PAY MORE ATTENTION TO TEXT IN HIGH-RESOLUTION MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Failures of high-resolution MLLMs are commonly attributed to a visual problem, motivating zooming, cropping, and related visual interventions to recover fine-grained evidence or suppress interference. Yet recent studies suggest that relevant visual evidence is already encoded in intermediate representations, indicating that visual-side improvements alone are insufficient. This raises a natural question: does the remaining bottleneck lie in the text that guides visual search? We identify a previously overlooked linguistic bottleneck: questions formulated for answering do not necessarily specify the visual evidence required for localization. To address this mismatch, we introduce EviSpec, a training-free compiler that derives complementary evidence specifications while preserving the original question for final reasoning. We further validate it through matched-control experiments that isolate the roles of evidence specification and localization. With the search budget fixed, structured evidence specifications yield an 11.9% relative gain over generic requests. With evidence geometry matched, the evidence localized by EviSpec yields a 14.8% relative gain over random evidence. Together, these controls isolate the benefit of specifying what evidence to seek rather than merely expanding visual access. Across all five MLLMs, EviSpec consistently improves upon the corresponding baseline on each of the three benchmarks, yielding average relative gains of **10.4%, 8.8%, and 12.4%** on V*Bench, HR-Bench-4K, and HR-Bench-8K, respectively. Beyond high-resolution reasoning, EviSpec also achieves state-of-the-art performance on VQA and hallucination-focused benchmarks.

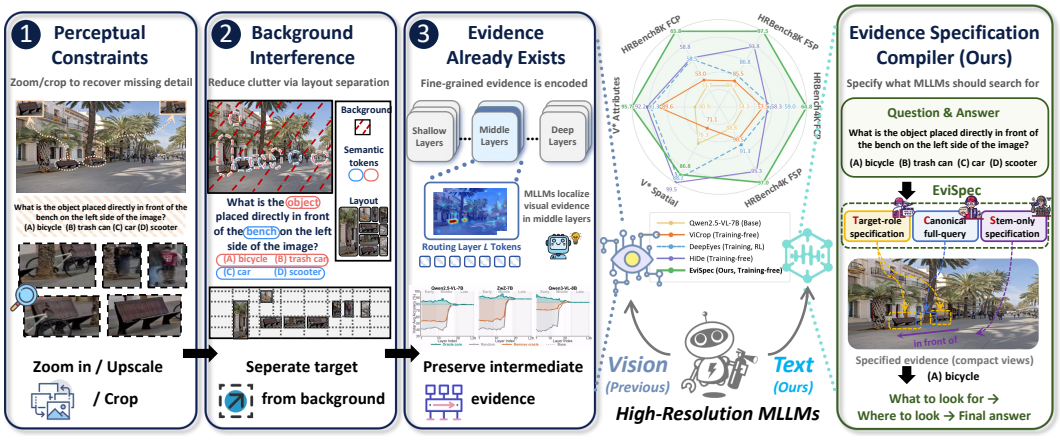


Figure 1: **Motivation and overview of EviSpec.** Prior work addresses high-resolution failures through **visual** side, from recovering missing details to suppressing interference and preserving already-encoded evidence. We instead focus on the **text** guiding visual search. EviSpec consistently outperforms visual baselines and compiles each question into complementary target-role (T), canonical full-query (C), and stem-only (S) evidence specifications. Their ordered combination, TCS, specifies what and where to search, while retaining the original question for final reasoning.

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

1 INTRODUCTION

Multimodal large language models (MLLMs) increasingly support native or dynamic-resolution inputs, exposing substantially more visual detail to the language model (Wang et al., 2024b; Jiang et al., 2024b; Bai et al., 2025a;b; Liu et al., 2025; Zhu et al., 2025; Zhao et al., 2026b;a). Yet high-resolution reasoning remains challenging when the answer depends on sparse evidence hidden within a large and cluttered scene, such as a small sign, a distant object, or a relation between spatially separated entities. Benchmarks such as V*Bench and HR-Bench make this challenge explicit by requiring models to recover fine-grained evidence from high-resolution images (Wu & Xie, 2024; Wang et al., 2025c). As shown in Fig. 1.❶–❸, existing work has predominantly addressed the failures on **the visual side**: zooming into promising regions, selecting crops, invoking visual tools, suppressing irrelevant content, or exploiting internal attention to recover evidence (Khayatkhoei et al., 2025; Liu et al., 2026; Xie et al., 2026; Zheng et al., 2026; Zhu et al., 2026). Despite their methodological differences, these approaches share a premise—high-resolution failure is primarily a problem of **where and how the model looks**. However, this visual-side view may be incomplete.

Recent analyses suggest that high-resolution MLLMs can already encode answer-relevant visual content in intermediate representations (Wang et al., 2025a; Mao et al., 2026) in Fig 1.❸. This raises a complementary question that has received far less attention: *what tells the model what visual evidence to look for?* Since visual search is guided by language, we ask whether part of the remaining high-resolution bottleneck lies not in the image itself, but **in the text used to guide visual search**. Our key observation is that text formulated for answering is not necessarily specified for localization. While answering preserves the user's decision problem, localization requires explicit cues about the target, its attributes, and relevant relations (Yu et al., 2018; Ma et al., 2023). These requirements can diverge substantially in high-resolution VQA: a question may support correct reasoning while leaving the required visual evidence implicit or ambiguously specified. We call this mismatch the **evidence-specification gap:** the question states what to decide, but not necessarily what to find.

To address this gap, we introduce **EviSpec**, a training-free question-to-evidence compiler that provides a dedicated language-side interface for visual search. As illustrated in Fig. 1, EviSpec compiles the original question into **three complementary evidence specifications**. The **T**arget-role specification (T) identifies the answer-bearing target and relevant reference entities; the **C**anonical full-query specification (C) rewrites the complete question with stable entity expressions; and the option-isolated **S**tem specification (S) applies the same canonicalization after removing candidate answers. Their ordered combination, TCS, specifies what evidence to seek and where to search for it. MLLMs then localize these specifications through prompt-calibrated cross-modal attention and adaptive region formation, producing compact evidence views alongside the original image. Rather than changing how the model sees, EviSpec changes the text that guides what it looks for.

(1) Importantly, to isolate this text-side effect from increased visual access, we conduct **two controlled experiments**. First, with the visual-search budget fixed, structured evidence specifications yield an **11.9%** improvement over generic requests, showing that the gain does not come from additional search. Second, with evidence geometry matched, regions localized by EviSpec yield a **14.8%** improvement over random evidence, ruling out gains from simply providing an extra crop or favorable region geometry. Together, these controls isolate the benefit of specifying what evidence to seek and retrieving question-relevant content. **(2)** This text-side intervention produces consistent gains across models and tasks. Across all five evaluated MLLMs, EviSpec improves upon the corresponding baseline on each of the three high-resolution benchmarks, with average relative gains of **10.4%, 8.8%, and 12.4%** on V*Bench, HR-Bench-4K, and HR-Bench-8K, respectively. The benefit further extends beyond high-resolution reasoning, where EviSpec achieves state-of-the-art (SOTA) performance on additional VQA and hallucination-focused benchmarks. In the end, these results indicate that high-resolution MLLM failures **cannot be reduced to insufficient visual access**: how the evidence is specified in text is an important part of the multimodal reasoning interface.

Our principal contributions are summarized as follows:

- ***Problem Reframing.*** We identify an overlooked linguistic bottleneck in high-resolution MLLMs: while prior work primarily improves visual access, **the text guiding visual search** fails to specify the evidence for localization. We formalize this mismatch as the evidence-specification gap.

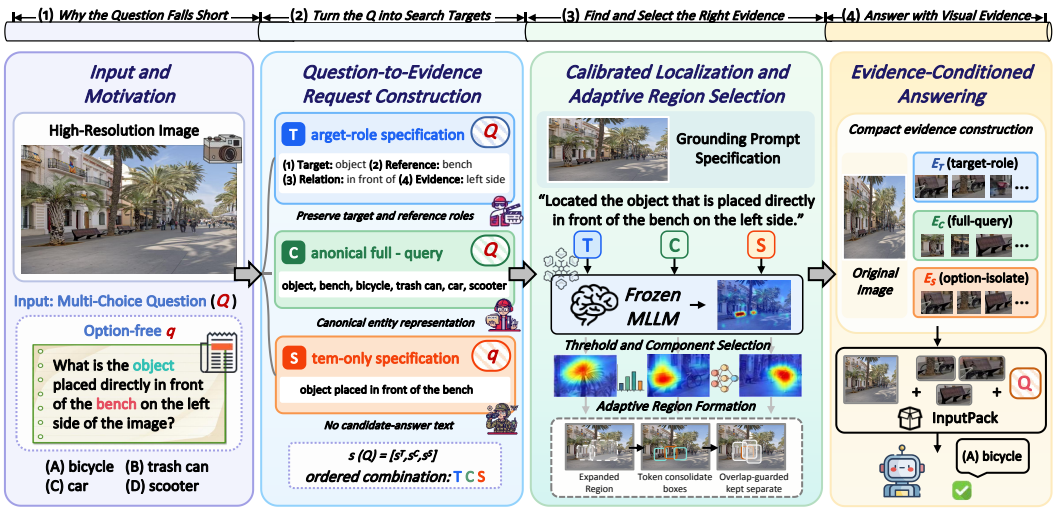


Figure 2: Overview of EviSpec: T preserves target roles, C captures the full query, and S isolates answer options. Prompt-calibrated attention extracts compact visual evidence, while adaptive region formation preserves relational context.

- ***Paradigm Reformulation.*** We introduce EviSpec, a training-free question-to-evidence method that derives complementary search specifications while preserving the question for final reasoning. It uses no external detector or model modification, with a single inference hyperparameter.
- ***Controlled and Comprehensive Validation.*** Controlled analyses reveal how evidence specification and question-relevant localization drive the gains beyond additional visual access. Across five frozen MLLMs and eight benchmarks, EviSpec consistently achieves SOTA performance on high-resolution VQA and hallucination-focused evaluations.

2 Related Work

Multimodal Large Language Models MLLMs integrate visual encoders with pretrained language models to support image understanding and visual reasoning (Li et al., 2023a; Dai et al., 2023; Driess et al., 2023; Liu et al., 2024a; Zhang et al., 2025a). Recent models, including Qwen2.5-VL, Qwen3-VL, and InternVL3, process images at native or dynamic resolutions and preserve fine-grained visual information (Bai et al., 2023; Chen et al., 2024a;b; Wang et al., 2024a; Bai et al., 2025a;b; Wang et al., 2025b; Zhu et al., 2025). However, a larger visual input does not ensure that question-relevant evidence is retained, localized, or used during inference. Most advances improve visual encoding, token allocation, or multi-modal alignment (Wen et al., 2026a;b).

High-Resolution Visual Question Answering High-resolution visual question answering (HR-VQA) requires models to recover small objects, fine attributes, and relations between spatially separated entities. V* introduced guided visual search and V*Bench, while HR-Bench extended evaluation to 4K and 8K images. Existing methods improve evidence access through regional decomposition, attention-guided cropping, relevance maps, grounding-aware head selection, adaptive localization, evidence routing, or crop policies (You et al., 2024; Li et al., 2024; Dong et al., 2024; Huang et al., 2025; Zhong et al., 2025; Li et al., 2026; Lin et al., 2026; Liu et al., 2026; Wei et al., 2026; Xie et al., 2026; Zheng et al., 2026; Wang et al., 2026b; Zhu et al., 2026; Yuan et al., 2026). Question-conditioned grounding and visual question decomposition connect linguistic structure with image regions (Shi et al., 2018; Kamath et al., 2021; Le et al., 2022; Li et al., 2022; Jiang et al., 2024a; Peng et al., 2024; Rasheed et al., 2024; Liu et al., 2024b; Zhang et al., 2024; Ma et al., 2025; Wu et al., 2025; Dong et al., 2026a;b; Huang et al., 2026; Wu et al., 2026). EviSpec separates evidence specification from localization and answering. It compiles the question into target-role, canonical, and option-isolated localization requests while preserving it for answering.

3 METHODOLOGY

Fig. 2 provides an overview of the EviSpec pipeline, including request construction, localization, region formation, and evidence-conditioned answering. Details are provided in the Appendix F.

3.1 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

As shown in Fig. 2 (1) , $I : \Omega_I \to \mathbb{R}^3$ denote an image on source-pixel domain Ω_I, with visual token grid $\Omega_V = \{1, \dots, H_V\} \times \{1, \dots, W_V\}$. A query $Q = (q, \mathcal{O})$ contains a question and an optional set of candidate answers, serialized as $\bar{Q} = \mathrm{Serialize}(q, \mathcal{O})$; when $\mathcal{O} = \varnothing$, $\bar{Q} = q$. The compiler, localization, and answer calls, denoted F_θ^{comp}, F_θ^{loc}, and F_θ^{ans}, share frozen parameters but use distinct fixed prompts. The evidence constructor G_θ invokes the compiler and localizer to produce ordered views $\mathbf{E} = G_\theta(I, Q)$. Final inference retains the original image and query:

$$\hat{a} = F_\theta^{\mathrm{ans}}(\mathrm{InputPack}([I] \| G_\theta(I, Q)), Q). \tag{1}$$

Evidence construction uses neither gold answers, annotation boxes, nor benchmark categories. Only the final call performs task-specific answer reasoning; no parameters are trained.

3.2 QUESTION-TO-EVIDENCE REQUEST CONSTRUCTION

As shown in Fig. 2 (2) , the option-free stem q_0 is the prefix of $\bar{Q}$ before its first option marker, or q when no options are present. Fixed compiler prompts produce the ordered specifications

$$\widetilde{\mathbf{s}}(\bar{Q}) = [s^T, s^C, s^S] = \big[\mathrm{Role}(\bar{Q}), \mathrm{Can}(\bar{Q}), \mathrm{Can}(q_0)\big]. \tag{2}$$

For direct questions, Role returns one answer-bearing target phrase with its identifying context. For relation questions, it returns comma-delimited subject and reference fields in question order and records relation mode. Canonicalization rewrites entities into head-first phrases: C uses the complete query, whereas S applies the same contract to the option-free stem. $\mathbf{s}(\bar{Q})$ removes empty parsed requests and exact duplicates, preserving $T/C/S$ order. Let $\mathcal{R}(Q) \subseteq T, C, S$ denote the surviving branch labels. Each retained request has the representation $s^r = (u_{r,1}, \dots, u_{r,K_r}; m_r)$ for $r \in \mathcal{R}(Q)$. Here, $u_{r,j}$ is an entity field and m_r an optional direct or relation-oriented mode. Token group $\mathcal{T} * r, j$ contains signal tokens aligned with $u * r, j$. All branches share the localization procedure below and terminate before final answering. Details are provided in Appendix B.1.

3.3 CALIBRATED LOCALIZATION AND ADAPTIVE REGION SELECTION

As shown in Fig. 2 (3) , each retained request is inserted into a shared grounding prompt and localized on the original image by F_θ^{loc}. $A_{r,t}^{(h)}(p)$ denote the attention of head h from token t to visual location p. With $\mathcal{H}_F$ indexing all attention heads, the response is

$$A_{r,t}(p) = \frac{1}{|\mathcal{H}_F|} \sum_{h \in \mathcal{H}_F} A_{r,t}^{(h)}(p). \tag{3}$$

Each nonconstant map is min–max normalized to place its responses on a common scale:

$$\mathrm{N}(X)(p) = \frac{X(p) - \min_{z \in \Omega_V} X(z)}{\max_{z \in \Omega_V} X(z) - \min_{z \in \Omega_V} X(z)}. \tag{4}$$

Constant maps are assigned zero, and we write $\bar{A}_{r,t} = \mathrm{N}(A_{r,t})$. Instruction-prefix tokens can activate salient regions Chen et al. (2026); their token set $\mathcal{P}_r$ defines a pointwise median prior:

$$B_r(p) = \mathrm{median}_{u \in \mathcal{P}_r} \mathrm{N}(A_{r,u})(p). \tag{5}$$

Prefix tokens end at the grounding instruction's colon; chat-template suffix tokens are excluded. Each signal-token response is calibrated separately:

$$R_{r,t}(p) = \mathrm{N}\big([\bar{A}_{r,t} - B_r]_+\big)(p). \tag{6}$$

Here, $[\cdot]+$ clips negative values to zero. If prior removal produces an identically zero map, we retain the original $\bar{A}r, t$. Signal-token maps are kept separate to avoid diluting informative entity- or

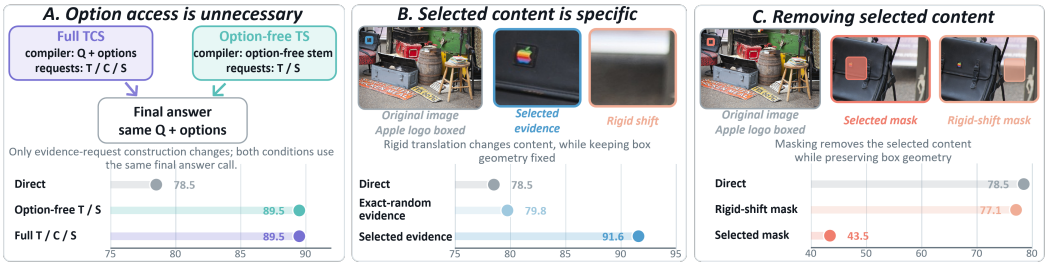


Figure 3: **Matched controls on V*Bench with Qwen2.5-VL-7B.** (A) Option-free TS matches full-input TCS while using the same final answer call. (B) Selected evidence outperforms exact-geometry random evidence (91.6% vs. 79.8%). (C) Masking selected regions causes a substantially larger accuracy drop than masking rigidly shifted regions (43.5% vs. 77.1%).

attribute-specific responses through averaging with unrelated tokens. For each calibrated response map $R_{r,t}$, we independently compute a threshold $\tau_{r,t} = \mathrm{Yen}(R_{r,t})$ using Yen's criterion (Yen et al., 1995). Thresholding $R_{r,t}$ at $\tau_{r,t}$ yields the foreground support $p \in \Omega_V : R_{r,t}(p) \geq \tau_{r,t}$, which is partitioned into eight-connected components collected in $\mathcal{C}r, t^+$. We then select as the seed the component with the largest total response mass:

$$Cr, t = \arg, \max C \in \mathcal{C}r, t^+ \sum_{p \in C} R_{r,t}(p). \tag{7}$$

Equal-mass ties are resolved in row-major order. Expansion is enabled for requests with multiple entity fields ($K_r > 1$), as in relation-mode T requests. Let $V^-_{r,t}$ denote response values below $\tau_{r,t}$; when nonempty, $\tau^-_{r,t} = \mathrm{Yen}(V^-_{r,t})$ defines

$$C^-_{r,t} = \mathrm{CC}_8\big(\{p \in \Omega_V : R_{r,t}(p) \geq \tau^-_{r,t}\}; C^*_{r,t}\big)\,. \tag{8}$$

$\mathrm{CC}_8(\cdot; C^*_{r,t})$ denotes the eight-connected component containing the original seed. If expansion is disabled or $V^-_{r,t}$ is empty, the seed is retained. Selected components are mapped to source-pixel boxes in Ω_I, and token groups associate these boxes with entities. An enclosing entity box replaces its token boxes only if it has no positive-area overlap with another entity's boxes. Otherwise, the original token boxes are preserved to avoid absorbing a reference entity. Detailed in Appendix F.1.

3.4 Evidence-Conditioned Answering

As shown in Fig. 2 (4) , for each surviving request, $\mathcal{B}_r$ contain its retained boxes in row-major order. Their source pixels are assembled on a transparent canvas:

$$E_r = \mathrm{Assemble}([\Pi(I, b)]_{b \in \mathcal{B}_r})\,, \qquad r \in \mathcal{R}(Q). \tag{9}$$

$\Pi(I, b)$ copies the pixels inside b; assembly removes empty gaps while preserving the retained pixels and region order. Exact duplicate views are removed in $T/C/S$ order to form $\mathbf{E}$. The model's native multi-image interface receives

$$\mathbf{X}_E = \mathrm{InputPack}\big([I] \| \mathrm{UniqueView}\big([E_r]_{r \in \mathcal{R}(Q)}\big)\big)\,. \tag{10}$$

The original image supplies the global spatial reference, while $\hat{a} = F^{\mathrm{ans}}_\theta(\mathbf{X}_E, Q)$ preserves the original query, task-specific answer prompt, and decoding rule. All benchmarks share these operators, without external detectors, learned fusion weights, or benchmark-conditioned branches.

4 Analyses of Evidence Specification

To isolate the effect of structured requests from increased visual access, we conduct **two controlled experiments**: one on localization and one on the utility of localized evidence for answering. These analyses control for request multiplicity, spatial coverage, and access to answer choices. All experiments use Qwen2.5-VL-7B on V*Bench, with protocols detailed in Appendix B and Appendix C.

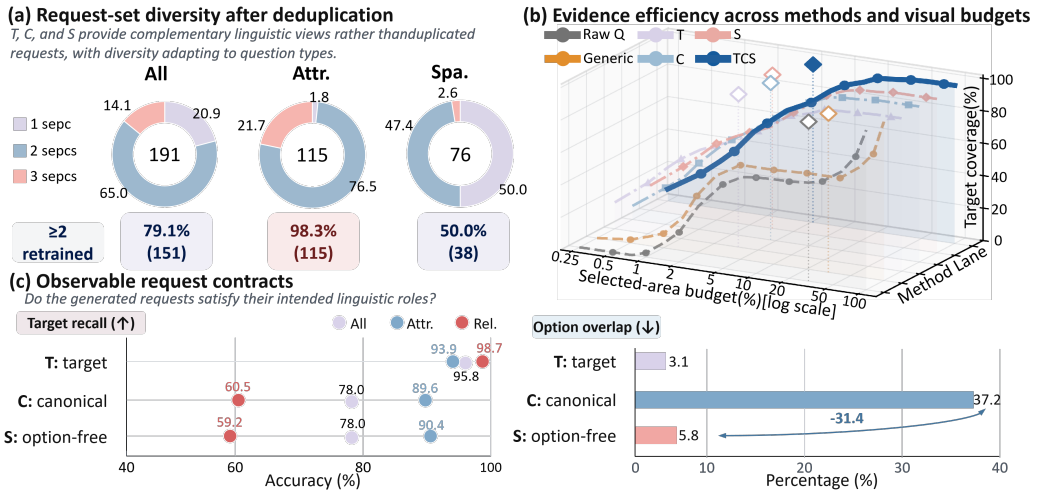


Figure 4: Analysis of evidence specifications: (a) diversity across T, C, and S, (b) evidence efficiency under varying visual budgets, and (c) semantic differences in their intended localization roles.

4.1 Disentangling Request Structure from Multiplicity

We first explore whether EviSpec benefits from generating more localization requests or from expressing the evidence need more effectively. Generic request generation and TCS each receive three generation attempts before deduplication and evidence assembly.TCS achieves an accuracy of 88.5%, compared with 79.1%, representing an absolute improvement of 9.4 percentage points and a relative gain of 11.9%. Yet TCS retains fewer distinct non-empty requests after deduplication, averaging 1.9 rather than 2.6. The combination of higher accuracy and fewer retained requests distinguishes request quality from request quantity. In this control, the improvement cannot be attributed to a larger retained request set. Instead, the result supports explicitly organizing the information supplied to the localizer, rather than treating additional formulations as inherently useful search instructions. This distinction motivates EviSpec's compiler interface: the objective is to specify the required visual evidence, not simply to produce more versions of the question.

4.2 Disentangling Request Roles from Spatial Extent

The request profiles in Fig. 4 (a,c) help interpret what this organization changes. The branches exhibit different degrees of target preservation and answer-option overlap, with different subsets retained for attribute and spatial questions. These properties reflect different treatments of the same evidence need: T foregrounds target roles, C canonicalizes the complete query, and S isolates the question stem. Exact-string deduplication removes duplicate outputs while preserving distinct formulations, yielding a question-dependent request set rather than a fixed three-request expansion. The spatial consequences are examined through target coverage under selected-area budgets in Fig. 4 (b). Coverage alone can reward broad selections that include the target with substantial surrounding content. The area constraint makes this distinction explicit by asking whether target coverage is retained without exceeding the permitted spatial extent. At intermediate budgets, TCS achieves higher coverage than raw questions and generic requests. Its advantage concerns the concentration of target-relevant content within compact selections, rather than coverage obtained through unrestricted enlargement. The request profiles and coverage curves provide complementary observations: the former characterize the linguistic inputs, while the latter reveal differences in spatial selectivity.

4.3 Disentangling Selected Content from Visual Access

Compact localization is useful only if it preserves information needed for the final answer. **First,** we examine whether constructing such evidence requires access to the candidate answers. In Fig. 3 (A), option-free TS and full-input TCS both achieve 89.5% accuracy, while the final answer call receives the same question and options in both settings. Thus, the question stem alone is sufficient to retain

Table 1: Answer accuracy (%) on V*Bench and HR-Bench. **Red** and **blue** entries mark the best and second-best distinct results within each open-source base-model block; Δ is EviSpec minus Base.

High-Resolution Visual Question Answering											
Method	**Base Model**	**Train Free**	**V***			**HRBench-4K**			**HRBench-8K**		
			Attr	**Spa.**	**Avg**	**FSP**	**FCP**	**Avg**	**FSP**	**FCP**	**Avg**
–	GPT-4o	–	–	–	66.0	70.0	48.0	59.0	62.0	49.0	55.5
Base	Qwen2.5-VL 3B	–	80.9	61.8	73.3	83.0	50.3	66.6	79.8	45.0	62.4
ViCrop		✓	81.7	**65.8**	75.4	86.3	49.8	68.0	80.3	45.5	62.9
HiDe		✓	**85.2**	**71.1**	**79.6**	**87.8**	**51.5**	**69.6**	**88.3**	**51.3**	**69.8**
EviSpec		✓	**90.4**	**71.1**	**82.7**	**92.3**	**55.3**	**73.8**	**89.3**	**53.8**	**71.5**
Δ (vs. Base)		–	+9.5	+9.3	+9.4	+9.3	+5.0	+7.2	+9.5	+8.8	+9.1
Base	Qwen2.5-VL 7B	–	80.9	76.3	79.1	88.8	54.3	71.5	84.3	51.5	67.9
ViCrop		✓	89.6	71.1	82.2	90.5	57.5	74.0	85.5	53.0	69.3
DeepEyes		✗	91.3	**88.2**	90.1	91.3	59.0	75.1	86.8	58.5	72.6
HiDe		✓	92.2	**89.5**	**91.1**	**95.3**	58.3	**76.8**	**93.8**	**58.8**	**76.3**
DeepScan		✓	**93.0**	86.8	90.6	90.1	**59.7**	75.0	87.2	57.6	72.4
EviSpec		✓	**95.7**	86.8	**92.1**	**97.0**	**64.8**	**80.9**	**97.5**	**65.8**	**81.6**
Δ (vs. Base)		–	+14.8	+10.5	+13.0	+8.2	+10.5	+9.4	+13.2	+14.3	+13.7
Base	Qwen2.5-VL 32B	–	87.8	88.1	87.9	89.8	58.0	73.9	84.5	56.3	70.4
ViCrop		✓	88.7	85.5	87.4	90.5	**59.5**	74.6	88.8	54.3	71.5
HiDe		✓	**91.3**	**89.5**	**90.6**	**92.0**	**60.0**	**76.2**	**90.5**	**57.5**	**74.0**
EviSpec		✓	**89.6**	**90.8**	**90.1**	**93.8**	**60.0**	**76.6**	**94.5**	**59.8**	**77.1**
Δ (vs. Base)		–	+1.8	+2.7	+2.2	+4.0	+1.5	+2.7	+10.0	+3.5	+6.7
Base	Qwen3-VL 8B	–	87.0	78.9	83.8	90.2	63.7	77.0	84.0	62.3	73.1
ViCrop		✓	89.6	75.0	83.8	91.8	64.3	78.0	88.0	63.3	75.6
HiDe		✓	88.7	84.2	86.9	**93.8**	**68.5**	**81.1**	**94.8**	**64.0**	**79.4**
DeepScan		✓	**92.2**	**89.5**	**91.1**	92.5	64.5	78.5	90.0	62.3	76.1
EviSpec		✓	**94.8**	**86.8**	**91.6**	**94.8**	**70.3**	**82.5**	**94.0**	**66.8**	**80.4**
Δ (vs. Base)		–	+7.8	+7.9	+7.8	+4.6	+6.6	+5.5	+10.0	+4.5	+7.3
Base	InternVL3 8B	–	81.7	**78.9**	80.6	82.8	**58.8**	70.8	80.0	**59.8**	69.9
ViCrop		✓	88.7	75.0	83.3	88.0	57.0	72.5	82.8	**54.8**	68.8
HiDe		✓	**92.2**	**88.2**	**90.6**	**91.8**	58.3	**75.0**	**86.3**	53.8	**70.0**
EviSpec		✓	**90.4**	**88.2**	**89.5**	**93.8**	**61.0**	**77.4**	**90.8**	**59.8**	**75.3**
Δ (vs. Base)		–	+8.7	+9.3	+8.9	+11.0	+2.2	+6.6	+10.8	+0.0	+5.4

the observed aggregate accuracy in this control, supporting a separation between the information used to guide visual search and the complete input used for answer selection.

Second, we test whether the benefit comes from the selected content rather than merely from receiving additional evidence views. A shared rigid translation preserves box count, dimensions, overlap, union area, and packed-view dimensions while altering the sampled image content. As shown in Fig. 3 (B), selected evidence achieves 91.6% accuracy, compared with 79.8% for geometry-matched random evidence, corresponding to an 11.8-point absolute gain and a 14.8% relative improvement. With geometry held fixed, this difference supports the value of selecting question-relevant content.

Third, we perform the complementary intervention by removing the selected content. In Fig. 3 (C), accuracy drops to 43.5% when selected regions are masked, compared with 77.1% for geometrically matched shifted regions. Thus, removing selected regions is substantially more disruptive than removing matched alternatives. Together, the provision and removal interventions show from opposite directions that the localized regions contain information relevant to the downstream answer.

5 Experiments

5.1 Experimental Setup

Benchmarks and metrics We evaluate V*Bench, HR-Bench-4K and HR-Bench-8K (Wu & Xie, 2024; Wang et al., 2025c). V*Bench contains 191 questions, including 115 direct-attribute and 76 relative-position questions. Each HR-Bench split contains 800 questions divided evenly between fine-grained single-region perception (FSP) and cross-region perception (FCP). We report exact-answer accuracy after normalizing each response to one option letter. To examine transfer beyond

Table 2: Additional-task accuracy (%) under a shared protocol across benchmarks. Within each model block, **red** and **blue** bold values indicate the best and second-best distinct results.

⇄ Transfer Across Visual Tasks														
Method	**POPE**				**ZoomBench**			**MME-RW-Lite**			**TreeBench**			**DocVQA**
	Adv	**Pop**	**Ran**	**Avg**	**MCQ**	**Blank**	**Avg**	**Per**	**Rea**	**Avg**	**Per**	**Rea**	**Avg**	**Acc**
Qwen2.5-VL-7B-Instruct														
Base	84.0	84.4	85.1	84.5	49.9	**37.1**	46.5	51.6	39.3	46.8	43.6	33.2	37.0	81.7
ViCrop	84.9	85.1	85.7	85.2	41.2	26.8	37.4	55.6	41.6	50.1	40.3	**34.4**	36.5	81.8
HiDe	85.1	85.4	86.3	85.6	**58.1**	**36.2**	**52.3**	**57.8**	**42.1**	**51.7**	**49.0**	34.0	**39.5**	81.8
DeepScan	**85.4**	**86.0**	**86.8**	**86.1**	53.5	34.4	48.4	49.0	37.2	44.4	47.0	32.0	37.5	**82.1**
EviSpec	**86.4**	**87.3**	**89.2**	**87.6**	**61.8**	**37.1**	**55.3**	**56.9**	**45.2**	**52.3**	**50.3**	**34.8**	**40.5**	**82.7**
Δ (vs. Base)	+2.4	+2.9	+4.1	+3.1	+11.9	+0.0	+8.8	+5.3	+5.9	+5.5	+6.7	+1.6	+3.5	+1.0
Qwen3-VL-8B-Instruct														
Base	87.2	**88.6**	90.9	88.9	46.4	34.4	43.2	**55.9**	43.6	51.1	56.4	**37.8**	**44.7**	88.7
HiDe	**87.5**	**88.6**	**91.8**	**89.3**	**57.2**	34.8	51.2	55.3	**45.9**	**51.6**	**57.7**	36.7	44.4	89.0
DeepScan	86.7	88.5	91.2	88.7	56.0	**41.5**	**52.2**	52.8	40.1	47.8	53.7	36.3	42.7	**89.7**
EviSpec	**87.8**	**88.7**	**92.1**	**89.6**	**62.0**	**45.1**	**57.5**	**58.3**	**47.9**	**54.2**	**58.4**	**39.8**	**46.7**	**90.0**
Δ (vs. Base)	+0.6	+0.1	+1.3	+0.7	+15.6	+10.7	+14.3	+2.4	+4.3	+3.1	+2.0	+2.0	+2.0	+1.3

Table 3: Controlled request ablation on Qwen2.5-VL-7B. Entries report answer accuracy (%). **Red** and **blue** bold values indicate the best and second-best distinct results, respectively.

Config.	**V***			**HRBench-4K**			**HRBench-8K**		
	Attr	**Spa.**	**Avg**	**FSP**	**FCP**	**Avg**	**FSP**	**FCP**	**Avg**
T	93.0	81.6	88.5	94.0	62.5	78.3	94.8	**62.5**	**78.6**
C	91.3	82.9	88.0	95.0	57.0	76.0	92.5	59.5	76.0
S	91.3	82.9	88.0	94.5	60.8	77.6	93.5	62.3	77.9
TC	93.0	82.9	89.0	95.0	**63.0**	**79.0**	94.3	62.0	78.1
TS	93.0	**84.2**	89.5	95.3	62.8	**79.0**	**95.8**	61.0	78.4
CS	**93.9**	**84.2**	**90.1**	**95.8**	60.5	78.1	93.3	61.8	77.5
TCS	**95.7**	**86.8**	**92.1**	**97.0**	**64.8**	**80.9**	**97.5**	**65.8**	**81.6**

these core benchmarks, we further evaluate POPE, ZoomBench, MME-RealWorld-Lite, TreeBench, and DocVQA. (Mathew et al., 2021; Li et al., 2023b; Zhang et al., 2025b; Wang et al., 2026a; Wei et al., 2026). POPE is reported on its adversarial, popular and random splits; ZoomBench separates multiple-choice (MCQ) and open-ended blank questions; MME-RealWorld-Lite and TreeBench report perception and reasoning accuracy; and DocVQA reports answer accuracy.

Models and inference The completed evaluation covers five frozen MLLMs: Qwen2.5-VL-3B, Qwen2.5-VL-7B, Qwen2.5-VL-32B (Bai et al., 2025a), Qwen3-VL-8B (Bai et al., 2025b), and InternVL3-8B (Zhu et al., 2025). The main-table Qwen executions and the branch ablation use a maximum pixel budget of $16{,}384 \times 28^2$. We enable FlashAttention-2 Dao (2024) for efficient attention computation. EviSpec requires no learned parameters. Compiler prompts, attention processing, region formation and evidence-conditioned answering are shared across datasets. Complete implementation settings and supplementary evaluations are reported in Appendix E.

5.2 Quantitative Results

Performance on Benchmarks. Table 1 compares EviSpec with Base Model, ViCrop, HiDe, DeepEyes (Zheng et al., 2026), and DeepScan (Li et al., 2026). All methods are evaluated under a unified protocol; GPT-4o Hurst et al. (2024) serves as a proprietary reference and is excluded from within-block ranking. EviSpec improves over Base in all 15 model–benchmark combinations, with average relative gains of 10.4%, 8.8%, and 12.4% on V*Bench, HR-Bench-4K, and HR-Bench-8K, respectively. It also outperforms DeepScan across all settings and HiDe across all HR-Bench evaluations.

Fine-Grained Gains and Generalization. The gains span both fine-grained perception and cross-region reasoning. On Qwen2.5-VL-7B, EviSpec improves over Base by 14.8 points on the V*Bench attribute split and 14.3 points on the HR-Bench-8K FCP split. Table 2 further extends the evaluation to five additional visual tasks, where EviSpec achieves the best aggregate performance across all benchmarks for both models, with particularly strong gains on ZoomBench. Improvements in

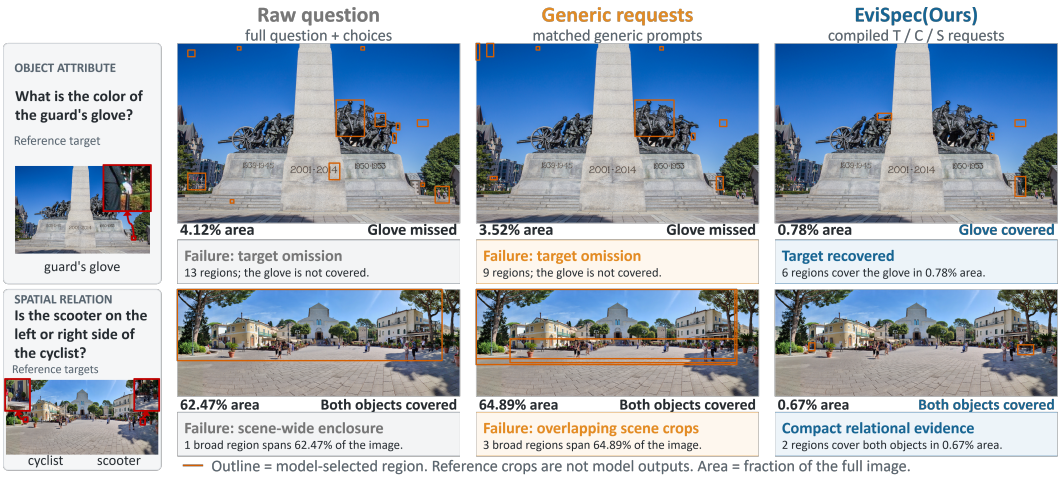


Figure 5: Qualitative comparison of evidence localization using raw questions, generic requests.

hallucination evaluation, real-world perception and reasoning, and document understanding further indicate generalization beyond high-resolution visual search.

5.3 Qualitative Results

Fig. 5 compares localization from raw multiple-choice questions, matched generic requests, and EviSpec. The examples highlight two aspects of evidence selection: recovering small answer-bearing objects and preserving spatial relations. In the upper example, the question asks for the color of a guard's glove in a cluttered monument scene. Raw-question and generic-request localization attend to several salient objects but miss the glove, whereas EviSpec retains it as the target and selects the evidence needed to determine its color. In the lower example, all three conditions cover both the cyclist and scooter, but the raw question and generic requests retain broader scene regions. EviSpec instead selects compact regions around the two entities while preserving their spatial arrangement, retaining the information needed for answering with less surrounding context.

5.4 Ablation Studies

We evaluate individual specifications and their combinations with Qwen2.5-VL-7B, holding the localization procedure, image budget, and final answering procedure fixed. This ablation assesses the request configuration used for the benchmark evaluation. Table 3 shows that the complete TCS configuration achieves the highest accuracy in all nine reported split and average metrics, reaching 92.1%, 80.9%, and 81.6% on V*Bench, HR-Bench-4K, and HR-Bench-8K, respectively. Among the single-request configurations, T gives the highest average accuracy on all three benchmarks, supporting the use of an explicit answer-bearing target together with its identifying context. Additional perturbation, option-isolation, and fusion analyses are provided in Appendix D.

6 Conclusion

This work identifies the evidence-specification gap in high-resolution multimodal reasoning: questions formulated for answering are not effective requests for locating the evidence needed to answer them. We introduce EviSpec, a training-free compiler that separates localization-facing evidence specification from final answering while preserving the original query. EviSpec consistently improves answer accuracy over the corresponding base models across all 15 benchmark combinations, covering five frozen MLLMs evaluated on three high-resolution benchmarks. More importantly, matched controls show that the gains do not rely on answer-option access or additional region geometry, but on selecting question-relevant visual content. Together, these results show that high-resolution reasoning failures are not solely a problem of visual access, and establish evidence specification as a distinct inference interface between question understanding and visual search.

AI USE STATEMENT

During the preparation of this manuscript, generative AI tools were used only for language editing and formatting assistance. No generative AI tools were used in the conception of the research, method development, experimental execution, or result analysis. All implementations and experimental verification were conducted by the authors, who assume full responsibility for the final content of the manuscript.

ETHICS STATEMENT

This work does not involve human subjects, sensitive data, personally identifiable information, or potentially harmful applications. All experiments are conducted on publicly available benchmarks and do not involve the collection or processing of private information. We do not anticipate any ethical concerns or direct negative societal impacts arising from this research.

REPRODUCIBILITY STATEMENT

We are committed to the reproducibility of EviSpec. Since EviSpec is training-free, all experiments use frozen MLLMs without model retraining. Upon publication, we will release the source code, evaluation and dataset preparation scripts, and all compiler and grounding prompts. The supplementary materials provide complete inference configurations and implementation details for reproducing the main experiments, ablations, and analyses.

REFERENCES

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

Shuai Bai, 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, 2025a.

Shuai Bai et al. Qwen3-vl technical report, 2025b.

Zengjie Chen, Yuxiang Cai, Jingcai Guo, Taotao Cai, Jianwei Yin, and Zhi Chen. Accelerating multimodal large language models with prior-corrected token reduction. *arXiv preprint arXiv:2606.24156*, 2026.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites, 2024a. URL https://arxiv.org/abs/2404.16821.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024b. URL https://arxiv.org/abs/2312.14238.

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.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, volume 2024, pp. 35549–35562, 2024.

Qihua Dong, Kuo Yang, Lin Ju, Handong Zhao, Yitian Zhang, Yizhou Wang, Huimin Zeng, Jianglin Lu, and Yun Fu. Ref-adv: Exploring MLLM visual reasoning in referring expression tasks. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=iEBgrepR9i.

Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. Mmtok: Multimodal coverage maximization for efficient inference of vlms. In *International Conference on Learning Representations*, volume 2026, pp. 48075–48099, 2026b.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *Advances in Neural Information Processing Systems*, 37:42566–42592, 2024.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Martin Nicolas Everaert, Xiruo Liu, Hiroyuki Takeda, Raja Bala, Vivek Yadav, and Vidya Narayanan. Visual grounding for object questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11966–11975, 2026.

Mingxin Huang, Yuliang Liu, Dingkang Liang, Lianwen Jin, and Xiang Bai. Mini-monkey: Alleviating the semantic sawtooth effect for lightweight mllms via complementary image pyramid. In *International Conference on Learning Representations*, volume 2025, pp. 85804–85823, 2025.

Yihong Huang, Fei Ma, Yihua Shao, Jingcai Guo, Zitong YU, Laizhong Cui, and Qi Tian. Nüwa: Mending the spatial integrity torn by VLM token pruning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=C9yclwdquU.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Xinyi Jiang, Guoming Wang, Junhao Guo, Juncheng Li, Wenqiao Zhang, Rongxing Lu, and Siliang Tang. DIEM: Decomposition-integration enhancing multimodal insights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27304–27313, 2024a.

Yao Jiang, Xinyu Yan, Ge-Peng Ji, Keren Fu, Meijun Sun, Huan Xiong, Deng-Ping Fan, and Fahad Shahbaz Khan. Effectiveness assessment of recent large vision-language models. *Visual Intelligence*, 2(1):17, 2024b.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1780–1790, 2021.

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.

Thao Minh Le, Vuong Le, Sunil Gupta, Svetha Venkatesh, and Truyen Tran. Guiding visual question answering with attention priors, 2022.

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.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10965–10975, 2022.

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

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. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.emnlp-main.20.

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, 2024.

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.

Che Liu, Yingji Zhang, Dong Zhang, Weijie Zhang, Chenggong Gong, Yu Lu, Shilin Zhou, Ziliang Gan, Ziao Wang, Haipang Wu, et al. Nexus-o: An omni-perceptive and -interactive model for language, audio, and vision. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10787–10796, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

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

Xianjie Liu, Yiman Hu, Yixiong Zou, Liang Wu, Jian Xu, and Bo Zheng. HiDe: Rethinking the zoom-in method in high resolution MLLMs via hierarchical decoupling. *arXiv preprint arXiv:2510.00054*, 2026. Accepted at ICML 2026; version 3.

Tianren Ma, Lingxi Xie, Yunjie Tian, Boyu Yang, and Qixiang Ye. Clawmachine: Learning to fetch visual tokens for referential comprehension. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TOtk9dTYGG.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5303–5315, 2023.

Zhongkuan Mao, Xianjie Liu, Tianyu Meng, Yidong Wang, Wenzhuo Zhao, Ronghao Xian, Yao Jiang, Fei Shen, Junfeng Fang, Yong Dai, et al. Thinking once is enough: Intermediate-layer evidence routing for high-resolution vqa. *arXiv preprint arXiv:2607.27830*, 2026.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021. doi: 10.1109/WACV48630.2021.00225.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lLmqxkfSIw.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13009–13018. IEEE, 2024.

Yang Shi, Tommaso Furlanello, Sheng Zha, and Animashree Anandkumar. Question type guided attention in visual question answering. In *Proceedings of the European Conference on Computer Vision*, 2018.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. In *International Conference on Learning Representations*, volume 2025, pp. 13712–13736, 2025a.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, and Zhaoxiang Zhang. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. In *International Conference on Learning Representations*, 2026a.

Jiacong Wang, Zijian Kang, Haochen Wang, Xiao Liang, Ya Wang, Jiawen Li, Bohong Wu, Jiao Ran, Haiyong Jiang, Chao Feng, et al. Vgr: Visual grounded reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 89506–89528, 2026b.

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

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

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025c.

Yidong Wang, Zhuohao Yu, Jindong Wang, Qiang Heng, Hao Chen, Wei Ye, Rui Xie, Xing Xie, and Shikun Zhang. Exploring vision-language models for imbalanced learning. *International Journal of Computer Vision*, 132(1):224–237, 2024b.

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

Liangjian Wen, Qun Dai, Jianzhuang Liu, Jiangtao Zheng, Yong Dai, Dongkai Wang, Zhao Kang, Jun Wang, Zenglin Xu, and Jiang Duan. Infmasking: Unleashing synergistic information by contrastive multimodal interactions. *Advances in Neural Information Processing Systems*, 38: 15529–15555, 2026a.

Liangjian Wen, Linjie Li, Jiang Duan, Yong Dai, Jianzhuang Liu, and Zhao Kang. Dependency, compression, and synergy: A unified information-theoretic view of multimodal learning. *arXiv preprint arXiv:2609.14421*, 2026b.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. In *International Conference on Learning Representations*, volume 2025, pp. 93334–93358, 2025.

Yixuan Wu, Yang Zhang, Jian Wu, Philip Torr, and Jindong Gu. Postalign: Multimodal grounding as a corrective lens for mllms. In *International Conference on Learning Representations*, volume 2026, pp. 78875–78894, 2026.

Junfei Xie, Peng Pan, and Xulong Zhang. Head-aware visual cropping: Enhancing fine-grained vqa with attention-guided subimage, 2026.

Jui-Cheng Yen, Fu-Juay Chang, and Shyang Chang. A new criterion for automatic multilevel thresholding. *IEEE Transactions on Image Processing*, 4(3):370–378, 1995.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, volume 2024, pp. 57153–57180, 2024.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1307–1315, 2018.

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

Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. In *International Conference on Learning Representations*, volume 2025, pp. 64669–64714, 2025a.

Haowei Zhang, Jianzhe Liu, Zhen Han, Shuo Chen, Bailan He, Volker Tresp, Zhiqiang Xu, and Jindong Gu. Visual question decomposition on multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1926–1949. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.107.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. MME-RealWorld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, 2025b.

Wenzhuo Zhao, Keren Fu, Jiahao He, Xiaohong Liu, Qijun Zhao, and Guangtao Zhai. Samba+: General and accurate salient object detection via a more unified mamba-based framework. *arXiv preprint arXiv:2602.01593*, 2026a.

Wenzhuo Zhao, Xiuzhi Li, Zhongkuan Mao, Ronghao Xian, Yao Jiang, Zhao Gao, Keren Fu, Qijun Zhao, and Jian Cheng. Is it time for the renaissance of salient object detection in the era of mllms? *arXiv preprint arXiv:2607.29222*, 2026b.

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

Liangyu Zhong, Fabio Philipp Rosenthal, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, and Leo Schwinn. FOCUS: Internal MLLM representations for efficient fine-grained visual question answering. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Zipeng Zhu, Zhanghao Hu, Qinglin Zhu, Yuxi Hong, Yijun Liu, Jingyong Su, Yulan He, and Lin Gui. Beyond static cropping: Layer-adaptive visual localization and decoding enhancement, 2026.

Appendix: Pay More Attention to Text in High-Resolution MLLMs

Contents

Appendix A **Relation to Prior Work and Experimental Overview** Page 16
Comparison with related approaches, matched baselines, and an overview of the supplementary evaluations.

Appendix B **Matched Question-to-Evidence Controls** Page 17
Prompt-family comparisons, request-interface controls, incremental specification analysis, and component diagnostics.

Appendix C **Matched Selected-Content Intervention** Page 20
Selected-evidence interventions and exact-geometry controls that isolate the contribution of question-relevant visual content.

Appendix D **Supplementary Perturbation and Fusion Results** Page 21
Query perturbations, parser audits, option-free controls, compiled-query diversity, fusion studies, and qualitative evidence localization.

Appendix E **Implementation Details** Page 22
Implementation settings, software and hardware environments, and end-to-end inference-efficiency measurements.

Appendix F **Supplementary Method and Proofs** Page 23
Complete operators, the adaptive-wrapper inference procedure, formal guarantees, and computational-complexity analysis.

Table 4: **Comparison with related approaches.** "Intermediate object" denotes the representation produced before final answering.

Work	Intermediate object	Implementatic	Relation to EviSpec
Visual Question Decomposition (Zhang et al., 2024)	Grounded reasoning subquestions	Fine-tuned	Decomposes questions into grounded reasoning steps, whereas EviSpec compiles localization requests while keeping the final reasoning query unchanged.
DIEM (Jiang et al., 2024a)	Matched subquestion–subimage pairs	Training-free	Pairs subquestions with subimages, while EviSpec produces complementary localization requests and preserves a single final query.
HiDe (Liu et al., 2026)	Token-decoupled attention regions and compact layout-preserving evidence	Training-free	Uses attention-guided evidence regions as a training-free comparator. EviSpec instead studies how the query is compiled into complementary requests before a shared localization stage.
FOCUS (Zhong et al., 2025)	Target-object relevance maps and crops	Training-free	Improves visual search after identifying target objects, whereas EviSpec intervenes on the linguistic request supplied to localization.
HAVC / LASER (Xie et al., 2026; Zhu et al., 2026)	Head- or layer-selected attention crops	Training-free	Improve which internal spatial signal is used; EviSpec instead changes the query representation evaluated through a shared spatial operator.
AdaptVision (Lin et al., 2026)	Learned crop-tool actions	Reinforcement learning	Learns an acquisition policy and token allocation, whereas EviSpec uses a training-free request compiler.
Visual Grounding for Object Questions (Everaert et al., 2026)	Answer-supporting evidence regions	Trained grounding model	Studies a broader grounding task, while EviSpec focuses on question-to-evidence compilation for high-resolution visual reasoning.

A Relation to Prior Work and Experimental Overview

Table 4 compares EviSpec with visual-question decomposition, question-conditioned cropping, evidence grounding and learned visual acquisition. EviSpec compiles one multiple-choice question into complementary localization requests and preserves the complete question for final answering. Matched comparisons use the same localization and evidence-view operators across request variants.

Complete-question canonicalization provides a matched baseline for evaluating role-separated requests and option isolation. The remaining experiments cover request structure, selected visual content, perturbation consistency, and the complete inference procedure.

The supplementary evaluation is organized around three evidence blocks. Prompt-family comparisons evaluate typed request construction. Selected-content experiments compare the resulting evidence with matched visual alternatives, while perturbation analyses examine behavior across modified inputs. Implementation and formal operator details are reported afterward.

B MATCHED QUESTION-TO-EVIDENCE CONTROLS

B.1 REQUEST INTERFACE AND MATCHED PROMPT-FAMILY CONTROLS

Three controlled evaluations distinguish request structure from downstream visual processing. All use the 191 V*Bench questions, Qwen2.5-VL-7B, layer 15, bfloat16, FlashAttention-2, a fixed 448×448 evidence-view resolution, the original image and unchanged final question. Request generation permits at most 1,024 new tokens; answer generation uses the same greedy, valid-letter contract in every condition.

PROMPT TEMPLATES AND REQUEST MAPPING

The T/C/S templates and matched generic templates are held constant throughout the matched comparison. Table 5 gives all six request instances. C and S share one instruction literal and differ only in whether the substituted input is the complete serialized question Q or the option-free prefix q_0.

Table 5: Request-by-request mapping in the primary matched prompt-family control. The two families have identical input scopes and output tags.

Family	Request	Input	Required tags
T/C/S	T	complete Q	`MODE`, `FINAL_OUTPUT`
T/C/S	C	complete Q	`FINAL_OUTPUT`
T/C/S	S	option-free q_0	`FINAL_OUTPUT`
Generic	G_1	complete Q	`MODE`, `FINAL_OUTPUT`
Generic	G_2	complete Q	`FINAL_OUTPUT`
Generic	G_3	option-free q_0	`FINAL_OUTPUT`

The operational lexical content of the five unique instruction literals is reproduced below, with Markdown emphasis removed and line wrapping normalized for typesetting; {`INPUT`} marks the runtime substitution. The complete prompt text and request pairing are therefore specified directly in this appendix.

Role-separated T literal.

> You are a visual grounding query parser. Do not answer the multiple-choice question and do not copy answer options.
>
> First decide whether the question is DIRECT or RELATION.
>
> DIRECT asks for an attribute, text, count, identity, or existence. Output one answer-bearing target phrase. Preserve identifying context inside the same phrase. Example: "What color is the backpack carried by the man wearing a yellow shirt?" → backpack carried by man wearing yellow shirt.
>
> RELATION asks to compare, order, match, or locate one visual object relative to another. Output both the subject and reference as two independently groundable noun phrases, in question order. Never omit the subject. Example: "Where is the person in relation to the recycle bin?" → person, recycle bin. Example: "What is the position of the red car compared to the black car?" → red car, black car.
>
> Return exactly two tagged fields and no other text: `<MODE>direct</MODE><FINAL_OUTPUT>one target phrase</FINAL_OUTPUT>` or `<MODE>relation</MODE><FINAL_OUTPUT>subject, reference</FINAL_OUTPUT>`.
>
> Use lowercase. A comma followed by a space is reserved for separating RELATION targets. Question: {`INPUT`}.

Canonical-form C/S literal.

> You are a highly precise language analysis engine. Your sole function is to extract entities (e.g., objects, people) from a user's question, and deconstruct them into a canonical, attribute-based format by strictly following a set of rules and a thinking process.
>
> **Thinking Process.** Before generating the final output, you must internally follow these steps in order:

1. Identify Core Entities: read the entire question and identify all key noun phrases, for example, "the green surfboard" and "the purple umbrella".

2. Deconstruct Attributes for Each Entity Individually: before considering the relationship between entities, look at each entity in isolation and apply Rules 2, 3 and 4 to fully deconstruct its attributes. For instance, first process "the green surfboard" using Rule 2 to obtain `surfboard with green color`, then process "the purple umbrella" to obtain `umbrella with purple color`.

3. Handle Relationships Between Entities: after all entities have been individually deconstructed, check for spatial or logical relationships between them (Rule 5). If a relationship exists, list the already deconstructed entities as separate items.

4. Assemble and Normalize: gather all canonical entity strings, convert all text to lowercase, and join all entities into a single line separated by a comma and a space.

5. Final Formatting: enclose the resulting single-line string within the `<FINAL_OUTPUT>` and `</FINAL_OUTPUT>` tags.

Extraction Rules. Rule 1 (Simple Entities): if a noun is not described by modifiers, extract the noun itself; for example, "the scooter" becomes `scooter`. Rule 2 (Adjective Attribute Deconstruction): if an entity is modified by one or more adjectives, use `noun with [property] [type]`, chaining multiple properties. Colors map to `color`, sizes to `size`, and materials to `material`; for example, "the large blue truck" becomes `truck with large size with blue color`. Rule 3 (Possessive Inversion): convert possessive forms such as `X's Y` to `Y of X`; for example, "the woman's handbag" becomes `handbag of woman`. Rule 4 (Attributive Prepositional Phrases): if a prepositional phrase describes a component of an entity, preserve the structure and recursively apply the rules to the entity within the phrase; for example, "the man in the green shirt" becomes `man in a shirt with green color`. Rule 5 (Relational Prepositional Phrases): if a prepositional phrase describes a relationship between two separate entities, extract the entities as separate items only after each has been fully deconstructed. Do not include relational words. Thus, "the dog on the left side of the scooter" becomes `dog, scooter`, and "Is the green surfboard on the left side of the purple umbrella?" becomes `surfboard with green color, umbrella with purple color`. Rule 6 (Compound Nouns): treat recognized compound nouns as one entity; for example, "the traffic light" becomes `traffic light`.

Output Format Rules. The final and only output content must be enclosed within `<FINAL_OUTPUT>` and `</FINAL_OUTPUT>`. It must be one continuous lowercase line. Separate multiple entities by a comma followed by a space. Exclude articles, question words and purely relational words.

Now, following all the rules above, extract the entities from the question below: {`INPUT`}.

Matched generic literals.

G_1: You are a visual search query rewriter. Do not answer the multiple-choice question. First decide whether the input is DIRECT (attribute, text, count, identity, or existence) or RELATION (comparison, order, match, or location of one object relative to another). Then write a concise visual search description that preserves every visual object, attribute, and relation without assigning target, context, or spatial roles or adding facts. For RELATION include both compared objects in question order. Return exactly `<MODE>direct</MODE><FINAL_OUTPUT>one search description</FINAL_OUTPUT>` or `<MODE>relation</MODE><FINAL_OUTPUT>one search description containing both objects</FINAL_OUTPUT>`. Input: {`INPUT`}.

G_2: Paraphrase the input below into one concise visual search description. Preserve every visual object, attribute, and relation; do not add facts or assign semantic roles. Return only `<FINAL_OUTPUT>one search description</FINAL_OUTPUT>`. Input: {`INPUT`}.

G_3: Convert the input below into one concise visual search description. Preserve every visual object, attribute, and relation; do not add facts or assign semantic roles. Return only `<FINAL_OUTPUT>one search description</FINAL_OUTPUT>`. Input: {`INPUT`}.

Table 6 illustrates the requests produced by the two prompt families. Both complete-input families can reproduce option text, and each matched variant explicitly specifies the text available to the request generator.

Table 6: Example requests generated by the structured and matched generic prompt families. Entries are ordered $T/C/S$ or $G_1/G_2/G_3$.

ID	T/C/S outputs	Matched generic outputs
0	`glove; glove with rubber material; glove`	`glove material; find glove material: rubber, cotton, kevlar, leather.; visual search for glove material`
1	`dustpan; dustpan with purple color; dustpan with color`	`the color of the dustpan; find a dustpan with a purple color.; visual search for dustpan with specific color`

Table 7: Matched request-interface comparison on 191 V*Bench questions. Accuracy values are percentages rounded to one decimal place; bold marks the higher value in each row.

Comparator	Comparator accuracy	T/C/S accuracy
Raw option-free stem	83.2	**88.5**
Canonical C	89.0	**90.1**

The experiments below use frozen Qwen2.5-VL-7B analysis archives with the same image processor, answer prompt and greedy decoding setup. The main table reports cross-model benchmark results, while these matched executions examine request construction and visual evidence. Accuracies are rounded to one decimal place.

"Direct" denotes the Base answer pipeline, "T/C/S" denotes the complete EviSpec request construction, and "Selected evidence" denotes the visual content supplied by its matched intervention. The main-table EviSpec accuracy is 92.1%, while the matched request-construction and selected-content executions reach 90.6% and 91.6%, respectively.

B.2 Request-interface comparisons

The first comparison holds the visual operator and answer procedure fixed while changing only the requests supplied to localization. Direct uses the complete question without an evidence request. The option-free stem removes the answer options, and canonical C rewrites the complete question into a structured entity description. The T/C/S compiler emits a role-separated target request, a canonical entity request and an option-isolated request. A union control packs these regions into one canvas, whereas the standard T/C/S pipeline keeps the views separate.

The one-canvas T/C/S union reaches 88.5% against the raw option-free stem and 90.1% against canonical C. The separate-view execution reaches 90.6% on the same 191 V*Bench questions. The one-canvas comparison evaluates request construction, while the EviSpec pipeline presents separate evidence views to the final answer stage.

The structured compiler uses fewer distinct requests and reaches higher accuracy under the same request budget. Both conditions receive three generation attempts; T/C/S assigns target, canonical and option-isolated roles to them.

The matched generic family retains more distinct requests and a larger selected area, whereas the structured family produces higher accuracy with fewer, more compact requests. Model capacity and the downstream visual operator are unchanged between the two prompt families.

B.3 Incremental specification and component diagnostics

The branch diagnostic derives each accuracy change from the controlled request ablation in the main paper. It records the effect of adding the canonical or option-isolated request to an existing request set.

Table 8: Equal-search-budget comparison. Both conditions attempt three generated requests before duplicate removal and evidence assembly. Bold marks higher accuracy or exclusive successes, and fewer distinct requests.

Quantity	Generic3	T/C/S
Accuracy	79.1	**88.5**
Exclusive successes	4	**22**
Average distinct non-empty requests	2.6	**1.9**

Table 9: Request-by-request prompt-family comparison on V*Bench. Bold marks higher accuracy or exclusive successes, fewer requests, and smaller selected area.

Quantity	Direct	Matched generic	T/C/S
Accuracy	78.5	77.0	**84.3**
Exclusive successes	–	7	**21**
Average distinct non-empty requests	–	2.9	**1.9**
Three distinct requests	–	176/191	27/191
Average selected area	–	69.6	**8.0**

Across all three benchmarks, the largest increase occurs when the final missing request is added to form TCS. This pattern complements the main ablation by showing that the three request roles provide their strongest result jointly.

The component grid separates attention-map processing from region formation. Calibrated maps with adaptive regions achieve the highest accuracy at 93.2 and raise mIoU from 12.9 to 17.8 relative to the legacy pairing. The legacy pairing retains the highest mIoG and coverage. EviSpec uses the calibrated-map and adaptive-region configuration in the request-construction comparisons.

C Matched Selected-Content Intervention

This intervention keeps the frozen model, answer prompt, layer and question set fixed while changing the visual content supplied to the model. Blank and resized views control the image budget, selected-only input measures evidence sufficiency, selected-region masking removes the selected appearance, and random controls preserve the number and geometry of selected boxes.

Table 12: Matched V*Bench evidence-intervention results. Values are observed accuracy percentages; bold marks the highest value.

Condition	Accuracy (%)
Direct	78.5
Blank evidence views	78.5
Global resized views	79.1
Random evidence	80.5
Selected evidence	**91.6**
Selected evidence only	**91.6**
Random mask	79.8
Selected-region mask	44.0

Selected evidence produces the highest observed accuracy among the content conditions, while masking the selected region removes the appearance cues needed for reliable answering. The selected-only condition matches selected evidence in this execution, so the compact evidence view alone carries the visual content used for answering.

C.1 Exact-geometry rigid-translation confirmation

The geometry-matched control translates the complete selected box configuration with one shared integer-pixel offset. It preserves box count, pixel dimensions, overlap, union area and packed-view dimensions while changing the source content.

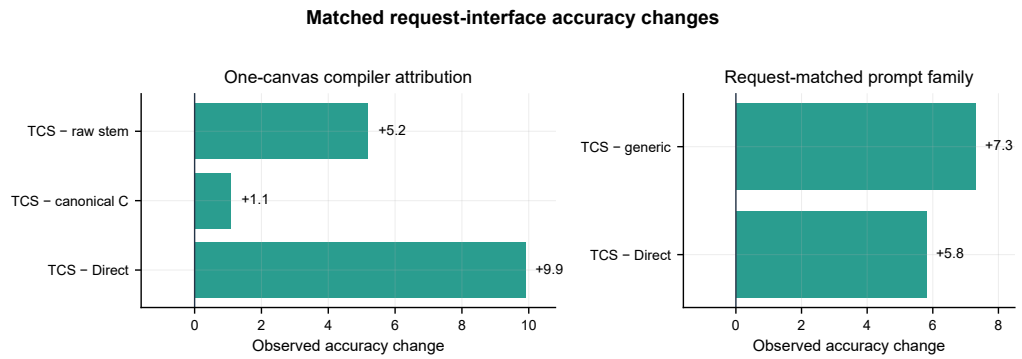


Figure 6: Matched request-interface accuracy changes. The left panel reports the one-canvas compiler-attribution contrasts; the right panel reports the request-matched prompt-family contrasts.

Table 10: Observed incremental specification changes in accuracy. Within each benchmark, bold marks the largest positive change.

Benchmark	Contrast	Accuracy change
V*Bench	$TC - T$	$+0.5$
	$TS - T$	$+1.0$
	$TCS - TC$	$+\mathbf{3.1}$
	$TCS - TS$	$+2.6$
HR-Bench-4K	$TC - T$	$+0.7$
	$TS - T$	$+0.7$
	$TCS - TC$	$+\mathbf{1.9}$
	$TCS - TS$	$+\mathbf{1.9}$
HR-Bench-8K	$TC - T$	-0.5
	$TS - T$	-0.2
	$TCS - TC$	$+\mathbf{3.5}$
	$TCS - TS$	$+3.2$

Table 13: Exact-geometry V*Bench conditions. Values are observed accuracy percentages; bold marks the highest value.

Condition	Accuracy (%)
Direct	78.5
Selected evidence	**91.6**
Exact-geometry random evidence	79.8
Selected-region mask	43.5
Exact-geometry random mask	77.1

All translated configurations preserve the assembled-view geometry by construction. The selected-evidence condition remains higher than the geometry-matched random evidence condition, while the selected-region mask remains lower than its random-mask counterpart. This isolates source content as the changing factor in the matched control.

D Supplementary Perturbation and Fusion Results

D.1 Query perturbations and parser audits

The perturbation suite applies option permutations, distractor replacements and target rewrites to the same V*Bench questions. Higher values are preferable for evidence IoU, worst-case accuracy and all-correct accuracy; lower values are preferable for answer changes.

Under both selectors, TCS gives higher evidence IoU, worst-case accuracy and all-correct accuracy, together with fewer answer changes than Single-C.

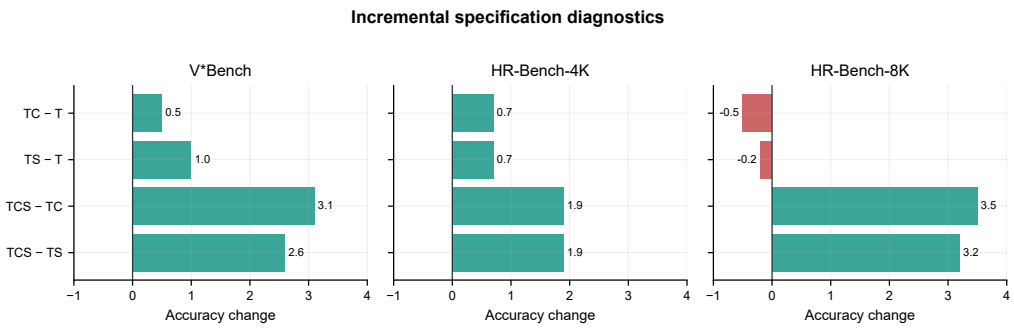


Figure 7: Observed incremental specification changes across the three benchmarks. Zero marks unchanged accuracy.

Table 11: Matched component diagnostic on 191 V*Bench questions. All entries are percentages; localization columns report region measurements. Bold marks the highest value in each metric column.

Attention map	Region formation	Accuracy	mIoU	mIoG	Cov.@0.5
Legacy	Legacy	89.5	12.9	**93.5**	**94.2**
Legacy	Adaptive	86.4	1.6	92.7	92.7
Calibrated	Legacy	85.3	**28.8**	50.3	48.2
Calibrated	Adaptive	**93.2**	17.8	92.6	93.7

The target-stem branch has the lowest empty-output rate and the lowest complete option-overlap rate. Canonical C retains the complete question and therefore shows the largest option overlap.

D.2 Strict option-free compiler intervention

The option-free intervention removes answer options before request construction while keeping the model, layer, image budget and answer prompt fixed.

The strict option-free compiler reaches the same aggregate accuracy as the full-input TCS condition in this matched execution. The correctness transition count is 170 correct-to-correct, 19 incorrect-to-incorrect, and one exchange in each direction. On HR-Bench-8K, adding C to TS raises accuracy by 3.2 in Table 10; the contribution of C therefore varies with the benchmark.

D.3 Compiled-query diversity and fusion

The query count depends on question type. Attribute questions usually retain two or three complementary requests, whereas spatial questions more often collapse to one or two requests.

For the fixed-threshold selector, TCS concatenation achieves the highest value in every reported column. For the adaptive selector, it ties the highest accuracy and achieves the highest attribute, spatial, mIoU, mIoG and coverage values.

D.4 Qualitative evidence localization

E Implementation Details

E.1 Implementation settings

All Qwen2.5-VL diagnostic controls in this appendix use `USE_ALLATT=False`, bfloat16, FlashAttention-2, greedy decoding and the same $16{,}384 \times 28^2$ maximum processor budget. The software environment uses Python 3.11.13, PyTorch 2.8.0+cu128 and Transformers 4.57.1, and experiments are run on NVIDIA A800 GPUs. The same compiler, attention-processing, region-formation and evidence-conditioned answering procedures are used across datasets unless otherwise stated.

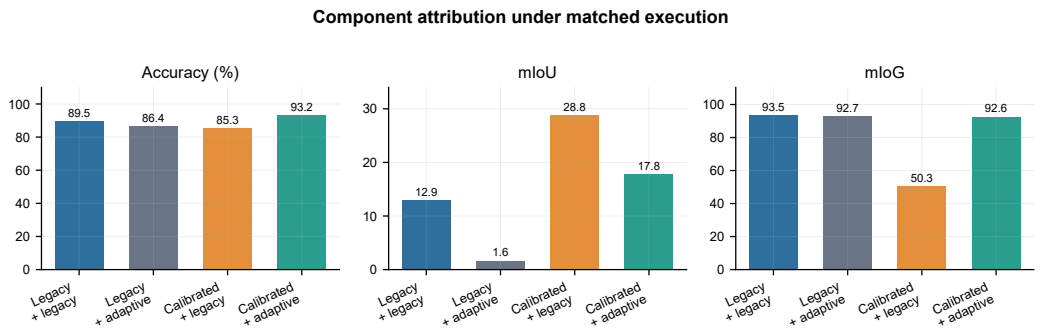


Figure 8: Component attribution under matched execution. The panels show accuracy, mIoU and mIoG for the four attention-map and region-formation pairings.

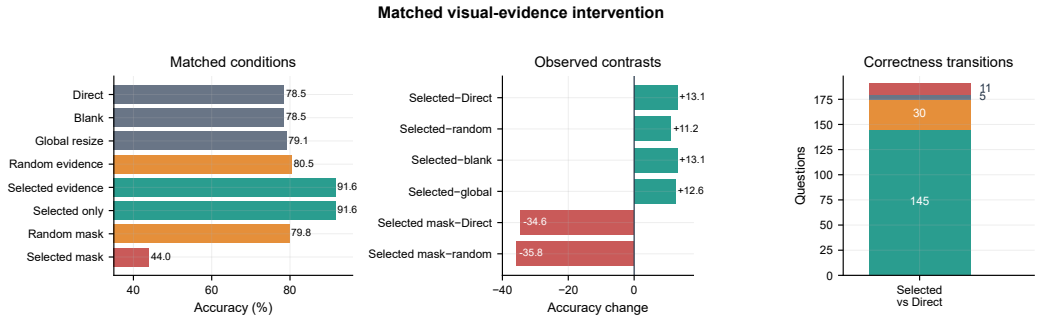


Figure 9: Matched visual-evidence intervention. The panels show condition accuracies, accuracy differences between matched conditions and correctness transitions.

E.2 Inference efficiency

We additionally report the average end-to-end inference time per question and peak GPU-memory usage on V*Bench. Table 19 compares the base model with representative training-free visual-search and thinking-with-images methods under our evaluation setup. All methods are measured under the same NVIDIA A800 hardware setup, and the reported latency is averaged over all benchmark questions.

EviSpec averages 6.43 s per question with a 22 GB peak memory footprint. Its memory use matches HiDe and remains close to the optimized ViCrop* and DeepEyes implementations. In latency, EviSpec incurs only a modest additional cost over HiDe and ViCrop*, while remaining substantially faster than DeepEyes. The original ViCrop implementation does not fit within 80 GB of GPU memory; ViCrop* therefore denotes our memory-optimized implementation, which reduces peak usage to 21 GB and enables the reported 5.98 s/question measurement.

F Supplementary Method and Proofs

This appendix expands the operators summarized in the main Method and proves the properties that follow from their construction. The results below are representational and algorithmic guarantees. They concern deterministic properties of the compiler and evidence operators; semantic relevance and answer accuracy are assessed by the matched experiments above.

The compiler and evidence-packing operators are shared by both localization wrappers evaluated above. For completeness, the equations below spell out the adaptive connected-region wrapper, which has the richer region operator; the fixed-threshold wrapper replaces only the thresholding and region-formation step. The request compiler, field separation, evidence assembly and final answer contract are otherwise identical. The formal compiler and packing guarantees therefore apply to both wrappers, while the expansion guarantee applies when adaptive expansion is enabled.

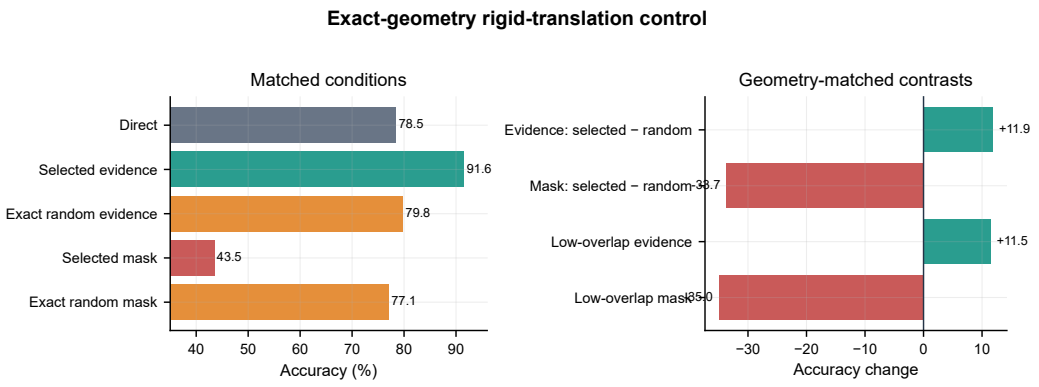


Figure 10: Exact-geometry rigid-translation control. The panels report condition accuracies and accuracy differences for the full set and the low-overlap subset.

Table 14: Question-perturbation results on V*Bench. All entries are percentages; bold marks the better value within each selector pair.

Selector	Query strategy	Evidence IoU	Answer change	Worst accuracy	All correct
Fixed-threshold	Single-C	82.2	1.7	89.7	87.4
Fixed-threshold	TCS	**84.0**	**0.9**	**91.6**	**90.1**
Adaptive	Single-C	81.1	2.2	86.9	83.8
Adaptive	TCS	**81.9**	**1.8**	**89.5**	**86.9**

F.1 Notation and Complete Operators

Let F_θ denote the frozen MLLM used by the main method. Superscripts comp, loc and ans identify calls with different fixed prompts but shared parameters θ. Let $Q = (q, \mathcal{O})$ denote a multiple-choice query, let $\bar{Q} = \mathrm{Serialize}(q, \mathcal{O})$ be its complete serialized input string, and let q_0 be the prefix of $\bar{Q}$ before the first option marker. Define $s^T = \mathrm{Role}(\bar{Q})$, $s^C = \mathrm{Can}(\bar{Q})$ and $s^S = \mathrm{Can}(q_0)$. Here Role and Can are calls to F_θ^{comp} with the fixed role-separated and canonical-form literals reproduced above; the deterministic parser retains only the required tagged field and maps malformed or empty fields to an empty request. The retained compiler sequence is

$$\mathbf{s}(\bar{Q}) = \mathrm{UniqueText}(\widetilde{\mathbf{s}}(\bar{Q})). \tag{11}$$

$\widetilde{\mathbf{s}}(\bar{Q}) = [s^T, s^C, s^S]$ is the ordered sequence before filtering. $\mathrm{UniqueText}$ removes empty requests and exact duplicates while preserving the first occurrence. Each retained request is represented by $s^r = (u_{r,1}, \ldots, u_{r,K_r}; m_r)$, where $u_{r,j}$ is a parsed entity field and m_r is an optional direct/relation tag. After tokenization, $\mathcal{T}_{r,j}$ contains the signal tokens assigned to field $u_{r,j}$.

Let Ω_I be the source-pixel domain and $\Omega_V = \{1, \ldots, H_V\} \times \{1, \ldots, W_V\}$ the original-image visual-token grid. For a scalar map $X : \Omega_V \to \mathbb{R}$, define

$$\mathrm{range}_{\Omega_V}(X) = \max_{p \in \Omega_V} X(p) - \min_{p \in \Omega_V} X(p), \tag{12}$$

and the sample-local normalization

$$\mathrm{N}(X) = \frac{X - \min_{\Omega_V} X}{\mathrm{range}_{\Omega_V}(X)}. \tag{13}$$

Equation equation 13 applies when $\mathrm{range}_{\Omega_V}(X) > 0$; for a constant map, we define $\mathrm{N}(X) \equiv 0$. Thus every non-constant map is normalized independently on its own visual grid, with per-question normalization for each dataset.

For request r, let $\mathcal{T}_{r,j}$ be the token group assigned to its jth parsed entity. The calibrated response of each signal token is computed independently. At the fixed localization layer $\ell^\star$, the all-head response is

$$A_{r,t} = \frac{1}{|\mathcal{H}_F|} \sum_{h \in \mathcal{H}_F} A_{r,t}^{(\ell^\star, h)}. \tag{14}$$

Table 15: Parser diagnostics over 1,348 original and perturbed records. Entries are percentages; bold marks the lowest value in each column.

Parser input	Empty output (%)	Complete option overlap (%)
T-full	4.4	2.5
T-stem	**0.3**	**1.8**
C	1.3	43.5
S	1.3	6.2

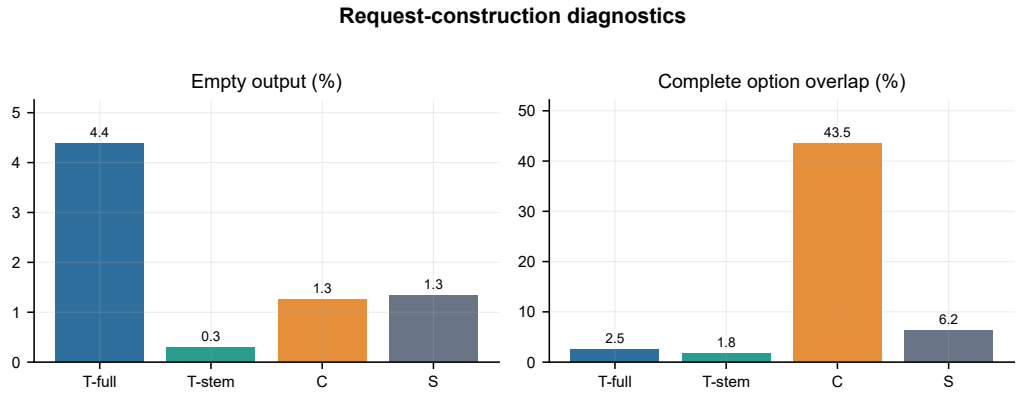


Figure 11: Request-construction diagnostics. The panels report empty-output rates and complete answer-option overlap for the four parser inputs.

Its normalized form is

$$\bar{A}_{r,t} = \mathrm{N}(A_{r,t}). \tag{15}$$

The instruction-prefix prior is

$$B_r(p) = \mathrm{median}_{u \in \mathcal{P}_r} \bar{A}_{r,u}(p). \tag{16}$$

The nonnegative residual is

$$D_{r,t}(p) = [\bar{A}_{r,t}(p) - B_r(p)]_+. \tag{17}$$

Here $\mathcal{P}_r$ contains only the fixed instruction-prefix tokens. The normalized residual is

$$R_{r,t} = \mathrm{N}(D_{r,t}). \tag{18}$$

When $D_{r,t} \equiv 0$, the implementation instead sets $R_{r,t} = \bar{A}_{r,t}$. This direct fallback uses the normalized token map. The implementation retains separate token maps through thresholding; $\mathcal{T}_{r,j}$ is used later to group the independently localized token boxes.

For each signal-token map, Yen's operator returns a threshold within the observed support:

$$\tau_{r,t} = \mathrm{Yen}(R_{r,t}). \tag{19}$$

The associated high-response mask is

$$\mathcal{M}^+_{r,t} = \{p \in \Omega_V : R_{r,t}(p) \geq \tau_{r,t}\}. \tag{20}$$

Its eight-connected components are

$$\mathcal{C}^+_{r,t} = \mathrm{CC}_8(\mathcal{M}^+_{r,t}). \tag{21}$$

The maximum-mass component supplies the seed:

$$C^*_{r,t} \in \arg\max_{C \in \mathcal{C}^+_{r,t}} \sum_{p \in C} R_{r,t}(p). \tag{22}$$

A fixed row-major tie rule makes the latter selection unique when several components have equal mass. Define the request-structural expansion flag

$$e_r = \begin{cases} 1, & K_r > 1, \\ 0, & K_r = 1. \end{cases} \tag{23}$$

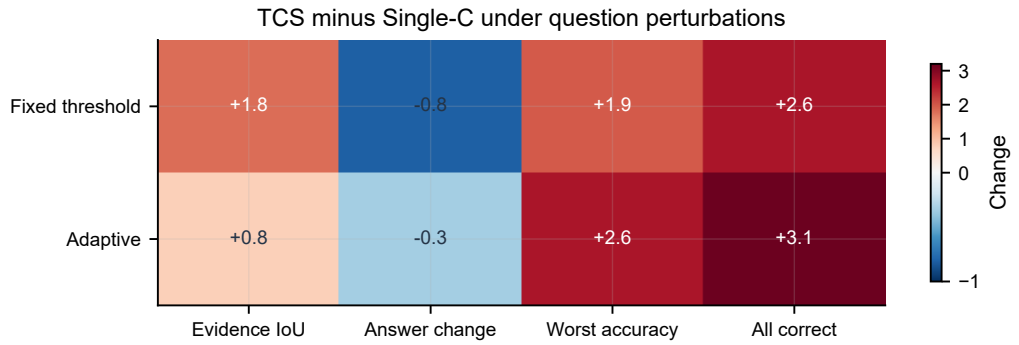


Figure 12: Changes from Single-C to TCS under question perturbations. Positive values indicate higher evidence stability or accuracy; a negative answer-change value indicates fewer changed answers.

Table 16: Matched strict option-free control on V*Bench. Accuracy values are percentages rounded to one decimal place; bold marks the best value in each column.

Method	Correct	Overall accuracy	Attribute	Spatial
Direct	150/191	78.5	80.0	76.3
Single-C	168/191	88.0	91.3	82.9
Single-T-stem	169/191	88.5	**93.0**	81.6
Single-S	168/191	88.0	91.3	82.9
Strict option-free TS	**171/191**	**89.5**	**93.0**	**84.2**
Full-input TCS	**171/191**	**89.5**	**93.0**	**84.2**

Under the compiler contracts above, relation-mode T requests contain separate subject and reference fields, while multi-entity C/S requests contain multiple comma-delimited entity fields. This rule matches the operational behavior of the original branch conditions, with parsed request structure determining expansion across the $T/C/S$ labels. If $e_r = 0$, the selected token box is $b_{r,t} = b(C^*_{r,t})$. If $e_r = 1$, let

$$V^-_{r,t} = \{R_{r,t}(p) : R_{r,t}(p) < \tau_{r,t}\}. \tag{24}$$

If $V^-_{r,t}$ is nonempty, define $\tau^-_{r,t} = \mathrm{Yen}(V^-_{r,t})$ and

$$\mathcal{M}^-_{r,t} = \{p \in \Omega_V : R_{r,t}(p) \geq \tau^-_{r,t}\}. \tag{25}$$

The expanded component is

$$C^-_{r,t} = \mathrm{CC}_8(\mathcal{M}^-_{r,t}; C^*_{r,t}). \tag{26}$$

The selected token box is $b_{r,t} = b(C^-_{r,t})$. The notation $\mathrm{CC}_8(S; C^*)$ means the component of S that contains seed C^*. The containment proof below applies whenever $V^-_{r,t}$ is nonempty. In the degenerate empty-support case, the implementation keeps the high-threshold seed.

Each selected token box is mapped from Ω_V to normalized coordinates on Ω_I and assigned to its parser field. Let $\mathcal{B}^0_{r,j}$ be the boxes assigned to field j, and let $h_{r,j} = b(\bigcup_{\beta \in \mathcal{B}^0_{r,j}} \beta)$ be their enclosing rectangle. Field j is fused only when $h_{r,j}$ is disjoint from every box assigned to another field; otherwise its boxes remain separate. All guards are evaluated against the pre-fusion boxes, so the output is independent of field processing order. The final sequence $\mathcal{B}_r$ concatenates fields in parser order and applies the fixed row-major tie rule within each field.

The evidence view is

$$E_r = \mathrm{Assemble}(\mathrm{sort}_{\text{row-major}}[\Pi(I, b) : b \in \mathcal{B}_r]), \tag{27}$$

where $\Pi(I, b)$ copies source pixels inside b without resizing. Assemble removes only horizontal and vertical source intervals that intersect no retained box, using monotone source-to-canvas coordinate maps. Overlapping source pixels are written once; source order and shared geometry are therefore preserved while empty gaps become transparent. Empty request branches are omitted. UniqueView

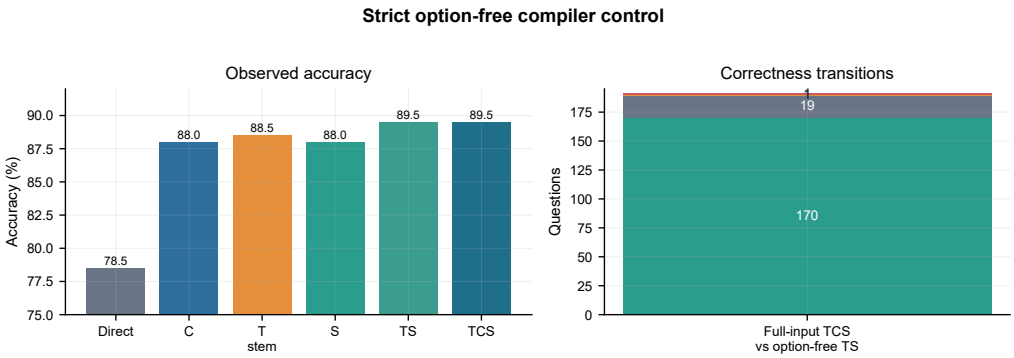


Figure 13: Strict option-free compiler control. The left panel reports observed accuracy and the right panel shows correctness transitions between full-input TCS and option-free TS.

Table 17: Number of distinct, non-empty compiled queries on 191 unperturbed V*Bench questions. The final three columns are percentages within each row.

Question group	N	One query	Two queries	Three queries
All	191	20.9	64.9	14.1
Attributes	115	1.7	76.5	21.7
Spatial	76	50.0	47.4	2.6

removes a view only when its shape, alpha support and visible pixel values exactly equal an earlier view, retaining the first occurrence in T, C, S order. Finally, $\mathrm{InputPack}$ places the unmodified source image first, appends the surviving evidence views in that order and supplies the unchanged question to the answer call:

$$\mathbf{E} = \mathrm{UniqueView}([E_r : r \in (T, C, S),\ E_r \text{ defined}]). \tag{28}$$

$$\hat{a} = F_\theta^{\mathrm{ans}}\big(\mathrm{InputPack}([I] \| \mathbf{E}), Q\big). \tag{29}$$

F.2 Complete Adaptive-Wrapper Inference Procedure

For clarity, the full inference procedure is listed in execution order.

1. Represent the multiple-choice query as $Q = (q, \mathcal{O})$, serialize it as $\bar{Q} = \mathrm{Serialize}(q, \mathcal{O})$, and split $\bar{Q}$ at the first standard option marker to obtain the option-free prefix q_0.
2. Apply the role-separated and canonical-form output instructions to $\bar{Q}$, and apply the same canonical-form instruction to q_0, using the three constructions defined in the main paper; remove exact textual duplicates.
3. For each remaining specification, insert it into the shared grounding prompt and run the frozen MLLM while recording the prescribed text-to-vision attention maps.
4. Aggregate and normalize the maps, estimate the median prefix prior, and compute each calibrated signal-token map in Eqs. equation 14–equation 18.
5. Apply the sample-local Yen threshold to each token map and select its maximum-mass connected seed. Apply the lower-threshold connected expansion according to Eq. equation 23 whenever the parsed request contains more than one entity field ($K_r > 1$).
6. Map the selected grid boxes back to the source image, apply the field-guarded fusion rule, and assemble one evidence view per request construction.
7. Remove duplicate evidence views, append the remaining views to the original image, restore the unchanged multiple-choice question, and decode the final option with the frozen MLLM.

The procedure uses frozen model weights and fixed compiler, localization and answering operators.

Table 18: Observed fusion comparison on 191 unperturbed V*Bench questions. All entries are percentages; bold marks the best value in each metric within each selector block.

Selector	Fusion strategy	Acc.	Attr	Spatial	mIoU	mIoG	Cov.@0.5
Fixed-threshold	Single-T	90.6	93.9	85.5	11.1	83.6	84.1
Fixed-threshold	Single-C	92.2	95.7	86.8	11.2	85.4	86.1
Fixed-threshold	Single-S	92.2	95.7	86.8	11.2	84.6	85.3
Fixed-threshold	Naive union	89.5	93.0	84.2	10.7	87.0	87.8
Fixed-threshold	Attention average	91.1	95.7	84.2	12.0	86.1	87.4
Fixed-threshold	Confidence selection	93.2	97.4	85.5	11.4	85.7	86.5
Fixed-threshold	TCS concat	**93.7**	**98.3**	**88.2**	**12.5**	**91.5**	**92.2**
Adaptive	Single-T	88.0	92.2	81.6	14.4	80.5	82.9
Adaptive	Single-C	88.0	91.3	82.9	10.9	84.7	85.7
Adaptive	Single-S	88.0	91.3	82.9	14.0	82.6	84.5
Adaptive	Naive union	88.0	91.3	82.9	10.7	86.4	87.4
Adaptive	Attention average	85.3	91.3	76.3	7.0	65.4	66.5
Adaptive	Confidence selection	**89.5**	93.0	81.6	12.8	84.7	86.5
Adaptive	TCS concat	**89.5**	**94.8**	**84.2**	**15.8**	**90.9**	**91.8**

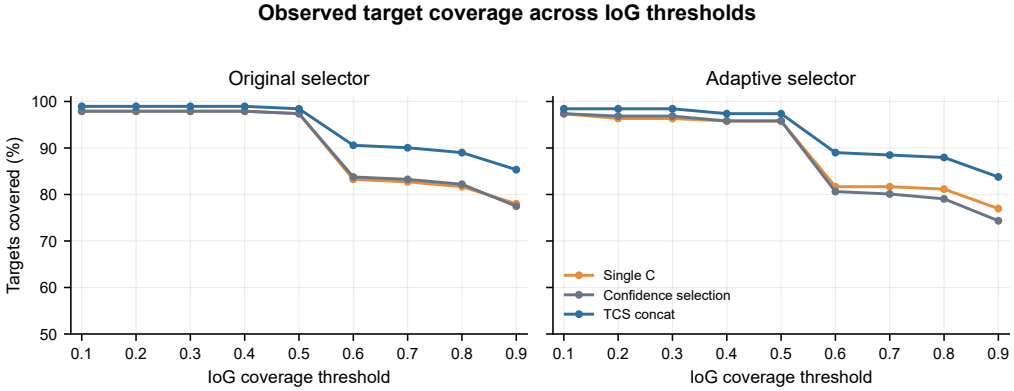


Figure 14: Observed target coverage across matched IoG thresholds.

F.3 Question-Compilation Guarantees

Proposition 1 (option-order invariance of the isolated branch). Let $Q = (q, \mathcal{O})$ and $Q' = (q, \pi(\mathcal{O}))$ differ only by a permutation π of text after the first option marker. Then

$$s^S(Q) = s^S(Q'), \qquad G_S(I, Q; F) = G_S(I, Q'; F), \tag{30}$$

provided that tokenization, frozen-model inference and the region operators are run with the same state and tie rules.

Proof. Write $P(Q)$ for the prefix of the serialized question ending immediately before the first option marker. By assumption, Q and Q' have the same text before this marker, and the permutation acts only on the suffix. Hence $P(Q) = P(Q')$. The option-isolated compiler is the composition $\mathrm{Can} \circ P$, so

$$s^S(Q) = \mathrm{Can}(P(Q)) = \mathrm{Can}(P(Q')) = s^S(Q').$$

The resulting grounding prompt and token sequence are identical. With the same image and frozen model, the recorded attention maps are identical. Equations equation 14–equation 27 therefore produce the same regions and evidence view for the option-isolated branch. The final answer stage receives the complete option sequence. □

Proposition 2 (separate grounding of successfully extracted fields). Suppose the role-separated full-query parser returns an ordered, comma-delimited list $s^T = [u_1, \ldots, u_M]$. Before evidence-view fusion, EviSpec assigns every u_i a distinct token group $\mathcal{T}_{T,i}$ and constructs its token maps independently by Eq. equation 18.



Figure 15: Qualitative comparison of evidence localization on V*Bench. Each row shows the original image, the selected regions and the assembled evidence view for representative attribute and relation questions.

Proof. The comma delimiter defines M disjoint fields. Tokenization assigns every token position to one field, so for $i \neq j$, $\mathcal{T}_{T,i} \cap \mathcal{T}_{T,j} = \varnothing$. Equation equation 18 constructs every signal-token map independently. Field interactions occur later during guarded box fusion and view construction. □

F.4 Calibration Guarantees

Proposition 3 (bounded calibrated response). For every specification r, token t and pixel p,

$$0 \leq \bar{A}_{r,t}(p) \leq 1, \qquad 0 \leq R_{r,t}(p) \leq 1. \tag{31}$$

Proof. If $A_{r,t}$ is non-constant, Eq. equation 13 subtracts its minimum and divides by its positive range. The minimum is mapped to zero, the maximum to one, and every intermediate value lies between them. If the map is constant, the definition returns zero, so $\bar{A}_{r,t} \in [0, 1]$ in both cases.

Table 19: Average inference efficiency on V*Bench. Time is reported in seconds per question and memory denotes peak GPU-memory usage. ViCrop* is our memory-optimized implementation of ViCrop. The original ViCrop implementation runs out of memory even on an 80 GB GPU under this setup.

Method	Avg. time (s/question)	Peak GPU memory
Base	2.20	–
HiDe	5.36	22 GB
DeepEyes	12.56	21 GB
ViCrop	–	OOM on 80 GB
ViCrop*	5.98	21 GB
EviSpec	6.43	22 GB

The positive-part operator makes $D_{r,t}$ nonnegative. When its range is positive, the same normalization argument gives $\mathrm{N}(D_{r,t}) \in [0,1]$. A nonzero constant residual is mapped to zero by the constant-map rule. When $D_{r,t} \equiv 0$, the fallback defined after Eq. equation 18 uses $\bar{A}_{r,t}$, which is already bounded. Thus $R_{r,t} \in [0,1]$. □

Proposition 4 (median robustness to minority prefix contamination). Fix a pixel p and suppose the prefix contains m responses. If fewer than $m/2$ responses are changed arbitrarily while every unchanged response lies in $[a,b]$, then the pointwise median prior remains in $[a,b]$.

Proof. Let $k < m/2$ responses be arbitrary and let the remaining $m-k > m/2$ responses lie in $[a,b]$. More than half of all values are at least a and at most b, so their median lies in $[a,b]$. Therefore $a \le B_r(p) \le b$. □

Proposition 5 (cancellation of a shared prompt response). Assume that, after token-map normalization, a spatial response $\delta : \Omega \to \mathbb{R}$ is added to every entity-token map and every prefix-token map:

$$\bar{A}'_{r,t} = \bar{A}_{r,t} + \delta, \qquad \bar{A}'_{r,u} = \bar{A}_{r,u} + \delta \quad \text{for all } u \in \mathcal{P}_r.$$

Then the pre-normalized calibrated residual is unchanged:

$$[\bar{A}'_{r,t} - \mathrm{median}_{u\in\mathcal{P}_r}\, \bar{A}'_{r,u}]_+ = [\bar{A}_{r,t} - \mathrm{median}_{u\in\mathcal{P}_r}\, \bar{A}_{r,u}]_+. \tag{32}$$

Proof. At each pixel p, adding the same scalar $\delta(p)$ to every member of a finite set shifts every ordered value by $\delta(p)$. In particular,

$$\mathrm{median}_u(\bar{A}_{r,u}(p) + \delta(p)) = \mathrm{median}_u\, \bar{A}_{r,u}(p) + \delta(p).$$

Substituting this identity gives

$$\bar{A}_{r,t}(p) + \delta(p) - \mathrm{median}_u\, \bar{A}_{r,u}(p) - \delta(p) = \bar{A}_{r,t}(p) - \mathrm{median}_u\, \bar{A}_{r,u}(p).$$

The two arguments of the positive-part operator are identical at every pixel, which proves Eq. equation 32. Subsequent normalization is applied to the same residual map. □

F.5 Region-Formation Guarantees

Proposition 6 (mass-optimal seed within the threshold mask). For a fixed signal-token map and threshold mask, the selected component satisfies

$$\sum_{p\in C^*_{r,t}} R_{r,t}(p) \ge \sum_{p\in C} R_{r,t}(p), \qquad \forall C \in \mathcal{C}^+_{r,t}. \tag{33}$$

Proof. The visual grid contains finitely many eight-connected components. Equation equation 22 selects $C^*_{r,t}$ by maximizing the displayed mass function. A fixed row-major rule resolves ties while preserving the inequality. □

Proposition 7 (enabled expansion contains the seed). Assume $V^-_{r,t}$ in Eq. equation 24 is nonempty and the threshold returned by Yen's operator lies within its input support. Then

$$C^*_{r,t} \subseteq C^-_{r,t}, \qquad b(C^*_{r,t}) \subseteq b(C^-_{r,t}). \tag{34}$$

Proof. Every value in $V^-_{r,t}$ is strictly smaller than $\tau_{r,t}$. A threshold chosen within that support therefore satisfies $\tau^-_{r,t} < \tau_{r,t}$. Every seed pixel belongs to the high mask, so for $p \in C^*_{r,t}$,

$$R_{r,t}(p) \geq \tau_{r,t} > \tau^-_{r,t}.$$

Hence every seed pixel also belongs to $\mathcal{M}^-_{r,t}$. The operator $\mathrm{CC}_8(\mathcal{M}^-_{r,t}; C^*_{r,t})$ selects the lower-mask component containing those pixels, which proves $C^*_{r,t} \subseteq C^-_{r,t}$. The axis-aligned bounding rectangle of a set cannot shrink when the set is enlarged: each coordinate minimum can only decrease or remain unchanged, and each maximum can only increase or remain unchanged. Therefore the bounding-box inclusion follows. □

Proposition 8 (parser-field separation invariant of guarded fusion). For every field j accepted for fusion, its enclosing rectangle is disjoint from every pre-fusion box assigned to a field $k \neq j$.

Proof. Let $h_{r,j}$ be the enclosing rectangle of the boxes in field j. By the definition of guarded fusion, $\mathcal{B}^0_{r,j}$ is replaced by $\{h_{r,j}\}$ only if $h_{r,j} \cap b = \varnothing$ for every $b \in \mathcal{B}^0_{r,k}$ and every $k \neq j$. Thus every accepted fused rectangle satisfies the stated separation property directly. Because all guards use the same pre-fusion box sets, the decision is independent of the order in which fields are inspected. □

F.6 Evidence-View Guarantees

Proposition 9 (source support of visible evidence and order preservation). Let $\mathrm{supp}_\alpha(E_r)$ be the pixels with nonzero alpha in the packed evidence view. Every visible evidence pixel is copied from a selected source-image box:

$$\forall p \in \mathrm{supp}_\alpha(E_r), \quad \exists b \in \mathcal{B}_r,\ z \in b \text{ such that } E_r(p) = I(z). \tag{35}$$

If two non-overlapping boxes satisfy $b_1 \prec_{\mathrm{rm}} b_2$, their extracted contents appear in that order in E_r.

Proof. By Eq. equation 27, Assemble receives the crops $\Pi(I, b)$ for $b \in \mathcal{B}_r$ and copies each retained source region to its compressed coordinate interval. Every visible output pixel therefore equals a source pixel inside a selected box, which proves Eq. equation 35.

For order preservation, the assembly operator sorts the source x- and y-boundaries and maps them to compressed coordinates with monotone coordinate maps. The row and column precedence defined by $\prec_{\mathrm{rm}}$ is preserved. The complete input in Eq. equation 29 also retains the original image for global context. □

F.7 Computational Complexity

Let $R = |\mathbf{s}(\bar{Q})|$, let $T = \sum_{r,j} |\mathcal{T}_{r,j}|$ be the total number of grounding tokens, let $N = |\Omega_V| = H_V W_V$, and let $H_a = |\mathcal{H}_F|$ be the number of aggregated attention heads at the fixed localization layer. Attention aggregation and calibration require

$$\mathcal{O}(TH_aN) \quad \text{time and} \quad \mathcal{O}(N + H_aN) \quad \text{working memory when token maps are streamed.} \tag{36}$$

Thresholding and connected-component extraction are linear in N for each token map. Sorting $|\mathcal{B}_r|$ box boundaries for view construction costs $\mathcal{O}(|\mathcal{B}_r| \log |\mathcal{B}_r|)$. The complete evidence stage is therefore bounded by

$$\mathcal{O}\left(TH_aN + TN + \sum_{r=1}^{R} |\mathcal{B}_r| \log |\mathcal{B}_r|\right), \tag{37}$$

in addition to the frozen-model forward passes and final generation.

The propositions establish deterministic properties of question compilation, prefix calibration, region expansion, guarded fusion and evidence-view assembly. The matched prompt-family comparisons evaluate request construction, and the exact-geometry experiment evaluates the contribution of selected visual content.