WHEN IS SEEING ENOUGH? CROSS-GRANULARITY EVIDENCE SEARCH WITH SUFFICIENCY VERIFICATION FOR HIGH-RESOLUTION UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Although multimodal large language models (MLLMs) have substantially advanced visual understanding, resolving fine-grained details in high-resolution images remains challenging under limited visual input budgets. Existing approaches recover such details through local cropping and iterative region search, but the relevance of individual regions does not establish whether the accumulated evidence is sufficient to support the correct answer. This raises two coupled challenges: *acquiring complementary evidence across granularities* and *determining when the accumulated evidence is sufficient to support an answer*. To address these challenges, we propose a novel framework named Cross-granularity Evidence Search with Sufficiency Verification (**CAREER**) for high-resolution visual question answering. The core idea is to couple complementary evidence acquisition with explicit sufficiency verification to guide continued observation and answer acceptance. Specifically, we first bootstrap a low-cost global-to-local evidence state and construct a complementary multi-scale candidate atlas that exposes potentially informative observations. Building on this atlas, we adaptively refine or expand views and explore alternative regions according to unresolved evidence requirements, accumulating observations to retain both local details and broader context. To assess evidence sufficiency, we introduce an option-wise verifier that evaluates visual support without access to the generator's provisional answer, and then incorporate its sufficiency verification to guide further observation and adaptive stopping. Additionally, experiments across three MLLM backbones and three high-resolution QA settings show that **CAREER** consistently outperforms state-of-the-art visual-search baselines, while also reducing object hallucination on POPE benchmark. The code is available at https://github.com/CAREER-ICLR2027/CAREER.git.

1 INTRODUCTION

Multimodal large language models (MLLMs) (Zhang et al., 2024; Jiang et al., 2025; Kil et al., 2024) advance vision-language tasks (Yang et al., 2026; Fu et al., 2026b), including visual question answering (Kuang et al., 2025; Khayatkhoei et al., 2025; Zhong et al., 2026), document understanding (Luo et al., 2024; Ding et al., 2026; Liu et al., 2026), and visual grounding (Song et al., 2026; Wu et al., 2026). Yet high-resolution details remain challenging (Khayatkhoei et al., 2025; Tong et al., 2024b). Encoding and cross-modal projection may lose information (Tong et al., 2024b; Li et al., 2025b), while reasoning can attenuate visual evidence (Li et al., 2025d). Effective perception requires complementing global views with question-relevant local observations to recover otherwise overlooked visual details (Wu & Xie, 2024; Khayatkhoei et al., 2025).

Existing approaches to high-resolution perception in MLLMs can be broadly grouped into two categories: resolution-aware representation (Bai et al., 2025b; Guo et al., 2024) and query-guided visual search (Li et al., 2025a; Shen et al., 2025). In particular, frameworks in the former category employ dynamic-resolution encoding (Wang et al., 2024) or multi-granularity semantic fusion (Bai et al., 2025a; Zhang et al., 2026; Li et al., 2025c) to preserve fine-grained visual information, whereas those in the latter adaptively identify and inspect question-relevant regions through diverse search strategies (Zhong et al., 2026; Wang et al., 2025b; Li et al., 2026). However, accessibility alone

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.

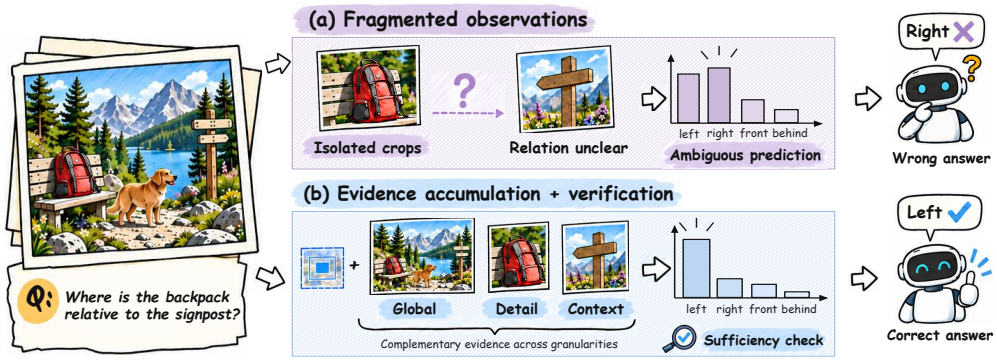


Figure 1: Motivation insights for high-resolution perception. (a) Query-relevant regions viewed in isolation may capture local details but leave cross-region relations unresolved, thus providing insufficient evidence for reliable prediction. (b) **CAREER** accumulates complementary evidence across granularities and verifies whether it is sufficient to distinguish among candidate answers.

does not ensure sufficiency: individually relevant observations may still lack the context needed to integrate evidence across granularities and distinguish among candidate answers.

Despite advances in high-resolution perception (Li et al., 2025a), these methods prioritize the relevance of individual observations, leaving multi-granularity evidence integration and sufficiency assessment underexplored. In Figure 1, separate crops reveal the backpack and signpost but lose the shared context needed to infer their relative positions. This motivates our first research question. ❶ *How can we adaptively acquire and integrate complementary visual evidence across granularities?* The model must then assess whether these observations jointly distinguish candidate answers, identify what evidence is still missing, and determine where to search if needed. This raises a second question. ❷ *How can we derive answer-level sufficiency verification to assess evidence sufficiency and determine whether further visual search is needed?* These challenges motivate **CAREER**, which couples multi-granularity candidate acquisition with verification-guided adaptive visual search.

To address these concerns, we propose Cross-granularity Evidence Search with Sufficiency Verification (**CAREER**), a training-free framework coupling multi-granularity evidence acquisition with answer-level sufficiency verification. We first assess whether the global view supports a confident answer. For unresolved queries, we combine query-directed SAM proposals and hierarchical regions into a multi-scale candidate atlas covering local details and broader context. An evidence bundle retains the global image and all valid local observations for joint reasoning across scales. Guided by this bundle and prior feedback, the MLLM proposes scope adjustments or alternative regions to explore. An independent verifier evaluates each option against accumulated evidence without receiving the MLLM's provisional answer, accepting the leading option only when its absolute support and separation from alternatives meet the thresholds. Otherwise, search continues, revisiting earlier states with untried observations when exploration stalls, thereby avoiding repeated uninformative visual observations while retaining accumulated evidence. If the budget or feasible views are exhausted, the saved global answer serves as an unverified fallback. This loop connects evidence acquisition with explicit acceptance criteria. The main contributions are summarized as follows:

❶ ***Novel Perspective.*** We formulate high-resolution perception as the coupled problem of acquiring complementary visual evidence across granularities and assessing whether the accumulated evidence is sufficient to support a reliable answer. This formulation distinguishes the relevance of individual observations from their collective sufficiency for answering. This perspective unifies decisions on where to look next and when evidence justifies accepting an answer.

❷ ***Verification-Guided Visual Search.*** To enhance high-resolution perception, we propose a natural yet effective framework **CAREER** which couples multi-granularity evidence acquisition with sufficiency verification for efficient visual search. Under explicit verifier-alignment and evidence-reachability assumptions, we derive finite-budget bounds on acceptance coverage and final prediction risk, including the branch that returns an unverified fallback answer.

❸ ***Comprehensive Evaluation.*** Extensive experiments show that **CAREER** is *(i) accurate in high-resolution QA*, improving mean overall accuracy by 4.03% relative to CVSearch on `LLaVA-OV-7B`, *(ii) effective across backbones*, outperforming competing baselines in all nine

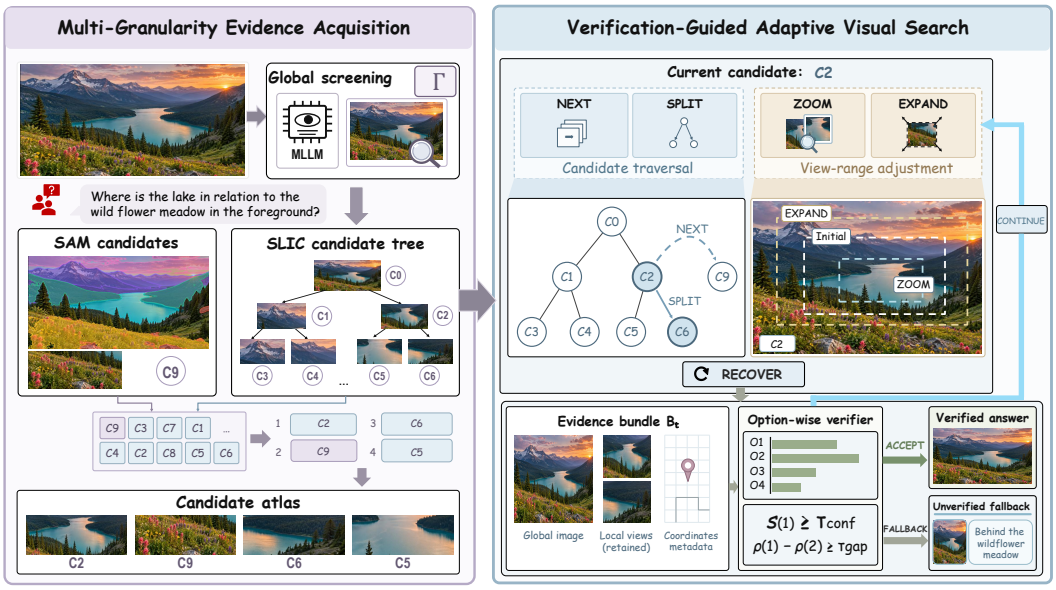


Figure 2: Overview of CAREER. The core of CAREER lies in coupling cross-granularity evidence acquisition with verification-guided adaptive visual search. The candidate ID C_i indexes a fine-grained region, while a_i denotes a candidate answer option in closed-ended visual question answering.

backbone–setting combinations by overall accuracy, and *(iii) less prone to object hallucination*, consistently reducing erroneous object-existence predictions across POPE-COCO evaluation splits.

2 Related Work

Multimodal Large Language Models. Multimodal large language models (MLLMs) advance visual understanding (Kuang et al., 2025; Fu et al., 2026a) by combining visual representations with language reasoning (Tong et al., 2024a). BLIP-2 (Li et al., 2023a) connects frozen image and language models through a lightweight Q-Former. InstructBLIP (Dai et al., 2023) makes visual feature extraction instruction-aware, while LLaVA (Liu et al., 2023) uses visual instruction tuning (Cocchi et al., 2025). Recent models, including Qwen3-VL (Bai et al., 2025a) and InternVL3 (Zhu et al., 2025), further improve understanding and reasoning through architecture and pre-training advances (McKinzie et al., 2024). However, visual input budgets still constrain fine-grained high-resolution perception (Jiang et al., 2026), motivating selective acquisition of details and their context.

High-Resolution Perception in MLLMs. Fine-grained processing of high-resolution images (Khayatkhoei et al., 2025; Zhang et al., 2025; Wang et al., 2025a; Li et al., 2024b) increases visual encoding and multimodal inference costs (Shi et al., 2026; Meng et al., 2024). Methods balance detail and efficiency through multi-scale representations (Shi et al., 2024) and adaptive visual search (Tran et al., 2026; Yi et al., 2026), including hierarchical exploration (Shen et al., 2025; Li et al., 2025a) and selective local inspection (Wang et al., 2025b; Li et al., 2026). Interleaved reasoning and local inspection (Shao et al., 2024) and visual tool use learned through reinforcement learning (Zheng et al., 2026; Su et al., 2026) guide observation, while grounding helps correct hallucinations (Yin et al., 2024; Chen et al., 2024a). Yet relevant regions need not jointly distinguish competing answers. We couple multi-granularity acquisition with option-wise sufficiency verification, using feedback on accumulated evidence to guide further observation and answer acceptance.

3 The Proposed CAREER

Problem Formulation. We consider a pre-trained MLLM $\mathcal{M}_\theta$ comprising a visual encoder $\mathcal{E}_\text{v}$, a multimodal projector $\mathcal{P}$, and an LLM $\mathcal{L}$. Given a high-resolution image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, and an input question q, we aim to generate an answer $\hat{y}$ supported by visual evidence from $\mathbf{I}$. In particular, representing the full image within a limited visual token budget may obscure fine-grained details required to answer q. Therefore, our objective is to acquire complementary visual evidence by

constructing a cross-granularity set $\mathcal{V} = \{\mathbf{I}\} \cup \{\mathrm{Crop}(\mathbf{I}, b_t)\}_{t=1}^{T}$, where b_t specifies an observed region and T denotes the number of local views. More precisely, we seek to obtain sufficient evidence for a reliable answer without altering the model parameters θ.

3.1 Framework Overview

In this work, we view high-resolution perception as an iterative process that alternates between acquiring visual evidence and assessing whether the accumulated evidence is sufficient to support an answer. Building on this insight, we propose an adaptive framework **CAREER** for acquiring sufficient visual evidence to enhance high-resolution MLLM understanding. At the beginning, **CAREER** maintains an evolving evidence bundle while organizing complementary multi-scale observations into a structured candidate atlas, from which additional views are acquired only when further evidence is needed. After each acquisition, **CAREER** verifies whether the updated evidence state is sufficient to support an answer, and uses this feedback to guide the next search decision. In this way, **CAREER** jointly addresses two complementary questions: where to look next and when the accumulated evidence is sufficient to answer reliably. The detailed framework of **CAREER** is illustrated in Figure 2.

3.2 Cross-Granularity Evidence Acquisition

We maintain an evidence bundle of observed views and a structured candidate atlas of potential observations (Shen et al., 2025; Lin et al., 2026). This separation lets search target missing evidence while retaining observations already acquired.

Global-to-Local Evidence Bootstrapping. We first assess whether the global observation alone provides sufficient evidence for a confident answer before acquiring additional local views. Given a high-resolution image $\mathbf{I}$, a question q, and a set of answer options $\mathcal{A} = \{a_i\}_{i=1}^{M}$, we initialize the evidence bundle with the global image, i.e., $\mathcal{B}_0 = \{\mathbf{I}\}$. The frozen generator $\mathcal{M}_g$ processes $(\mathcal{B}_0, q, \mathcal{A})$. From this forward pass, we extract the logits $\mathbf{z} = \{z_i\}_{i=1}^{M}$ corresponding to the answer tokens, and normalize them with softmax to obtain the probabilities $\mathbf{p} = \{p_i\}_{i=1}^{M}$. Accordingly, the predicted answer is expressed as $\hat{y}_0 = a_{\arg\max_i p_i}$. Subsequently, we perform global screening using two complementary signals:

$$\Gamma = \mathbb{I}\left[p_{(1)} \geq \tau_{\mathrm{conf}} \ \wedge \ p_{(1)} - p_{(2)} \geq \tau_{\mathrm{gap}}\right], \tag{1}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function, and $p_{(1)}$ and $p_{(2)}$ represent the largest and second-largest probabilities among the candidate answers, respectively. The confidence criterion ensures that the leading answer receives sufficient absolute support, while the margin criterion requires it to be clearly separated from the strongest alternative. If $\Gamma = 1$, we directly return the top-ranked answer and terminate visual search. Otherwise, we proceed to bootstrap complementary local observations.

Complementary Multi-Scale Candidate Atlas. For queries unresolved by global screening, we construct a complementary multi-scale candidate atlas that combines query-directed localization with scene-structured partitioning. In particular, we extract a set of localization phrases $\mathcal{O}_q = \{o_\ell\}_{\ell=1}^{L}$ from the query q, and then leverage SAM 3 (Carion et al., 2026) to identify the corresponding objects. Subsequently, we retain their bounding boxes as $\mathcal{C}_{\mathrm{SAM}}$ together with the associated dense visual features $\mathbf{H}_{\mathrm{SAM}}$. Since query-directed localization may miss contextual evidence not explicitly specified by the query, we complement these targeted regions with scene-structured candidates. Following prior work (Fu et al., 2025; Li et al., 2026), we apply Simple Linear Iterative Clustering (SLIC) (Achanta et al., 2012) to partition the visual feature space into spatially coherent regions. We recursively apply SLIC partitioning to construct a hierarchical region tree $\mathcal{T}$, whose nodes correspond to candidate regions at different spatial scales. Finally, these resulting regions form $\mathcal{C}_{\mathrm{SLIC}}$, and the final multi-scale candidate atlas is given by $\mathcal{C} = \mathcal{C}_{\mathrm{SAM}} \cup \mathcal{C}_{\mathrm{SLIC}}$.

Cross-Granularity Candidate Ranking. To prioritize observations in the candidate atlas, we rank candidate regions by two complementary factors: their relevance to the query and the visual variation they contain. For each candidate $c \in \mathcal{C}$, let b_c denote its bounding box in the original image and $\mathbf{I}_c = \mathrm{Crop}(\mathbf{I}, b_c)$ denote the corresponding image patch. To quantify query relevance, we leverage CLIP (Radford et al., 2021) to map each candidate patch and the query-related text into a shared embedding space. Specifically, let $\mathbf{v}_c$ denote the ℓ_2-normalized embeddings for patch $\mathbf{I}_c$, and let $\{\mathbf{t}_\ell\}_{\ell=0}^{L}$ denote the normalized text embeddings, where $\mathbf{t}_0$ encodes the original query q and $\mathbf{t}_\ell$ for

$\ell \geq 1$ encodes the corresponding localization phrase. Therefore, the query relevance for the candidate c is estimated as:

$$S_q(c) = \frac{\mathbf{v}_c^\top \bar{\mathbf{t}}_q}{\|\bar{\mathbf{t}}_q\|_2}, \quad \text{where} \quad \bar{\mathbf{t}}_q = \frac{1}{L+1} \sum_{\ell=0}^{L} \mathbf{t}_\ell. \tag{2}$$

Beyond semantic relevance, we characterize the visual variation contained within each candidate region. Following the feature-dispersion prior of CVSearch (Li et al., 2026), we first measure representation-level variation using the cached dense SAM features. Let Ω_c denote the set of feature locations whose spatial centers fall within b_c after being mapped to the original-image coordinate system. With $\mathbf{h}_j = \mathbf{H}_{\text{SAM}}[j]$ and $\bar{\mathbf{h}}_c = |\Omega_c|^{-1} \sum_{j \in \Omega_c} \mathbf{h}_j$, we define its feature dispersion as:

$$D_f(c) = \max\left(0,\, 1 - \frac{1}{|\Omega_c|} \sum_{j \in \Omega_c} \frac{\mathbf{h}_j^\top \bar{\mathbf{h}}_c}{\|\mathbf{h}_j\|_2 \|\bar{\mathbf{h}}_c\|_2}\right). \tag{3}$$

This score measures within-region feature dispersion, with lower values reflecting stronger directional alignment among the local features. To complement this representation-level signal with image-space variation, we additionally compute the mean Sobel gradient magnitude (Sobel et al., 1968):

$$D_e(c) = \frac{\sum_x M_c^-(x)\, \|\nabla_{\text{Sobel}} Y_c(x)\|_2}{\sum_x M_c^-(x)}, \tag{4}$$

where Y_c denotes the grayscale patch resized to a fixed resolution with aspect-ratio-preserving padding, ∇_{Sobel} is the Sobel gradient operator, and M_c^- is an eroded valid-pixel mask that excludes padded pixels and padding-induced boundary gradients. Furthermore, we combine these complementary signals into the final candidate-priority score:

$$R(c) = w S_q(c) + (1-w)\left[\lambda D_f(c) + (1-\lambda) D_e(c)\right], \tag{5}$$

where $w, \lambda \in [0,1]$, w balances query relevance against regional variation, while λ controls the relative contributions of feature-space dispersion and image-space edge density. Importantly, $R(c)$ serves only as a prior for candidate traversal, and subsequent navigation and verification dynamically determine which views to acquire and when the accumulated evidence is sufficient to answer.

3.3 Verification-Guided Adaptive Visual Search

To determine whether the accumulated evidence is sufficient to distinguish one candidate answer from its alternatives, we couple evidence-conditioned visual navigation with option-wise verification for search control. Conditioned on the current evidence, the generator proposes what to observe next to resolve uncertainty about the candidate answers, while the verifier determines whether to accept the current answer, continue acquiring evidence, or redirect it to a previous state.

Evidence-Conditioned Visual Navigation. Throughout the search, we maintain an evidence bundle containing the global image and all valid local observations. The search is initialized from the highest-ranked candidate according to Eq. 5. After acquiring t local observations, we denote the accumulated evidence bundle by $\mathcal{B}_t$. Conditioned on $\mathcal{B}_t$ and feedback retained from earlier rounds, the generator $\mathcal{M}_g$ updates the answer and proposes the next observation before current-round verification. While retaining the current candidate identity, `ZOOM` inspects fine-grained details, whereas `EXPAND` incorporates broader surrounding context. To move beyond the current view, `SPLIT` descends to a finer child region in the hierarchy $\mathcal{T}$, whereas `NEXT` switches to another candidate region in the joint atlas $\mathcal{C}$. For `SPLIT` and `NEXT`, the precomputed priority score $R(c)$ selects the highest-ranked feasible unvisited region from the current node's children and the joint atlas, respectively. Thus, the evidence-conditioned proposal determines how to navigate, while the candidate priorities determine which feasible region to visit. The proposed action is executed only when the verification controller described below returns `CONTINUE`. Once a valid new view has been acquired, we augment the evidence bundle as follows:

$$\mathcal{B}_{t+1} = \mathcal{B}_t \cup \{(\mathbf{I}_{t+1}, b_{t+1})\}, \qquad \mathbf{I}_{t+1} = \text{Crop}(\mathbf{I}, b_{t+1}). \tag{6}$$

This update preserves all previously acquired views as the search focus changes, enabling joint reasoning over evidence collected across regions and scales.

Verification-Guided Search Control. To determine whether another observation is necessary, we assess whether the accumulated evidence sufficiently supports one candidate answer while distinguishing it from its alternatives. At step t, the frozen verifier $\mathcal{M}_v$ independently evaluates each option $a_i \in \mathcal{A}$, conditioned on the evidence bundle $\mathcal{B}_t$ and the question q. For each option, it assigns one of three judgments: *Support*, *Refute*, or *Insufficient*, corresponding to supportive, contradictory, and inconclusive visual evidence, respectively. Applying a softmax over the three judgment logits yields the support probability for each option, which we further normalize across the candidate set to obtain its relative support share:

$$s_{t,i} = P_{\mathcal{M}_v}\left(\textit{Support} \mid \mathcal{B}_t, q, a_i\right), \qquad \rho_{t,i} = \frac{s_{t,i}}{\sum_{j=1}^{M} s_{t,j}}. \tag{7}$$

Here, $s_{t,i}$ measures the absolute visual support assigned to option a_i given the current evidence bundle, whereas $\rho_{t,i}$ captures its relative dominance within the candidate set. These signals are complementary because a large relative share may arise even when all candidate answers receive weak absolute support. We identify the leading and runner-up options as $i_t^\star = \arg\max_i \rho_{t,i}$ and $j_t^\star = \arg\max_{j \neq i_t^\star} \rho_{t,j}$, respectively. We denote the leading option's absolute support and relative share by $c_t = s_{t,i_t^\star}$ and $r_t = \rho_{t,i_t^\star}$. Mirroring the screening rule in Eq. 1, the controller returns `ACCEPT` when $c_t \geq \tau_{\text{conf}}$ and $\rho_{t,i_t^\star} - \rho_{t,j_t^\star} \geq \tau_{\text{gap}}$. Otherwise, search continues with the proposed observation or invokes `RECOVER` under the conditions below. To detect search stagnation, we monitor the stepwise improvements in c_t and r_t. If both remain below their respective progress thresholds for two consecutive steps, the controller triggers `RECOVER` and returns to a previous search state with an untried legal observation. Recovery is also triggered when the current state offers no remaining legal observation. This backtracking mechanism prevents the search from becoming trapped in uninformative regions while preserving valid observations accumulated in the evidence bundle.

3.4 Overall Pipeline

`CAREER` first screens the global view, returning the leading option when both criteria hold. Otherwise, it accumulates candidate views until the verifier's absolute-support and separation criteria are satisfied. Stalled search recovers to a previous state with untried legal observations. If the budget or feasible views are exhausted before acceptance, the saved global answer is returned as an unverified fallback (Appendix B.3). Algorithm 1 in Appendix A gives the complete procedure.

3.5 Theoretical Analysis

Under the stated assumptions, we derive upper bounds on the error rates of accepted answers, together with a lower bound on acceptance coverage. These bounds clarify how score reliability and evidence acquisition govern answer acceptance, while separately accounting for errors introduced by unverified fallback predictions. Detailed proofs are provided in Appendix B.

Definition 3.1. Let Y and $\widehat{Y}$ be the correct and returned option indices. The events $G = \{\Gamma = 1\}$, V, and $F = (G \cup V)^c$ form a mutually exclusive partition corresponding to global acceptance, verifier acceptance, and unverified fallback, respectively. With masses C_g, C_v, C_f, define

$$C_{\text{acc}} = C_g + C_v, \qquad r_{\text{acc}} = \Pr(\widehat{Y} \neq Y \mid G \cup V), \qquad R_K = \Pr(\widehat{Y} \neq Y), \tag{8}$$

where K is the observation budget and r_{fb} denotes conditional fallback error. Conditional risks are used only on nonempty branches, with $C_f r_{\text{fb}} = 0$ when $C_f = 0$.

Assumption 3.1. *On G, generator scores approximate the global-screening posterior within ε_g in sup norm, except on trajectories of population mass at most δ_g. On G^c, verifier shares approximate their history-conditioned posteriors within ε_v at every successfully scored round, except on trajectories of mass at most δ_v. Appendix B.1 formalizes these events. Normalization alone does not ensure alignment (Guo et al., 2017).*

Theorem 3.1. *Under finite termination and Assumption 3.1, let $N \leq K$ be the terminal observation count and define the following error bounds for the global and verifier branches:*

$$B_g = \min\left\{C_g,\ \mathbb{E}[\mathbf{1}_G \min\{1, 1 - p_{(1)} + \varepsilon_g\}] + \delta_g\right\}, \tag{9}$$

$$B_v = \min\left\{C_v,\ \mathbb{E}[\mathbf{1}_V \min\{1, 1 - \rho_{N,i_N^\star} + \varepsilon_v\}] + \delta_v\right\}, \tag{10}$$

Table 1: Performance of CAREER and existing methods on high-resolution benchmarks. **FSP** and **FCP** denote fine-grained single-instance and cross-instance perception, respectively.

Datasets	V*Bench			HR-Bench 4K			HR-Bench 8K			Avg.
Method	Attr.	Spatial	Overall	FSP	FCP	Overall	FSP	FCP	Overall	Overall
InternVL2.5-8B										
Backbone	67.8	71.1	69.1	76.0	56.0	66.0	62.0	53.0	57.5	64.2
ZoomEye	83.5$_{\uparrow 15.7}$	86.8$_{\uparrow 15.7}$	84.8$_{\uparrow 15.7}$	89.0$_{\uparrow 13.0}$	62.0$_{\uparrow 6.0}$	75.5$_{\uparrow 9.5}$	89.0$_{\uparrow 27.0}$	58.0$_{\uparrow 5.0}$	73.5$_{\uparrow 16.0}$	77.9$_{\uparrow 13.7}$
RAP	**90.4**$_{\uparrow 22.6}$	85.5$_{\uparrow 14.4}$	88.5$_{\uparrow 19.4}$	93.0$_{\uparrow 17.0}$	59.0$_{\uparrow 3.0}$	76.0$_{\uparrow 10.0}$	92.0$_{\uparrow 30.0}$	56.0$_{\uparrow 3.0}$	74.0$_{\uparrow 16.5}$	79.5$_{\uparrow 15.3}$
CVSearch	88.7$_{\uparrow 20.9}$	88.2$_{\uparrow 17.1}$	88.5$_{\uparrow 19.4}$	92.0$_{\uparrow 16.0}$	63.0$_{\uparrow 7.0}$	77.5$_{\uparrow 11.5}$	92.0$_{\uparrow 30.0}$	61.0$_{\uparrow 8.0}$	76.5$_{\uparrow 19.0}$	80.8$_{\uparrow 16.6}$
CAREER (Ours)	**90.4**$_{\uparrow 22.6}$	**92.1**$_{\uparrow 21.0}$	**91.1**$_{\uparrow 22.0}$	**94.0**$_{\uparrow 18.0}$	**64.0**$_{\uparrow 8.0}$	**79.0**$_{\uparrow 13.0}$	**94.0**$_{\uparrow 32.0}$	**62.0**$_{\uparrow 9.0}$	**78.0**$_{\uparrow 20.5}$	**82.7**$_{\uparrow 18.5}$
LLaVA-OV-7B										
Backbone	75.7	75.0	75.4	72.0	54.0	63.0	67.0	53.0	60.0	66.1
ZoomEye	93.9$_{\uparrow 18.2}$	**85.5**$_{\uparrow 10.5}$	90.6$_{\uparrow 15.2}$	84.0$_{\uparrow 12.0}$	55.0$_{\uparrow 1.0}$	69.5$_{\uparrow 6.5}$	85.0$_{\uparrow 18.0}$	53.0$_{\uparrow 0.0}$	69.0$_{\uparrow 9.0}$	76.4$_{\uparrow 10.2}$
RAP	80.9$_{\uparrow 5.2}$	77.6$_{\uparrow 2.6}$	79.6$_{\uparrow 4.2}$	85.0$_{\uparrow 13.0}$	57.0$_{\uparrow 3.0}$	71.0$_{\uparrow 8.0}$	82.0$_{\uparrow 15.0}$	53.0$_{\uparrow 0.0}$	67.5$_{\uparrow 7.5}$	72.7$_{\uparrow 6.6}$
CVSearch	93.9$_{\uparrow 18.2}$	82.9$_{\uparrow 7.9}$	89.5$_{\uparrow 14.1}$	90.0$_{\uparrow 18.0}$	62.0$_{\uparrow 8.0}$	76.0$_{\uparrow 13.0}$	88.0$_{\uparrow 21.0}$	58.0$_{\uparrow 5.0}$	73.0$_{\uparrow 13.0}$	79.5$_{\uparrow 13.4}$
CAREER (Ours)	**96.5**$_{\uparrow 20.8}$	82.9$_{\uparrow 7.9}$	**91.1**$_{\uparrow 15.7}$	**92.0**$_{\uparrow 20.0}$	**68.0**$_{\uparrow 14.0}$	**80.0**$_{\uparrow 17.0}$	**92.0**$_{\uparrow 25.0}$	**62.0**$_{\uparrow 9.0}$	**77.0**$_{\uparrow 17.0}$	**82.7**$_{\uparrow 16.6}$
Qwen2.5-VL-7B										
Backbone	73.9	67.1	71.2	85.0	52.0	68.5	78.0	52.0	65.0	68.2
ZoomEye	85.2$_{\uparrow 11.3}$	85.5$_{\uparrow 18.4}$	85.3$_{\uparrow 14.1}$	88.0$_{\uparrow 3.0}$	57.0$_{\uparrow 5.0}$	72.5$_{\uparrow 4.0}$	87.0$_{\uparrow 9.0}$	53.0$_{\uparrow 1.0}$	70.0$_{\uparrow 5.0}$	75.9$_{\uparrow 7.7}$
RAP	88.7$_{\uparrow 14.8}$	78.9$_{\uparrow 11.8}$	84.8$_{\uparrow 13.6}$	**92.0**$_{\uparrow 7.0}$	57.0$_{\uparrow 5.0}$	74.5$_{\uparrow 6.0}$	**93.0**$_{\uparrow 15.0}$	59.0$_{\uparrow 7.0}$	76.0$_{\uparrow 11.0}$	78.4$_{\uparrow 10.2}$
CVSearch	**92.2**$_{\uparrow 18.3}$	84.2$_{\uparrow 17.1}$	89.0$_{\uparrow 17.8}$	90.0$_{\uparrow 5.0}$	63.0$_{\uparrow 11.0}$	76.5$_{\uparrow 8.0}$	91.0$_{\uparrow 13.0}$	61.0$_{\uparrow 9.0}$	76.0$_{\uparrow 11.0}$	80.5$_{\uparrow 12.3}$
CAREER (Ours)	**92.2**$_{\uparrow 18.3}$	**90.8**$_{\uparrow 23.7}$	**91.6**$_{\uparrow 20.4}$	91.0$_{\uparrow 6.0}$	**65.0**$_{\uparrow 13.0}$	**78.0**$_{\uparrow 9.5}$	92.0$_{\uparrow 14.0}$	**67.0**$_{\uparrow 15.0}$	**79.5**$_{\uparrow 14.5}$	**83.0**$_{\uparrow 14.8}$

with the verifier integrand set to zero outside V. For $C_{\mathrm{acc}} > 0$,

$$r_{\mathrm{acc}} \le \frac{B_g + B_v}{C_{\mathrm{acc}}}, \qquad R_K \le B_g + B_v + C_f r_{\mathrm{fb}}. \tag{11}$$

Under the additional compatibility and reachability conditions in Assumptions B.1 and B.2,

$$C_{\mathrm{acc}} \ge \max\left\{ C_g,\ 1 - (1 - C_g) \prod_{t=1}^{K} (1 - a_t \kappa_t) - \delta_v \right\}. \tag{12}$$

Here, $a_t \kappa_t$ lower-bounds the chance that an unresolved trajectory acquires and successfully scores a posterior-concentrated observation at round t.

Remark 3.1. *The bounds separate acceptance risk from fallback risk and accommodate adaptive navigation and recovery through the realized history. Coverage depends on actual opportunities to acquire sufficient evidence. These conditional guarantees imply neither unconditional improvement over global inference nor monotone accuracy in K.*

4 Experiments

4.1 Experimental Setup

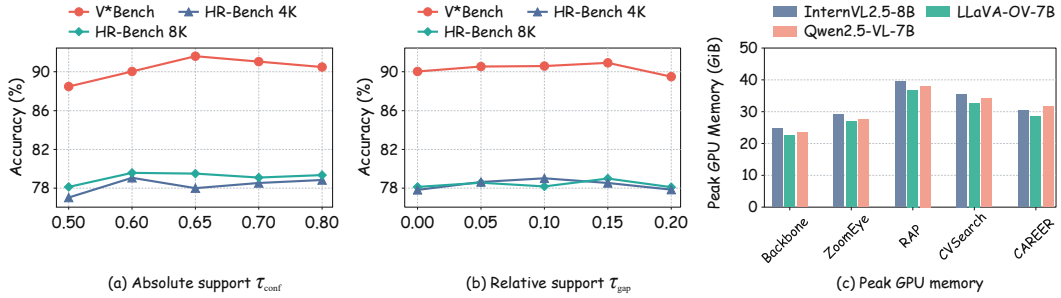
Benchmark Datasets. We evaluate CAREER on three complementary benchmarks. V*Bench (Wu & Xie, 2024) and HR-Bench (Wang et al., 2025a) assess high-resolution visual understanding; HR-Bench uses all 800 released instances per resolution (4K/8K), except for the verifier-choice subset. POPE-COCO (Li et al., 2023b) tests object hallucination through object-existence questions under three negative sampling settings. Appendix C.1 details the datasets and evaluation protocols.

Compared Baselines. For high-resolution QA, we compare CAREER with the corresponding backbone MLLM using only the global image, together with ZoomEye (Shen et al., 2025), RAP (Wang et al., 2025b), and CVSearch (Li et al., 2026). For POPE-COCO, the baselines include the same global-image model and four hallucination mitigation methods, VCD (Leng et al., 2024), MFCD (Liu et al., 2025), AllPath (Qian et al., 2026), and PAI (Liu et al., 2024). Each comparison uses the same backbone within a benchmark setting. More details are illustrated in Appendix C.2.

Implementation Details. We use the frozen `InternVL2.5-8B` (Chen et al., 2024b), `LLaVA-OV-7B` (Li et al., 2024a), and `Qwen2.5-VL-7B` (Bai et al., 2025b) as backbone MLLMs. A shared frozen `Qwen3-VL-4B-Instruct` (Bai et al., 2025a) verifier scores each option independently of the generator's provisional answer. We keep the controller configuration fixed across backbones and use `Qwen2.5-VL-7B` with the same verifier for all ablations. Appendix C.3 provides the controller hyperparameters, implementation details, and complete evaluation settings.

Table 2: POPE-COCO object-hallucination results using `Qwen2.5-VL-7B` for **CAREER** and baselines. **Random** samples absent objects at random. **Popular** selects frequently occurring absent objects, while **Adversarial** selects absent objects that often co-occur with those present in the image.

Datasets	Random			Popular			Adversarial		
Metrics	Acc.	Rec.	F1	Acc.	Rec.	F1	Acc.	Rec.	F1
Backbone	84.40	95.47	85.95	83.47	95.47	85.24	81.83	95.47	84.01
VCD	88.67 $_{\uparrow 4.27}$	94.80 $_{\downarrow 0.67}$	89.32 $_{\uparrow 3.37}$	87.63 $_{\uparrow 4.16}$	94.80 $_{\downarrow 0.67}$	88.46 $_{\uparrow 3.22}$	86.47 $_{\uparrow 4.64}$	94.80 $_{\downarrow 0.67}$	87.51 $_{\uparrow 3.50}$
MFCD	85.33 $_{\uparrow 0.93}$	95.13 $_{\downarrow 0.34}$	86.64 $_{\uparrow 0.69}$	84.47 $_{\uparrow 1.00}$	95.13 $_{\downarrow 0.34}$	85.96 $_{\uparrow 0.72}$	84.13 $_{\uparrow 2.30}$	95.13 $_{\downarrow 0.34}$	85.71 $_{\uparrow 1.70}$
AllPath	90.60 $_{\uparrow 6.20}$	95.93 $_{\uparrow 0.46}$	91.08 $_{\uparrow 5.13}$	88.60 $_{\uparrow 5.13}$	95.93 $_{\uparrow 0.46}$	89.38 $_{\uparrow 4.14}$	86.60 $_{\uparrow 4.77}$	95.93 $_{\uparrow 0.46}$	87.74 $_{\uparrow 3.73}$
PAI	86.43 $_{\uparrow 2.03}$	94.27 $_{\downarrow 1.20}$	87.42 $_{\uparrow 1.47}$	85.60 $_{\uparrow 2.13}$	94.27 $_{\downarrow 1.20}$	86.75 $_{\uparrow 1.51}$	84.17 $_{\uparrow 2.34}$	94.27 $_{\downarrow 1.20}$	85.62 $_{\uparrow 1.61}$
CAREER (Ours)	**92.77** $_{\uparrow 8.37}$	**97.47** $_{\uparrow 2.00}$	**93.09** $_{\uparrow 7.14}$	**91.23** $_{\uparrow 7.76}$	**97.47** $_{\uparrow 2.00}$	**91.75** $_{\uparrow 6.51}$	**88.33** $_{\uparrow 6.50}$	**97.47** $_{\uparrow 2.00}$	**89.31** $_{\uparrow 5.30}$



(a) Absolute support τ_{conf} (b) Relative support τ_{gap} (c) Peak GPU memory

Figure 3: (a) Absolute-support sweep with $\tau_{\text{gap}} = 0.15$. (b) Normalized-support-margin sweep with $\tau_{\text{conf}} = 0.65$. (c) Peak GPU memory across three backbone MLLMs.

4.2 Performance Comparison

High-resolution Visual Question Answering. Table 1 compares **CAREER** with global-image inference and visual-search baselines across three backbones. The results show that **Obs. ❶** ***CAREER*** *consistently improves high-resolution understanding across backbones and benchmarks.* **CAREER** achieves the highest overall accuracy across all nine backbone-benchmark combinations, with relative gains of up to 5.48% over the strongest search baseline. On HR-Bench 8K, relative gains over global-image inference range from 22.31% to 35.65%. Improvements extend to both single-instance and cross-instance perception, although individual subtask scores do not always surpass every baseline. Together, these results support acquiring complementary observations to recover fine-grained visual details and contextual relationships that a global view alone may miss.

Object Hallucination Evaluation. We assess **CAREER**'s object-existence judgments on the POPE-COCO benchmark. Table 2 reports results for `Qwen2.5-VL-7B` and supports **Obs. ❷** ***CAREER*** *consistently reduces object hallucination across POPE settings.* **CAREER** leads all nine combinations of sampling splits and metrics for this backbone and 17 of the 18 settings for the other two backbones. On Popular and Adversarial, accuracy reaches 91.23% and 88.33%, exceeding AllPath by 2.97% and 2.00% in relative terms, respectively. F1 also improves by 2.65% and 1.79% relative to AllPath, respectively, reflecting a better balance between precision and recall. Recall remains 97.47% across all three splits, indicating that improved rejection of absent objects preserves detection of present objects even under harder negatives. Appendix D.1 provides the complete cross-backbone comparisons, while Appendix D.4 analyzes answer acceptance and the remaining sources of error.

4.3 Ablation Study

As shown in Table 3, five variants assess the core components of **CAREER**. V1 disables evidence-conditioned navigation, and V2 removes prior local observations. V3 and V4 retain only the absolute-support and normalized-support-margin gates, respectively. V5 ranks candidates using query relevance alone.

Table 3: Ablation accuracy (%) on Qwen2.5-VL-7B. Arrows show drops in percentage points from full **CAREER**.

	High-resolution QA			POPE-COCO		
Var.	V*Bench	HR-4K	HR-8K	Random	Popular	Adver.
V1	87.4 $_{\downarrow 4.20}$	74.5 $_{\downarrow 3.50}$	73.0 $_{\downarrow 6.50}$	91.10 $_{\downarrow 1.67}$	89.30 $_{\downarrow 1.93}$	86.10 $_{\downarrow 2.23}$
V2	88.5 $_{\downarrow 3.10}$	75.0 $_{\downarrow 3.00}$	74.0 $_{\downarrow 5.50}$	90.90 $_{\downarrow 1.87}$	89.10 $_{\downarrow 2.13}$	85.90 $_{\downarrow 2.43}$
V3	90.1 $_{\downarrow 1.50}$	77.0 $_{\downarrow 1.00}$	77.0 $_{\downarrow 2.50}$	92.10 $_{\downarrow 0.67}$	90.30 $_{\downarrow 0.93}$	87.20 $_{\downarrow 1.13}$
V4	89.0 $_{\downarrow 2.60}$	76.0 $_{\downarrow 2.00}$	75.5 $_{\downarrow 4.00}$	91.30 $_{\downarrow 1.47}$	89.20 $_{\downarrow 2.03}$	85.80 $_{\downarrow 2.53}$
V5	89.0 $_{\downarrow 2.60}$	75.5 $_{\downarrow 2.50}$	74.5 $_{\downarrow 5.00}$	91.20 $_{\downarrow 1.57}$	89.50 $_{\downarrow 1.73}$	86.50 $_{\downarrow 1.83}$
CAREER	**91.6**	**78.0**	**79.5**	**92.77**	**91.23**	**88.33**

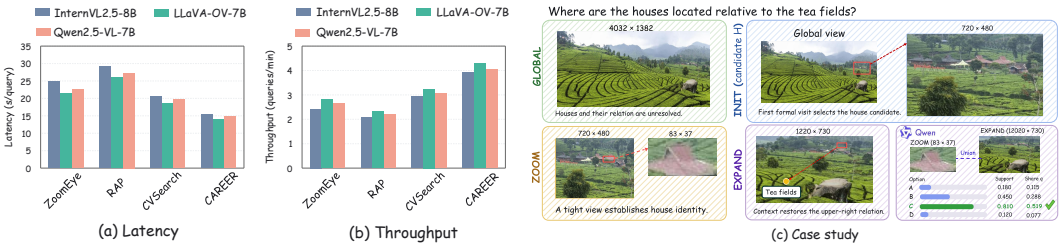


Figure 4: (a,b) Latency and throughput on H200. (c) Detail identifies the houses, while context establishes their spatial relation. Both views support option C.

These ablations yield two insights: **Obs.** ❸ *Adaptive observation, retained evidence, and visual ranking cues improve evidence acquisition.* Removing adaptive navigation or evidence retention reduces HR-Bench 8K accuracy by 6.5 and 5.5 percentage points, respectively. Query-only ranking reduces accuracy by 5.0 points on HR-Bench 8K and underperforms the full ranking in all six settings, supporting the additional value of visual cues beyond query relevance when selecting informative regions for further inspection.. **Obs.** ❹ *Combining absolute support with answer separation improves acceptance decisions.* Both single-threshold variants underperform the full gate in all six settings. On HR-Bench 8K, absolute support alone and normalized-support margin alone reduce accuracy by 2.5 and 4.0 percentage points, respectively. Appendix D.2 provides matched controls, additional mechanism ablations, and the complete candidate-ranking comparison.

4.4 Further Analysis

Sensitivity Analysis. Examining the two thresholds separately helps distinguish the effects of absolute visual support from those of answer separation, clarifying their roles in the acceptance decision. Using `Qwen2.5-VL-7B`, we vary each acceptance threshold while fixing the other at its default. As depicted in the left and middle subfigures of Figure 3, we conclude that **Obs.** ❺ *Higher acceptance thresholds do not consistently improve accuracy.* In the absolute-support sweep, V*Bench accuracy rises from 88.48% at 0.50 to 91.6% at the default 0.65, then falls to 90.58% at 0.80. In the normalized-support-margin sweep, accuracy peaks at 0.15 and drops by 2.29% relative to that peak at 0.20. Stricter thresholds alter stopping decisions and subsequent observations, so increasing a threshold need not improve final accuracy. These results favor the moderate default settings.

Efficiency Analysis. Figure 4(a,b) reports latency and throughput on a single H200, and Figure 3 (right) compares peak memory. **Obs.** ❻ ***CAREER*** *improves accuracy while reducing latency relative to the other visual-search methods.* Across the three backbones, it reduces latency by 24.6% to 25.0% and peak memory by 2.5 to 5.2 GiB relative to CVSearch while achieving higher overall accuracy (Table 1). Global backbone inference remains the fastest option, and ZoomEye uses less peak memory. Backbone-only results are reported in Appendix C.3, and budget analyses in Appendix D.3.

Case Study. Figure 4(c) illustrates why *recognizing a target and establishing its spatial relation require complementary evidence*. The tight crop identifies the houses but excludes the tea fields needed as a spatial reference. The expanded view restores this context and places the houses in the upper right relative to the tea fields. Retaining both observations in the evidence bundle resolves these two sources of uncertainty. With the target identity and reference region established, the verifier can compare the candidate spatial relations using both observations and determine whether they jointly support one answer over its alternatives. Option C then reaches an absolute support of 0.810 and a leading–runner-up normalized-support margin of 0.231, exceeding $\tau_{\text{conf}} = 0.65$ and $\tau_{\text{gap}} = 0.15$, respectively. The verifier accepts option C and terminates the search.

5 Conclusion

In this paper, we introduce **CAREER**, a training-free framework for reliable high-resolution understanding in multimodal large language models. It combines complementary multi-scale observations with evidence sufficiency verification, linking where to look and when to stop through accumulated evidence. Experiments across three MLLM backbones show consistent gains in high-resolution question answering and reduced object hallucination, with lower latency than the evaluated visual-search baselines. Future work could extend this framework to open-ended visual reasoning.

AI USE STATEMENT

Large language models were used as writing assistants during manuscript preparation. They supported English-language polishing, sentence restructuring, grammar correction, and consistency checks for terminology and LaTeX formatting. They also assisted with improving the presentation of several figures and tables. Every suggested change was reviewed and revised by the authors before inclusion.

For technical passages, model-generated suggestions were treated only as editable drafts and were checked against the method, equations, implementation, and experimental results. LLM assistance did not generate the reported measurements or determine which results were included. The authors take full responsibility for the scientific content, factual accuracy, citations, and final wording of the paper. The MLLMs evaluated as part of **CAREER** are research components of the proposed framework and are distinct from the writing assistance disclosed here.

ETHICS STATEMENT

This work follows the ICLR Code of Ethics. It does not involve human-subject research, animal experiments, or the collection of new personal data. Our evaluation uses publicly released benchmarks, including V*Bench, HR-Bench, and POPE-COCO, together with publicly available pretrained models. We use these resources for research purposes under their stated terms and cite their original sources.

High-resolution visual question answering systems can still produce incorrect or biased outputs. Although **CAREER** is designed to improve evidence acquisition and reduce object hallucination, its verifier does not provide a general safety guarantee. The method should therefore be evaluated carefully before use in sensitive or high-stakes settings. We report both verified answers and unverified fallback behavior to make this limitation explicit.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the reported experiments. The main paper and appendix describe the complete inference algorithm, backbone and verifier models, candidate construction and ranking, controller thresholds, observation budgets, prompts, evaluation metrics, and baseline settings. They also report ablations, threshold sensitivity, efficiency measurements, cross-backbone results, and verifier comparisons.

The experiments use the public V*Bench, HR-Bench, and POPE-COCO benchmarks and publicly released model checkpoints. Dataset statistics and evaluation protocols are documented in Appendix C. The implementation is available through the repository linked in the abstract, and the appendix provides the prompts and default configurations required to reproduce the method and its main evaluations.

REFERENCES

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

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 Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024a.

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*, 2024b.

Federico Cocchi, Nicholas Moratelli, Davide Caffagni, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. Llava-more: A comparative study of llms and visual backbones for enhanced visual instruction tuning. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4337–4347. IEEE, 2025.

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.

Yihao Ding, Siwen Luo, Yue Dai, Yanbei Jiang, Zechuan Li, Qiang Sun, Geoffrey Martin, Wei Liu, and Yifan Peng. A survey on mllm-based visually rich document understanding: Methods, challenges, and emerging trends. *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13319–13340, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026a.

Yiwei Fu, Hui Wan, Xiao Luo, and Minghua Deng. Star: Test-time adaptation can enhance universal prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31482–31492, 2026b.

Yuqian Fu, Runze Wang, Bin Ren, Guolei Sun, Biao Gong, Yanwei Fu, Danda Pani Paudel, Xuanjing Huang, and Luc Van Gool. Objectrelator: Enabling cross-view object relation understanding across ego-centric and exo-centric perspectives. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6530–6540. IEEE, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pp. 390–406. Springer, 2024.

Jiachen Jiang, Jinxin Zhou, Bo Peng, Xia Ning, and Zhihui Zhu. Analyzing fine-grained alignment and enhancing vision understanding in multimodal language models. *Advances in Neural Information Processing Systems*, 38:874–899, 2026.

Shixin Jiang, Jiafeng Liang, Jiyuan Wang, Xuan Dong, Heng Chang, Weijiang Yu, Jinhua Du, Ming Liu, and Bing Qin. From specific-mllms to omni-mllms: a survey on mllms aligned with multi-modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8617–8652, 2025.

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.

Jihyung Kil, Zheda Mai, Justin Lee, Arpita Chowdhury, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, and Wei-Lun Chao. Mllm-compbench: A comparative reasoning benchmark for multimodal llms. *Advances in Neural Information Processing Systems*, 37:28798–28827, 2024.

Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8):1–36, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

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*, 2024a.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9098–9108. IEEE, 2025a.

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, 2023a.

Liupeng Li, Haoqian Kang, Zhenyu Lu, Jinpeng Wang, Bin Chen, Ke Chen, and Yaowei Wang. Cvsearch: Empowering multimodal llms with cognitive visual search for high-resolution image perception. *arXiv preprint arXiv:2605.23655*, 2026.

Wenyan Li, Raphael Tang, Chengzu Li, Caiqi Zhang, Ivan Vulic, and Anders Søgaard. Lost in embeddings: Information loss in vision-language models. In *EMNLP (Findings)*, pp. 22676–22693, 2025b.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025c.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26753–26763. IEEE, 2024b.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *arXiv preprint arXiv:2502.03628*, 2025d.

Zichuan Lin, Yicheng Liu, Yang Yang, Lvfang Tao, and Deheng Ye. Adaptvision: efficient vision-language models via adaptive visual acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11923–11932, 2026.

Bingqian Liu, Fu Zhang, Guoqing Chen, and Jingwei Cheng. Multi-frequency contrastive decoding: Alleviating hallucinations for large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28556–28572, 2025.

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

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pp. 125–140. Springer, 2024.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *IEEE transactions on pattern analysis and machine intelligence*, 2026.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15630–15640. IEEE, 2024.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pp. 304–323. Springer, 2024.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *Advances in Neural Information Processing Systems*, 37:23464–23487, 2024.

Jiaye Qian, Ge Zheng, Yuchen Zhu, and Sibei Yang. Intervene-all-paths: Unified mitigation of lvlm hallucinations across alignment formats. *Advances in neural information processing systems*, 38: 142168–142198, 2026.

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

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. *Advances in Neural Information Processing Systems*, 37:8612–8642, 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*, pp. 6613–6629, 2025.

Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. When do we not need larger vision models? In *European Conference on Computer Vision*, pp. 444–462. Springer, 2024.

Yuheng Shi, Xiaohuan Pei, Minjing Dong, and Chang Xu. Catching the details: Self-distilled roi predictors for fine-grained mllm perception. In *International Conference on Learning Representations*, volume 2026, pp. 136133–136155, 2026.

Irwin Sobel, Gary Feldman, et al. A 3x3 isotropic gradient operator for image processing. *a talk at the Stanford Artificial Project in*, 1968:271–272, 1968.

Shezheng Song, Shasha Li, Shan Zhao, Xiaopeng Li, Qian Wan, Chengyu Wang, Tianwei Yan, Ma Jun, and Jie Yu. Where does vision meet language? understanding and refining visual fusion in mllms via contrastive attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10051–10060, 2026.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. *Advances in Neural Information Processing Systems*, 38:8222–8251, 2026.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37: 87310–87356, 2024a.

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

Huyen TT Tran, Van-Quang Nguyen, Farros Alferro, Kang-Jun Liu, and Takayuki Okatani. 360° image perception with mllms: A comprehensive benchmark and a training-free method. In *European Conference on Computer Vision*, pp. 296–313. Springer, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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, 2025a.

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*, 2025b.

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

Penghao Wu, Yushan Zhang, Haiwen Diao, Bo Li, Lewei Lu, and Ziwei Liu. Visual jigsaw post-training improves mllms. In *International Conference on Learning Representations*, volume 2026, pp. 93221–93245, 2026.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in llms, mllms, and beyond: Methods, theories, applications, and opportunities. *ACM Computing Surveys*, 58(8):1–41, 2026.

Xiaoyang Yi, Jing Chen, Li Peng, Yuru Bao, and Jian Zhang. Beyond the panorama: Training-free hierarchical perception-reasoning for fine-grained vision in mllms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30534–30549, 2026.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *Findings of the association for computational linguistics: ACL 2024*, pp. 12401–12430, 2024.

YiFan 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 *International Conference on Learning Representations*, volume 2025, pp. 89655–89701, 2025.

Yipeng Zhang, Yifan Liu, Zonghao Guo, Yidan Zhang, Xuesong Yang, Xiaoying Zhang, Chi Chen, Jun Song, Yuan Yao, Tat-Seng Chua, et al. Llava-uhd v2: Exploiting hierarchical vision granularity in mllms via inverse semantic pyramid. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12934–12942, 2026.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026.

Liangyu Zhong, Fabio Philipp Rosenthal, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, and Leo Schwinn. Focus: Internal mllm representations for efficient fine-grained visual question answering. *Advances in Neural Information Processing Systems*, 38:96498–96535, 2026.

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 CONTENTS

A Complete Inference Algorithm **16**

B Theoretical Analysis **16**

B.1 Notation and Assumptions 16

B.2 Proof of Theorem 3.1 18

B.3 Acceptance Risk and Fallback Behavior 19

B.4 Theoretical Quantities and Decision Mechanisms 20

C Experimental Setup **20**

C.1 Datasets and Evaluation 21

C.2 Baselines 21

C.3 Implementation 22

D Additional Experiments **23**

D.1 Cross-Backbone POPE Results 23

D.2 Matched Controls and Search Mechanisms 25

D.3 Accuracy and Computation across Observation Budgets 27

D.4 Acceptance, Error Sources, and Verifier Choice 27

E Detailed Prompts **29**

E.1 Question-Only Planning Prompt 29

E.2 MLLM Generator Prompts 30

E.3 Independent Verifier Prompt 32

F Case Studies **32**

F.1 Spatial Relation via ZOOM and EXPAND 33

F.2 Cross-Candidate Evidence Accumulation 34

F.3 Single-View Evidence Sufficiency 35

F.4 Hierarchical SPLIT for Fine Text 36

F.5 Flagged Unverified Fallback 37

A Complete Inference Algorithm

Algorithm 1 gives the complete inference procedure summarized in Section 3.4. It makes the global screen, evidence-bundle update, verifier gate, recovery rule, and bounded fallback explicit.

Algorithm 1 `CAREER`: Verification-Guided Multi-Granularity Visual Search

Require: Image $\mathbf{I}$, question q, options $\mathcal{A}$, generator $\mathcal{M}_g$, verifier $\mathcal{M}_v$, and local-observation budget K

Compute $\mathbf{p}$, save $\hat{y}_0 \leftarrow a_{\arg\max_i p_i}$, and evaluate Eq. 1 // Global screening
if $\Gamma = 1$ **then return** $\hat{y}_0$
Construct $\mathcal{C} = \mathcal{C}_{\mathrm{SAM}} \cup \mathcal{C}_{\mathrm{SLIC}}$ and rank candidates using Eq. 5
// Multi-scale candidate atlas
Initialize $\mathcal{B}_0 \leftarrow \{\mathbf{I}\}$, $t \leftarrow 0$, and select the highest-ranked feasible candidate
while $t < K$ and a feasible observation remains **do**
    Acquire $(\mathbf{I}_{t+1}, b_{t+1})$, set $t \leftarrow t+1$, and update $\mathcal{B}_t$ using Eq. 6
    // Accumulate the evidence bundle
    Update the answer and propose an action from $\mathcal{B}_t$ and earlier verifier feedback
    Evaluate every option independently and compute $(s_{t,i}, \rho_{t,i})$ using Eq. 7
    // Option-wise verification
    if $c_t \geq \tau_{\mathrm{conf}}$ and $\rho_{t,i_t^\star} - \rho_{t,j_t^\star} \geq \tau_{\mathrm{gap}}$ **then return** the leading option
    if progress in both c_t and r_t stagnates for two steps, or no local action remains, **then** `RECOVER`
    // Redirect unproductive search
    else execute the proposal, replanning after failure while a legal observation remains
end while
return $\hat{y}_0$ with an unverified-fallback flag // Bounded termination

B Theoretical Analysis

Following the three terminal branches in Algorithm 1, the analysis separates global acceptance, verifier acceptance after search, and fallback to the saved global answer. Score–posterior alignment yields accepted-answer and full-system risk bounds. Additional reachability and threshold-compatibility conditions yield an acceptance-coverage bound. These guarantees apply to the executed adaptive policy. They do not equate softmax confidence with correctness or certify visual grounding.

B.1 Notation and Assumptions

Execution and Terminal Events. All probabilities are over the task population and inference randomness, with $M \geq 2$ answer options. Let Y denote the correct option index. The generator produces scores $\mathbf{p}$ and saves their leading option, with a fixed tie rule. Its index is $\widehat{Y}_0 = \arg\max_i p_i$, corresponding to $\hat{y}_0 = a_{\widehat{Y}_0}$ in Algorithm 1. The global screen accepts the same option, so $\widehat{Y}_g = \widehat{Y}_0$ and $G = \{\Gamma = 1\}$. On G^c, the executed policy acquires at most K valid local observations. Let $N \leq K$ be the terminal number of acquired observations.

Let $I_t = 1$ when the t-th local observation is acquired and option-wise verification returns finite logits. Only such rounds define Support probabilities $s_{t,i} > 0$ and normalized relative-support shares

$$S_t = \sum_i s_{t,i}, \qquad \rho_{t,i} = s_{t,i}/S_t. \tag{B.1}$$

These are the same support shares as in Eq. 7. Let $i_t^\star = \arg\max_i \rho_{t,i}$ and $j_t^\star = \arg\max_{i \neq i_t^\star} \rho_{t,i}$, with a fixed tie rule. The verifier accepts only when $s_{t,i_t^\star} \geq \tau_{\mathrm{conf}}$ and $\rho_{t,i_t^\star} - \rho_{t,j_t^\star} \geq \tau_{\mathrm{gap}}$. This is the normalized leading–runner-up support margin used by the controller. Let V denote verifier acceptance and let $F = (G \cup V)^c$ denote fallback. Their masses C_g, C_v, and C_f sum to one, with $C_{\mathrm{acc}} = C_g + C_v$.

Let $\mathcal{G}_0$ contain all information exposed to the prediction models at global screening, so G, $\widehat{Y}_g$, and $\widehat{Y}_0$ are $\mathcal{G}_0$-measurable. Let $\mathcal{F}_t$ be the nested history after t acquired local observations. Its visual information consists of the observations actually exposed to the prediction models up to round t, after the corresponding model-specific preprocessing, together with their metadata, model outputs, and controller decisions used to continue or stop. The underlying source image $\mathbf{I}$ may be retained by the

controller for localization and crop extraction, but is not itself treated as an observed variable in $\mathcal{F}_t$ unless it is exposed through such an observation. Every control phase is assumed to terminate after finitely many attempts. Set $N = 0$ on G. After termination, extend the history constantly and set $I_t = 0$ for $t > N$. With the continuation or termination decision included in $\mathcal{F}_t$, N is a bounded stopping time with respect to $(\mathcal{F}_t)_{t=0}^K$.

Population Posteriors. Define $\pi_i^g = \Pr(Y = i \mid \mathcal{G}_0)$ and $\pi_{t,i} = \Pr(Y = i \mid \mathcal{F}_t)$ for $0 \le t \le K$. The scores $\mathbf{p}$ and $\boldsymbol{\rho}_t$ are not assumed to be posteriors. The following condition states the branchwise alignment needed by the risk bound.

To formalize Assumption 3.1, take $\varepsilon_g, \varepsilon_v \in [0, 1]$ and define

$$E_g = \{\|\mathbf{p} - \boldsymbol{\pi}^g\|_\infty \le \varepsilon_g\}, \tag{B.2}$$

$$E_v = \{\|\boldsymbol{\rho}_t - \boldsymbol{\pi}_t\|_\infty \le \varepsilon_v \text{ for every } 1 \le t \le N \text{ with } I_t = 1\}. \tag{B.3}$$

An empty search trajectory satisfies E_v. The branchwise alignment condition requires $\delta_g, \delta_v \ge 0$ such that

$$\Pr(G \cap E_g^c) \le \delta_g, \qquad \Pr(G^c \cap E_v^c) \le \delta_v. \tag{B.4}$$

These are population failure masses, rather than conditional error rates. Generator alignment is needed on its accepted branch, while verifier alignment is required along the executed search trajectory.

Assumption 3.1, together with bounded stopping, is sufficient for the accepted-answer and full-system risk bounds in Theorem 3.1. It requires neither reachability nor independence among observations.

Additional Conditions for Acceptance Coverage. The coverage bound additionally requires that posterior-concentrated histories be reachable and that the verifier thresholds accept them.

Definition B.1 (Posterior concentration)**.** *For $\eta \in [0, 1/2)$, let $H_t = \{\max_i \pi_{t,i} \ge 1 - \eta\}$ on a successfully scored round $I_t = 1$. This is discriminative sufficiency relative to the population. High concentration can also reflect task priors, so H_t does not establish that a particular target, attribute, or relation is visibly resolved.*

Assumption B.1 (Compatibility at concentrated histories)**.** *The verifier thresholds are in $(0, 1]$. There is $S_- > 0$ such that, on $G^c \cap E_v$, every successfully scored round satisfying H_t has $S_t \ge S_-$, and*

$$S_-(1 - \eta - \varepsilon_v) \ge \tau_{\text{conf}}, \qquad 1 - 2\eta - 2\varepsilon_v \ge \tau_{\text{gap}}. \tag{B.5}$$

Assumption B.2 (Reachability through actual observations)**.** *Using the successfully scored observation indicator I_t defined above, let $V_{\le t}$ denote verifier acceptance by round t, and define*

$$U_0 = G^c, \qquad U_t = G^c \cap V_{\le t}^c \cap \bigcap_{j=1}^{t} (\{I_j = 1\} \cap H_j)^c. \tag{B.6}$$

When $\Pr(U_{t-1}) > 0$, set

$$a_t = \Pr(I_t = 1 \mid U_{t-1}). \tag{B.7}$$

If $\Pr(U_{t-1} \cap \{I_t = 1\}) > 0$, a number $\kappa_t \in [0, 1]$ satisfies

$$\Pr(H_t \mid U_{t-1}, I_t = 1) \ge \kappa_t. \tag{B.8}$$

When either conditioning event is empty, set $a_t \kappa_t = 0$. No claim of positive reachability is made without an actual observation opportunity.

The factors a_t and κ_t are policy-dependent. In particular, a_t absorbs exhausted candidates, control limits, input-capacity limits, and scoring failures. Selection and localization may remain adaptive and dependent. No independence among observations is assumed.

Together, branchwise alignment controls the error mass of the two accepted routes, while reachability and threshold compatibility turn posterior-concentrated observations into verifier acceptance. These are exactly the two implications used in Theorem 3.1.

B.2 Proof of Theorem 3.1

Step 1. Use the Appropriate Posterior on Each Acceptance Branch. Because generator acceptance is decided using $\mathcal{G}_0$,

$$\Pr(\widehat{Y}_g \neq Y, G) = \mathbb{E}[\mathbf{1}_G(1 - \pi^g_{\widehat{Y}_g})]. \tag{B.9}$$

Put $d_g = \min\{1, 1 - p_{(1)} + \varepsilon_g\}$ and $D_g = \mathbb{E}[\mathbf{1}_G d_g]$. On $G \cap E_g$, the conditional error is at most d_g. On $G \cap E_g^c$ it is at most one. Therefore,

$$\Pr(\widehat{Y}_g \neq Y, G) \leq \mathbb{E}[\mathbf{1}_G d_g] + \mathbb{E}[\mathbf{1}_{G \cap E_g^c}(1 - d_g)] \tag{B.10}$$

$$\leq D_g + \delta_g, \tag{B.11}$$

and the probability is also at most C_g. Thus the global accepted-error bound used in Theorem 3.1 is

$$B_g = \min\left\{C_g,\, \mathbb{E}[\mathbf{1}_G \min\{1, 1 - p_{(1)} + \varepsilon_g\}] + \delta_g\right\}. \tag{B.12}$$

Later search information on other trajectories does not change this identity or require generator scores to align with a later posterior.

For search acceptance, the conditional-expectation identity at bounded stopping gives

$$\Pr(Y = i \mid \mathcal{F}_N) = \sum_{t=0}^{K} \mathbf{1}\{N = t\}\pi_{t,i} = \pi_{N,i}. \tag{B.13}$$

Indeed, intersecting any event in the stopped σ-algebra with $\{N = t\}$ gives an event in $\mathcal{F}_t$, to which the defining conditional-expectation identity applies. Summing over t proves the equation. The events V, E_v, and the terminal prediction are $\mathcal{F}_N$-measurable. On $V \cap E_v$,

$$1 - \pi_{N,i^\star_N} \leq d_v := \min\{1, 1 - \rho_{N,i^\star_N} + \varepsilon_v\}. \tag{B.14}$$

Extend d_v by zero off V. With $D_v = \mathbb{E}[\mathbf{1}_V d_v]$, the same calculation yields

$$\Pr(\widehat{Y} \neq Y, V) \leq \min\{C_v, D_v + \delta_v\} = B_v. \tag{B.15}$$

Together with Eq. B.11, this proves

$$\Pr(\widehat{Y} \neq Y, G \cup V) \leq B_g + B_v. \tag{B.16}$$

Dividing by $C_{\mathrm{acc}} > 0$ proves the selective-risk inequality. The disjoint fallback contribution is exactly $C_f r_{\mathrm{fb}}$, proving the full-system inequality. A zero-probability branch contributes zero, regardless of the undefined conditional risk on that branch.

Step 2. Concentration Triggers the Verifier Gate under Compatibility. At a successfully scored round on $G^c \cap E_v \cap H_t$, choose j with $\pi_{t,j} = \max_i \pi_{t,i} \geq 1 - \eta$. Since the posterior probabilities sum to one, $\pi_{t,i} \leq \eta$ for every $i \neq j$. Alignment therefore gives

$$\rho_{t,j} \geq 1 - \eta - \varepsilon_v, \qquad \rho_{t,i} \leq \eta + \varepsilon_v \quad (i \neq j). \tag{B.17}$$

Consequently,

$$\rho_{t,j} - \max_{i \neq j} \rho_{t,i} \geq 1 - 2\eta - 2\varepsilon_v \geq \tau_{\mathrm{gap}} > 0. \tag{B.18}$$

Thus j is the unique verifier leader and its leading–runner-up margin passes the actual controller gate. The absolute-support gate also passes because

$$s_{t,j} = S_t \rho_{t,j} \geq S_-(1 - \eta - \varepsilon_v) \geq \tau_{\mathrm{conf}}. \tag{B.19}$$

Hence every posterior-concentrated, aligned, successfully scored round on the search branch is accepted. This implication concerns only executed rounds with valid scores. It makes no claim about unseen regions or failed scoring.

Step 3. Count Only Actual Opportunities to Resolve Search. The events U_t satisfy

$$U_t \subseteq U_{t-1} \setminus (\{I_t = 1\} \cap H_t). \tag{B.20}$$

For a nonempty conditioning event, reachability therefore gives

$$\Pr(U_t) \le \Pr(U_{t-1})\big[1 - \Pr(I_t = 1, H_t \mid U_{t-1})\big] \tag{B.21}$$

$$\le (1 - a_t\kappa_t)\Pr(U_{t-1}). \tag{B.22}$$

The inequality is also valid when $\Pr(U_{t-1}) = 0$. Since $\Pr(U_0) = 1 - C_g$, induction yields

$$\Pr(U_K) \le (1 - C_g)\prod_{t=1}^{K}(1 - a_t\kappa_t). \tag{B.23}$$

On $F \cap E_v$, Step 2 excludes H_t at every successfully scored verifier round, hence $F \cap E_v \subseteq U_K$. Also $F \subseteq G^c$, so

$$C_f \le (1 - C_g)\prod_{t=1}^{K}(1 - a_t\kappa_t) + \delta_v. \tag{B.24}$$

Together with $C_{\mathrm{acc}} = 1 - C_f \ge C_g$, this proves the coverage bound. For $K = 0$, the product is empty and equals one. The actual coverage is exactly C_g. □

Threshold-Only Consequences. Suppose the global-screen thresholds lie in $(0, 1]$. Its gate implies

$$p_{(1)} \ge \beta_g := \max\left\{\tau_{\mathrm{conf}}, \frac{1 + (M-1)\tau_{\mathrm{gap}}}{M}\right\}. \tag{B.25}$$

The second bound follows because every nonleading probability is at most $p_{(1)} - \tau_{\mathrm{gap}}$ and all probabilities sum to one. At verifier acceptance, the margin gate implies

$$\rho_{N,i_N^\star} - \rho_{N,i} \ge \tau_{\mathrm{gap}} \quad \text{for every } i \ne i_N^\star. \tag{B.26}$$

Since the shares sum to one, summing these $M - 1$ inequalities gives

$$\rho_{N,i_N^\star} \ge \frac{1 + (M-1)\tau_{\mathrm{gap}}}{M}. \tag{B.27}$$

The absolute-support gate gives $\rho_{N,i_N^\star} \ge \tau_{\mathrm{conf}}/S_N$. Because $S_N \le M$ for option-wise probabilities, one may take $S_+ = M$. A smaller valid upper bound can be used if established separately on aligned accepted histories. Define

$$\beta_v = \min\left\{1, \max\left\{\frac{1 + (M-1)\tau_{\mathrm{gap}}}{M}, \frac{\tau_{\mathrm{conf}}}{S_+}\right\}\right\}, \tag{B.28}$$

$$\alpha_g = \min\{1, 1 - \beta_g + \varepsilon_g\}, \qquad \alpha_v = \min\{1, 1 - \beta_v + \varepsilon_v\}. \tag{B.29}$$

If $\tau_{\mathrm{conf}} > S_+$, aligned verifier acceptance is impossible. Separating aligned and exceptional trajectories yields, for $b \in \{g, v\}$,

$$\Pr(\widehat{Y} \ne Y, \text{branch } b) \le \alpha_b C_b + (1 - \alpha_b)\min\{C_b, \delta_b\}. \tag{B.30}$$

Either this bound or B_b may be used, taking their minimum when both apply. The score-dependent bound can exploit stronger realized confidence, whereas the threshold-only form depends only on the gate. For example, with four options, $\tau_{\mathrm{conf}} = 0.65$, $\tau_{\mathrm{gap}} = 0.15$, and $S_+ = 4$, one obtains $\beta_v = 0.3625$. The default thresholds alone therefore do not provide a low-error certificate.

B.3 Acceptance Risk and Fallback Behavior

For nonempty branches, let $r_g, r_v, r_{\mathrm{fb}}$ be their respective conditional errors. The exact decomposition is

$$R_K = C_g r_g + C_v r_v + C_f r_{\mathrm{fb}}, \tag{B.31}$$

$$C_{\mathrm{acc}} r_{\mathrm{acc}} = C_g r_g + C_v r_v. \tag{B.32}$$

Separating generator fast-path acceptance from verifier acceptance is necessary to estimate the branchwise terms. A combined accepted-error rate cannot establish which branch is reliable. The acceptance/fallback table in Appendix D.4 reports pooled non-fallback acceptance and its conditional error, corresponding to C_{acc} and r_{acc}. It does not separately identify C_g, C_v or certify either branch's alignment. Retaining a justified fallback-risk estimate can be much tighter than replacing its contribution by C_f. If no justified information about fallback risk is available, use $r_{\mathrm{fb}} \leq 1$.

The saved answer is unchanged on F, but search selects the fallback subset adaptively. Alignment of generator scores with $\boldsymbol{\pi}^g$ on G does not bound error conditional on this different subset, and those scores need not align with $\boldsymbol{\pi}_N$. Let $\widehat{Y}_A$ denote the output on $A = G \cup V$. The exact accuracy change relative to returning $\widehat{Y}_0$ on every task is

$$\begin{aligned} \mathrm{Acc}_{\mathrm{full}} - \mathrm{Acc}_0 = \; & \Pr(A, \widehat{Y}_0 \neq Y, \widehat{Y}_A = Y) \\ & - \Pr(A, \widehat{Y}_0 = Y, \widehat{Y}_A \neq Y). \end{aligned} \tag{B.33}$$

Because the saved global answer is the generator's leading option, the fast path contributes zero to this difference. Only verifier-accepted outputs can change the baseline answer. There is no unconditional guarantee of improvement.

B.4 Theoretical Quantities and Decision Mechanisms

The effective resolution rate is $h_t = a_t \kappa_t$. Since $1 - x \leq e^{-x}$ for $x \in [0, 1]$, the coverage result also gives

$$C_f \leq (1 - C_g) \exp\left(- \sum_{t=1}^{K} h_t \right) + \delta_v. \tag{B.34}$$

For example, if an executed policy satisfies $h_t \geq h > 0$ for the first $L \leq K$ rounds, then $C_f \leq (1 - C_g)(1 - h)^L + \delta_v$. Such a positive bound must include the loss of opportunities from exhaustion. It cannot be assigned to padded rounds. A finite candidate set or input limit may leave a nonzero fallback mass at every larger budget.

An asymptotic coverage statement needs additional assumptions on a family of policies. If budgets truncate one common policy, alignment failure is uniformly bounded by δ_v for every truncation, and the cumulative effective hazard diverges, then $\liminf_{K \to \infty} C_{\mathrm{acc}}(K) \geq 1 - \delta_v$. The implemented generator receives the remaining budget, so different budgets need not induce such common prefixes. Neither this conditional limit nor monotonic accuracy is claimed for the reported budget sweep.

The executed policy constructs and ranks the joint SAM–SLIC atlas before any local observation. Each local round proposes an action from current evidence and earlier feedback, then verifies the current bundle before executing that proposal. Recovery monitors improvements in c_t and r_t and preserves accumulated evidence. These choices are part of $\mathcal{F}_t$ and may change both a_t and κ_t. Both proofs accommodate this policy through its executed history and policy-dependent reachability factors. The bound does not establish that one component increases either quantity. Thresholds change acceptance and therefore subsequent trajectories. These effects require matched empirical comparisons. Posterior concentration can arise from distributional priors as well as newly observed details, and is not a substitute for testing evidence grounding.

The branch indicators, accepted scores, numbers of actual observations, and branchwise correct/error counts are measurable from execution logs with task labels. For fixed tolerances, empirical averages estimate D_g, D_v, C_g, C_v. Finite-sample confidence bounds are needed before interpreting such estimates as population bounds. In contrast, one labeled answer per task does not reveal $\boldsymbol{\pi}_t$, certify the uniform alignment events, or identify the latent-history reachability rates. An empirical score-calibration plot does not establish Assumption 3.1 after adaptive stopping. The experiments do not estimate or certify numerical values of $\varepsilon_g, \varepsilon_v, \delta_g, \delta_v, \kappa_t$, or the support-scale conditions. Accordingly, these results are conditional characterizations, not certified risk levels for the evaluated system.

C Experimental Setup

This section details the evaluation protocol and inference configuration of **CAREER**, supplementing Section 4.1 and the additional experiments in Appendix D.

C.1 Datasets and Evaluation

The evaluation covers high-resolution visual understanding and object-existence judgments:

- **V*Bench** (Wu & Xie, 2024) contains 191 questions on attribute recognition and spatial reasoning, with average image resolution 2246×1582. We report task-specific and overall accuracy.
- **HR-Bench** (Wang et al., 2025a) contains 200 base questions per resolution (4K and 8K), covering fine-grained single-instance perception (FSP) and cross-instance perception (FCP). Except for the verifier-choice study in Table S8, we evaluate all four cyclic option permutations, yielding 800 evaluation instances per resolution, and average their accuracies following the released protocol. The 4K images are question-relevant crops of the 8K images. We report FSP, FCP, and overall accuracy.
- **POPE-COCO** (Li et al., 2023b) contains Random, Popular, and Adversarial splits, each with 3,000 object-existence questions and equal numbers of positive and negative examples. Negative questions concern randomly selected absent objects, frequent absent objects, or absent objects that commonly co-occur with present objects, respectively. We report accuracy, recall, and F1.

Each model configuration is evaluated once per released evaluation instance, without excluding outputs or averaging independent runs. Metrics are reported as percentages. For POPE-COCO, `yes` is positive and metrics are derived from integer confusion counts. Overall accuracy includes accepted and fallback answers, whose outcomes are separated in Appendix D.4.

C.2 Baselines

For high-resolution QA, we compare **CAREER** with the corresponding backbone MLLM and three training-free visual-search approaches:

- **Backbone** directly answers the question from the global image under its native input resizing, without cropping or additional visual search. It serves as the reference for measuring the gains brought by each visual-search method under the same MLLM.
- **ZoomEye** (Shen et al., 2025) models a high-resolution image as a hierarchical tree, where the root is the full image and each child node is a zoomed-in sub-region of its parent. The algorithm enables MLLMs to simulate human-like zooming behavior by navigating from root to leaf nodes in search of task-relevant visual evidence.
- **RAP** (Wang et al., 2025b) partitions the high-resolution image into crops and retrieves those most relevant to the query using a visual retriever. It preserves the relative positions of the retrieved crops with a spatial-awareness layout and adaptively determines how many crops to retain.
- **CVSearch** (Li et al., 2026) first assesses whether the global view is sufficient and, if not, attempts expert-assisted localization before broader search. It decomposes the image into semantically coherent regions via adaptive patching and explores local details with a dynamic bottom-up search driven by visual complexity.

For POPE-COCO, the corresponding backbone is retained and compared with four inference-time hallucination-mitigation methods:

- **VCD** (Leng et al., 2024) contrasts the output distribution conditioned on the original image with that conditioned on a noise-distorted image. Since the distorted input amplifies language priors, subtracting its distribution under an adaptive plausibility constraint suppresses hallucinations.
- **MFCD** (Liu et al., 2025) attributes object hallucination to insufficient frequency information, which amplifies language priors. It performs contrastive decoding against outputs obtained when the model neglects high- or low-frequency image components, thereby removing the resulting hallucination distribution.
- **AllPath** (Qian et al., 2026) shows that hallucination arises from the interplay of image-to-input-text, image-to-output-text, and text-to-text pathways. It identifies critical hallucination heads within each pathway and intervenes on them, with strategies tailored to discriminative and generative question formats.
- **PAI** (Liu et al., 2024) amplifies the attention assigned to image tokens during decoding so that generation relies more on visual evidence. It further subtracts the logits obtained from a text-only input to counteract the bias induced by language priors.

The controls in Appendix D.2 share the generator, verifier, prompts, decoding settings, and maximum local-observation budget. CVSearch + same verifier retains CVSearch's candidate construction and traversal while applying the shared verification gate and saved-answer fallback after each observation. The fixed five-observation control supplies the global image and four non-overlapping quadrants together. Each controller ablation changes only its designated component, as specified in Appendix D.2.

C.3 Implementation

Table S1 lists the shared controller configuration used across all evaluations.

- **Models and Evidence Inputs.** The frozen generators are `InternVL2.5-8B`, `LLaVA-OV-7B`, and `Qwen2.5-VL-7B`. Controller analyses use `Qwen2.5-VL-7B`, and the default frozen verifier is `Qwen3-VL-4B-Instruct`. SAM 3 supplies phrase-conditioned localization and cached dense features. CLIP measures query–region relevance. Each MLLM receives the evidence bundle in a fixed order, beginning with the global image and followed by retained local observations with source coordinates and metadata. Crops use original-resolution pixels with model-specific input resizing. Duplicate observations are merged, recovery preserves the evidence bundle, and the option-wise verifier does not receive the generator's provisional answer.
- **Configuration and Budget.** Settings are selected on development images disjoint from the test data and held fixed across the three generators and six evaluation settings, except where a parameter or verifier is explicitly varied. Candidate ranking follows Section 3 and Appendix C.3. The default budget permits eight valid local observations after the global image. Failed localization, duplicate crops, and recovery do not add observations. The budget sweep uses 0, 1, 2, 4, and 8 local observations.
- **Multi-Scale Candidate Atlas.** SAM proposals form the object-aware component. The SLIC component uses 600 atoms, a position weight of 3.5, split and keep thresholds of 0.30 and 0.15, a maximum hierarchy depth of 3, and between 4 and 8 splits. Local normalization inside tree construction and silhouette scoring are enabled. Both candidate sources are merged and ranked before the first local observation.
- **Verification and Recovery.** The initial global pass applies Eq. 1. After each valid local observation, the verifier applies the absolute-support and normalized-support-margin gates in Section 3, with $\tau_{\text{conf}} = 0.65$ and $\tau_{\text{gap}} = 0.15$. Recovery instead monitors the leading absolute support $c_t = s_{t,i_t^\star}$ and relative share $r_t = \rho_{t,i_t^\star}$. It triggers when both $c_t - c_{t-1} < \delta_s$ and $r_t - r_{t-1} < \delta_r$ hold for two consecutive steps, with $\delta_s = \delta_r = 0.01$. No unchanged-leader condition is imposed. Recovery also triggers when the current state has no remaining legal observation, returning to a previous state with an untried legal observation while retaining $\mathcal{B}_t$. A failed proposal removes that attempt from the legal choices and permits replanning before this exhaustion condition is tested. The generator proposes only `ZOOM`, `EXPAND`, `SPLIT`, or `NEXT`. Input-capacity limits are included in observation feasibility. If the budget or feasible observations are exhausted without acceptance, the controller returns the saved global answer as an unverified fallback.

Appendix E provides the prompt templates.

Edge density is not compared directly on each candidate's native pixel grid. For candidate box b_c, we crop the corresponding region, convert it to grayscale in $[0, 1]$, and resize it with preserved aspect ratio and padding to a fixed resolution $H_s \times W_s$:

$$(Y_c, M_c) = \text{ResizePad}(\text{Gray}(\mathbf{I}[b_c]); H_s, W_s)\,. \tag{C.1}$$

Here Y_c is the resized grayscale image and $M_c \in \{0, 1\}^{H_s \times W_s}$ marks valid crop pixels with one and padding with zero. All candidates use the same resizing, interpolation, and padding rules. To exclude artificial gradients at the padding boundary, we erode M_c once with a 3×3 structuring element and compute gradients only within the resulting mask M_c^-.

Using the standard Sobel kernels

$$K_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \qquad K_y = K_x^\top, \tag{C.2}$$

Table S1: Default **CAREER** controller configuration. The local-observation budget excludes the initial global image.

Setting	Symbol	Value
Query relevance weight	w	0.70
Visual information weight	$1-w$	0.30
Feature dispersion share of visual score	λ	0.50
Absolute-support threshold	τ_{conf}	0.65
Normalized-support-margin threshold	τ_{gap}	0.15
Minimum support improvement	δ_s	0.01
Minimum relative-support improvement	δ_r	0.01
Stagnation patience	P	2
Maximum valid local observations	K	8

we compute gradient magnitude and define edge density as its masked mean:

$$G_c(x) = \sqrt{((K_x * Y_c)(x))^2 + ((K_y * Y_c)(x))^2}, \qquad D_e(c) = \frac{\sum_x M_c^-(x) G_c(x)}{\sum_x M_c^-(x)}. \tag{C.3}$$

Eq. C.3 requires $\sum_x M_c^-(x) > 0$. The valid-size check must ensure that pixels remain after mask erosion. An empty mask cannot be used in this average. Fixed output resolution and masked averaging reduce the direct effects of native crop size, valid-pixel count, and padding, but do not guarantee strict scale invariance. The raw values $S_q(c)$, $D_f(c)$, and $D_e(c)$ enter Eq. 5 directly, without min–max normalization across candidates.

Main experiments use two NVIDIA H200 GPUs. Efficiency evaluation uses one H200. Latency is measured in seconds per question, throughput in questions per minute, and peak GPU memory in GiB. Budget analyses report acquired observations and visual tokens. Latency includes computation spent on failed attempts. Acceptance coverage includes both global-screened and verifier-accepted answers. Coverage and overall wrong acceptance use all questions as their denominator. Accepted error and fallback accuracy use only their respective branches (Appendix D.4).

Backbone-only efficiency. On a single H200, global-image inference with `InternVL2.5-8B`, `LLaVA-OV-7B`, and `Qwen2.5-VL-7B` takes 1.68, 1.44, and 1.56 seconds per query, respectively. The corresponding throughputs are 35.71, 41.67, and 38.46 queries per minute. These backbone-only results provide a reference for the visual-search methods in Figure 4(a,b).

D Additional Experiments

This section extends the main results with cross-backbone comparisons, matched controls, observation-budget analyses, and verifier diagnostics. Unless explicitly varied, all experiments use the configuration in Appendix C.

D.1 Cross-Backbone POPE Results

We first test whether the POPE gains transfer to other base MLLMs. Table S2 complements the `Qwen2.5-VL-7B` results in Table 2 with complete results for `InternVL2.5-8B` and `LLaVA-OV-7B` under the same evaluation protocol.

CAREER obtains the best result on all nine metrics with `InternVL2.5-8B`. With `LLaVA-OV-7B`, it is best on eight of nine metrics: relative to AllPath, its Random accuracy is 0.11% lower, while recall and F1 are 2.68% and 0.12% higher, respectively. These results show that the object-existence gains transfer across base MLLMs while preserving the expected accuracy–recall trade-offs.

Table S2: POPE-COCO object-existence results with `InternVL2.5-8B` and `LLaVA-OV-7B`. Acc., Rec., and F1 are percentages recomputed from the integer confusion counts in Table S3. Bold marks the best result for each base MLLM, including ties.

	Random			Popular			Adversarial		
Method	Acc.↑	Rec.↑	F1↑	Acc.↑	Rec.↑	F1↑	Acc.↑	Rec.↑	F1↑
`InternVL2.5-8B`									
Backbone	78.43	92.73	81.13	76.33	92.73	79.67	72.33	92.73	77.02
+ VCD	81.27	90.87	82.91	79.47	90.87	81.57	75.63	90.87	78.85
+ MFCD	80.57	91.93	82.55	78.83	91.93	81.28	74.90	91.93	78.55
+ AllPath	84.17	93.20	85.48	82.37	93.20	84.09	79.47	93.20	81.95
+ PAI	82.63	92.47	84.19	80.70	92.47	82.73	77.27	92.47	80.27
+ `CAREER` (Ours)	**86.37**	**95.53**	**87.51**	**84.63**	**95.53**	**86.14**	**80.73**	**95.53**	**83.22**
`LLaVA-OV-7B`									
Backbone	85.13	94.47	86.40	83.43	94.47	85.08	83.17	94.47	84.88
+ VCD	87.33	93.07	88.02	85.77	93.07	86.74	84.63	93.07	85.83
+ MFCD	86.57	95.27	87.64	84.90	95.27	86.32	84.27	95.27	85.83
+ AllPath	**91.47**	94.73	91.74	87.83	94.73	88.62	86.53	94.73	87.55
+ PAI	87.93	93.87	88.61	86.37	93.87	87.32	85.40	93.87	86.54
+ `CAREER` (Ours)	91.37	**97.27**	**91.85**	**89.23**	**97.27**	**90.03**	**88.73**	**97.27**	**89.62**

Table S3: Integer confusion counts underlying Table S2. Each cell reports TP/TN/FP/FN. Every split contains 1,500 positive and 1,500 negative questions ($N = 3{,}000$). No outputs are excluded and no results are averaged across runs.

Method	Random	Popular	Adversarial
`InternVL2.5-8B`			
Backbone	1391/962/538/109	1391/899/601/109	1391/779/721/109
+ VCD	1363/1075/425/137	1363/1021/479/137	1363/906/594/137
+ MFCD	1379/1038/462/121	1379/986/514/121	1379/868/632/121
+ AllPath	1398/1127/373/102	1398/1073/427/102	1398/986/514/102
+ PAI	1387/1092/408/113	1387/1034/466/113	1387/931/569/113
+ `CAREER` (Ours)	**1433/1158/342/67**	**1433/1106/394/67**	**1433/989/511/67**
`LLaVA-OV-7B`			
Backbone	1417/1137/363/83	1417/1086/414/83	1417/1078/422/83
+ VCD	1396/1224/276/104	1396/1177/323/104	1396/1143/357/104
+ MFCD	1429/1168/332/71	1429/1118/382/71	1429/1099/401/71
+ AllPath	1421/1323/177/79	1421/1214/286/79	1421/1175/325/79
+ PAI	1408/1230/270/92	1408/1183/317/92	1408/1154/346/92
+ `CAREER` (Ours)	**1459/1282/218/41**	**1459/1218/282/41**	**1459/1203/297/41**

Table S4: Matched controls and mechanism ablations with `Qwen2.5-VL-7B`. Accuracy is reported in percent. Observations and latency are averages per question. Bold marks the highest accuracy in each column, including ties.

Variant	High-resolution acc. ↑			POPE acc. ↑			Cost/query ↓	
	V*	4K	8K	Rand.	Pop.	Adv.	Obs.	Lat. (s)
Global and fixed-view controls								
Backbone	71.2	68.5	65.0	84.40	83.47	81.83	1.00	1.56
Verifier direct scoring	77.0	70.0	67.5	86.03	87.70	82.97	1.00	2.94
Generator + verifier	81.2	76.0	76.5	87.33	86.13	85.17	1.00	4.18
Fixed 5-observation grid + verifier	86.9	75.0	74.0	89.60	88.07	85.73	5.00	6.73
Matched search baseline								
CVSearch + same verifier	88.5	74.5	75.0	91.07	89.53	86.70	4.86	21.84
Controller ablations								
Candidate-atlas traversal only	87.4	74.5	73.0	91.10	89.30	86.10	3.72	12.91
Without evidence-bundle accumulation	88.5	75.0	74.0	90.90	89.10	85.90	4.47	13.88
Absolute-only gate	90.1	77.0	77.0	92.10	90.30	87.20	4.10	13.41
Margin-only gate	89.0	76.0	75.5	91.30	89.20	85.80	3.96	13.07
Without tree candidates	**91.6**	75.0	76.5	91.20	89.43	86.90	5.39	16.72
Without ZOOM/EXPAND	89.0	75.5	77.0	90.97	89.17	86.70	3.86	13.02
Without RECOVER	90.1	**78.5**	78.0	**92.77**	90.07	87.30	4.08	13.58
Random candidate order	89.0	78.0	75.5	91.37	89.73	86.80	4.55	14.46
Full `CAREER`	**91.6**	78.0	**79.5**	**92.77**	**91.23**	**88.33**	4.41	14.81

D.2 Matched Controls and Search Mechanisms

We examine how the search controller changes accuracy and computation when its generator and verifier are fixed. Unless otherwise specified, `Qwen2.5-VL-7B` generates answers and `Qwen3-VL-4B-Instruct` verifies the evidence. The variants share prompts, decoding settings, and a maximum of eight local observations. Table S4 groups global and fixed-view controls, the matched search baseline, and controller ablations. Table S5 isolates the candidate-ranking signals. Both tables report the three high-resolution benchmarks and all three POPE splits.

Matched controls. The global and fixed-view controls separate verification from acquiring new evidence. Verifier direct scoring evaluates every option using only the global image and returns the highest-scoring option without calling the generator. Generator + verifier generates one global answer and checks every option without cropping or further search. Fixed 5-observation grid supplies the global image and its four non-overlapping quadrants together. It uses no query-conditioned localization, dynamic ranking, or action selection. The matched CVSearch control retains CVSearch's candidates and search order, then applies the same verifier, acceptance thresholds, and fallback rule after each valid new observation.

Controller ablations. The controller ablations isolate how **`CAREER`** visits and revisits evidence. Candidate-atlas traversal only disables ZOOM, SPLIT, and EXPAND, visiting the jointly ranked candidate queue through NEXT while retaining evidence-bundle accumulation, the two-threshold gate, and fallback. Without evidence-bundle accumulation retains only the global image and current local observation for generation, verification, and replanning. Absolute-only and Margin-only retain one acceptance condition each. These four variants correspond to V1–V4 in Table 3; their observation counts and latencies are included here to clarify the accompanying computational changes. Without tree candidates uses only SAM candidates and therefore has no SPLIT action. Without

ZOOM/EXPAND removes both within-candidate adjustments. Without RECOVER advances to the next unvisited global candidate when progress stalls or local actions are exhausted. Random candidate order keeps the candidate set and budget, but samples uniformly from unvisited candidates at each NEXT.

Accuracy and computation. Table S4 yields three further insights: **Obs. ❼** ***CAREER*** *improves accuracy while using fewer observations than the matched CVSearch controller.* Its relative accuracy gains over the matched controller on V*Bench, HR-4K, and HR-8K are 3.50%, 4.70%, and 6.00%, respectively. The corresponding relative gains on POPE Random, Popular, and Adversarial are 1.87%, 1.90%, and 1.88%. Mean observations fall from 4.86 to 4.41, and latency from 21.84 to 14.81 seconds. The shared gate and fallback lower CVSearch's accuracy on all three high-resolution settings relative to original CVSearch. The benefit therefore depends on how verification interacts with evidence acquisition and fallback. Fixed five-view input improves V*Bench and POPE over both global-image controls, yet falls below Generator + verifier on HR-4K and HR-8K. More spatial coverage alone does not reproduce the gains of adaptive observation.

Turning from the matched comparison to individual search components, we obtain the second insight. **Obs. ❽** *Fine detail and retained context contribute together.* The largest losses from removing evidence-conditioned navigation or evidence-bundle accumulation occur on HR-8K. Relative to the full method, candidate-atlas traversal only reduces accuracy by 4.59%, 4.49%, and 8.18% on the three high-resolution settings. Removing accumulation reduces accuracy by 3.38%, 3.85%, and 6.92%, respectively. Without accumulation, the controller still uses 4.47 views, slightly more than Full `CAREER`. Its lower accuracy therefore accompanies nearly unchanged observation effort. Removing ZOOM/EXPAND also reduces all six accuracies. These results favor refining promising regions while keeping evidence from earlier observations available.

The candidate-construction ablations provide the third insight: **Obs. ❾** *Candidate coverage affects the observations needed for a decision.* Removing the hierarchical region tree raises views to 5.39 and latency to 16.72 seconds. It matches V*Bench accuracy but reduces HR-4K and HR-8K accuracy by 3.85% and 3.77%, respectively, relative to the full method. Random order matches HR-4K but reduces HR-8K accuracy by 5.03% relative to the full method. RECOVER improves five comparisons or ties them, whereas removing it raises HR-4K from 78.0% to 78.5%. Recovery is most useful here on V*Bench, HR-8K, and the harder POPE splits. The full acceptance gate exceeds both single-condition gates across all six evaluations.

Candidate-ranking ablations. Table S5 isolates the signals used to rank candidate regions while retaining the candidate atlas, adaptive navigation, evidence accumulation, and full acceptance gate. Visual only uses feature dispersion and edge density without query relevance. Query only uses query relevance alone and corresponds to V5 in Table 3. Query + edge and Query + dispersion add one visual cue at a time; the full method combines query relevance with both cues.

Table S5: Candidate-ranking ablations with `Qwen2.5-VL-7B`. Accuracy is reported in percent. Bold marks the highest accuracy in each column.

Ranking variant	High-resolution acc. ↑			POPE acc. ↑		
	V*Bench	HR-4K	HR-8K	Random	Popular	Adversarial
Visual only	87.4	73.5	72.5	90.50	88.70	85.70
Query only (V5)	89.0	75.5	74.5	91.20	89.50	86.50
Query + edge	90.6	77.5	77.5	**92.80**	90.40	87.40
Query + dispersion	91.1	77.0	76.5	91.90	90.10	87.10
Full `CAREER`	**91.6**	**78.0**	**79.5**	92.77	**91.23**	**88.33**

The full ranking outperforms Visual only and Query only in all six settings. On HR-Bench 8K, accuracy rises from 72.5% and 74.5%, respectively, to 79.5%. Adding either visual cue to Query only improves every evaluation; combining both yields the highest accuracy on all three high-resolution benchmarks and the Popular and Adversarial POPE splits. Query + edge is slightly higher on POPE

Random (92.80% versus 92.77%). These results support complementary query and visual signals without implying that the full score dominates every setting.

All controls use the fixed configuration in Table S1, including the verifier comparison below.

D.3 Accuracy and Computation across Observation Budgets

A budget of K permits at most K valid local observations after the initial global image, so the total-input limit is $K + 1$. Failed localization, duplicate crops, and restoration of a historical state do not add an observation. Their generator, verifier, and SAM calls still count toward computation. The matched CVSearch control and **CAREER** share the generator, verifier, thresholds, fallback rule, input resolution, decoding settings, hardware, and timing boundary. Table S6 reports their actual average consumption after early stopping, alongside the original CVSearch.

Table S6: Accuracy and computation at each local-observation budget. Accuracy is in percent, visual tokens are in thousands, and latency is in seconds per question. Bold marks the highest accuracy at each nonzero budget.

Method	**Budget**	**Accuracy ↑**			**Mean computation ↓**		
	K	V*	HR-8K	Adv.	Obs.	Tokens	Lat. (s)
Backbone	0	71.2	65.0	81.83	1.00	0.90	1.56
CVSearch	1	82.7	68.0	83.47	1.87	1.95	6.04
CVSearch + verifier	1	83.8	69.0	84.17	1.82	2.36	6.88
CAREER	1	**85.3**	**71.5**	**85.67**	1.79	2.18	5.51
CVSearch	2	85.9	71.5	84.77	2.69	3.76	9.71
CVSearch + verifier	2	86.9	73.0	85.33	2.63	4.45	10.74
CAREER	2	**88.0**	**75.0**	**86.93**	2.58	3.86	8.36
CVSearch	4	88.0	74.5	85.93	4.12	7.18	14.96
CVSearch + verifier	4	89.0	78.0	86.27	3.96	7.75	16.20
CAREER	4	**91.6**	**80.0**	**87.67**	3.69	6.38	11.87
CVSearch	8	89.0	76.0	86.60	6.42	13.60	19.67
CVSearch + verifier	8	88.5	75.0	86.70	4.86	10.45	21.84
CAREER	8	**91.6**	**79.5**	**88.33**	4.41	8.90	14.81

As shown in Table S6, the budget sweep yields the following insight: **Obs. ⑩** *Most high-resolution gains arrive within four local observations.* Across all four nonzero budgets, **CAREER** has higher accuracy on every benchmark and lower latency than both CVSearch variants. Increasing the budget from one to four raises V*Bench accuracy from 85.3% to 91.6% and HR-8K from 71.5% to 80.0%. Increasing it further to eight leaves V*Bench unchanged and lowers HR-8K accuracy by 0.63% relative to the four-observation budget, while POPE Adversarial rises from 87.67% to 88.33%. This last increase costs 2.94 additional seconds per question. The larger budget helps the difficult existence judgments more than the two high-resolution tasks in this sweep.

Early stopping keeps actual consumption below the allowed observation count. At budget eight, **CAREER** uses 4.41 of at most nine image inputs on average. Compared with matched CVSearch, latency falls by 19.9%, 22.2%, 26.7%, and 32.2%, respectively. At the largest budget, **CAREER** also reduces visual tokens from 10.45k to 8.90k while improving all three accuracies. The controller uses the available budget selectively, and the computational difference grows as the allowed search becomes longer.

D.4 Acceptance, Error Sources, and Verifier Choice

Table S7 partitions `Qwen2.5-VL-7B` outcomes into non-fallback acceptance and fallback. Coverage uses all test questions as its denominator. Accepted error uses only accepted questions, whereas

wrong acceptance over all samples measures the fraction of the entire test set that the gate accepts incorrectly. These quantities correspond to C_{acc}, r_{acc}, and $C_{\text{acc}} r_{\text{acc}}$ in Appendix B.3. Overall risk is their acceptance/fallback-weighted decomposition. These empirical statistics describe decision behavior, while the parameters of Theorem 3.1 characterize generator alignment at global screening, trajectory-wise verifier alignment, exceptional probability masses, feasible-observation reachability, and the scale of summed Support scores.

Table S7: Acceptance and fallback outcomes with `Qwen2.5-VL-7B` and the default verifier. Counts are total/correct, with 800 evaluation instances per HR-Bench resolution. All rates and accuracies are percentages. The two panels describe the same predictions.

(a) Counts and conditional accuracy

Benchmark	Accepted total/correct	Accepted acc. ↑	Fallback total/correct	Fallback acc. ↑	Overall acc. ↑
V*Bench	125/121	96.80	66/54	81.82	91.62
HR-Bench 4K	440/400	90.91	360/224	62.22	78.00
HR-Bench 8K	416/384	92.31	384/252	65.63	79.50
POPE Random	2160/2117	98.01	840/666	79.29	92.77
POPE Popular	2040/1979	97.01	960/758	78.96	91.23
POPE Adversarial	1740/1657	95.23	1260/993	78.81	88.33

(b) Coverage and incorrect acceptance

Benchmark	Accepted coverage	Error among accepted ↓	Wrong acceptance over all samples ↓
V*Bench	65.45	3.20	2.09
HR-Bench 4K	55.00	9.09	5.00
HR-Bench 8K	52.00	7.69	4.00
POPE Random	72.00	1.99	1.43
POPE Popular	68.00	2.99	2.03
POPE Adversarial	58.00	4.77	2.77

As shown in Table S7, comparing accepted and fallback predictions yields the following insight: **Obs. ⓫** *The gate accepts a majority of questions at higher conditional accuracy.* Coverage ranges from 52.00% to 72.00%, and accepted accuracy ranges from 90.91% to 98.01%. Accepted accuracy is 18.31% to 46.10% higher than fallback accuracy across benchmarks. HR-4K has the highest error among accepted predictions at 9.09%, corresponding to 5.00% of all questions. On POPE, coverage falls from 72.00% on Random to 58.00% on Adversarial as accepted error rises from 1.99% to 4.77%. Harder negative questions both reduce non-fallback acceptance and increase the errors among the questions still accepted.

The POPE confusion matrices locate the remaining errors more precisely. Each split contains 3,000 questions with equal numbers of positive and negative examples. Figure S1 reports specificity, false-positive rate, and the proportion of `yes` predictions alongside accuracy and recall.

To examine the source of the remaining POPE errors, we compare the split-level diagnostics in Figure S1 and obtain the following insight: **Obs. ⓬** *The accuracy loss on harder POPE splits comes from false positives.* Recall remains at 97.47% across the three splits. Specificity falls from 88.07% to 79.20%, while the false-positive rate rises from 11.93% to 20.80%. Yes-rate increases from 54.70% to 59.13% despite the balanced labels. The split-dependent loss is therefore concentrated in rejecting absent objects. Frequent categories and co-occurring objects remain harder to rule out even when positive-object recall stays high.

We rerun the complete search-and-verification pipeline separately for each verifier, holding the generator, prompts, acceptance thresholds, decoding configuration, and maximum observation budget fixed. Each run determines its own acquired observations, stopping time, and fallback decisions. Table S8 compares six Qwen checkpoints on all 191 V*Bench questions and the first 200 HR-8K

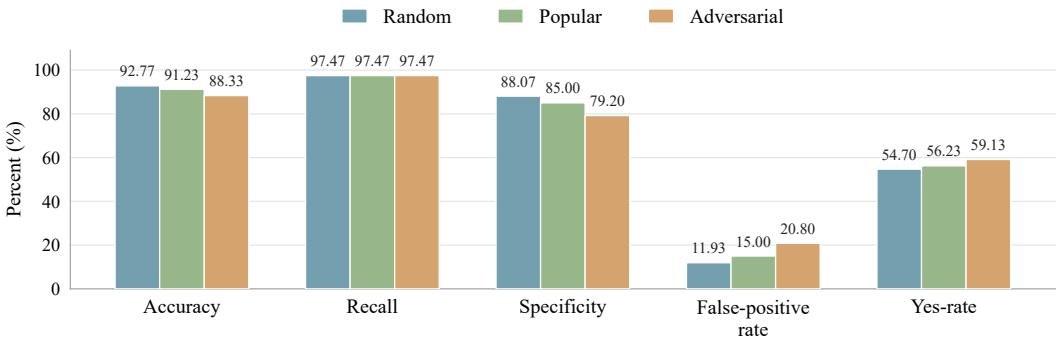


Figure S1: POPE diagnostics with `Qwen2.5-VL-7B`. Bars show percentages, with exact values above each bar. Specificity is the true-negative rate, and yes-rate is the fraction of predictions that answer `yes`. Each split contains 3,000 balanced positive and negative questions.

evaluation instances in the released dataset order. All accuracies and acceptance statistics in this table use this 391-instance subset. All other HR-Bench experiments use the full 800 evaluation instances per resolution.

Table S8: End-to-end verifier comparison within the Qwen family. All checkpoints are Instruct variants. Acceptance counts are total/wrong over 391 instances: all 191 V*Bench questions and the first 200 HR-8K evaluation instances. Rates and accuracies are percentages. The default verifier is shaded, and bold marks the best metric, including ties.

Verifier	V* ↑	HR-8K ↑	Accepted total/wrong	Accepted coverage	Accepted error ↓	Wrong/ all ↓
`Qwen2.5-VL-3B`	89.5	77.5	205/18	52.43	8.78	4.60
`Qwen3-VL-2B`	91.1	78.0	226/20	57.80	8.85	5.12
`Qwen2.5-VL-7B`	91.1	78.5	221/17	56.52	7.69	4.35
`Qwen3-VL-4B` (default)	91.6	79.5	229/12	58.57	5.24	3.07
`Qwen3-VL-8B`	92.1	**80.0**	238/11	60.87	4.62	**2.81**
`Qwen3-VL-32B`	**92.7**	79.0	251/11	**64.19**	**4.38**	**2.81**

As shown in Table S8, varying the verifier yields the following insight: **Obs. ⓭** *Larger Qwen3 verifiers increase coverage, while accuracy gains depend on the benchmark.* Across Qwen3 sizes from 2B to 32B, coverage rises from 57.80% to 64.19% and V*Bench accuracy from 91.1% to 92.7%. HR-8K instead peaks at 80.0% with 8B and falls to 79.0% with 32B. Both larger models accept 11 incorrect predictions, so the lower conditional error at 32B accompanies more accepted correct predictions rather than fewer false acceptances.

Parameter count also does not determine behavior across model generations. `Qwen3-VL-2B` accepts more questions than `Qwen2.5-VL-7B`, but its accepted error is higher at 8.85% versus 7.69%. The default `Qwen3-VL-4B` reaches 91.6% on V*Bench and 79.5% on HR-8K, with 58.57% coverage and 5.24% accepted error. Relative to 4B, 8B improves V*Bench and HR-8K accuracy by 0.55% and 0.63%, respectively, and increases coverage by 3.93%, while reducing accepted errors from 12 to 11.

E Detailed Prompts

E.1 Question-Only Planning Prompt

This call is made only after the global screen fails and a local observation is permitted. Its entire task-dependent input is the question text. No answer options or image are provided.

Question-Only Planning

Given the question, list short localization phrases and the visual requirements needed to answer it. Use only entities, parts, or attributes explicitly stated in the question. Do not answer the question or infer an unknown queried attribute, relation, or object presence. For a relation, include the target, reference entity, and required correspondence or spatial context. For existence, require a visible instance for presence and adequate scene coverage and discernibility for absence. Assign each requirement a unique identifier. Return one JSON object with two arrays: `localization_phrases` (strings) and `requirements` (objects with string fields `id` and `description`).

For "What color is the luggage?", an output has phrase `luggage` and separate requirements for identifying the queried luggage and observing its color. It must not contain a guessed color. Empty phrase lists are allowed when the question supplies no usable entity. Parse failure is an output-error fallback according to the acceptance-control protocol.

E.2 MLLM Generator Prompts

Answer Mode. The generator receives the question, all options, and $\mathcal{B}_t$. For local rounds it also receives the question-derived requirements. The answer token is constrained to a supplied, tokenizer-checked single-token option identifier. The controller maintains its mapping to the option text.

MLLM Generator: Answer Mode

Answer the question using the supplied visual evidence. Return exactly one supplied option identifier. No reference answer is available. Do not treat option wording, likely object properties, or absent observations as visual evidence. If the evidence is incomplete, still choose your best provisional option. The controller determines its status.

The global answer-token logits define Eq. 1, and their highest-probability option is saved for global acceptance and fallback. In each local round, the generator updates its answer and proposes an action before the current verifier call. The proposal is executed only after the controller returns `CONTINUE`.

MLLM Generator: Navigation and Replanning

Navigation/replanning mode. You receive the question and options, numbered evidence requirements, the current bundle with view identifiers and original-image coordinates, current focus, remaining observation budget, failed attempts, and supplied legal action–candidate pairs. Any verifier feedback supplied here comes from previously completed rounds and is marked with its source round and view identifiers. It may contain each option's Support score, decision code, grounded facts, and missing requirements. No verifier result for the current round is available when this proposal is made. No reference answer or correctness label is available.

Choose one decision-critical unresolved requirement using the current evidence and any available earlier feedback. Reassess earlier missing requirements against newly acquired views. With no earlier feedback, use the question-derived requirements and current evidence. State what a new observation must establish, not the answer you hope to confirm. If option explanations disagree or all claim support while the normalized-support margin remains small, identify the requirement needing disambiguation. Do not treat either explanation as observed truth. Reference the relevant earlier option feedback, when available, and requirement identifier. If an explanation is unavailable, use the available scores and requirements and do not invent a verifier finding.

Choose exactly one supplied legal pair when an observation is permitted. ZOOM/EXPAND preserve the current candidate identifier and require a short, nonempty `sam_prompt`. For SPLIT/NEXT, the controller supplies only the highest-ranked feasible unvisited destination for each action under $R(c)$. Use that destination identifier and a null phrase. A localization phrase may use an entity, part, or attribute stated in the question or grounded in a supplied view. Do not insert an unknown queried attribute or relation. Never output RECOVER, new candidate identifiers, or your own box. Use null for all action fields if no legal pair or no remaining budget is supplied. Replanning after failed execution uses the feedback attached to the pending proposal and updated legal pairs. A failed attempt does not add a local observation. The controller recovers only on stagnation or when the current state has no legal observation.

Return one JSON object and no other text:

```
{
"requirement_id": "r2",
"feedback_option_ids": ["B"],
"evidence_gap": "<fact to establish>",
"action": "ZOOM",
"candidate_id": "c7",
"sam_prompt": "<localization phrase>"
}
```

This is a ZOOM structure example. Identifiers must come from the actual inputs. Use JSON null, not the string "null", for inactive fields. The requirement and evidence-gap fields may also be null in a terminal state. The option-id array may be empty when no semantic feedback is available.

Figure S2: Generator navigation prompt, invoked before current-round verification using current evidence and any earlier verifier feedback.

E.3 Independent Verifier Prompt

One completion is requested per option. The first decision token supplies the score. The following JSON supplies semantic feedback. An explanation is a continuation of that same inference call, not an unreported second verification call.

Independent Verifier: Option-wise Evaluation

Role and isolated inputs. Verify one candidate option using the supplied question, numbered requirements, and visual bundle with view identifiers and original-image coordinates. You receive exactly one option. You do not receive the generator's provisional answer, competing options, reference answers, ground-truth labels, or correctness annotations.

Your first output token must be exactly one of:

- `A` (SUPPORT): the required identity, when applicable, and queried attribute, relation, or existence state are visually established and support this option.
- `B` (REFUTE): established visible evidence directly contradicts this option.
- `C` (INSUFFICIENT): required identity, detail, correspondence, or context is absent, unreadable, occluded, ambiguous, or unresolved.

Judge the option independently. Do not eliminate alternatives. Missing evidence is not refutation. Ground each factual statement in a supplied view and requirement. Infer spatial relationships from scene content and original-image coordinates, never from the displayed arrangement or size of crops. Option wording and likely properties are not observations.

For an existence question, a visible matching instance establishes presence. Failure to localize or see an object in a limited crop does not establish absence. Support "no" or refute "yes" only if the visible scene provides adequate coverage and discernibility for the queried object. Identify that evidence. Otherwise return INSUFFICIENT and state the missing coverage or detail. Do not assume that an unobserved region has been searched.

After the decision token, output a newline and one JSON object. Include at most two short grounded facts and two missing requirements. A grounded record has `requirement_id`, `view_ids` (a nonempty array), and `fact`. A missing record has `requirement_id` and `needed_evidence`. Use only supplied identifiers. Missing evidence is a request, not a claim about unseen content. Empty arrays are allowed. Do not invent a gap for decisive evidence.

```
C
{"grounded": [{"requirement_id": "r1",
"view_ids": ["v1"], "fact": "<visible fact>"}],
"missing": [{"requirement_id": "r2",
"needed_evidence": "<unresolved fact or context>"}]}
```

The example shows the structure, not a prescribed label, fact, or identifier. Use actual JSON double quotes. Output nothing before the decision token or after the JSON object.

Figure S3: Option-wise verifier prompt with first-token scoring and a grounded semantic continuation.

F Case Studies

Figures S4–S8 present recorded evidence and decisions on HR-Bench 4K examples. They illustrate evidence sufficiency and complement the aggregate results in Appendix D. Figures S5 and S6 show evidence bundles and their verifier scores rather than a complete replay of candidate selection. Boxes, where shown, use original-image coordinates $[x, y, w, h]$. The local verifier shares are $\rho_{t,i} = s_{t,i} / \sum_j s_{t,j}$, and the reported gap is the difference between the largest and second-largest share. These runs use the default gates $\tau_{\text{conf}} = 0.65$ and $\tau_{\text{gap}} = 0.15$ from Table S1. Both candidate sources are merged and ranked before any local observation. INIT denotes the first candidate visit from that joint atlas and is not a NEXT action.

F.1 SPATIAL RELATION VIA ZOOM AND EXPAND

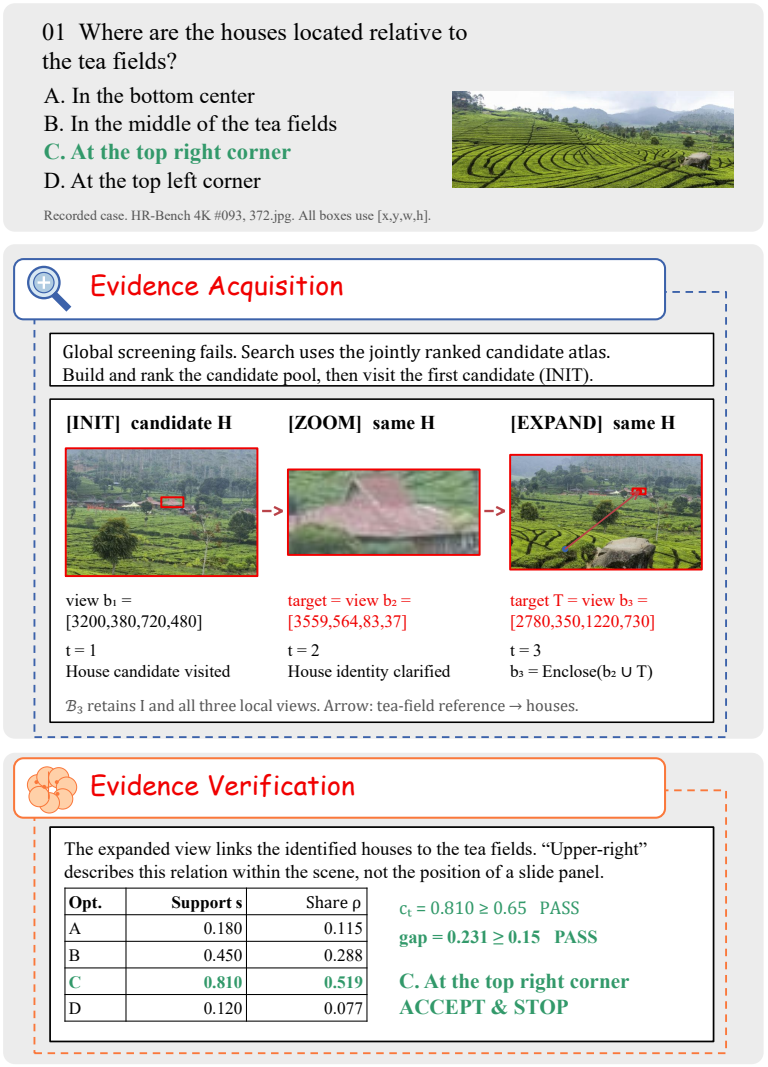


Figure S4: Complementary detail and context for spatial reasoning. The tight view identifies the houses, while the expanded view establishes their upper-right relation to the tea fields. The retained local observations in the evidence bundle support option C.

Evidence Requirement. The answer requires both object identity (the small structures are houses) and relational context (where those houses lie with respect to the tea fields). A tight target box alone cannot establish the second fact.

Search and Decision. The tight view clarifies house identity but omits the tea-field reference, while the expanded view restores the context needed to establish the spatial relation. Both views belong to candidate H, consistent with the within-candidate adjustments in the scale-control procedure. Their joint evidence gives option C absolute support 0.810 and a leading–runner-up share margin of 0.231, so both gates pass. Retaining complementary observations matters because detail resolves what the target is, and context resolves where it is relative to the reference. The "upper-right" relation is measured within the scene.

F.2 Cross-Candidate Evidence Accumulation

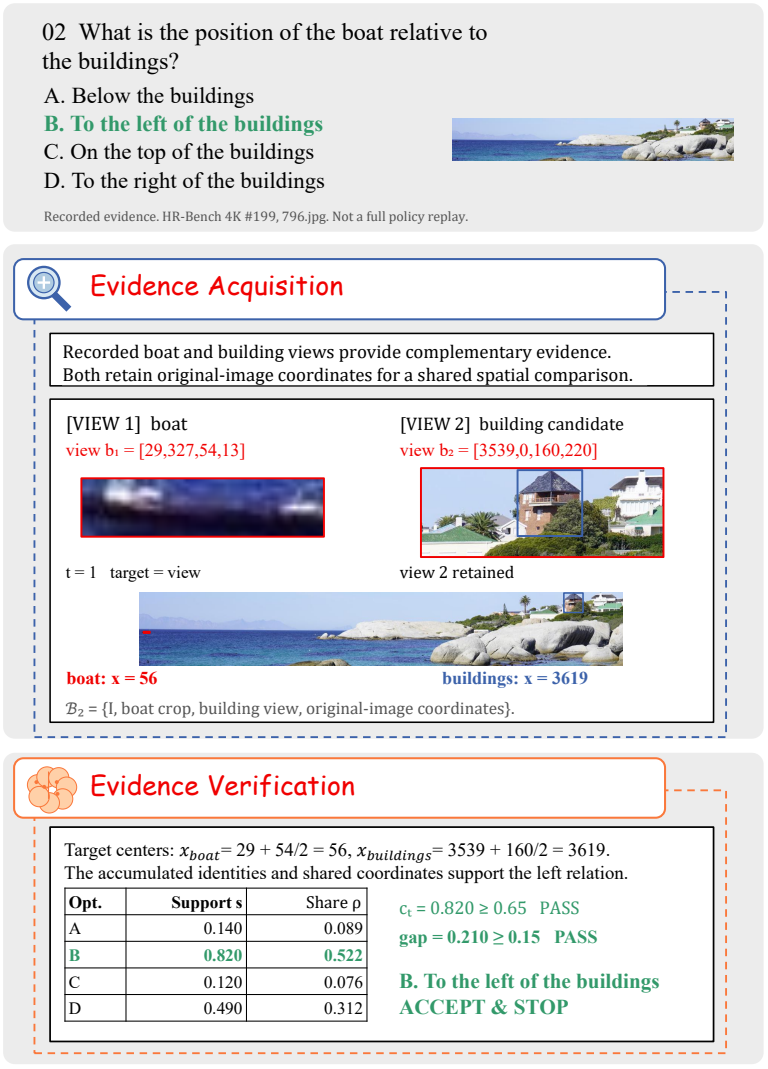


Figure S5: Complementary evidence from two candidate regions. The boat and building views retain their original-image coordinates, making their left–right relation directly comparable.

Evidence Requirement. A crop of either object is insufficient because the question asks for the boat's position relative to the buildings. The system must preserve both identities and compare them in a shared coordinate frame.

Evidence and Decision. The boat view places its center at $x = 56$, while the building view places the reference center at $x = 3619$. The bundle $\mathcal{B}_2$ contains the global image, both crops, and their coordinates. With this bundle, option B has absolute support 0.820 and a leading–runner-up share margin of 0.210, satisfying both acceptance conditions. These recorded views show how retaining observations in a shared coordinate frame supports relational verification.

F.3 SINGLE-VIEW EVIDENCE SUFFICIENCY



Figure S6: One localized view resolves the umbrella color. The displayed global image and local crop form an evidence bundle that satisfies both verifier gates.

Evidence Requirement. The global proposal is option A, but the full-resolution scene does not make the small canopy color sufficiently visible. A single target-specific crop is enough to reveal the requested attribute.

Evidence and Decision. The displayed local umbrella observation is retained with its original-image box in $\mathcal{B}_1$. Its recorded verifier scores give option A absolute support 0.999992 and a leading–runner-up share margin of 0.909597, passing both gates. This demonstrates that one local view can supply sufficient evidence. When the first acquired view supplies the missing attribute, verifier acceptance can end observation at this point, with the joint candidate atlas already constructed.

F.4 HIERARCHICAL SPLIT FOR FINE TEXT

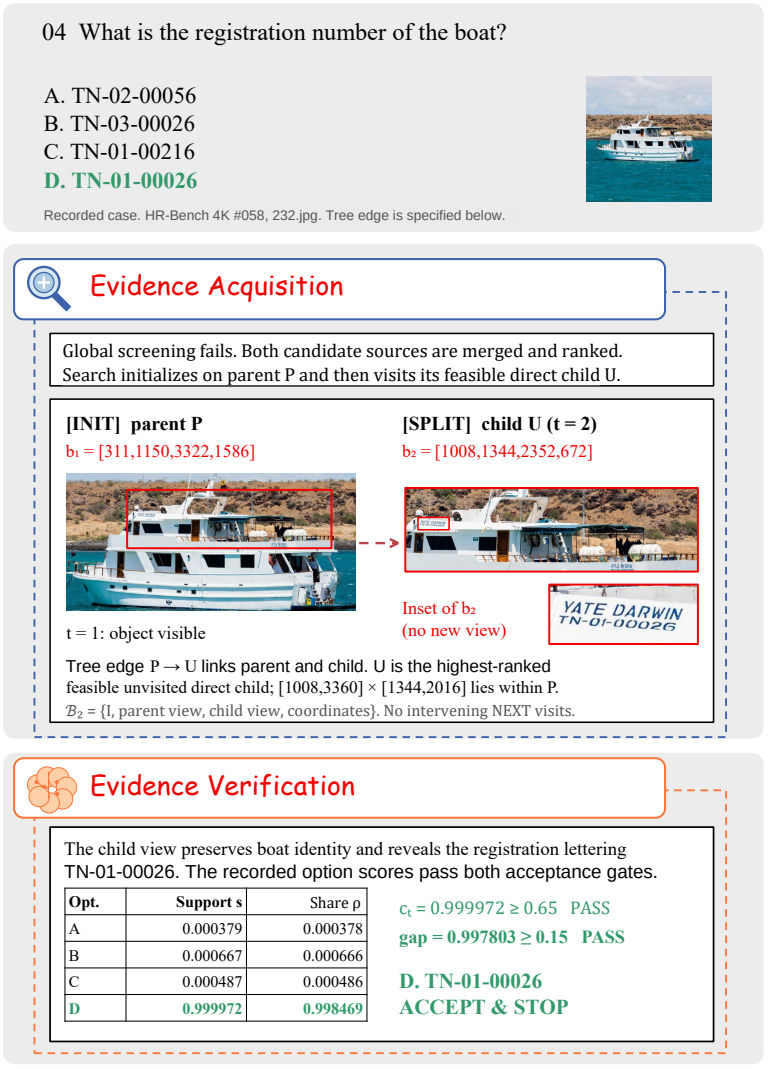


Figure S7: A legal parent-to-child SPLIT for fine-grained text recognition. Formal search initializes on tree parent P and then visits its highest-ranked feasible unvisited direct child U, where the registration number becomes readable.

Evidence Requirement. The full boat view establishes object identity, but the registration lettering requires a finer observation. In the recorded hierarchical region tree, $U = [1008, 1344, 2352, 672]$ is a direct child contained within $P = [311, 1150, 3322, 1586]$.

Search and Decision. After joint-atlas construction, INIT visits P at $t = 1$ and SPLIT visits U at $t = 2$. The inset is only a magnified rendering of U, not an additional observation. The child view preserves boat identity while exposing "TN-01-00026". Option D obtains absolute support 0.999972 and a leading–runner-up share margin of 0.997803, so it is accepted. The sequence satisfies the direct-child constraint in the scale-control procedure without inserting unrelated NEXT visits.

F.5 FLAGGED UNVERIFIED FALLBACK

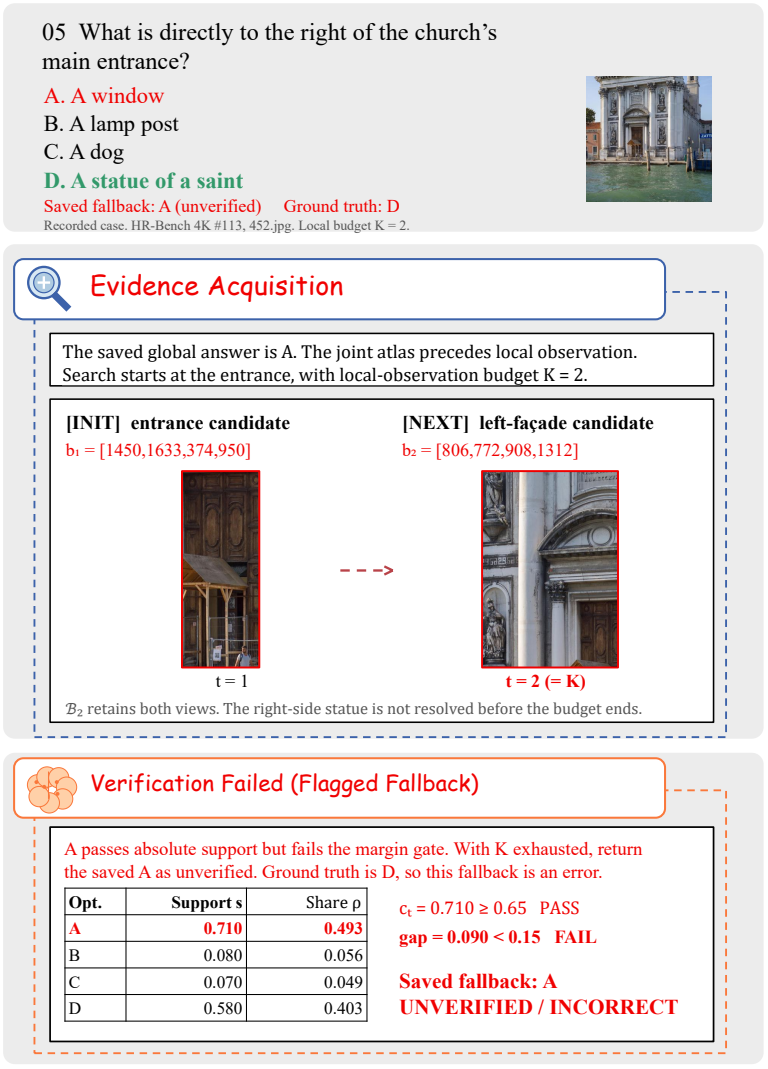


Figure S8: Failure under an exhausted local-observation budget. The saved global answer A passes the absolute-support gate but fails the normalized-support-margin gate after two local observations. It is therefore returned with an *unverified* status. The ground-truth answer is D.

Evidence Requirement. The question requires identifying the object immediately to the right of the church entrance. The global generator saves A, while the two available local observations cover the entrance and a left-façade candidate but never resolve the right-side statue.

Search and Outcome. With a local-observation budget of $K = 2$, INIT consumes the first local observation and NEXT consumes the second. The leading absolute support is 0.710, but the leading–runner-up share margin is only 0.090, so the answer is not accepted. Once the budget is exhausted, the controller returns the saved global answer A and explicitly marks it as unverified, as required by the acceptance-control procedure. Because the ground truth is D, this case is an exposed failure rather than a successful or "safe" prediction. It illustrates why coverage and fallback risk are reported separately in Appendix B.3.