm{mr}} \subseteq \mathcal{D}_{\mathrm{cal}}$, so no test score participates in threshold selection.

A.4 Collaborative Grounding

Entity decomposition. The base VLM is prompted with a rule-based template Liu et al. (2025) to decompose the question Q into a list of canonical entities $\mathcal{E} = \{e_1, \ldots, e_M\}$. The decomposition prompt is shwon as Fig. 8.

Visual expert branch. Each entity in $\mathcal{E}$ is sent to SAM3 as an independent text prompt. We retain at most $k = 10$ boxes per entity and apply cross-entity NMS with IoU threshold 0.7 to remove duplicate detections of the same instance.

Attention branch. For each entity token we reshape the attention to the visual-token grid, apply a Gaussian blur with $\sigma = 3$, and linearly normalize the map to $[0, 1]$. Connected components above the relative threshold $\tau = 0.5$ are converted into the attention boxes $\mathcal{B}_{\mathrm{att}}$.

Per-VLM attention layers. The cross-modal attention consumed by the attention branch is read from the self-attention modules of the decoder, averaged over all attention heads. Since different VLMs expose their most discriminative grounding signal at different depths, we select the attention source per VLM. For Qwen2.5-VL-7B and InternVL3-8B (both 28 decoder layers), we extract the attention of a single mid-level decoder layer, namely the 15-th layer (1-indexed). For Qwen3-VL-8B (36 decoder layers), we instead aggregate the attention of all decoder layers, because its DeepStack multi-layer visual injection distributes grounding cues across depths so a single layer is insufficient. When more than one layer is used, the per-layer maps are first averaged over heads and then averaged across the selected layers to form the aggregated saliency map $\mathcal{A}(I)$.

B Two-Stage Refinement Details

This section elaborates on the second stage of Collaborative Grounding introduced in the main paper, where every attention box that is not covered by the visual expert is cropped from the image on a slightly enlarged region and then re-queried with SAM3. We first describe how the uncovered attention boxes are selected, then how each selected box is enlarged and cropped, how the visual expert is re-queried on the crop, how the newly discovered detections are mapped back to the original image, and finally how the detections are filtered and assembled into the refined evidence pool.

B.1 Coverage Test for Uncovered Attention Boxes

The second stage only processes attention boxes whose evidence has not already been captured by the visual expert. For every expert box in $\mathcal{B}_{\mathrm{exp}}$ we compute the IoU against all attention boxes in

Decomposition Prompt

Prompt =
"### Thinking Process
Before generating the final output, you must internally follow these steps in order:

1. **Identify Core Entities**: Read the entire question and identify all key noun phrases. For example, "the green surfboard," "the purple umbrella."
2. **Deconstruct Attributes for Each Entity Individually**: **This is the critical step. Before considering the relationship between entities, look at each entity in isolation and apply Rules 2, 3, and 4 to fully deconstruct its attributes.**
* For instance, first process "the green surfboard" using Rule 2 to get `surfboard with green color`.
* Then, process "the purple umbrella" using Rule 2 to get `umbrella with purple color`.
3. **Handle Relationships Between Entities**: After all entities have been individually deconstructed, check for spatial or logical relationships between them (Rule 5). If a relationship exists, you will list the **already deconstructed** entities as separate items.
4. **Assemble and Normalize**:
* Gather all the canonical entity strings you have transformed.
* Convert all text to lowercase.
* Join all entities into a single line, separated by a comma and a space (", ").
5. **Final Formatting**: Enclose the resulting single-line string within the `<FINAL_OUTPUT>` and `</FINAL_OUTPUT>` tags.

Extraction Rules

Rule 1: Simple Entities
If a noun is not described by any modifiers, extract the noun itself.
* *Example*: "the scooter" becomes `scooter`.

Rule 2: Adjective Attribute Deconstruction
If an entity is modified by one or more adjectives, the format must be `noun with [property] [type]`. Chain multiple properties consecutively.
* **Format**: `noun with [property1] [type1] with [property2] [type2]`
* **Common Type Mappings**:
* Colors (red, blue, black, silver) -> `color`
* Sizes (large, small, big) -> `size`
* Materials (wooden, metal, plastic) -> `material`
* *Example*: "the large blue truck" becomes `truck with large size with blue color`.

Rule 3: Possessive Inversion
Convert all possessive forms (e.g., `X's Y`) uniformly into the `Y of X` format.
* *Example*: "the woman's handbag" becomes `handbag of woman`.

Rule 4: Attributive Prepositional Phrases
If a prepositional phrase describes a component of an entity (e.g., "in a shirt"), preserve the structure and recursively apply the rules to the entity within the phrase.
* *Example*: "the man in the green shirt" becomes `man in a shirt with green color`.

Rule 5: Relational Prepositional Phrases
If a prepositional phrase describes a relationship between two separate entities (e.g., `next to`, `on the left of`), **extract the entities as separate items, but only after each entity has been fully deconstructed according to the other rules.** Do not include the relational words in the output.
* *Example 1 (Simple)*: "the dog on the left side of the scooter" becomes `dog, scooter`.
* **Example 2 (With Attributes)**: *"Is the green surfboard on the left side of the purple umbrella?" becomes `surfboard with green color, umbrella with purple color`.** (Note: The surfboard and umbrella are first deconstructed individually, then listed as separate entities).

Rule 6: Compound Nouns
Recognized compound nouns should be treated as a single entity.
* *Example*: "the traffic light" becomes `traffic light`.

Output Format Rules
1. **Sole Output**: The final and only output content must be enclosed within `<FINAL_OUTPUT>` and `</FINAL_OUTPUT>` tags.
2. **Single-Line Format**: The output must be a single continuous line of text.
3. **Delimiter**: Multiple entities must be separated by a comma followed by a space (", ").
4. **Lowercase**: All output characters must be in lowercase.
5. **Content Exclusion**: The final entity string must not include articles (a, an, the), question words (what, which, is), or purely relational words (side, next to, of).

Now, following all the rules above, extract the entities from the question below:
{input_text}"

Figure 8: Decomposition prompt template

$\mathcal{B}_{\text{att}}$. An attention box is regarded as covered when its maximum IoU with any expert box reaches a threshold, and it is treated as an exclusive region otherwise. We adopt a low IoU threshold of 0.1 so that only attention regions that barely overlap with the expert boxes are forwarded to refinement. To bound the computational cost, the exclusive attention boxes are sorted by area in descending order and at most four regions per image are retained for re-querying, while each entity contributes at most two exclusive boxes.

B.2 Box Enlargement

Directly cropping a tight attention box often removes the contextual pixels around the target, which lowers the recall of the visual expert on the crop. We therefore enlarge each exclusive box before cropping. Given a normalized box $b = (x_0, y_0, x_1, y_1)$ with center (c_x, c_y), width $w = x_1 - x_0$, and height $h = y_1 - y_0$, the enlarged box under an expansion ratio $r = 0.3$ is:

$$\begin{aligned} x_0' &= \max\big(0,\, c_x - (1+r)\tfrac{w}{2}\big), \\ x_1' &= \min\big(1,\, c_x + (1+r)\tfrac{w}{2}\big), \\ y_0' &= \max\big(0,\, c_y - (1+r)\tfrac{h}{2}\big), \\ y_1' &= \min\big(1,\, c_y + (1+r)\tfrac{h}{2}\big). \end{aligned} \tag{6}$$

The enlargement grows the box by thirty percent around its center, and clipping to $[0, 1]$ keeps the enlarged box inside the image. Boxes that remain very small after enlargement, namely with normalized width or height below 0.02, are skipped, because such regions rarely enclose a reliable target and re-querying them wastes computation.

B.3 Cropping and Expert Re-query

The enlarged box is converted from normalized coordinates to pixel coordinates through $(\lfloor x_0' W \rfloor, \lfloor y_0' H \rfloor, \lfloor x_1' W \rfloor, \lfloor y_1' H \rfloor)$, where W and H denote the image width and height in pixels. The resulting pixel box is clamped to the valid image range, and any degenerate crop whose left edge meets its right edge or whose top edge meets its bottom edge is discarded. We then crop the corresponding sub image and encode it for the SAM3 service. The same entity set $\mathcal{E}$ that guided the first stage is reused as the text prompt, so the expert now searches for the queried entities inside a magnified local view in which small or occluded targets occupy a far larger portion of the field of view than in the full image.

B.4 Coordinate Mapping and Filtering

Each detection returned by SAM3 on the crop is expressed in the local normalized frame of the enlarged region. Let $b_{\text{loc}} = (u_0, v_0, u_1, v_1)$ be such a detection, and let $w' = x_1' - x_0'$ and $h' = y_1' - y_0'$ be the width and height of the enlarged crop. The detection is mapped back to the global image frame by:

$$\begin{aligned} o_{x_0} &= x_0' + u_0\, w', \quad o_{x_1} = x_0' + u_1\, w', \\ o_{y_0} &= y_0' + v_0\, h', \quad o_{y_1} = y_0' + v_1\, h'. \end{aligned} \tag{7}$$

The mapped boxes then pass through a cascade of filters before joining the evidence pool. A box is dropped when its normalized area falls below 0.0005, which removes spurious fragments. A box is regarded as a duplicate and dropped when its IoU with any existing box, or with any box already accepted during the current stage, reaches 0.8. To keep the localized panel display compact, we retain at most two new boxes per entity and at most four new boxes per image, again ranked by area. Any expert box that spans more than eighty percent of the image is excluded from cropping and recorded only in the legend, since such a large box carries little localization value.

B.5 Evidence Assembly

The surviving boxes are merged with the first stage pool to form the refined set $\mathcal{B}^{(2)}$. A box that is spatially contained within a larger accepted box, judged with a small tolerance of three percent along each side, is preserved only as a visual overlay rather than added again, which prevents redundant panels for the same instance. The refined set is finally rendered on the image with a distinct color

border and a textual legend for each region, where the legend font scales with the image height so that it stays readable across resolutions. This rendering yields the localized panel display that is fed back to the VLM for a single re-query.

C Dataset Information

We evaluate LazyMCoT on four challenging high-resolution multiple-choice benchmarks. All benchmarks adopt accuracy as the primary evaluation metric.

C.1 V^* Bench

V^* Bench Wu & Xie (2024) is introduced together with the V^* visual search algorithm to evaluate the ability of multimodal LLMs to localize and reason over small or visually inconspicuous targets in high-resolution natural scenes. The benchmark contains 191 images with an average resolution of 2246×1582, sourced primarily from the SA-1B collection. Each image is paired with a multiple-choice question that targets one of two abilities: (i) *Direct Attribute Recognition* (Att.), which asks about color, material, shape, or other intrinsic attributes of a small object, and (ii) *Spatial Relationship Reasoning* (Spa.), which requires inferring relative positions between two or more objects. Because the queried targets occupy only a tiny fraction of the image, V^* Bench is widely used as a stress test for fine-grained visual grounding.

C.2 HR-Bench-4K and HR-Bench-8K

HR-Bench Wang et al. (2025b) is the first benchmark explicitly designed to evaluate MLLMs at 4K and 8K resolutions, addressing the gap that previous benchmarks rarely exceed 2K. HR-Bench provides two splits, HR-Bench-4K and HR-Bench-8K, each consisting of high-resolution images that are evenly partitioned into two task types: *Single-Instance Perception* (Sin.), where the question concerns a single fine-grained object that may be small or inconspicuous, and *Cross-Instance Perception* (Cro.), where the question requires comparing or relating multiple instances across the image. Human accuracy is 87% Wang et al. (2025b), highlighting the difficulty of high-resolution understanding and making HR-Bench an ideal testbed for visual grounding.

C.3 TreeBench

TreeBench Wang et al. (2026) is a recent benchmark that probes both fine-grained perception and high-order reasoning under traceable visual evidence. It is constructed by sampling $1{,}000$ object-dense images from SA-1B and, after a three-stage manual quality control by eight LMM experts, retains 405 challenging multiple-choice VQA pairs. Each question is annotated with both the answer and the corresponding ground-truth bounding boxes, ensuring that the evaluation rewards genuine localization rather than language priors. The benchmark is organized into ten fine-grained categories grouped under two competencies: **Perception** (*Attributes*, *Material*, *Physical State*, *Object Retrieval*, *OCR*) and **Reasoning** (*Perspective Transformation*, *Ordering*, *Contact & Occlusion*, *Spatial Containment*, *Comparison*).

D More statistical features

To verify that the proposed statistics in Sec. 3 generalize beyond a single VLM backbone, we report the same two diagnostic plots on all three evaluated VLMs in Fig. 9. The left panel of each row shows the distribution of the GBDT routing score $s(x)$ for ori-correct and ori-wrong samples, while the right panel shows the joint distribution of $s(x)$ versus the option entropy $H(\tilde{p})$. Across Qwen2.5-VL-7B, InternVL3-8B, and Qwen3-VL-8B, the routing score consistently produces well-separated modes between the two classes, and is monotonically correlated with answer entropy with ori-wrong samples concentrated in the high-score, high-entropy region. This consistency confirms that the proposed first-token statistics are not VLM-specific artifacts, and that the same lightweight router can be calibrated for any frozen VLM without architectural assumptions.

E Additional Related Work

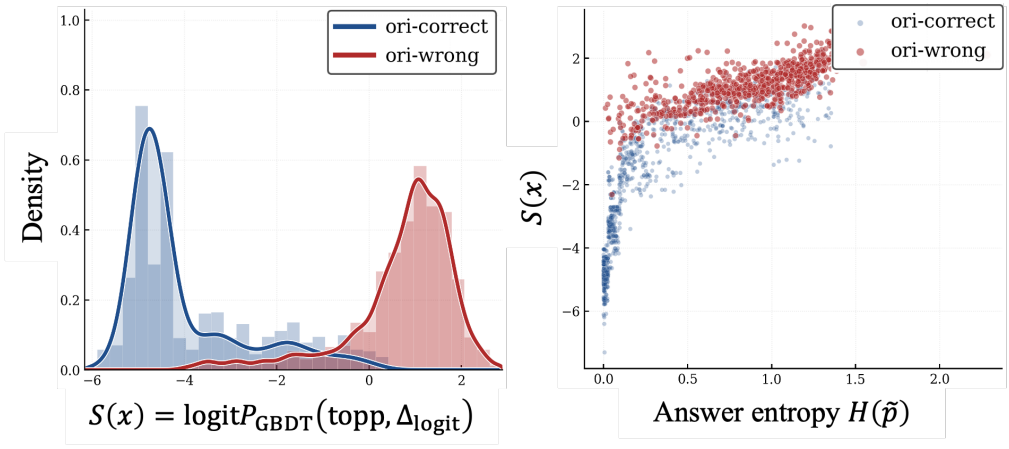
Routing Mechanisms in Large Language Models. Existing frameworks like Route LLM Ong et al. (2024) and Hybrid LLM Ding et al. (2024) balance computational cost and response quality by employing external classifiers to evaluate prompts prior to execution. However, these conventional paradigms face critical limitations in multimodal tasks. They introduce unavoidable computational overhead and operate passively. Specifically, they assess static text without perceiving the actual visual complexity or the internal cognitive state of the base model during execution. In contrast, our Adaptive Routing shifts the paradigm from passive allocation to active perception. LazyMCoT utilizes the predictive uncertainty of the first generated token as a zero cost intrinsic probe to directly measure spatial semantic binding and perceptual bottlenecks. By evaluating internal statistical state, our method acts as an active perception trigger. It dynamically alters the visual input distribution through collaborative grounding only when visual evidence is insufficient.

F More qualitative results

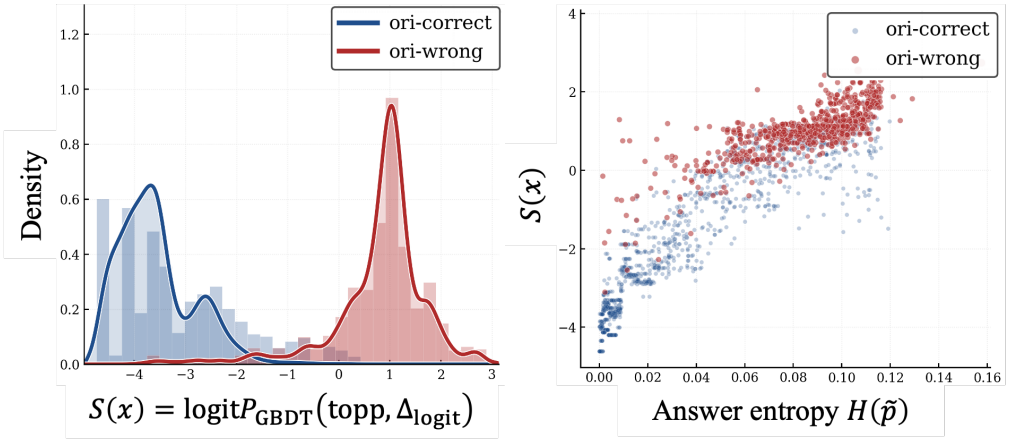
We provide additional qualitative comparisons on the three high-resolution benchmarks to illustrate how Collaborative Grounding helps the base VLM recover the correct answer on hard samples. Fig. 10, Fig. 11, and Fig. 12 respectively present cases drawn from V* Bench, TreeBench, and HR-Bench. For each case we show, from left to right, the original image with the question, the Localized Panel Display (LPD) generated by HiDe, and the LPD produced by our LazyMCoT, accompanied by the predicted answers. Across all three benchmarks, HiDe frequently misses the queried small or co-occurring targets because its attention-only cropping is sensitive to noise, while LazyMCoT couples cross-modal attention with a SAM3 visual expert and further refines the evidence pool through Stage 2 re-querying. The resulting LPDs cover the question-relevant entities more completely and are decorated with per-entity color borders and textual legends, which guide the VLM to faithfully read off the correct answer in a single re-query.

G Some Inference Cases

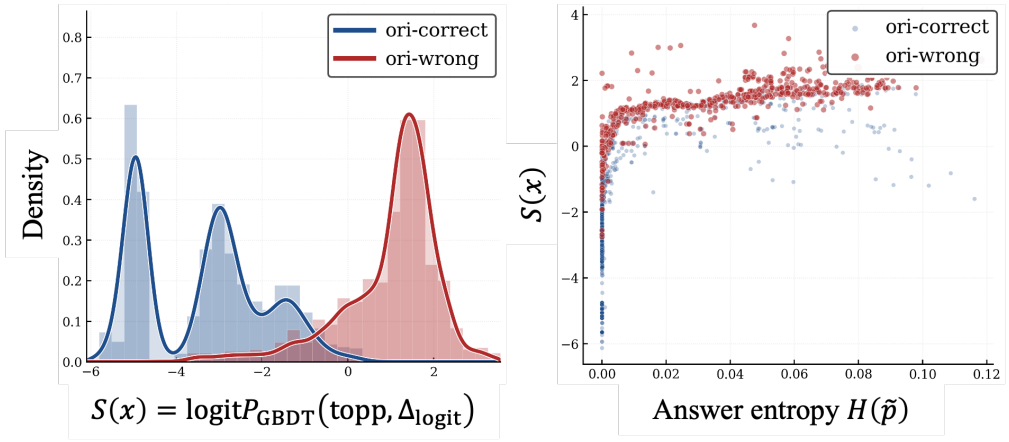
Fig. 13 presents end-to-end inference traces of LazyMCoT alongside the base VLM and HiDe on five representative samples drawn from V* Bench and HR-Bench. Each row shows, from left to right, the prediction of the original VLM on the raw image, the prediction of HiDe on its attention-only LPD, and the prediction of LazyMCoT on its Collaborative Grounding output. The first three rows correspond to hard samples on which the base Qwen2.5-VL-7B fails. Adaptive Routing identifies them as uncertain and dispatches them to Collaborative Grounding, which produces LPDs in which the queried entities are precisely highlighted by SAM3 colored boxes and textual legends. With this faithful evidence, the same VLM corrects its answer in a single re-query. The last two rows correspond to easy samples on which the base InternVL3-8B already answers correctly. Here HiDe's forced grounding truncates the global context and misleads the VLM into wrong predictions, whereas Adaptive Routing tags the samples as confident and emits the direct answer immediately, as marked by the **Direct Answer** flag in the rightmost column. Together, these traces illustrate the complementary roles of the two components, Collaborative Grounding rescues hard samples that the VLM cannot solve alone, while Adaptive Routing prevents Collaborative Grounding from interfering with samples that are already well solved.



(a) Qwen2.5-VL-7B-Instruct



(b) InternVL3-8B-Instruct



(c) Qwen3-VL-8B-Instruct

Figure 9: **Statistical features for sample routing across VLM backbones.** For each backbone, the left panel plots the routing score $s(x)$ distribution, and the right panel plots $s(x)$ versus the option entropy $H(\tilde{p})$. The same separability pattern holds across all backbones, validating that the proposed first-token statistics are VLM-agnostic.

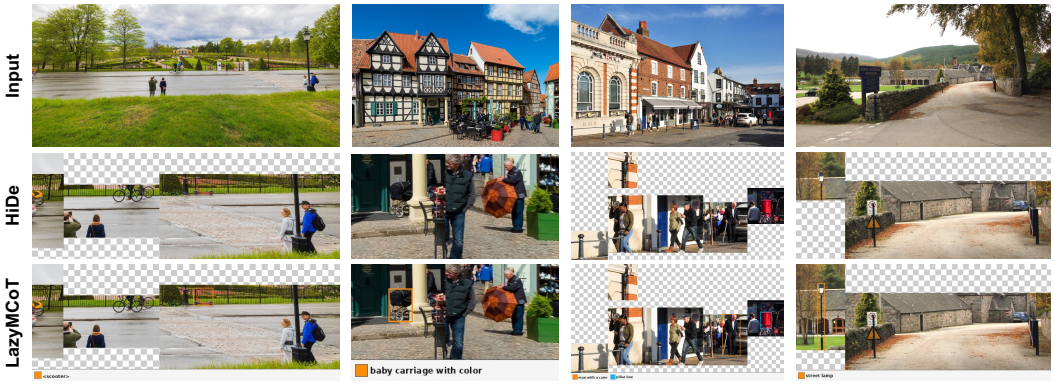


Figure 10: Additional qualitative comparison on V* Bench.

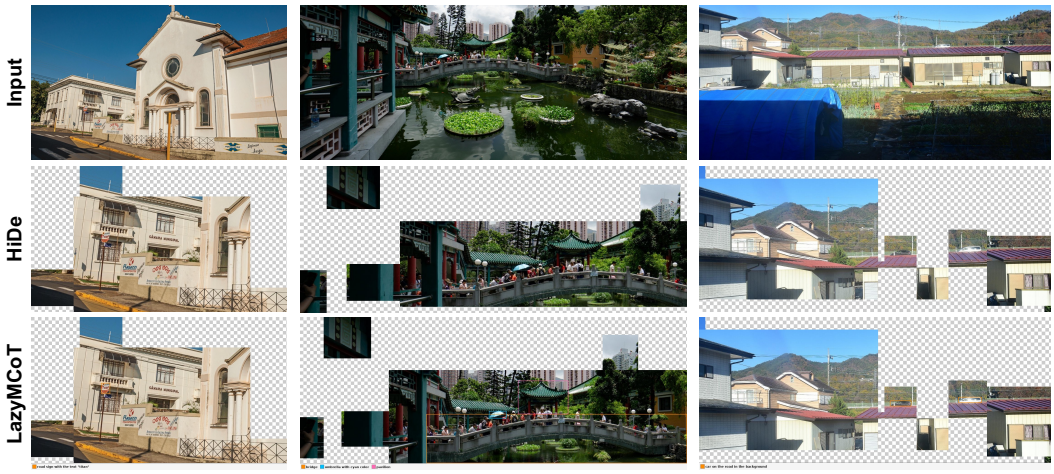


Figure 11: Additional qualitative comparison on TreeBench.

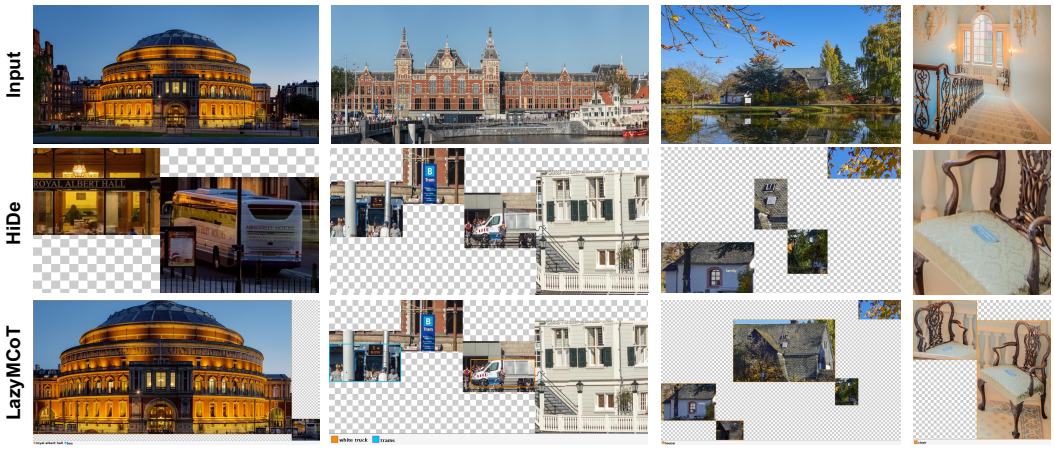


Figure 12: Additional qualitative comparison on HR-Bench.

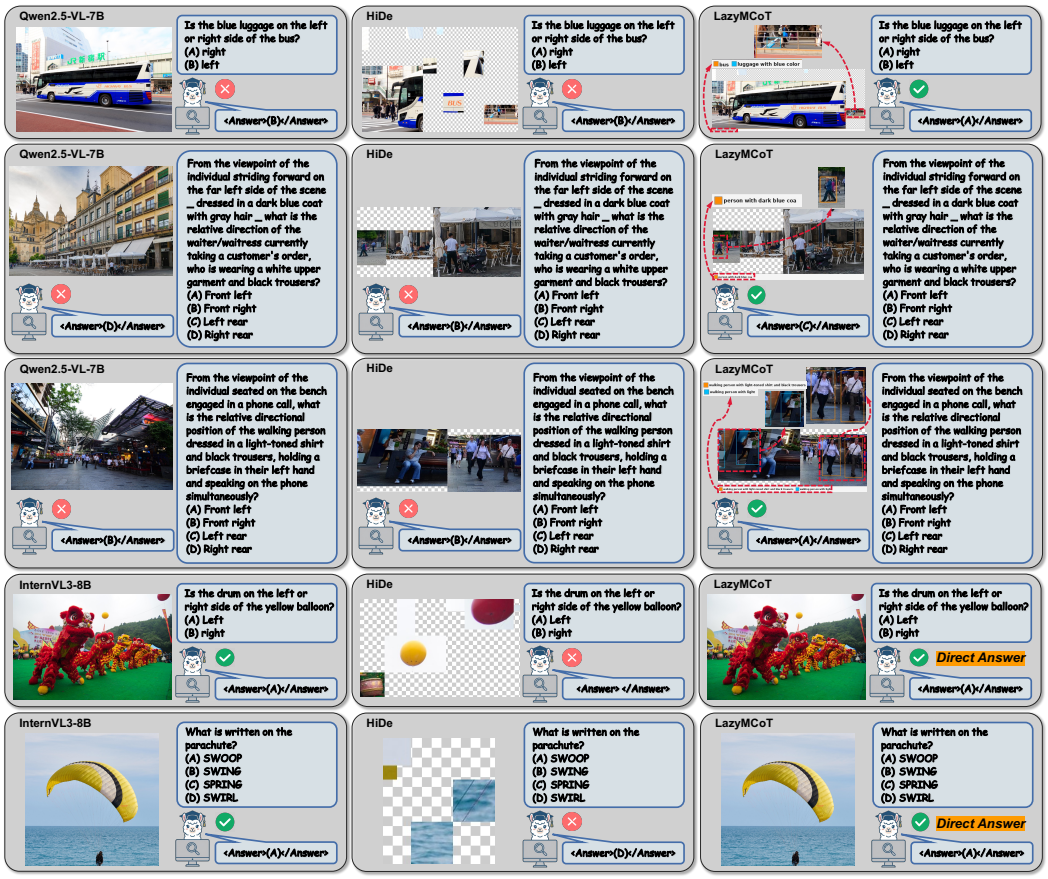


Figure 13: **End-to-end inference cases comparing the base VLM, HiDe, and LazyMCoT.** The top three rows show hard samples where Adaptive Routing triggers Collaborative Grounding and LazyMCoT produces a clean LPD that recovers the correct answer. The bottom two rows show easy samples where the router emits a **Direct Answer** and bypasses grounding.