THINKING ONCE IS ENOUGH: INTERMEDIATE-LAYER EVIDENCE ROUTING FOR HIGH-RESOLUTION VQA

Anonymous authors
Paper under double-blind review

ABSTRACT

High-resolution visual question answering (HR-VQA) is often treated as a problem of insufficient evidence acquisition, where failing multimodal large language models must inspect images again through cropping, re-encoding, or multi-round search. We show that this view is incomplete: in many cases, fine-grained evidence has already survived visual encoding and become identifiable and influential within an intermediate-layer routing window, but is later diluted before answer generation. We propose Thinking-Once, a **training-free, single-visual-pass** evidence-routing method that reconstructs question-conditioned attention at this window, preserves core entity tokens and compact background context, and routes this evidence to later layers without extra visual encoding. Across five base models, Thinking-Once consistently improves or matches the corresponding base setting, increasing the average scores on V*Bench, HRBench-4K, and HRBench-8K by *+3.1*, *+3.0*, and *+2.7* points while reducing the average peak memory by about 4 GB. On Qwen2.5-VL-7B, it improves the three benchmarks by *+9.9*, *+4.6*, and *+5.5* points, raising the cross-benchmark mean from 72.5 to 79.1. With the ZwZ-8B base model, Thinking-Once reaches a mean score of 82.7. Against 11 open-source HR-VQA baselines, it obtains the best or tied-best score on all three benchmark averages and the best overall mean; for example, compared with DeepScan, it reduces V*Bench inference time by **97.2%** while improving the cross-benchmark mean from 77.8 to 79.1. These results show that HR-VQA can be improved by routing already encoded evidence rather than repeatedly acquiring new visual inputs. Code is available in the appendix.

1 INTRODUCTION

Multimodal large language models (MLLMs) have achieved strong performance on standard visual understanding benchmarks (Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2023; Dai et al., 2023; Zhu et al., 2024; Wang et al., 2024; Liu et al., 2025a; Bai et al., 2025b;a; 2023), yet they continue to struggle with high-resolution and fine-grained tasks where the model must recognize small objects, read localized text, or deduce spatial relationships. The core difficulty lies in the need to precisely localize subtle, question-relevant visual evidence amid overwhelming background interference and then leverage these localized cues for subsequent perception and reasoning. When high-resolution images are encoded directly, the resulting large volume of visual tokens dilutes attention to question-relevant evidence, causing fine-grained details to be overshadowed by redundant global context.

Strategies follow two directions. *External evidence reacquisition* obtains extra evidence through multi-turn search, crop/zoom-in, region re-encoding, or training-time zooming distillation (Bolya et al., 2023; Khayatkhoei et al., 2025; Zhang et al., 2025b; Shen et al., 2025; Zheng et al., 2026; Liu et al., 2026b; Li et al., 2026a; Wei et al., 2026; Jiang et al., 2026; Li et al., 2026b). They recover missing details but do so by changing the visual input pipeline, leading to substantial inference cost. In contrast, *compression-first token reduction* methods improve efficiency by pruning or merging visual tokens (Chen et al., 2024; Yang et al., 2025; Cai et al., 2025; Huang et al., 2025; Shang et al., 2025; Li et al., 2025b; Chen et al., 2026a; Liu et al., 2026a; Han et al., 2026; Zhang et al., 2026a; 2025d; Chen et al., 2025). Their generic redundancy criteria can often remove entity or contextual tokens that remain necessary for relation-sensitive HR-VQA. These limitations motivate a different perspective: rather than relying solely on additional observation or token reduction, we more effec-

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

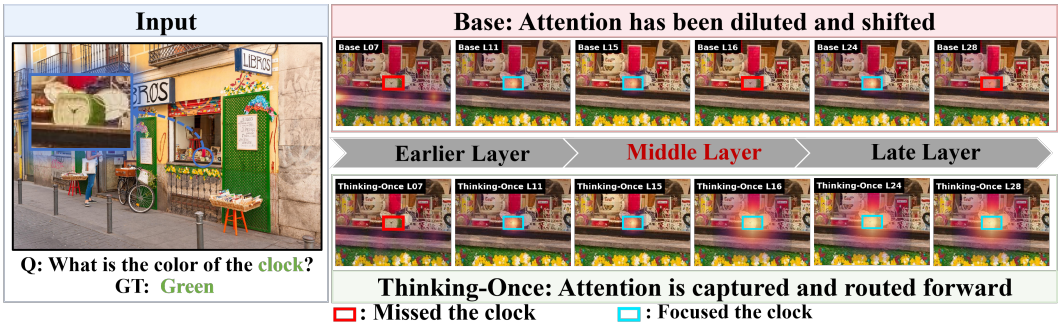


Figure 1: The base model loses focus on the question-relevant region, whereas Thinking-Once captures the relevant evidence at an intermediate layer and preserves it throughout subsequent reasoning.

tively route the critical question-relevant evidence that has already emerged in intermediate layers within a visual pass.

Our observations in the *Observations and Analyses* section collectively support a different view. As shown in Fig. 1, MLLMs can already activate question-relevant regions at intermediate layers, but this evidence may become diluted or displaced during subsequent reasoning. Meanwhile, visual embeddings produced by the vision encoder can retain fine-grained information from the original image (Team, 2026). These observations suggest that HR-VQA failures are not always caused by missing visual evidence; they can potentially also arise from the ineffective preservation and transmission of evidence that has been encoded and localized. We therefore shift the focus from repeatedly acquiring new visual inputs or indiscriminately compressing tokens to routing question-relevant evidence through later layers, leading to the *Thinking-Once evidence-routing* paradigm.

Thinking-Once emphasizes evidence filtering and delivery within a single visual processing flow, relying on one full-image visual encoding and one multimodal forward pass. We extract visual tokens from an intermediate layer and use cross-modal attention distributions at this stage to isolate core entity tokens highly aligned with the current question, while retaining contextual background tokens supporting relation inference, attribute discrimination, and scene parsing, and route the filtered evidence set toward deeper network layers. Through this design, we reformulate visual information selection in HR-VQA as *query-conditioned reasoning-evidence routing*, which explicitly preserves the integrity of visual evidence for downstream reasoning within a single-visual-pass framework.

The main contributions of this paper are as follows:

- **Insightful Observations:** We identify the ineffective transmission of question-relevant visual evidence that has already emerged in intermediate layers as a vital and underexplored source of HR-VQA failure, complementing the conventional evidence-acquisition view.
- **Streamlined Paradigm:** We propose the Thinking-Once evidence routing mechanism, which enables high-resolution reasoning with one full-image encoding and one forward pass by routing intermediate-layer visual evidence without constructing extra visual inputs.
- **Systematic Method and Validation:** We design a query-conditioned evidence routing mechanism to transmit core and background tokens. Experiments and visualizations demonstrate that our approach significantly outperforms representative methods from existing paradigms within a simpler single-visual-pass reasoning process.

2 Related Work

2.1 High-Resolution VQA and Existing Paradigms

High-resolution visual question answering (HR-VQA) evaluates whether MLLMs can perceive and reason over fine-grained evidence in high-resolution scenes, including small objects, localized text, detailed attributes, and spatial relations. Benchmarks such as V*Bench (Wu & Xie, 2024) and HRBench (Wang et al., 2025) show that existing MLLMs struggle to capture question-relevant local evidence when high-resolution inputs produce long visual token sequences. Existing HR-VQA

methods enhance local perception through additional visual operations. *Visual-search* methods locate relevant regions with tool use or multi-step observation (Zhang et al., 2025b; Li et al., 2025a); *crop/zoom-in* methods crop or re-encode selected sub-images for local details (Khayatkhoei et al., 2025; Shen et al., 2025; Zheng et al., 2026; Liu et al., 2026b; Li et al., 2026a; Jiang et al., 2026; Li et al., 2026b); and *ZwZ* internalizes the ability to zoom in through training-time distillation (Wei et al., 2026). Although effective, these methods require extra computation, additional visual inputs, or complex training, and overlook evidence formed in intermediate layers. In contrast, we directly route such internal evidence with one full-image visual encoding and one multimodal forward pass.

2.2 Token Reduction and Evidence Routing

Efficiency-oriented methods reduce the cost of high-resolution MLLMs by pruning or merging visual tokens according to attention scores, token similarity, or saliency (Bolya et al., 2023; Cao et al., 2023; Liu et al., 2026a; Yang et al., 2025; Chen et al., 2024; Shang et al., 2025; Ye et al., 2025; Alvar et al., 2025; Zhang et al., 2025a; Endo et al., 2025; Li et al., 2025b; Han et al., 2026; Zhang et al., 2026a; Wen et al., 2026). Although effective, these methods identify redundant tokens for removal, and their generic selection criteria discard background or contextual tokens needed for attribute recognition, spatial reasoning, and relation understanding. Thinking-Once is related to this line of work at the implementation level because it shortens later-layer sequences, but differs in objective and selection principle: instead of optimizing token reduction itself, it constructs a question-conditioned evidence set through independent minimum-coverage selection for each routing query and preserves structured contextual support with background summaries. Thus, compression emerges as a consequence of evidence routing, rather than being the sole design objective.

3 Observations and Analysis

HR-VQA is treated as an evidence-acquisition problem: when a model fails, it should see the image again at a finer scale. This explanation is incomplete if the required evidence has already entered the model but is later diluted, misallocated, or discarded. So we diagnose HR-VQA with three questions: *Is the evidence available after visual encoding? Where and when does it become useful? What information must remain around the evidence core?* The answers distinguish evidence availability from evidence utilization and provide direct design constraints for Thinking-Once.

3.1 Encoded Visual Representations Retain Fine-Grained Evidence

Single-visual-pass routing is viable if the initial encoding retains information needed for later reasoning. LongCat-Next (Team, 2026) shows that lightweight decoders can reconstruct image structure, object boundaries, and spatial layout from frozen vision-encoder outputs, which indicates that visual embeddings retain recoverable fine-grained structure. Reconstructability establishes a narrower but critical condition, namely that local information can survive visual encoding and remain available to subsequent layers. Complementary evidence from interleaved visual reasoning, vision tool-use training, and recent analyses of RL-induced reasoning changes (Yang et al., 2026; Ma et al., 2026) points to the same, and we summarize it in Appendix B.

These observations identify an underexplored failure mode, consistent with recent evidence that large models may internally encode relevant evidence even when it is not surfaced in the final prediction (Liu et al., 2025b; 2026c). In a high-resolution image, a few relevant tokens compete with a larger set of unrelated tokens throughout language-side reasoning, so that an encoded target can be diluted before answer generation. The resulting question is therefore whether the model can allocate sufficient downstream computation to evidence that is already available, in addition to whether it can acquire that evidence. Thinking-Once targets this utilization-limited regime, and it does not claim to replace visual tools when the original encoding genuinely lacks the required information.

3.2 Intermediate Layers Define an Evidence-Routing Window

We localize the evidence within the visual sequence using three layer-wise oracle interventions. *Random Tokens* retain a random set with the same budget as the ground-truth-box tokens; *Oracle Core Tokens* retain only tokens inside the box; and *Remove Oracle Tokens* retain all tokens except those inside the box. Applying these interventions at transformer layers separates factors otherwise mixed: token budget, token identity, and intervention depth. This design allows us to ask whether the identity of the retained evidence influences the answer, which goes beyond asking whether fewer tokens can work. Figure 2 reveals an asymmetric evidence distribution. Under the same budget, Oracle

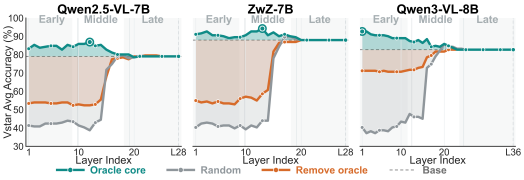


Figure 2: Layer-wise oracle token intervention on V*Bench. Oracle tokens outperform random tokens across most layers, while removing them degrades performance, especially in early and intermediate layers.

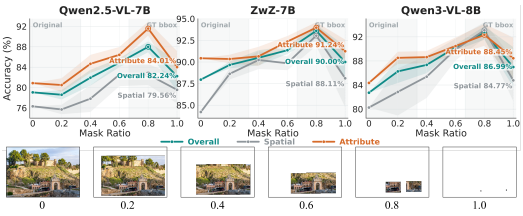


Figure 3: Effect of progressive background masking on HR-VQA performance. We gradually vary the background mask ratio from the original full image to the GT-bbox-only input.

Core Tokens outperform Random Tokens and can even exceed the base model, whereas removing the oracle tokens cannot be compensated for by retaining the much larger available background set. Entity tokens are therefore consistently information-dense and hard to replace: HR-VQA depends substantially more on preserving the right tokens than simply preserving many tokens.

The two oracle comparisons provide complementary evidence. The advantage over Random Tokens supports ***sufficiency***, since a compact entity-centered subset can support strong reasoning when it contains the relevant evidence. The failure of Remove Oracle Tokens supports ***necessity***, since abundant non-target information cannot reconstruct evidence removed from the entity region. Because Random Tokens use the same budget as Oracle Core Tokens and Remove Oracle Tokens keep more tokens overall, the result cannot be explained by token count alone. What matters is the identity of the retained tokens and the precise layer at which their influence is most effectively redirected. This necessity signal has an input-space counterpart: masking only the pixels inside the ground-truth boxes, while keeping the rest of the image fully intact, lowers full-image accuracy from 150/191 to 80/191 on V*Bench, and accuracy recovers as the target-region are restored (Appendix F.4).

The intervention effect is also layer-dependent, which lets us locate where routing should happen. Its advantage is strongest when question–entity correspondence has emerged while it can still influence the remaining computation, and interventions at deeper layers converge toward the base model as evidence becomes absorbed into later hidden states. This more clearly defines an *evidence-routing window*. At early layers, visual detail is available while question-conditioned correspondence may still be weak. At late layers, the model has formed richer multimodal states, while too little computation remains for an intervention to redirect reasoning. Intermediate layers provide the useful overlap, where evidence has become identifiable and sufficiently influential under intervention, which is the operating regime required by single-visual-pass routing (Neo et al., 2025; Jiang et al., 2025; Fan et al., 2025; Song et al., 2026; Kan et al., 2026). The earliest oracle optimum should not be read as the practical routing layer, because ground-truth boxes provide perfect localization that is unavailable at inference time. A usable routing layer must jointly provide reliable question-conditioned localization and sufficient remaining depth for the routed evidence to affect the answer.

3.3 Core Evidence Requires Contextual Support

Entity tokens form the evidence core, but retaining this core is insufficient. Figure 3 shows a non-monotonic effect of background removal: moderate masking improves accuracy, whereas aggressive masking degrades it. This indicates that background tokens contain both interference and supporting context, including spatial references, relational cues, and attribute-comparison evidence. The optimal representation is selective rather than minimal. This also explains why object-only cropping can be brittle: many HR-VQA questions refer to an entity but ask about relations, relative positions, or comparisons whose evidence lies outside the entity box. Removing all surrounding tokens preserves the referent but destroys part of the structure needed to interpret it.

The attention-sink intervention in Figure 4 further shows that localization and utility are not equivalent. Although such tokens may distort attention-based cropping (Liu et al., 2026b), suppressing them reduces accuracy across MLLMs and benchmarks. This suggests that spatially non-discriminative tokens can carry contextual, aggregation, or information-transfer signals, consistent with prior findings on attention-sink and register-like tokens in Transformers (Xiao et al., 2024; Darcet et al., 2024; Kang et al., 2025). Token utility should be judged by its contribution to later reasoning, not only by whether it maps onto a visible object.

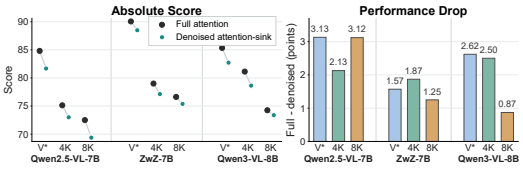


Figure 4: Effect of attention-sink suppression on HR-VQA performance. The consistent degradation across models and benchmarks suggests that these tokens contribute useful contextual or aggregation signals during downstream reasoning.

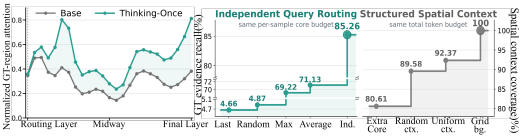


Figure 5: Left: GT-evidence recall across query-routing strategies under the same per-sample core-token budget. Right: spatial-context coverage across variants under the same total-token budget. Ind. denotes Independent.

3.4 From Analysis to Design Principles

The evidence-allocation view clarifies the applicability boundary of Thinking-Once. Routing cannot recover information that the vision encoder has irreversibly removed, nor can it solve cases where the relevant region is never represented in the initial encoding. It is therefore complementary to cropping, zooming, or visual search, which remain useful when new visual observations are needed. Thinking-Once instead targets the utilization-limited regime, where the evidence is encoded but not effectively preserved or allocated. We provide a more detailed discussion of these boundaries and their testable predictions in Appendices B and F.

The above analyses are translated into four design requirements. Routing should operate on internal representations, be conditioned on the question, occur at intermediate layers where evidence is identifiable and actionable, and preserve compact contextual support rather than aggressively pruning every non-entity token. Figure 5 further validates the two key design choices of Thinking-Once under matched token budgets. Under the same per-sample core-token budget, independent query routing achieves the highest GT-evidence recall, indicating that preserving evidence separately for different routing queries avoids the information suppression caused by query aggregation. Under the same total-token budget, structured grid background summaries provide the highest spatial-context coverage, showing that the spatial organization of contextual tokens is more effective than simply adding core tokens or allocating context randomly or uniformly. These results support independent evidence selection and structured background preservation as complementary components of the proposed routing mechanism. Together, these requirements define Thinking-Once: a single-visual-pass mechanism that preserves the structure of already encoded, question-aligned evidence and routes later-layer computation toward it.

4 Methodology

As illustrated in Figure 6, Thinking-Once performs question-guided evidence routing at inference time with one full-image visual encoding and one multimodal forward pass. More specifically, given a high-resolution image and a question, an MLLM with D transformer layers first runs the full multimodal sequence to an intermediate routing layer $L < D$:

$$H^L = \mathrm{SeqMerge}(V^L; T^L), \qquad V^L = \{v_i^L\}_{i=1}^N, \tag{1}$$

where V^L and T^L denote the visual and text hidden states, and $\mathrm{SeqMerge}$ preserves their prompt order. At layer L, Thinking-Once replaces the visual subsequence with a compact evidence representation E^L while keeping the text sequence unchanged. The image is encoded once, with no crop, sub-image re-encoding, or additional visual observation. Derivations are provided in Appendix A.

4.1 Question-Conditioned Evidence Distribution

Different question parts may require diverse visual evidence. We therefore form a routing-query set $\mathcal{Q}_r = \mathcal{Q}_{\mathrm{ent}} \cup \{q_{\mathrm{glb}}\}$, where $\mathcal{Q}_{\mathrm{ent}}$ contains anchors of entity spans explicitly mentioned in the question, and q_{glb} is a global question query instantiated by the last non-punctuation question token. Under causal self-attention, this token provides a compact representation of the preceding question context. Entity queries localize explicit objects and regions, whereas the global question query captures complementary attribute, relation, and broader contextual requirements. By default, the anchor of an entity span is its last non-punctuation token. For each $q_j \in \mathcal{Q}_r$, we reconstruct its

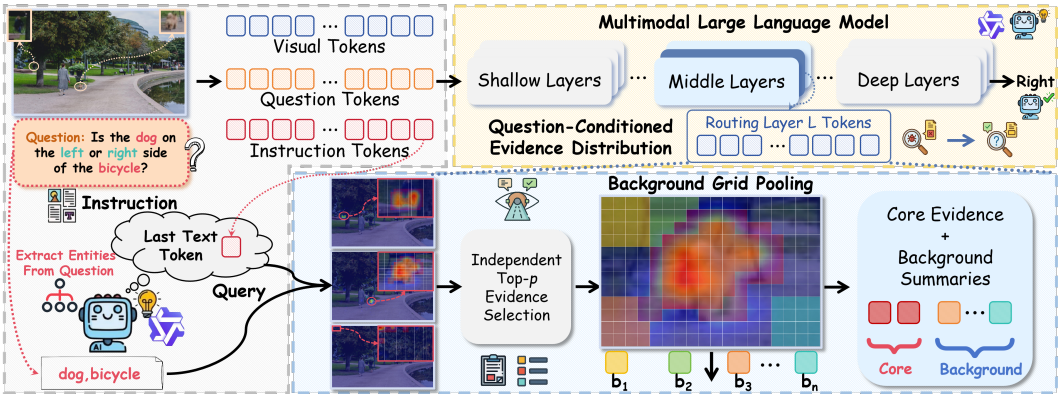


Figure 6: Overview of Thinking-Once. At a routing layer, question-conditioned queries independently select core evidence, while background grid pooling summarizes visual context before routing both to deeper layers within a single visual pass.

visual attention row from the MRoPE-transformed Su (2021; 2025) query and key states at layer L, and average the resulting probabilities across the H attention heads:

$$a^L_{j,i} = \frac{1}{H}\sum_{h=1}^{H} A^{L,h}_{q_j,i}, \qquad a^L_{j,i} \geq 0. \tag{2}$$

The forward pass uses FlashAttention-2 (Dao, 2024); this selective reconstruction is applied to the small routing-query set and does not materialize the full attention matrix. Since the visual attention mass $Z_j = \sum_{n=1}^{N} a^L_{j,n}$ is positive, we normalize the visual slice into an evidence distribution:

$$p_j(i) = \frac{a^L_{j,i}}{Z_j}, \qquad \sum_{i=1}^{N} p_j(i) = 1. \tag{3}$$

p_j is computed after visual and textual tokens have interacted, and is conditioned on the question.

4.2 Minimum-Coverage Evidence Routing

Averaging evidence distributions can suppress tokens that are vital to one key semantic aspect of the question. Thinking-Once routes evidence independently for each query. Let $[N] = \{1, \ldots, N\}$. For a coverage threshold $\rho \in (0, 1]$, $S_j(\rho)$ is the canonical fixed-tie-breaking solution of

$$S_j(\rho) \in \arg\min_{S \subseteq [N]} |S| \quad \text{s.t.} \quad \sum_{i \in S} p_j(i) \geq \rho. \tag{4}$$

This set is obtained by sorting $p_j(i)$ in descending order, using fixed tie-breaking, and taking the shortest valid prefix whose total mass reaches ρ. Focused queries retain few tokens, whereas diffuse queries automatically preserve more evidence. The final core set is $S_{\text{core}} = \bigcup_{q_j \in \mathcal{Q}_r} S_j(\rho)$, which guarantees at least ρ evidence coverage separately for every individual routing query. The union is not necessarily the global minimum-cardinality set satisfying all query constraints, but fixed tie-breaking makes S_{core} monotone in ρ.

4.3 Context-Preserved Single-Visual-Pass Inference

The core set preserves question-relevant tokens at full resolution, while the remaining visual tokens are summarized to retain coarse spatial contextual information. We partition the visual grid into disjoint cells $\mathcal{G}$ and define the unselected tokens in each cell as $U_g = g \cap ([N] \setminus S_{\text{core}})$. Each non-empty set is summarized by parameter-free mean pooling:

$$b_g = \frac{1}{|U_g|}\sum_{i \in U_g} v^L_i, \qquad |U_g| > 0. \tag{5}$$

Let $\mathrm{OrdMerge}$ place retained core tokens and background tokens according to the original visual raster order. A background token is inserted at the position of the last token it summarizes, so routing does not expose future visual states to earlier routed positions. The routed visual evidence is

$$E^L = \mathrm{OrdMerge}\big(V^L_{S_{\text{core}}}; \{b_g : g \in \mathcal{G}, |U_g| > 0\}\big). \tag{6}$$

Base Model	Setting	V*Bench			HRBench-4K			HRBench-8K			Mean	Mem. GB↓
		Attr	Spa.	Avg	FSP	FCP	Avg	FSP	FCP	Avg		
Qwen2.5-VL-7B	Base	80.9	76.3	79.1	86.3	55.5	70.9	83.5	51.3	67.4	72.5	∼20
	Ours	**89.6**	**88.2**	**89.0**	**89.8**	**61.3**	**75.5**	**87.5**	**58.3**	**72.9**	**79.1**	∼**16**
Qwen3-VL-8B	Base	84.4	80.3	82.7	91.3	65.8	78.5	84.0	62.8	73.4	78.2	∼20
	Ours	**87.8**	**81.6**	**85.3**	**91.8**	**70.5**	**81.1**	**84.3**	**64.3**	**74.3**	**80.2**	∼**16**
ZwZ-4B	Base	90.4	86.8	89.0	89.5	68.0	78.8	86.8	65.5	76.1	81.3	∼20
	Ours	**90.4**	**86.8**	**89.0**	**90.0**	**73.0**	**81.5**	**88.0**	**68.5**	**78.3**	**82.9**	∼**16**
ZwZ-7B	Base	90.4	84.2	88.0	90.0	59.5	74.8	88.5	58.3	73.4	78.7	∼20
	Ours	**90.4**	**89.5**	**90.1**	**90.8**	**67.3**	**79.0**	**89.8**	**63.5**	**76.6**	**81.9**	∼**16**
ZwZ-8B	Base	86.6	88.2	87.4	90.8	69.8	80.3	87.0	66.8	76.9	81.5	∼20
	Ours	**87.0**	**90.8**	**88.5**	**91.8**	**70.5**	**81.1**	**88.3**	**69.0**	**78.6**	**82.7**	∼**16**
Average Δ		**+2.5**	**+4.2**	**+3.1**	**+1.3**	**+4.8**	**+3.0**	**+1.6**	**+3.8**	**+2.7**	**+2.9**	**-4**

Table 1: Comprehensive results across five representative MLLMs. Mean averages the three benchmark averages. Average Δ is measured over the five base models against the corresponding base setting; higher accuracy and lower memory cost are better.

Replacing the original visual subsequence with E^L while keeping the textual states unchanged produces the final routed multimodal hidden-state representation for decoding.

$$\widetilde{H}^L = \mathrm{SeqMerge}(E^L; T^L). \tag{7}$$

This construction preserves question-relevant evidence at full resolution while retaining coarse scene context. Under the decoder's causal mask, the raster and prompt order preserve the causal direction of information during routing.

We keep the MRoPE coordinates for retained visual tokens. Each background token takes the integer-rounded centroid of coordinates in U_g. Let M be the maximum coordinate across all MRoPE axes in the routed prompt. Generated text tokens then use the shared scalar position during decoding

$$\mathbf{p}_{\mathrm{gen},n} = (M+n,\ M+n,\ M+n), \qquad n \geq 1. \tag{8}$$

Thus, routing shortens the KV cache of layers deeper than L without losing the original spatial coordinate system or breaking generation continuity in later layers. Let M_T be the number of text tokens and let $G_{\mathrm{act}} \leq G = |\mathcal{G}|$ be the number of non-empty background cells. The routed sequence length is $|\widetilde{H}^L| = M_T + |S_{\mathrm{core}}| + G_{\mathrm{act}}$. Since each background token replaces at least one unselected visual token,

$$|\widetilde{H}^L| \leq M_T + \min\{N,\ |S_{\mathrm{core}}| + G\} \leq M_T + N. \tag{9}$$

If $n_0 = N + M_T$ and $n_1 = |\widetilde{H}^L|$, the cost of each layer changes from $\mathcal{O}(n_0 d^2 + n_0^2 d)$ to $\mathcal{O}(n_1 d^2 + n_1^2 d)$. The routed length, the attention cost, and the KV cache decrease whenever at least one background token summarizes multiple unselected tokens. Thinking-Once treats efficiency as a consequence of routing compact, question-conditioned evidence, rather than as token removal.

5 Experiments

5.1 Experimental Setup

Models and Protocol. We evaluate Thinking-Once on multiple MLLMs, including Qwen2.5-VL-7B-Instruct, Qwen3-VL-8B-Instruct, and the ZwZ series (Bai et al., 2025b;a; Wei et al., 2026). Hereafter, we abbreviate the two instruction-tuned models as Qwen2.5-VL-7B and Qwen3-VL-8B. These models span different generations, parameter scales, and training recipes, allowing us to more broadly assess the generalizability of the proposed routing strategy across diverse models. All experiments follow the same single-visual-pass inference protocol, using one full-image visual encoding and one multimodal forward pass. **Benchmarks**. We evaluate our method on three HR-VQA benchmarks: V*Bench (Wu & Xie, 2024), HRBench-4K, and HRBench-8K (Wang et al., 2025). V*Bench measures fine-grained attribute and spatial perception, while HRBench evaluates single-instance perception (FSP) and cross-instance perception (FCP) under challenging high-resolution settings, following prior work (Liu et al., 2026b; Zheng et al., 2026). Implementation details and results on MME-RealWorld-Lite (Zhang et al., 2025c), ZoomBench (Wei et al., 2026), TreeBench (Wang et al., 2026), and POPE (Li et al., 2023b) are provided in Appendices C and D.

Method	Train.-free	Extra Visual	Time (min)↓	V*Bench			HRBench-4K			HRBench-8K			Mean	Mem. GB↓
				Attr	Spa.	Avg	FSP	FCP	Avg	FSP	FCP	Avg		
GPT-4o-1120	–	–	–	–	–	66.0	70.0	48.0	59.0	62.0	49.0	55.5	60.2	–
External Evidence Reacquisition: extra visual inputs or re-encoding														
VLM-R^3	✗	✓	∼30	86.1	71.1	80.1	81.0	57.5	69.3	69.5	51.5	60.5	70.0	∼23
DeepEyes	✗	✓	∼37	87.0	80.3	84.3	89.3	56.3	72.8	85.5	56.0	70.8	76.0	∼20
TreeVGR	✗	✓	∼**5**	85.2	85.5	85.3	89.5	57.0	73.3	84.5	57.3	70.9	76.5	∼18
DyFo	✓	✓	∼31	80.9	86.8	83.3	89.3	54.0	71.6	86.5	53.3	69.9	74.9	∼86‡
DeepScan	✓	✓	∼180	**89.6**	**88.2**	**89.0**	89.0	58.0	73.5	85.0	56.5	70.8	77.8	∼40‡
ViCrop	✓	✓	∼28	**89.6**	71.1	82.2	**90.5**	57.5	74.0	85.5	53.0	69.3	75.2	∼20
Compression-First Token Reduction: generic pruning or merging														
BTP	✓	✗	∼90	72.2	69.7	71.2	88.8	53.0	70.9	82.5	46.8	64.6	68.9	∼17
TRIO	✓	✗	∼14	84.3	71.1	79.1	88.3	53.3	70.8	82.5	49.5	66.0	72.0	∼50‡
HiPrune	✓	✗	∼12	82.6	77.6	80.6	88.8	55.0	71.9	85.0	53.5	69.3	73.9	∼70‡
VisionZip	✓	✗	∼11	81.7	77.6	80.1	88.8	54.8	71.8	85.5	53.5	69.5	73.8	∼36‡
V^{2}Drop	✓	✗	∼8	80.0	77.6	79.1	88.0	54.8	71.4	84.8	50.5	67.6	72.7	∼20
Ours: Single-Visual-Pass Evidence Routing without extra visual encoding														
Thinking-Once	✓	✗	∼**5**	**89.6**	**88.2**	**89.0**	89.8	**61.3**	**75.5**	**87.5**	**58.3**	**72.9**	**79.1**	∼**16**
Thinking-Once†	✓	✗	∼5	87.0	90.8	88.5	91.8	70.5	81.1	88.3	69.0	78.6	82.7	∼16

Table 2: Fine-grained comparison across HR-VQA paradigms. Open-source methods use Qwen2.5-VL-7B unless marked with † (ZwZ-8B). Compression-first pruning or merging methods retain 33.3% of visual tokens. Mean averages three benchmarks; time and average peak memory are measured on V*Bench. ‡ denotes two-GPU inference for OOM cases, with memory summed across GPUs. **Boldface** and underlining indicate the best and second-best Qwen2.5-VL-7B-based results.

5.2 Main Results

Table 1 evaluates Thinking-Once on five MLLM base models, including Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), and ZwZ-series models (Wei et al., 2026). Across these models, Thinking-Once consistently improves or matches the Base setting on the three benchmark averages. On average, it improves V*Bench by +3.1 points, HRBench-4K by +3.0 points, and HRBench-8K by +2.7 points, while reducing memory by about 4 GB; the routed visual-token ratios of the models are reported in Table 9. On the matched Qwen2.5-VL-7B base model, Thinking-Once improves V*Bench by +9.9 points, HRBench-4K by +4.6 points, and HRBench-8K by +5.5 points, increasing the mean score from 72.5 to 79.1 while substantially reducing peak memory from about 20 GB to 16 GB. With the stronger ZwZ-8B base model, Thinking-Once reaches a mean score of 82.7, indicating that the same internal evidence-routing principle remains broadly useful when the base model already has stronger high-resolution perception capability.

Gains are pronounced on relation-sensitive metrics. On V*Bench, Spa. accuracy improves by +4.2 points on average, versus +2.5 points for Attr accuracy. On HRBench, FCP improves by +4.8 points on 4K and +3.8 points on 8K, exceeding the corresponding FSP gains of +1.3 and +1.6 points. This shows Thinking-Once is useful when the answer depends on relations, comparisons, and cross-region context, rather than isolated object attributes. The results therefore support our central view: HR-VQA failures are not only caused by missing visual evidence, but also by ineffective transmission of already encoded evidence to later reasoning layers. Layer-wise attention diagnostics in Appendix F support this view: substantial target-centered attention has already emerged at intermediate layers, deteriorates during deeper processing in the base models, and is retained after routing. A pixel-level restoration experiment in Appendix F.4 provides the input-space counterpart of this link: masking only the annotated target regions lowers full-image accuracy by 36.6 points, and accuracy recovers as the region content is restored.

5.3 Comparison with Existing Methods

Table 2 compares Thinking-Once with representative HR-VQA methods under the original high-resolution setting. The comparison shows two distinct limitations of existing paradigms. First, external evidence reacquisition methods can recover missing local details by constructing extra visual inputs, re-encoding selected regions, or performing multi-step visual search. This confirms that additional observation is useful when the required evidence is absent from the current repre-

Variant	V^*			HRBench-4K			HRBench-8K			Mean
	Attr	Spa.	Avg	FSP	FCP	Avg	FSP	FCP	Avg	
Base	80.9	76.3	79.1	86.3	55.5	70.9	83.5	51.3	67.4	72.5
w/o bg.	82.6	84.2	83.3	88.8	59.5	74.1	83.8	56.5	70.1	75.8
Denoise	81.7	81.6	81.7	88.3	57.8	73.0	82.8	56.0	69.4	74.7
Entity only	82.6	84.2	83.3	88.8	59.0	73.9	83.8	56.0	69.9	75.7
Global only	81.7	75.0	79.1	88.3	50.0	69.1	82.8	48.3	65.5	71.2
Full	**89.6**	**88.2**	**89.0**	**89.8**	**61.3**	**75.5**	**87.5**	**58.3**	**72.9**	**79.1**

Table 3: Ablations on Qwen2.5-VL-7B. Mean averages benchmarks.

sentation. However, the gains are obtained by changing the visual input pipeline: methods such as DeepEyes (Zheng et al., 2026), DeepScan (Li et al., 2026b), and ViCrop (Khayatkhoei et al., 2025) introduce extra visual processing, and several runs require higher latency or memory.

Compression-first methods show that shortening the visual sequence is not sufficient for HR-VQA. BTP (Chen et al., 2026b), TRIO (Zhang et al., 2026b), HiPrune (Liu et al., 2026a), VisionZip (Yang et al., 2025), and V^2Drop (Chen et al., 2026a) avoid extra visual inputs, but their results are less stable under the same high-resolution setting. All evaluations guarantee that the input images are fed at their original resolution; under this setting, several methods run out of memory on a single NVIDIA A800 GPU, and the runnable compression baselines remain below the full Thinking-Once result. A controlled comparison under a matched budget and token-retention ratio is provided in Appendix E.2. More importantly, the gap to Thinking-Once is most visible on spatial and cross-instance metrics, where the answer depends on relations among small entities and surrounding context.

Thinking-Once is complementary to both paradigms. Like token-reduction methods, it shortens the later-layer sequence, but it does not select tokens according to generic redundancy alone. Instead, it more directly and effectively routes a compact, question-conditioned evidence set from the model's own intermediate representation through independent per-query coverage and background summaries, so compression emerges as a consequence of evidence allocation. Unlike external evidence reacquisition, it requires no additional visual input or re-encoding. The comparison thus supports a more specific conclusion: HR-VQA benefits when already encoded evidence is preserved and reallocated according to the current question, especially for relation-sensitive reasoning. Qualitative examples of entity-conditioned routing on single/multi-object cases are provided in Appendix E.

5.4 Ablation Studies

Table 3 reports component ablations on Qwen2.5-VL-7B, while the same removals on other base models are provided in Appendix E. Every variant is worse than the full method, indicating that the gain comes from preserving a structured evidence set rather than from a single component. Removing background summaries lowers the mean score by 3.3 points, suggesting that entity tokens alone are not enough for many HR-VQA questions. Spatial relations, relative positions, and attribute comparisons require surrounding regions as contextual references. Denoising attention sinks causes an even larger 4.4-point drop, showing that non-core tokens cannot be treated as uniform noise; some may carry global context or aggregation signals needed by later layers. The query ablations show that entity queries and the global question query are complementary. Entity-only routing loses 3.4 points, while global-only routing loses 7.9 points. Precise entity anchors provide the main localization signal, whereas the global question query supplies broader attribute, relation, and contextual demand. The full method is the best because it effectively preserves both types of evidence.

6 Conclusion

We have shown that fine-grained evidence can remain available after initial visual encoding. Ineffective evidence allocation can cause errors when key signals are diluted by redundant background during later reasoning. Question-relevant entities become identifiable at intermediate layers, defining an evidence-routing window in which they can influence downstream computation. These core tokens require compact contextual support for relational reasoning and attribute comparison. Thinking-Once routes this structured evidence to deeper layers within a single visual pass, improving accuracy and efficiency across five MLLMs. The results support evidence routing as a practical complement to additional visual acquisition in utilization-limited HR-VQA settings.

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 Thinking-once. Since Thinking-once 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

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

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401. IEEE, 2025.

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 46254–46272, 2025.

Qingqing Cao, Bhargavi Paranjape, and Hannaneh Hajishirzi. Pumer: Pruning and merging tokens for efficient vision language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12890–12903, 2023.

Junjie Chen, Xuyang Liu, Zichen Wen, Yiyu Wang, Siteng Huang, and Honggang Chen. Variation-aware vision token dropping for faster large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3489–3499, 2026a.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 19–35, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73004-7.

Xiaoyue Chen, Chen Gao, Yong Li, Xinlei Chen, et al. Balanced token pruning: Accelerating vision language models beyond local optimization. *Advances in Neural Information Processing Systems*, 38:62740–62764, 2026b.

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.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

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

Yingqi Fan, Anhao Zhao, Jinlan Fu, Junlong Tong, Hui Su, Yijie Pan, Wei Zhang, and Xiaoyu Shen. Visipruner: Decoding discontinuous cross-modal dynamics for efficient multimodal llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18896–18913, 2025.

Yuhang Han, Xuyang Liu, Zihan Zhang, Pengxiang Ding, Junjie Chen, Honggang Chen, Donglin Wang, Qingsen Yan, and Siteng Huang. Filter, correlate, compress: Training-free token reduction for mllm acceleration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4601–4609, 2026.

Wenxuan Huang, Zijie Zhai, Yunhang Shen, Shaosheng Cao, Fei Zhao, Xiangfeng Xu, Zheyu Ye, and Shaohui Lin. Dynamic-llava: Efficient multimodal large language models via dynamic vision-language context sparsification. In *International Conference on Learning Representations*, volume 2025, pp. 69927–69955, 2025.

Chaoya Jiang, Yongrui Heng, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. Vlm-r^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *Advances in Neural Information Processing Systems*, 38:63841–63869, 2026.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014. IEEE, 2025.

Zhehan Kan, Xin Li, Yanlin Liu, Xiaochen Yang, Xinghua Jiang, Yinsong Liu, Deqiang Jiang, Xing Sun, Qingmin Liao, and Wenming Yang. RAR: Reversing visual attention re-sinking for unlocking potential in multimodal large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xTbFeDhdRG.

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

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.

Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9098–9108, 2025a.

Haobin Li, Yutong Yang, Yijie Lin, Xiang Dai, Mouxing Yang, and Xi Peng. Reliable thinking with images. In *International Conference on Machine Learning*, 2026a.

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.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, 133(10):6794–6812, 2025b.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19155–19164, 2026b.

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

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, 2025a.

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

Jizhihui Liu, Guangdao Zhu, and Feiyi Du. Hiprune: Training-free visual token pruning via hierarchical attention in vision-language models (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 41275–41277, 2026a.

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. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=v6Tpe2LRrW.

Zhining Liu, Rana Ali Amjad, Ravinarayana Adkathimar, Tianxin Wei, and Hanghang Tong. Selfelicit: Your language model secretly knows where is the relevant evidence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9153–9173, 2025b.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, et al. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. In *International Conference on Learning Representations*, volume 2026, pp. 72849–72876, 2026c.

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

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 57172–57189, 2025.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867, 2025.

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

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

Jianlin Su. Transformer upgrade roadmap: 2, rotary position embedding. `https://www.spaces.ac.cn/archives/8265`, March 2021. Scientific Spaces, accessed September 25, 2026.

Jianlin Su. Transformer upgrade roadmap: 19, the second type of rotary position embedding. `https://www.spaces.ac.cn/archives/10862`, April 2025. Scientific Spaces, accessed September 25, 2026.

Meituan LongCat Team. Longcat-next: Lexicalizing modalities as discrete tokens, 2026. URL `https://arxiv.org/abs/2603.27538`.

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

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7907–7915, 2025.

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, 2024.

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, 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*, 2026.

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*, pp. 13084–13094, 2024.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025.

Wenjie Yang, Siqi Zhu, and Zengfeng Huang. Position: Your vlm may not be thinking with interleaved images. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026.

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. Atp-llava: Adaptive token pruning for large vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24972–24982. IEEE, 2025.

Ce Zhang, Kaixin Ma, Tianqing Fang, Wenhao Yu, Hongming Zhang, Zhisong Zhang, Haitao Mi, and Dong Yu. Vscan: Rethinking visual token reduction for efficient large vision-language models. *Transactions on Machine Learning Research*, 2026a. URL https://openreview.net/forum?id=KZYhyilFnt.

Haokui Zhang, Congyang Ou, Dawei Yan, Peng Wang, Qingsen Yan, Ying Li, Rong Xiao, and Chunhua Shen. Pio-fvlm: Rethinking training-free visual token reduction for vlm acceleration from an inference-objective perspective. *arXiv preprint arXiv:2602.04657*, 2026b.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20857–20867. IEEE, 2025a.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. Adaptive chain-of-focus reasoning via dynamic visual search and zooming for efficient vlms. *arXiv preprint arXiv:2505.15436*, 2025b.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, volume 2025, pp. 89655–89701, 2025c.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *Forty-second International Conference on Machine Learning*, 2025d. URL https://openreview.net/forum?id=80faIPZ67S.

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.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024.

Technical Appendix: Thinking-Once Evidence Routing for High-Resolution Visual Question Answering

A Formal Properties of the Routing Operator

This section provides the detailed routing specification omitted from the main paper and establishes the formal properties of the routing operator used there. The guarantees concern selective attention reconstruction, normalization, minimum-cardinality coverage, per-query coverage, background pooling, sequence length, causal ordering at the routing operation, generation-position continuity, and downstream computational cost. They do not assume or claim that attention mass is a perfect measure of semantic sufficiency.

A.1 Detailed Routing Specification and Preliminaries

Let an MLLM contain D transformer layers, and let $L < D$ be the routing layer. At layer L, the multimodal hidden state is

$$H^L = \mathrm{SeqMerge}(V^L; T^L), \qquad V^L = \{v_i^L\}_{i=1}^N, \tag{10}$$

where V^L and T^L are the visual and textual hidden states, respectively, and $\mathrm{SeqMerge}$ follows their original prompt positions. Throughout this appendix, H^L denotes the output hidden state produced after transformer layer L has completed, whereas $A^{L,h}_{q_j,i}$ denotes the attention probability computed within layer L from that layer's input states. Thinking-Once uses $A^{L,h}_{q_j,i}$ to select and summarize the output visual states V^L, and replaces only the visual subsequence after layer L; all textual states and their relative order are retained. Consequently, the first transformer layer affected by routing is layer $L+1$.

The routing-query set is

$$\mathcal{Q}_r = \mathcal{Q}_{\mathrm{ent}} \cup \{q_{\mathrm{glb}}\}. \tag{11}$$

Here, $\mathcal{Q}_{\mathrm{ent}}$ contains the last non-punctuation token of each extracted entity span, and q_{glb} is a global question query instantiated by the last non-punctuation question token. Under causal self-attention, this token provides a compact representation of the preceding question context. The entity queries provide explicit object- or region-level anchors, while the global question query captures complementary attribute, relation, and contextual demand.

For each routing query $q_j \in \mathcal{Q}_r$, we selectively reconstruct the corresponding visual attention row from the MRoPE-transformed query and key states. If $A^{L,h}_{q_j,i}$ denotes the attention probability from q_j to visual token i in head h, the head-averaged score is

$$a^L_{j,i} = \frac{1}{H}\sum_{h=1}^{H} A^{L,h}_{q_j,i}, \qquad i \in [N]. \tag{12}$$

The ordinary model forward pass continues to use FlashAttention-2 (Dao, 2024). The reconstruction is applied only to the small set $\mathcal{Q}_r$ and therefore does not materialize the complete sequence-by-sequence attention matrix.

Throughout, $A^{L,h}_{q_j,i}$ denotes the exact causal-softmax probability. For each head h, the reconstruction computes the attention logits of q_j against *all* keys visible to q_j under the causal mask, including the preceding textual keys, applies the softmax over this complete visible key set, and only then restricts the resulting probability row to the visual positions. Hence the full attention matrix is never materialized, yet the softmax normalizer of every reconstructed row is exact. The order of operations matters: because the full-row softmax denominators differ across heads, restricting each head to the visual slice and renormalizing *before* head averaging would in general yield a distribution different from the one defined by Eqs. equation 12 and equation 13. The visual-slice normalization in Eq. equation 13 is therefore applied exactly once, after averaging the exact per-head probabilities.

We assume a single-image, image-before-text prompt layout under full causal attention: all N visual tokens form a visible visual prefix before every $q_j \in \mathcal{Q}_r$, and no visual position in this prefix is masked from a routing query. Multi-image and interleaved layouts are handled by restricting each routing query to its visible visual-token set; this generalization is made precise in Remark A.5.

Define the visual attention mass and its normalized evidence distribution as

$$Z_j = \sum_{i=1}^{N} a_{j,i}^{L}, \qquad p_j(i) = \frac{a_{j,i}^{L}}{Z_j}. \tag{13}$$

Under the prompt-layout assumption, every visual token has a finite unmasked attention logit. Softmax therefore assigns positive probability to each visible visual token, head averaging preserves positivity, and $Z_j > 0$.

Let π_j be a permutation of $[N] = \{1, \dots, N\}$ that sorts p_j in non-increasing order. Ties are broken by a fixed deterministic rule: tokens with equal probability are ordered by ascending token index, i.e., by their original visual raster position. All statements below that invoke fixed tie-breaking refer to this rule. Define

$$C_j(k) = \sum_{t=1}^{k} p_j(\pi_j(t)), \qquad C_j(0) = 0, \tag{14}$$

and, for $\rho \in (0, 1]$,

$$k_j(\rho) = \min\{k \in [N] : C_j(k) \geq \rho\}, \qquad S_j(\rho) = \{\pi_j(1), \dots, \pi_j(k_j(\rho))\}. \tag{15}$$

The core set is

$$S_{\text{core}} = \bigcup_{q_j \in \mathcal{Q}_r} S_j(\rho). \tag{16}$$

Operationally, Thinking-Once therefore (i) runs the unmodified sequence to layer L, (ii) constructs entity and global-question routing queries, (iii) reconstructs only their visual attention rows, (iv) selects a minimum-coverage set independently for each query, (v) preserves the union at full token resolution while summarizing the unselected background, and (vi) resumes the remaining layers with the routed sequence.

A.2 Evidence Distribution and Minimum Coverage

Lemma A.1 (Normalized evidence distribution)**.** *For every routing query with $Z_j > 0$, p_j is a probability distribution over $[N]$.*

$$p_j(i) \geq 0, \qquad \sum_{i=1}^{N} p_j(i) = 1. \tag{17}$$

Proof. Non-negativity follows from $a_{j,i}^{L} \geq 0$ and $Z_j > 0$. Summing Eq. equation 13 over i gives $\sum_i p_j(i) = Z_j / Z_j = 1$. □

Proposition A.2 (Minimum-cardinality coverage)**.** *For any $\rho \in (0, 1]$, the canonical set $S_j(\rho)$ in Eq. equation 15 solves*

$$\min_{S \subseteq [N]} |S| \quad \text{s.t.} \quad \sum_{i \in S} p_j(i) \geq \rho. \tag{18}$$

Proof. The set is feasible because its mass is $C_j(k_j(\rho)) \geq \rho$. For any set S of cardinality m, the sum of its probabilities cannot exceed the sum of the m largest probabilities:

$$\sum_{i \in S} p_j(i) \leq C_j(m). \tag{19}$$

If $m < k_j(\rho)$, the definition of $k_j(\rho)$ gives $C_j(m) < \rho$; hence no set with fewer than $k_j(\rho)$ elements is feasible. Therefore $S_j(\rho)$ has minimum cardinality. The ascending-index tie-breaking rule selects one canonical minimum-cardinality solution when several exist. □

Corollary A.3 (Nested routing and per-query coverage)**.** *If $0 < \rho_1 \le \rho_2 \le 1$, then the canonical fixed-order solutions satisfy $S_j(\rho_1) \subseteq S_j(\rho_2)$. Moreover, for every $q_j \in \mathcal{Q}_r$,*

$$\sum_{i \in S_{\mathrm{core}}} p_j(i) \ge \rho, \tag{20}$$

where ρ is the threshold used to construct S_{core}.

Proof. Because $C_j(k)$ is non-decreasing, reaching a larger threshold cannot require a shorter sorted prefix, so $k_j(\rho_1) \le k_j(\rho_2)$. Both sets use the same fixed ordering π_j, which proves nesting for the canonical solution. Since $S_j(\rho) \subseteq S_{\mathrm{core}}$ and $p_j(i) \ge 0$, the mass of the union for query j is at least the mass of $S_j(\rho)$, which is at least ρ. □

Remark A.4 (Scope of the minimum guarantee). Proposition A.2 establishes minimum cardinality separately for each query. The union S_{core} guarantees every query's coverage but is not claimed to be the globally smallest set satisfying all query constraints jointly.

Remark A.5 (Multi-image and interleaved layouts). For multi-image or interleaved prompts, let $\mathcal{V}_j \subseteq [N]$ denote the set of visual positions visible to q_j under the attention mask; under the single-image, image-before-text assumption above, $\mathcal{V}_j = [N]$ for every routing query. In the general case, the sums in Eq. equation 13 range over $\mathcal{V}_j$ instead of $[N]$, so that $Z_j = \sum_{i \in \mathcal{V}_j} a^L_{j,i}$ and each p_j is a probability distribution supported on $\mathcal{V}_j$; masked visual positions are not eligible for query q_j and receive zero mass by convention. The minimum-coverage selection of Eq. equation 15 is then performed within $\mathcal{V}_j$, the core set remains the union of the per-query selections, and visual tokens visible to no routing query are treated as unselected background. All results transfer directly: Lemma A.1, Proposition A.2, and Corollary A.3 hold verbatim with $[N]$ replaced by $\mathcal{V}_j$, while Propositions A.6–A.9 and Corollary A.11 are unaffected because they depend only on the partition $\mathcal{G}$ and the physical raster positions, not on which queries selected the core tokens.

A.3 Context Preservation and Sequence-Length Bound

This subsection formalizes the background-downsampled routing operator. The selected core evidence remains at full token resolution, while every unselected visual token contributes to a coarse background summary.

Let $\mathcal{G}$ be a disjoint partition of the visual-token grid. For each cell $g \in \mathcal{G}$, define

$$U_g = g \cap ([N] \setminus S_{\mathrm{core}}). \tag{21}$$

Only non-empty cells emit summaries. Let

$$\mathcal{G}_{\mathrm{act}} = \{g \in \mathcal{G} : |U_g| > 0\}, \qquad G_{\mathrm{act}} = |\mathcal{G}_{\mathrm{act}}|, \qquad G = |\mathcal{G}|. \tag{22}$$

In our implementation, each active cell is summarized with parameter-free mean pooling:

$$b_g = \frac{1}{|U_g|} \sum_{i \in U_g} v^L_i, \qquad g \in \mathcal{G}_{\mathrm{act}}. \tag{23}$$

Mean pooling introduces no additional learned parameters and ensures that every unselected token contributes to exactly one background summary. The sequence-length and ordering results below require only that each active cell produces one token derived from its own source set U_g.

The routed visual evidence is

$$B^L = \{b_g : g \in \mathcal{G}_{\mathrm{act}}\}, \qquad E^L = \mathrm{OrdMerge}\big(\{v^L_i : i \in S_{\mathrm{core}}\}; B^L\big). \tag{24}$$

The routed multimodal hidden state is then

$$\widetilde{H}^L = \mathrm{SeqMerge}(E^L; T^L), \tag{25}$$

where the original textual states and their relative prompt order are unchanged.

The operator $\mathrm{OrdMerge}$ orders retained core tokens and background summaries according to the original visual raster order. A summary corresponding to U_g is assigned the representative insertion position

$$m_g = \max U_g. \tag{26}$$

Thus, a summary is inserted no earlier than any original visual token that contributes to it. This physical insertion position is used for routed sequence order and is distinct from the MRoPE coordinate used for positional encoding.

Proposition A.6 (Routed sequence-length bound)**.** *For M_T textual tokens, the routed sequence length n_1 satisfies*

$$n_1 = M_T + |S_{\text{core}}| + G_{\text{act}}. \tag{27}$$

Moreover,

$$n_1 \le M_T + \min\{N,\ |S_{\text{core}}| + G\} \le M_T + N. \tag{28}$$

Furthermore, $n_1 < M_T + N$ if and only if at least one active background cell summarizes more than one token.

Proof. The routed visual sequence contains exactly $|S_{\text{core}}|$ retained core tokens and one background summary token for each active cell, so Eq. equation 27 follows directly.

Since only non-empty cells emit summary tokens, we have $G_{\text{act}} \le G$, which gives

$$|S_{\text{core}}| + G_{\text{act}} \le |S_{\text{core}}| + G. \tag{29}$$

This proves the second candidate in the minimum term.

Next, because $\mathcal{G}$ is a partition of $[N]$, the non-empty sets U_g partition the unselected visual tokens $[N] \setminus S_{\text{core}}$. Therefore,

$$\sum_{g \in \mathcal{G}_{\text{act}}} |U_g| = N - |S_{\text{core}}|. \tag{30}$$

Each active cell is non-empty, so

$$G_{\text{act}} \le \sum_{g \in \mathcal{G}_{\text{act}}} |U_g| = N - |S_{\text{core}}|. \tag{31}$$

Hence

$$|S_{\text{core}}| + G_{\text{act}} \le N. \tag{32}$$

Combining the two upper bounds gives Eq. equation 28.

Finally, equality $|S_{\text{core}}| + G_{\text{act}} = N$ holds exactly when $G_{\text{act}} = \sum_{g \in \mathcal{G}_{\text{act}}} |U_g|$, which happens exactly when every active U_g is a singleton. Therefore the inequality is strict exactly when some active cell contains at least two unselected tokens, i.e., when at least one background summary merges multiple tokens. □

Proposition A.7 (Causal safety of ordered merging)**.** *Under a causal decoder mask, suppose retained core tokens preserve their original visual raster positions, each summary token of U_g is inserted at $m_g = \max U_g$, and all textual states retain their original relative order. Then the routing operation does not place information from a later original sequence position into an earlier routed position.*

Proof. For each routed visual token e, define its representative position $\text{rep}(e)$ and source set $\text{src}(e)$ as follows. If e is a retained core token v_i^L, then $\text{rep}(e) = i$ and $\text{src}(e) = \{i\}$. If e is a background summary token b_g, then $\text{rep}(e) = m_g = \max U_g$ and $\text{src}(e) = U_g$. Since the cells in $\mathcal{G}$ are disjoint, different active cells have disjoint U_g sets. Moreover, $m_g \in U_g \subseteq [N] \setminus S_{\text{core}}$, so a summary insertion position cannot coincide with a retained core-token position. Thus OrdMerge is well defined.

For a retained core token, clearly $\text{src}(e) \subseteq \{1, \dots, \text{rep}(e)\}$. For a background summary token, every $i \in U_g$ satisfies $i \le m_g$, and therefore

$$\text{src}(e) \subseteq \{1, \dots, \text{rep}(e)\}.$$

Ordered merging sorts retained core tokens and background summaries by their representative positions, so the routed visual order is monotone in $\text{rep}(e)$. Under a causal decoder mask, a routed state at representative position $\text{rep}(e)$ can therefore depend only on source positions no later than $\text{rep}(e)$. No information from a later original sequence position is moved to an earlier routed position, and no future-to-past information path is introduced by the routing operation. □

Remark A.8. Proposition A.7 is deliberately limited to the routing operation. It does not claim that background downsampling preserves every pairwise attention relation or leaves the model output unchanged.

Proposition A.9 (Generation-position continuity)**.** *The representative insertion position m_g in Eq. equation 26 is a physical raster-order position used for ordered merging and causal masking. It is distinct from the MRoPE coordinate assigned to the summary token for positional encoding. Retained visual tokens keep their original MRoPE coordinates. Each background summary token uses the integer-rounded centroid of the MRoPE coordinates of the tokens in U_g. Let M be the maximum coordinate value over all MRoPE axes in the routed prompt. If generated token $n \geq 1$ is assigned*

$$\mathbf{p}_{\mathrm{gen},n} = (M + n)\mathbf{1}_3, \tag{33}$$

then every generated coordinate is later than every routed-prompt coordinate on all three axes, and consecutive generated coordinates differ by one on each axis.

Proof. By definition, every routed-prompt coordinate component is at most M. For generated token $n \geq 1$, every component of $\mathbf{p}_{\mathrm{gen},n}$ equals $M + n$, which is strictly larger than M. Moreover,

$$\mathbf{p}_{\mathrm{gen},n+1} - \mathbf{p}_{\mathrm{gen},n} = \mathbf{1}_3. \tag{34}$$

Thus generation positions continue strictly and uniformly after the routed prompt. □

Remark A.10 (MRoPE coordinates of background summaries). The centroid-based MRoPE coordinate of a background summary is used only for positional encoding, while causal safety is determined by the physical routed sequence order and the decoder mask. Therefore, possible non-monotonicity of summary MRoPE coordinates along the physical sequence does not affect the causal-order argument in Proposition A.7. A separate consequence of integer rounding is that routed MRoPE coordinates need not be distinct: the rounded centroid of a background summary may coincide with the coordinate of a retained core token, or with that of another summary. Such coincidences are likewise harmless. MRoPE coordinates enter the computation only through relative rotary phases in the attention scores, which remain well defined when coordinates repeat, whereas sequence order, the causal mask, and KV-cache indexing are governed by the distinct physical insertion positions established in Proposition A.7. Moreover, by Proposition A.9, every generated-token coordinate is strictly larger than every routed-prompt coordinate, so any duplication is confined to the routed prompt and never involves generated positions.

A.4 Downstream Computational Cost

Corollary A.11 (Non-increasing downstream cost)**.** *Let $n_0 = M_T + N$ be the original prompt length at layers deeper than the routing layer, and let n_1 satisfy Eq. equation 28. For hidden dimension d, write the per-layer FLOPs of a standard Transformer layer as*

$$F(n) = c_1 n d^2 + c_2 n^2 d, \tag{35}$$

where $c_1, c_2 > 0$ are constants depending only on the architecture. Then the per-layer FLOPs change from $F(n_0)$ to $F(n_1)$, and the physical KV-cache length changes from n_0 to n_1. Both quantities are non-increasing, and they strictly decrease whenever $n_1 < n_0$, i.e., whenever at least one active background cell summarizes multiple unselected tokens.

Proof. Proposition A.6 gives $n_1 \leq n_0$, with strict inequality when at least one active background cell summarizes multiple unselected tokens. For positive d, c_1, c_2, both nd^2 and $n^2 d$ are strictly increasing in n, so $F(n)$ is strictly increasing in n. Therefore $F(n_1) \leq F(n_0)$, with strict inequality when $n_1 < n_0$. KV-cache storage is linear in physical sequence length, so the same non-increasing and strict decrease statements hold for the physical KV-cache length. □

Remark A.12 (Scope of the efficiency guarantee). The formal reduction applies to transformer layers deeper than L; the first L layers still process the original sequence. Selective attention-row reconstruction, sorting, background pooling, and ordered merging introduce additional routing overhead. However, reconstruction is restricted to $|\mathcal{Q}_r|$ query rows rather than the full attention matrix, and the remaining operations are performed once at layer L. The resulting end-to-end latency and memory changes are therefore measured empirically in the main experiments rather than inferred solely from the asymptotic bound.

A.5 What the Formal Properties Do Not Establish

The results above establish formal properties of the routing operator conditional on its attention-derived evidence distributions and background pooling rule. They show that the evidence distributions are normalized, the per-query minimum-coverage sets are well-defined, every routing query receives at least the prescribed coverage before union, background downsampling does not increase sequence length, ordered merging is causally safe under the stated insertion rule, generation positions remain continuous, and downstream physical sequence length, per-layer FLOPs, and KV-cache size are non-increasing.

These properties do not establish that $p_j(i)$ exactly measures causal importance, that a background summary token is semantically equivalent to all tokens it summarizes, that the routed representation is sufficient for every question, or that the answer distribution is unchanged by routing. Those claims require empirical validation, which is provided through controlled interventions, ablations, and benchmark results in the main paper and appendix.

B Supporting Evidence from Prior Work

B.1 Recoverability of Visual Encoder Representations

The main paper argues that high-resolution visual evidence does not always need to be reacquired through additional crops, zoom-in operations, or sub-image re-encoding. This argument relies on a necessary condition: the visual representations produced by the initial encoder should still retain recoverable information about the original image. Prior evidence from LongCat-Next (Team, 2026) supports this condition by showing that frozen visual encoder representations can preserve recoverable low-level and structural information.

Figure 7 shows visual reconstructions from several frozen vision encoders using a lightweight pixel decoder. Although the encoders are not trained specifically for pixel-level reconstruction in this analysis, the reconstructed images still preserve recognizable object contours, spatial layout, and local visual structures. This suggests that vision encoder outputs are not merely abstract semantic labels detached from the input image; instead, they may retain latent pathways through which fine-grained visual information survives the encoding process.

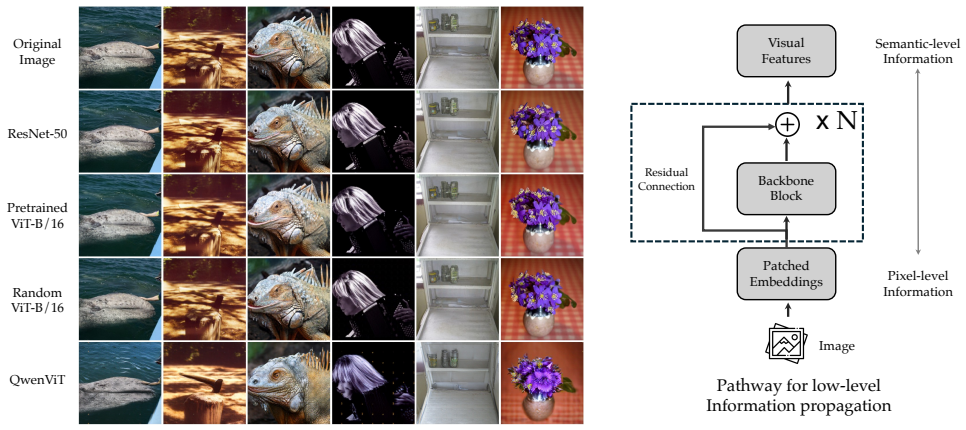


Figure 7: Supporting evidence from prior work on visual reconstruction from frozen vision encoders. A lightweight pixel decoder can recover recognizable visual structures from different encoder outputs, suggesting that visual embeddings may preserve image-level information beyond high-level semantic abstraction. Figure adapted from LongCat-Next (Team, 2026).

Table 4 provides the corresponding quantitative reconstruction results. Across different visual encoder architectures, the reconstruction metrics indicate that a non-trivial amount of visual information remains recoverable from frozen encoder outputs. In particular, the strong reconstruction fidelity of the randomly initialized ViT-B/16 suggests that part of this recoverability may arise from architectural pathways, such as residual connections, rather than solely from semantic pretraining.

The QwenViT result is especially relevant to MLLMs, as it suggests that the visual representations used by language-side reasoning can still carry recoverable image structure even after visual encoding.

Metric	ResNet50	ViT-B/16 (Pretrained)	ViT-B/16 (Random)	QwenViT (w/o merger)
PSNR (↑)	20.88 ± 3.44	21.86 ± 3.14	30.52 ± 3.42	18.16 ± 2.61
SSIM (↑)	0.509 ± 0.174	0.581 ± 0.139	0.887 ± 0.051	0.46 ± 0.14
rFID (↓)	0.4619	0.8850	0.5847	0.987

Table 4: Quantitative reconstruction performance across visual encoder architectures. PSNR and SSIM measure reconstruction fidelity (↑), while rFID measures perceptual discrepancy (↓). Results are adapted from LongCat-Next (Team, 2026).

These prior results do not imply that every high-resolution question can be answered from the initial visual encoding, nor do they prove that the encoded representation is semantically sufficient for all downstream reasoning tasks. Rather, they support a more specific claim used in our analysis: fine-grained visual evidence can remain available after visual encoding. Therefore, when an MLLM fails on HR-VQA, the failure may not always be caused by the complete absence of visual evidence. It may also arise because the evidence that already exists in the visual embeddings or intermediate hidden states is diluted by irrelevant high-resolution tokens or is not effectively routed to later reasoning layers. This motivates Thinking-Once to operate on internal visual representations and route question-conditioned evidence within a single visual pass, instead of assuming that the model must always reacquire evidence through additional visual inputs.

B.2 Interleaved Images Are Not Always Necessary

The main paper discusses that the gains of thinking-with-images or interleaved-image methods are not always attributable to the additional visual inputs themselves. The position paper by Yang et al. (Yang et al., 2026) provides direct supporting evidence by ablating interleaved images from several visual CoT models while keeping the evaluation protocol fixed. Their results, reproduced in Table 5, show that removing interleaved images often leads to only small changes in performance, and in some benchmarks even slightly improves the score.

The ablation results support a distinction between *visual evidence reacquisition* and *reasoning or utilization improvement*. If the interleaved images were consistently the dominant source of performance gains, removing them would be expected to cause large and systematic drops. Instead, the observed differences are often small, with several near-zero or positive changes. This suggests that part of the benefit of visual CoT or thinking-with-images training may come from improved language-side reasoning, task alignment, or better use of information already present in the original image representation, rather than solely from newly introduced visual observations.

These results should not be interpreted as evidence that interleaved images are never useful. Some settings, such as Pixel-Reasoner on V*Bench and HRBench-8K, do show noticeable degradation after removing them. The more relevant conclusion for our work is narrower: additional images are not always necessary, and their presence is not always the primary explanation for improved performance. This motivates Thinking-Once to ask whether high-resolution visual evidence that has already been encoded and aligned inside the model can be routed more effectively within a single visual pass, before resorting to extra visual inputs.

B.3 Tool Use Does Not Always Explain the Gain

Prior tool-use analyses provide complementary evidence that improvements from thinking-with-images or visual tool-use training may not come solely from the additional visual information obtained by tool calls. In particular, improvements can also arise from language-side reasoning alignment, better task adaptation, intrinsic capability changes, or reduced harmful tool interactions. This section summarizes two pieces of prior evidence that support our distinction between visual evidence reacquisition and visual evidence utilization.

Table 6 reports an ablation from DeepEyes (Zheng et al., 2026), where a model trained with text-only chain-of-thought data is compared with the full interleaved multimodal chain-of-thought setting.

Although the full iMCoT model achieves the strongest result on some benchmarks, text-only CoT training already brings a substantial improvement over the original Qwen2.5-VL-7B baseline on V*Bench and HRBench-4K. Since text-only CoT does not introduce additional interleaved visual observations, this result suggests that part of the improvement of thinking-with-images systems may come from changes in reasoning behavior or task alignment, rather than from newly acquired visual inputs alone.

A more direct decomposition is provided by Ma et al. (Ma et al., 2026), which analyzes crop-and-zoom tool-use reinforcement learning through the MED framework. Instead of only evaluating the final tool-available performance, the MED analysis evaluates training checkpoints under both tool-free and tool-available settings, and decomposes the observed performance change into tool-free intrinsic capability drift and tool-induced effects. Figure 8 summarizes this analysis. The reported decomposition shows that a large fraction of the learning progress can be explained by tool-free intrinsic improvement, while the additional tool-induced component is comparatively smaller.

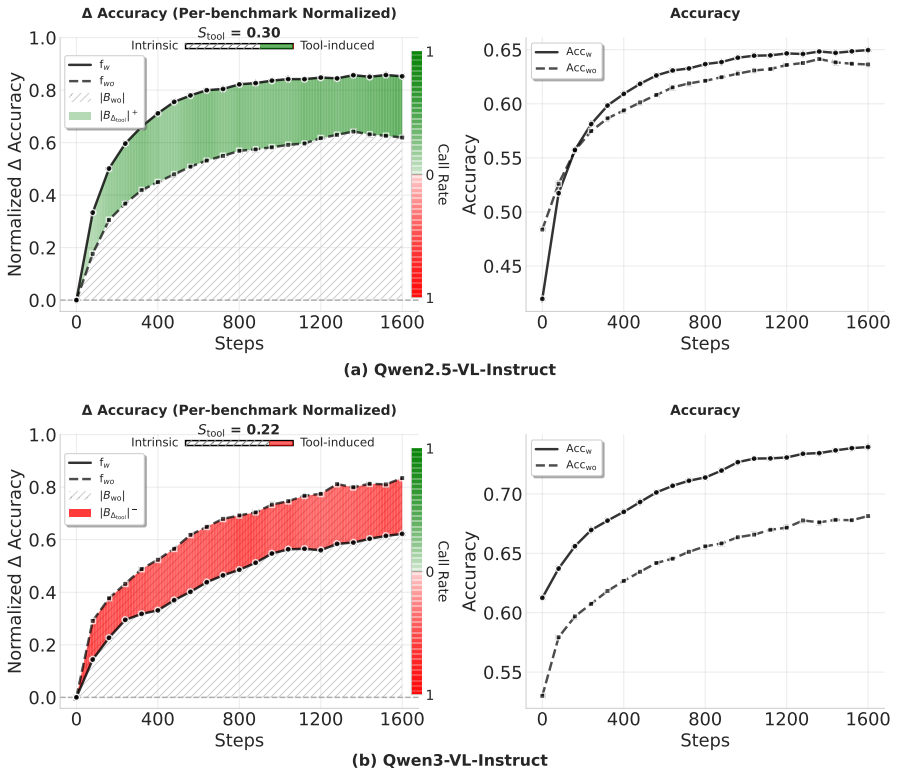


Figure 8: Supporting evidence from the MED analysis of vision tool-use reinforcement learning. The analysis decomposes performance changes into intrinsic tool-free capability drift and tool-induced effects, showing that much of the observed gain can be attributed to intrinsic improvement rather than the additional crop-and-zoom tool effect alone. Figure adapted from Ma et al. (Ma et al., 2026).

These two pieces of evidence are complementary. The DeepEyes ablation indicates that language-side training without additional interleaved images can already produce strong gains, while the MED analysis suggests that even tool-use reinforcement learning may improve performance partly through intrinsic model changes rather than through tool-induced visual reacquisition alone. These findings do not imply that interleaved images or crop-and-zoom tools are unnecessary in all cases. Instead, they show that attributing the gains of visual tool-use systems solely to newly acquired visual evidence is incomplete.

This distinction motivates the design of Thinking-Once. If a substantial part of the improvement in prior systems comes from better use of existing visual representations, improved reasoning behavior, or reduced interference, then high-resolution VQA should not be framed only as a problem of

Method	V* Bench	HRBench-4K	HRBench-8K	MME-Real-Lite
Qwen2.5-VL-7B	76.4	68.1	65.5	44.5
DeepEyes-7B	84.3	72.8	69.3	53.9
w/o Interleaved images	84.3 (0.0)	72.3 (-0.5)	69.5 (+0.2)	53.5 (-0.4)
Pixel-Reasoner-7B	85.3	72.5	68.9	50.0
w/o Interleaved images	83.8 (-1.5)	72.5 (0.0)	67.9 (-1.0)	49.3 (-0.7)
Thyme-7B	83.8	78.3	72.3	53.8
w/o Interleaved images	83.2 (-0.6)	78.5 (+0.2)	72.4 (+0.1)	53.9 (+0.1)
Method	**ChartQA**	**OCRBench**	**LogicVista**	**MathVision**
Qwen2.5-VL-7B	86.1	88.2	46.1	26.6
DeepEyes-7B	86.1	85.3	44.7	26.0
w/o Interleaved images	86.1 (0.0)	85.0 (-0.3)	43.4 (-1.3)	26.6 (+0.6)
Pixel-Reasoner-7B	87.1	82.1	42.1	27.6
w/o Interleaved images	87.1 (0.0)	82.5 (+0.4)	41.6 (-0.5)	27.3 (-0.3)
Thyme-7B	87.7	86.7	50.6	25.7
w/o Interleaved images	87.7 (0.0)	86.8 (+0.1)	50.3 (-0.3)	25.7 (0.0)

Table 5: Ablation study of interleaved images within the visual CoT of "Thinking with Images" models. Values in parentheses indicate the performance change after removing interleaved images relative to the corresponding full model. The results are adapted from Yang et al. (Yang et al., 2026).

Model	V* Bench	HRBench-4K	HRBench-8K
Qwen2.5-VL-7B	71.2	68.8	65.3
RL w. Text-only CoT	88.5	75.4	60.8
DeepEyes (iMCoT)	**90.1**	75.1	**72.6**

Table 6: Ablation on iMCoT from DeepEyes. Text-only CoT training already improves high-resolution VQA performance over the original Qwen2.5-VL-7B baseline, indicating that the benefit of thinking-with-images training is not entirely attributable to additional interleaved visual inputs. Results are adapted from DeepEyes (Zheng et al., 2026).

repeatedly acquiring new visual inputs. In the utilization-limited regime considered by Thinking-Once, the original full-image encoding and intermediate hidden states may already contain useful visual evidence, and the central problem is how to preserve, select, and route this evidence to later reasoning layers within a single visual pass.

B.4 Target-Instance Evidence Is Necessary and Actionable

The main paper argues that high-resolution visual reasoning depends on a small set of question-relevant visual evidence, and that these key evidence tokens should be preserved and routed rather than treated as interchangeable with generic background tokens. Prior work on traceable visual grounded reasoning provides supporting evidence for this view from the input-space perspective. In particular, TreeVGR (Wang et al., 2026) studies whether annotated target instances are genuinely necessary for visual grounded reasoning through two complementary interventions: masking target instances and providing ground-truth bounding boxes as textual evidence hints.

Table 7 reports these two interventions on TreeBench. In the target-masking setting, removing all annotated target instances causes a consistent and substantial performance drop across different MLLMs. This indicates that the target instances are not merely high-quality annotations or convenient visual explanations, but contain indispensable evidence for answering the questions. In other words, the visual information associated with these target instances cannot be reliably replaced by the remaining background or global image context.

Complementarily, the textual-box-hint setting shows that explicitly providing ground-truth bounding boxes as evidence hints improves performance for all evaluated models. This result suggests that making the target evidence explicit helps models focus their reasoning on the relevant visual regions and reduces reliance on global image impressions or language priors. Together, the two interventions provide both a necessity and usefulness signal: removing target evidence hurts performance, while explicitly identifying it improves performance.

Panel A: Masked target instances

Masking	Qwen2.5-VL-7B	InternVL3-8B	GPT-4o	o3	Gemini-2.5-Flash	Gemini-2.5-Pro
–	37.0	38.8	46.9	54.8	45.9	54.1
✓	31.8 (↓ 5.2)	29.6 (↓ 9.2)	29.1 (↓ 17.8)	33.8 (↓ 21.0)	29.9 (↓ 16.0)	33.1 (↓ 21.0)

Panel B: Explicit bounding-box-based textual hints

Textual Boxes	Qwen2.5-VL-7B	InternVL3-8B	GPT-4o	o3	Gemini-2.5-Flash	Gemini-2.5-Pro
–	37.0	38.8	46.9	54.8	45.9	54.1
✓	43.7 (↑ 6.7)	43.5 (↑ 4.7)	49.4 (↑ 2.5)	58.3 (↑ 3.5)	51.9 (↑ 6.0)	61.0 (↑ 6.9)

Table 7: Target-instance evidence on TreeBench. Panel A shows that masking all annotated target instances causes consistent performance degradation across models. Panel B shows that providing ground-truth target boxes as textual evidence hints produces consistent gains. Results are adapted from TreeVGR (Wang et al., 2026).

These results support the central assumption behind our evidence-routing analysis. Although TreeVGR operates at the level of annotated input-space instances, the conclusion naturally connects to token-level routing in MLLMs: visual tokens corresponding to question-relevant target instances are likely to carry high-value reasoning evidence. If such instances are removed, the remaining visual context cannot reliably compensate for the missing evidence; if they are explicitly identified, model reasoning becomes more accurate.

This prior evidence therefore complements our layer-wise oracle intervention in the main paper. The oracle intervention shows that target-region tokens are highly informative inside intermediate model representations, while TreeVGR shows that the corresponding target instances are necessary and actionable in visual grounded reasoning. Together, they motivate Thinking-Once to preserve and route question-relevant core evidence tokens within a single visual pass, rather than applying query-agnostic compression that may discard the visual evidence most responsible for the answer.

C Implementation and Evaluation Details

We provide additional implementation details for Thinking-Once Evidence Routing. Unless otherwise specified, all methods are evaluated under the same high-resolution visual question answering setting.

Evaluation protocol. All methods are evaluated on high-resolution visual question answering benchmarks, including V*Bench, HRBench-4K, and HRBench-8K. For V*Bench, we report attribute accuracy, spatial accuracy, average accuracy, inference time, and GPU memory usage. For HRBench-4K and HRBench-8K, we report FSP, FCP, and average accuracy. Additional supplementary evaluations are conducted on MME-RealWorld-Lite, ZoomBench, TreeBench, and POPE to examine real-world high-resolution perception, fine-grained zoom-oriented perception, traceable visual grounded reasoning, and object hallucination robustness.

C.1 Benchmark and Dataset Details

V*Bench. V*Bench is a high-resolution visual question answering benchmark designed to test whether MLLMs can identify and reason over fine-grained visual evidence in crowded images (Wu & Xie, 2024). In our evaluation, it contains 191 multiple-choice questions, including 115 direct-attribute questions and 76 relative-position questions. The attribute split mainly evaluates whether a model can recognize localized visual properties of small or subtle objects, while the spatial split focuses on relative position and relation judgment. We report Attribute, Spatial, and their average accuracy.

HRBench-4K and HRBench-8K. HRBench is constructed for high-resolution multimodal perception and contains two complementary versions, HRBench-8K and HRBench-4K (Wang et al., 2025). The 8K version preserves high-resolution images with an average resolution around 8K, while the 4K version is obtained by cropping question-relevant target regions from the corresponding 8K images using human annotations. Both versions contain two sub-tasks: Fine-grained Single-instance Perception (FSP), which evaluates localized attributes, OCR, and visual prompting around a single target, and Fine-grained Cross-instance Perception (FCP), which evaluates cross-object or cross-region reasoning such as spatial relationships, map analysis, and chart analysis. We report FSP, FCP, and their average accuracy for both HRBench-4K and HRBench-8K.

MME-RealWorld-Lite. MME-RealWorld is a manually annotated benchmark for evaluating MLLMs in high-resolution real-world scenarios (Zhang et al., 2025c). The full benchmark contains diverse real-world images and question-answer pairs covering multiple practical scenarios and subtasks. We use its lightweight evaluation subset, MME-RealWorld-Lite, for supplementary evaluation. It measures whether a model can perceive fine details and reason over realistic high-resolution content under practical visual conditions. We report Perception, Reasoning, and Overall scores.

ZoomBench. ZoomBench is a fine-grained multimodal perception benchmark introduced together with the Zooming-without-Zooming framework (Wei et al., 2026). It contains 845 hybrid-annotated VQA examples spanning six fine-grained perceptual dimensions and is designed to measure whether a model can recognize small, localized, or detail-sensitive evidence that often motivates zoom-in operations. The benchmark also supports a dual-view protocol for analyzing the gap between global-image understanding and regional visual evidence. We report its multiple-choice score, blank-answer score, and overall score.

TreeBench. TreeBench is a diagnostic benchmark for traceable visual grounded reasoning, introduced by TreeVGR (Wang et al., 2026). It focuses on subtle targets in complex scenes, requires evidence to be traceable through bounding-box annotations, and emphasizes second-order reasoning over object interactions and spatial hierarchies. The benchmark contains 405 challenging VQA pairs annotated by LMM experts. We use TreeBench to evaluate whether evidence routing helps not only answer prediction, but also reasoning over visually grounded and interaction-dependent evidence. We report Perception, Reasoning, and Overall scores.

POPE. POPE is a polling-based benchmark for evaluating object hallucination in large vision-language models (Li et al., 2023b). It converts hallucination evaluation into binary object-existence questions, such as whether a queried object is present in the image. The benchmark constructs negative object queries under three settings: Random, Popular, and Adversarial. These splits test whether a model falsely predicts non-existent objects under increasingly difficult negative sampling strategies. We report the Random, Popular, Adversarial, and average scores.

Benchmark	Main focus	Reported metrics	Evaluation role
V* Bench	Attribute and spatial reasoning in high-resolution scenes	Attr / Spatial / Avg	Main benchmark
HRBench-4K	Cropped 4K fine-grained perception from question-relevant regions	FSP / FCP / Avg	Main benchmark
HRBench-8K	Full high-resolution 8K perception and reasoning	FSP / FCP / Avg	Main benchmark
MME-RealWorld-Lite	Real-world high-resolution perception and reasoning	Perception / Reasoning / Overall	Supplementary benchmark
ZoomBench	Fine-grained zoom-oriented visual perception	MCQ / Blank / Overall	Supplementary benchmark
TreeBench	Traceable visual grounded reasoning with evidence boxes	Perception / Reasoning / Overall	Supplementary benchmark
POPE	Object hallucination under object-existence probing	Random / Popular / Adversarial / Avg	Supplementary benchmark

Table 8: Summary of the evaluation benchmarks used in the main paper and appendix.

Inference settings. Unless otherwise noted, Thinking-Once uses independent top-p routing with $\rho = 0.7$; $L = 15$ is used for Qwen2.5-VL-7B and ZwZ-7B, and $L = 19$ for the remaining models. For background preservation, the visual-token map is uniformly partitioned into an 8×8 grid, and the unselected tokens within each non-empty cell are summarized into one background token through mean pooling.

Training-free setting and inference stages. Thinking-Once Evidence Routing is a training-free inference-time method and requires no additional supervised fine-tuning on the target benchmarks. The term *single-visual-pass* refers specifically to processing and encoding the input image only once; it does not mean that the complete pipeline performs only one model invocation. For each sample, the pipeline sequentially invokes the same MLLM twice. The first invocation operates in text-only mode to extract question-relevant entity mentions from the cleaned question. The second invocation performs multimodal inference, during which the original full-resolution image is encoded once, intermediate-layer evidence routing is applied, and the final answer is generated. Entity extraction therefore introduces an additional text-only model call, but no additional visual input, crop, or image encoding.

Routing query construction. For entity-side routing, the same MLLM used for multimodal routing and answer generation is prompted in text-only mode to extract question-relevant entity mentions. Entity extraction is performed independently for each sample, without batching or caching, and extracted entities are not reused across samples. The resulting mentions are used only to construct internal routing queries; they do not introduce additional visual inputs, crops, or image encodings. Before extraction, we remove the dataset-specific answer-format suffix so that the text-only prompt focuses on the semantic content of the question rather than the multiple-choice output instruction.

```
ENTITY_PROMPT_TEMPLATE =
You are a highly precise language analysis engine. Your sole function is to extract entities
(e.g., objects, people) from a user's question.
Return a single line wrapped in <FINAL_OUTPUT> and </FINAL_OUTPUT>.
Lowercase only. Separate multiple entities with a comma and a space.
Question:

Answer Instruction =
Answer with the option's letter from the given choices directly.
```

Solution for empty entity extraction. If the text-only extractor returns an empty string or the designated no-entity response, we do not disable evidence routing or revert to the unmodified base model. Instead, we deterministically derive fallback entity phrases from the cleaned question stem. For relation templates, we retain the two compared mentions; for attribute templates, we retain the referenced object phrase; and when no template matches, we use the complete cleaned question stem. The resulting phrases are used as entity-side routing queries, while the global question query, instantiated by the last non-punctuation question token, is retained as usual. The same independent

minimum-coverage selection and background summarization are then applied without an additional model call, visual encoding, or learned fallback component. Consequently, an empty extraction degrades to deterministic question-derived routing rather than unconditioned pruning. The *Global only* setting reported in the ablation tables is a deliberately constructed component ablation and is not the default failure-handling path.

Routed token ratio. Table 9 reports the routed visual-token ratios of Thinking-Once on the five base models used in the main experiments. The reported values exclude ablation variants and measure the proportion of visual tokens kept after evidence routing.

Base Model	V*Bench	HRBench-4K	HRBench-8K	Avg.
Qwen2.5-VL-7B	28.37	14.39	22.49	21.75
ZwZ-7B	30.52	16.16	16.93	21.20
Qwen3-VL-8B	23.60	30.05	27.31	26.99
ZwZ-4B	18.76	27.39	23.54	23.23
ZwZ-8B	6.18	13.43	21.96	13.86
Average	21.49	20.28	22.45	21.41

Table 9: Routed visual-token ratios of Thinking-Once. Values indicate the retained visual-token ratio (%) after evidence routing; ablation variants are excluded.

The average routed ratio is 21.41%, meaning that Thinking-Once forwards only about one fifth of the visual-token sequence to the later layers after routing. The benchmark-level averages are also close, ranging from 20.28% on HRBench-4K to 22.45% on HRBench-8K, which suggests that the reduction is not specific to a single benchmark. At the same time, the ratio varies across MLLMs, from 13.86% on ZwZ-8B to 26.99% on Qwen3-VL-8B. This adaptive behavior is consistent with our routing objective: the method does not enforce a fixed compression rate, but preserves the amount of core evidence and compact background context required by the current model and input.

Hardware and efficiency reporting. All benchmark experiments are conducted on NVIDIA A800 GPUs. Efficiency results are reported on V*Bench, with inference time measured in minutes and GPU memory measured in GB; lower values indicate better efficiency. The reported inference time is the end-to-end wall-clock time of the complete evaluation pipeline. For Thinking-Once, it includes both the unbatched and uncached text-only entity-extraction invocation and the subsequent multimodal routing-and-answer invocation, which are executed sequentially using the same MLLM. Thus, the approximately five-minute runtime reported for Thinking-Once represents entity extraction plus multimodal inference over the complete V*Bench, rather than the cost of a single model invocation. Whenever possible, time and memory are measured on a single A800 GPU. For methods that encounter out-of-memory errors on a single A800 but can run with two A800 GPUs, we report measurements from the two-GPU execution. Memory denotes the dataset-averaged peak GPU usage; for a two-GPU run, we first average the peak usage of each GPU over V*Bench and then sum the two GPU-wise averages.

D Supplementary Evaluation Results

This section reports supplementary evaluation results beyond the main V*Bench, HRBench-4K, and HRBench-8K comparisons. Table 10 shows results on MME-RealWorld-Lite, ZoomBench, TreeBench, and POPE using the ZwZ-7B model. These benchmarks complement the main evaluation by covering real-world high-resolution perception, zoom-oriented fine-grained recognition, traceable visual grounded reasoning, and object hallucination robustness.

The results show that Thinking-Once consistently improves the ZwZ-7B model across all four additional benchmarks. The gains are especially visible on MME-RealWorld-Lite and TreeBench, suggesting that evidence routing benefits both general real-world visual reasoning and traceable grounded reasoning. The improvement on POPE further indicates that routing question-relevant evidence does not increase object hallucination; instead, it slightly improves robustness across random, popular, and adversarial splits.

Setting	MME-RealWorld-Lite			ZoomBench			TreeBench			POPE			
	Perception	Reasoning	Overall	MCQ	Blank	Overall	Perception	Reasoning	Overall	Random	Popular	Adv.	Avg
GPT-4o	49.1	42.1	46.4	–	–	–	–	–	–	–	–	–	–
Base	56.8	42.7	51.3	61.9	34.8	54.7	27.0	34.8	31.9	88.6	87.7	87.0	87.8
ViCrop	57.8	42.1	51.7	57.6	35.3	51.7	28.9	33.2	31.6	86.3	85.4	85.1	85.6
Ours	**59.2**	**43.9**	**53.2**	**62.9**	**35.7**	**55.7**	**30.4**	**35.2**	**33.4**	**90.4**	**88.6**	**87.7**	**88.9**
Δ	+2.4	+1.2	+1.9	+1.0	+0.9	+1.0	+3.4	+0.4	+1.5	+1.8	+0.9	+0.7	+1.1

Table 10: Additional benchmark results on ZwZ-7B. We report MME-RealWorld-Lite, ZoomBench, TreeBench, and POPE, with all scores rounded to one decimal place. Δ denotes the absolute percentage-point improvement over Base. Higher values indicate better performance.

Base Model	Setting	V*Bench			HRBench-4K			HRBench-8K			Mean	Time (min)↓	Mem. GB↓
		Attr	Spa.	Avg	FSP	FCP	Avg	FSP	FCP	Avg			
ZwZ-7B	Base	90.4	84.2	88.0	90.0	59.5	74.8	88.5	58.3	73.4	78.7	∼**5**	∼20
	HiDe‡	**94.8**	85.5	**91.1**	**94.8**	58.5	76.6	**93.5**	57.3	75.4	81.0	∼20	∼22
	ViCrop‡	90.4	85.5	88.5	91.0	59.8	75.4	88.3	56.8	72.5	78.8	∼20	∼22
	Ours	90.4	**89.5**	90.1	90.8	**67.3**	**79.0**	89.8	**63.5**	**76.6**	**81.9**	∼**5**	∼**16**

Table 11: Comparison with HiDe and ViCrop under the same ZwZ-7B model. Mean averages the three benchmark averages. ‡ denotes external evidence reacquisition using additional visual inputs or visual re-encoding, whereas Base and Ours use a single visual encoding pass. Boldface indicates the better result between Base and Ours; tied values are both highlighted. Time denotes end-to-end wall-clock runtime on V*Bench. For Ours, it includes the unbatched and uncached text-only entity-extraction invocation and the subsequent multimodal routing-and-answer invocation, both executed using the same MLLM. Memory denotes dataset-averaged peak GPU usage; for two-GPU runs, it is the sum of the two GPU-wise average peak usages.

E Additional Experiments and Qualitative Examples

E.1 Additional Experiments

E.2 Controlled Comparison of Compression-First Methods

Table 11 compares representative training-free methods on ZwZ-7B under the original high-resolution setting, providing complementary evidence on a different base model. We further conduct a controlled comparison on Qwen2.5-VL-7B with `max_pixel`=1280 to reduce differences caused by the input budget and token-retention configuration. All generic pruning and merging baselines retain 33.3% of the visual tokens. Under the same base model, visual-input budget, and retention ratio, every method completes without sample-level OOM. We consequently compare accuracy and end-to-end runtime under matched conditions.

Under this controlled setting, Thinking-Once achieves the highest average accuracy on V*Bench, HRBench-4K, and HRBench-8K, while remaining in the fastest runtime group at approximately three minutes. Generic token-reduction methods occasionally improve individual subsets, but their gains do not transfer consistently across the three high-resolution benchmarks. These results complement Table 11 by showing that the advantage of Thinking-Once is not caused by competing methods failing under the original high-resolution setting or requiring a different device configuration.

E.3 Additional Ablations

Tables 13 and 14 report additional component ablations on ZwZ-7B and Qwen3-VL-8B, respectively. They complement the main ablation study by evaluating whether background context, denoising, entity-only routing, and global-only routing contribute consistently across different MLLMs.

Method	V*Bench			HR-4K			HR-8K			Time
	Attr	Spa.	Avg	FSP	FCP	Avg	FSP	FCP	Avg	
Base	72.2	65.8	69.6	71.0	58.0	64.5	57.3	53.0	55.1	∼**3**
HiPrune	73.0 (↑ 0.8)	71.1 (↑ 5.3)	72.3 (↑ 2.7)	70.5 (↓ 0.5)	58.5 (↑ 0.5)	64.5 (± 0.0)	55.3 (↓ 2.0)	**54.3** (↑ 1.3)	54.8 (↓ 0.3)	∼**3**
VisionZip	72.2 (± 0.0)	71.1 (↑ 5.3)	71.7 (↑ 2.1)	70.8 (↓ 0.2)	57.5 (↓ 0.5)	64.1 (↓ 0.4)	56.5 (↓ 0.8)	52.8 (↓ 0.2)	54.6 (↓ 0.5)	∼4
V2Drop	73.0 (↑ 0.8)	75.0 (↑ 9.2)	73.8 (↑ 4.2)	70.0 (↓ 1.0)	55.0 (↓ 3.0)	62.5 (↓ 2.0)	56.8 (↓ 0.5)	51.0 (↓ 2.0)	53.9 (↓ 1.2)	∼**3**
BTP	64.3 (↓ 7.9)	60.5 (↓ 5.3)	62.8 (↓ 6.8)	69.3 (↓ 1.7)	51.3 (↓ 6.7)	60.3 (↓ 4.2)	56.8 (↓ 0.5)	51.3 (↓ 1.7)	54.0 (↓ 1.1)	∼8
TRIO	70.4 (↓ 1.8)	69.7 (↑ 3.9)	70.2 (↑ 0.6)	69.5 (↓ 1.5)	58.5 (↑ 0.5)	64.0 (↓ 0.5)	57.0 (↓ 0.3)	53.0 (± 0.0)	55.0 (↓ 0.1)	∼**3**
Ours	**75.7** (↑ 3.5)	**76.3** (↑ 10.5)	**75.9** (↑ 6.3)	**74.3** (↑ 3.3)	**61.0** (↑ 3.0)	**67.6** (↑ 3.1)	**58.5** (↑ 1.2)	**54.3** (↑ 1.3)	**56.4** (↑ 1.3)	∼**3**

Table 12: Controlled comparison on Qwen2.5-VL-7B with `max_pixel`=1280. All compression-first pruning and merging baselines retain 33.3% of the visual tokens. Green and red arrows denote absolute percentage-point improvements and decreases relative to Base, respectively; gray denotes no change. The reduced input budget allows all methods to complete without sample-level OOM. Time denotes end-to-end wall-clock runtime on V*Bench in minutes. For Ours, it includes the unbatched and uncached text-only entity-extraction invocation and the subsequent multimodal routing-and-answer invocation, both executed using the same MLLM.

Variant	V*Bench			HR-4K			HR-8K			Mean	Δ
	Attr	Spa.	Avg	FSP	FCP	Avg	FSP	FCP	Avg		
Base	90.4	84.2	88.0	90.0	59.5	74.8	88.5	58.3	73.4	78.7	-3.2
Full	**90.4**	**89.5**	**90.1**	**90.8**	**67.3**	**79.0**	**89.8**	**63.5**	**76.6**	**81.9**	–
w/o bg.	90.4 (± 0.0)	86.8 (↓ 2.7)	89.0 (↓ 1.1)	89.8 (↓ 1.0)	66.8 (↓ 0.5)	78.3 (↓ 0.7)	88.8 (↓ 1.0)	63.3 (↓ 0.2)	76.0 (↓ 0.6)	81.1 (↓ 0.8)	↓ 0.8
Denoise	90.4 (± 0.0)	85.5 (↓ 4.0)	88.5 (↓ 1.6)	90.0 (↓ 0.8)	64.3 (↓ 3.0)	77.1 (↓ 1.9)	88.8 (↓ 1.0)	62.0 (↓ 1.5)	75.4 (↓ 1.2)	80.3 (↓ 1.6)	↓ 1.6
Entity only	90.4 (± 0.0)	82.9 (↓ 6.6)	87.4 (↓ 2.7)	89.5 (↓ 1.3)	64.5 (↓ 2.8)	77.0 (↓ 2.0)	88.3 (↓ 1.5)	62.8 (↓ 0.7)	75.7 (↓ 0.9)	80.0 (↓ 1.9)	↓ 1.9
Global only	88.7 (↓ 1.7)	85.5 (↓ 4.0)	87.4 (↓ 2.7)	90.0 (↓ 0.8)	60.5 (↓ 6.8)	75.3 (↓ 3.7)	87.8 (↓ 2.0)	58.3 (↓ 5.2)	73.0 (↓ 3.6)	78.6 (↓ 3.3)	↓ 3.3

Table 13: Component ablations on ZwZ-7B. Green and red arrows denote absolute percentage-point improvements and decreases relative to Full, respectively; gray denotes no change. Mean averages the three benchmark averages, and Δ reports the corresponding Mean change relative to Full.

E.4 Qualitative Examples

This subsection provides qualitative examples showing how Thinking-Once routes question-relevant evidence in high-resolution scenes. The enlarged regions and heatmaps in the figures are used only for visualization; the method itself does not introduce crops, re-encoded sub-images, or a second visual input during inference.

Figure 9 shows a single-object attribute case. The base model fails to identify the color of a small motorcycle in a crowded high-resolution image. Thinking-Once first extracts the queried entity "motorcycle" from the question, uses it as an internal routing query, and localizes the corresponding evidence at intermediate layers. By routing this entity-centered evidence to later layers, the model answers the attribute question correctly.

Figure 10 shows a multi-object relational case. The question requires comparing the positions of two entities, namely the motorcycle and the dog. Thinking-Once extracts both entities and performs query-wise evidence routing so that each object can retain its own core evidence. This supports cross-instance spatial reasoning and leads to the correct relative-position answer.

The following qualitative examples further visualize layer-wise evidence localization from Layer 1 to Layer 28. They illustrate how question-guided responses evolve across depth and become more concentrated on the queried visual evidence in intermediate layers, supporting the evidence-routing window discussed in the main paper.

These qualitative examples show that question-relevant regions are not uniformly localized across all layers. Early layers preserve visual details but may lack stable question-conditioned focus, while late layers can absorb the evidence into downstream hidden states. The clearest localization typically

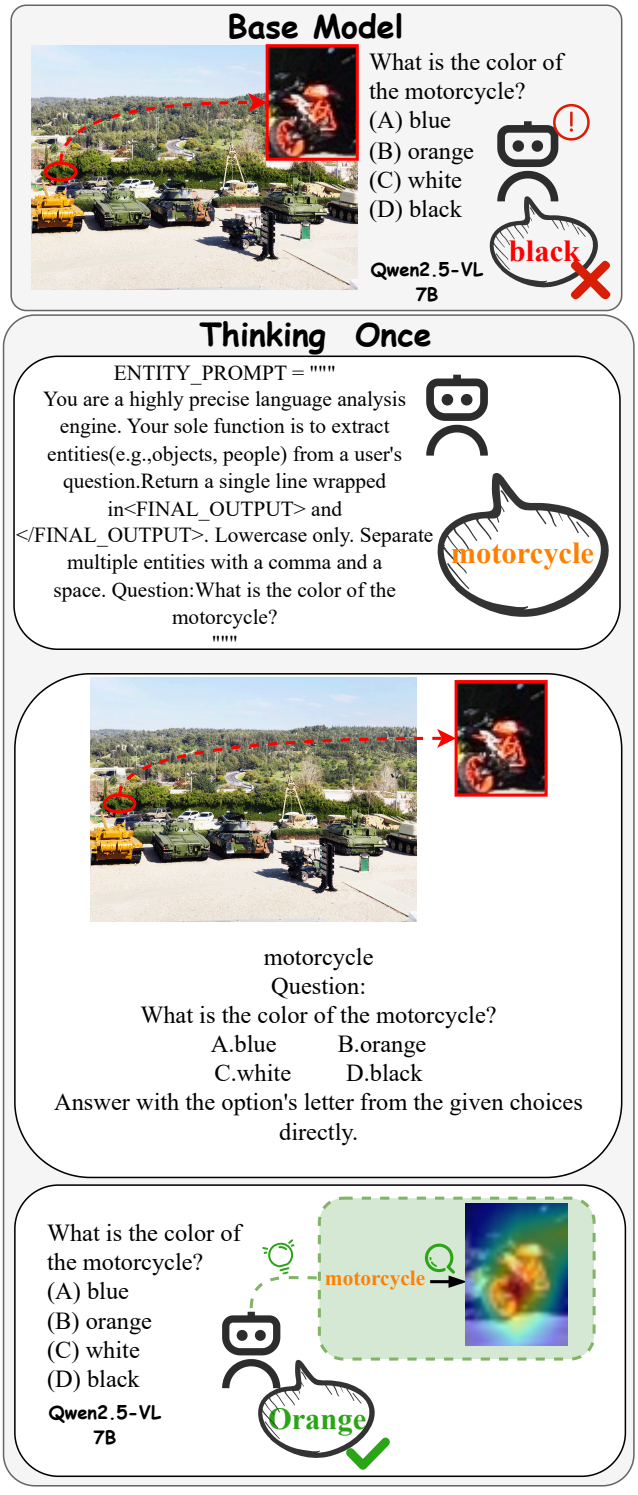


Figure 9: Single-object attribute case. The base model predicts the wrong color for a small motorcycle in a high-resolution scene, while Thinking-Once extracts the entity query, localizes the relevant evidence internally, and routes it to later layers for the correct answer. The zoomed patch and heatmap are visualization aids rather than additional visual inputs.

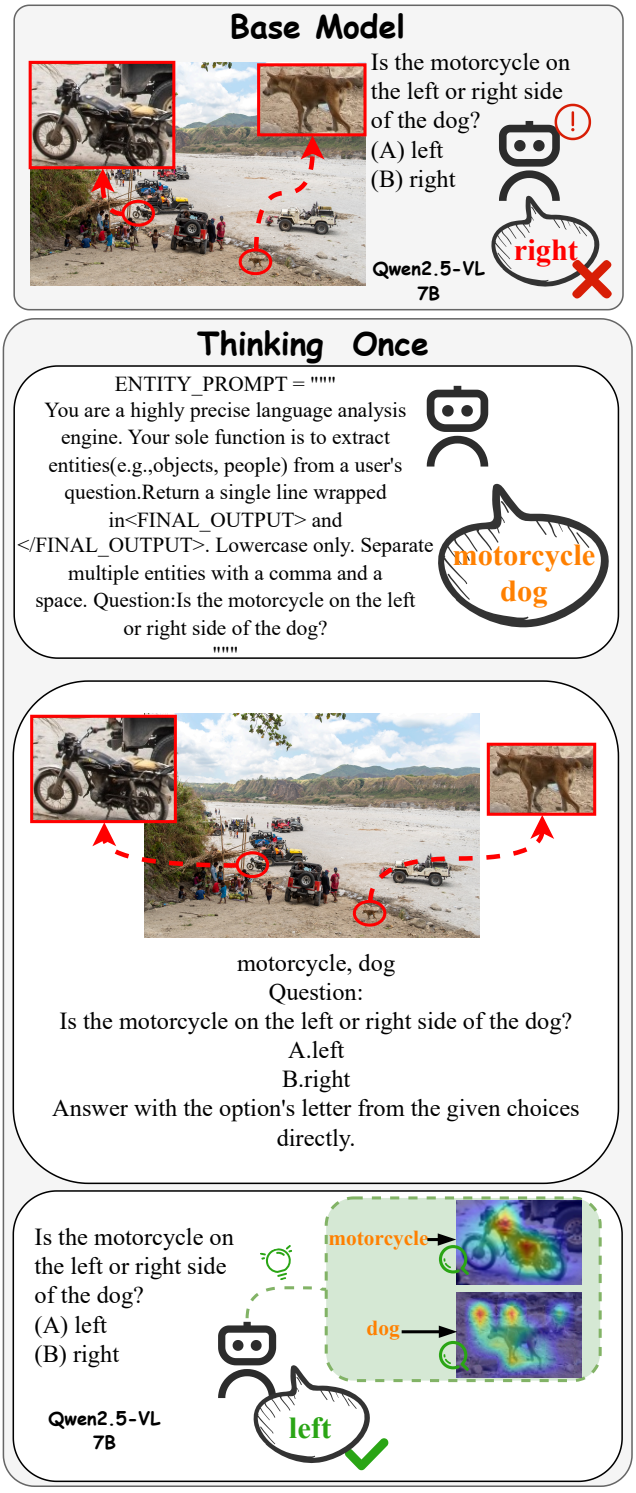


Figure 10: Multi-object relational case. The base model gives an incorrect spatial answer, whereas Thinking-Once extracts both queried entities, routes their question-relevant evidence independently, and preserves the evidence needed for relative-position reasoning. The visualized patches and heatmaps are for explanation only and are not used as extra inputs.



Figure 11: Layer-wise qualitative example of question-guided evidence localization. The visualization shows how the model response to the queried entity evolves across layers and becomes concentrated around the relevant visual evidence.

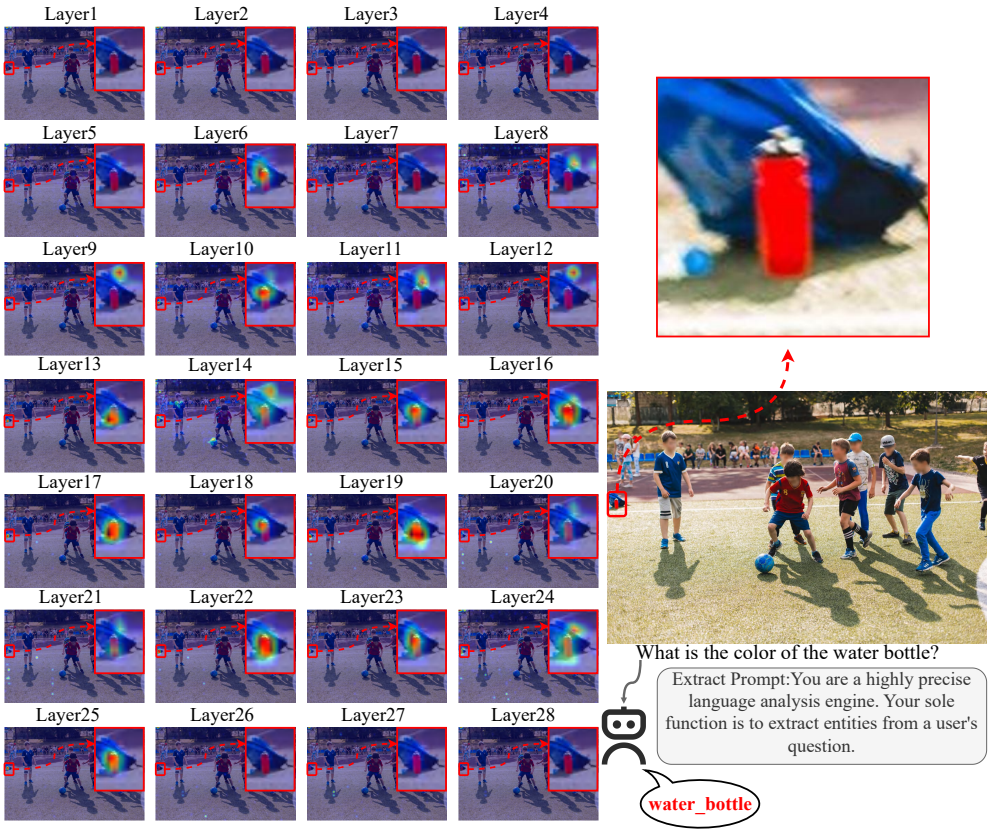


Figure 12: Additional layer-wise qualitative example of question-guided evidence localization. The example further illustrates that intermediate layers can expose question-relevant visual evidence before final answer generation.

Variant	V*Bench			HR-4K			HR-8K			Mean	Δ
	Attr	Spa.	Avg	FSP	FCP	Avg	FSP	FCP	Avg		
Base	84.4	80.3	82.7	91.3	65.8	78.5	84.0	62.8	73.4	78.2	-2.0
Full	**87.8**	**81.6**	**85.3**	**91.8**	**70.5**	**81.1**	**84.3**	**64.3**	**74.3**	**80.2**	–
w/o bg.	84.4 (↓ 3.4)	80.3 (↓ 1.3)	82.7 (↓ 2.6)	91.5 (↓ 0.3)	67.3 (↓ 3.2)	79.4 (↓ 1.7)	83.8 (↓ 0.5)	63.3 (↓ 1.0)	73.5 (↓ 0.8)	78.5 (↓ 1.7)	↓ 1.7
Denoise	84.4 (↓ 3.4)	80.3 (↓ 1.3)	82.7 (↓ 2.6)	91.3 (↓ 0.5)	66.0 (↓ 4.5)	78.6 (↓ 2.5)	84.0 (↓ 0.3)	62.8 (↓ 1.5)	73.4 (↓ 0.9)	78.2 (↓ 2.0)	↓ 2.0
Entity only	87.0 (↓ 0.8)	80.3 (↓ 1.3)	84.3 (↓ 1.0)	91.3 (↓ 0.5)	66.3 (↓ 4.2)	78.8 (↓ 2.3)	84.3 (± 0.0)	63.8 (↓ 0.5)	74.0 (↓ 0.3)	79.0 (↓ 1.2)	↓ 1.2
Global only	82.6 (↓ 5.2)	73.7 (↓ 7.9)	79.1 (↓ 6.2)	91.3 (↓ 0.5)	65.0 (↓ 5.5)	78.1 (↓ 3.0)	81.8 (↓ 2.5)	63.3 (↓ 1.0)	72.5 (↓ 1.8)	76.6 (↓ 3.7)	↓ 3.7

Table 14: Component ablations on Qwen3-VL-8B. Green and red arrows denote absolute percentage-point improvements and decreases relative to Full, respectively; gray denotes no change. Mean averages the three benchmark averages, and Δ reports the corresponding Mean change relative to Full.

appears in intermediate layers, which is consistent with the evidence-routing window identified by the oracle intervention analysis in the main paper.

F Extended Analysis

This appendix expands the evidence-allocation analysis behind Thinking-Once. The main paper presents the core empirical observations: visual information can survive encoding, entity-centered tokens have high causal value, useful routing is layer-dependent, and non-core tokens may still provide necessary context. Here we first define the cross-layer evidence-retention diagnostic used in the main paper, introduce a layer-wise object-attention enrichment analysis that directly measures whether target-centered attention survives deeper processing, and provide a matched-budget analysis of the two central routing choices. We then connect the broader observations more explicitly and state the boundaries under which single-visual-pass evidence routing is expected to help.

F.1 Cross-Layer GT-Evidence Retention

We quantify how much downstream attention remains allocated to the same routed GT evidence after the routing layer. For sample s, let G_s be the set of original visual tokens associated with its annotated GT region, let $S_{\text{core},s}$ be the core selected by Thinking-Once, and define the fixed target set

$$T_s = G_s \cap S_{\text{core},s}. \tag{36}$$

The same target-token identities are used for both Base and Thinking-Once. For Thinking-Once, their indices are remapped after sequence routing while their original spatial identities are retained. Background-summary tokens are included in the routed visual sequence but are not counted as members of T_s.

At layer $\ell \geq L$, we reconstruct the complete causal attention row of the global-question routing query. Let $\mathcal{K}^{\ell}_{c,s}$ be the key positions present in condition $c \in \{\text{Base}, \text{TO}\}$ at that layer and let $z^{\ell}_{c,s}(i)$ be the corresponding attention logit. The layer-wise retention mass is

$$M^{\ell}_{c,s} = \sum_{i \in T^{\ell}_{c,s}} \frac{\exp z^{\ell}_{c,s}(i)}{\sum_{k \in \mathcal{K}^{\ell}_{c,s}} \exp z^{\ell}_{c,s}(k)}, \tag{37}$$

where $T^{\ell}_{c,s}$ denotes the layer-specific indices of the fixed target tokens. Thus, the metric tracks the share of the query's complete causal attention assigned to the same full-resolution GT-core evidence, rather than changing the target region independently for the two conditions.

For the cross-model visualization, we first average $M^{\ell}_{c,s}$ over valid samples within each MLLM. Both trajectories of a model are then divided by that model's maximum Base mass, and post-routing depth is mapped to $d = (\ell - L)/(D - L)$. We linearly interpolate the trajectories to a common depth grid before averaging across Qwen2.5-VL-7B, ZwZ-7B, and Qwen3-VL-8B. The plotted three-point weighted moving average is used only for readability. To test whether retention is associated with

task success, we compute each sample's mean downstream Thinking-Once-minus-Base mass, standardize it within model, and compare correct ($n = 505$) and incorrect ($n = 68$) answers using a two-sided Mann–Whitney test over all 573 model–sample pairs ($U = 23{,}793$, $p \approx 2.4 \times 10^{-7}$, rank-biserial $r = 0.3857$).

Because Thinking-Once changes the set of keys after routing, Eq. equation 37 measures relative attention allocation within each condition; it should not by itself be interpreted as a token-count-invariant causal effect. We therefore use it as a mechanism diagnostic and rely on the matched-budget controls below to separate evidence identity and context organization from the number of retained tokens.

F.2 Layer-Wise Object-Attention Enrichment

The fixed-token retention metric above tracks attention assigned to the same full-resolution GT-core tokens after routing. We additionally introduce a spatial enrichment diagnostic that asks a complementary question: at each language-model layer, how concentrated is the model's visual attention inside the annotated target region relative to the fraction of the visual grid occupied by that region? This diagnostic is computed on all 191 V*Bench examples with ground-truth object bounding boxes. Suppressing the sample index for readability, the enrichment at language-model layer l is

$$E_l = \frac{\sum_{i \in B} a_{l,i} \big/ \sum_{i \in V} a_{l,i}}{|B|/|V|}, \tag{38}$$

where V is the set of original visual tokens, $B \subseteq V$ contains the visual tokens overlapping the ground-truth bounding box, and $a_{l,i}$ denotes the fused routing-query-to-visual attention assigned to token i. Thus, $E_l = 1$ corresponds to spatially uniform visual attention, whereas $E_l > 1$ indicates preferential concentration inside the target region. For example, $E_l = 100$ means that the attention density inside the target box is 100 times the density expected under a uniform spatial allocation; it does not mean that the box receives 100% of the total visual attention. We compute enrichment separately for each example and report the sample mean, with shaded regions showing ± 1 standard error of the mean. Because the values span several orders of magnitude, the vertical axis is logarithmic.

Controlled routing comparison. The Base and Thinking-Once runs use exactly the same image, cleaned question, routing prompt, and routing-query tokens. The only difference is that sequence routing is disabled for the Base control. Their trajectories therefore overlap exactly before the routing point; in Figure 13, the red Base curve is visually occluded by the green Thinking-Once curve in these layers. The dashed lines and yellow markers denote the outputs of the routing layers: L15 for Qwen2.5-VL-7B and L19 for Qwen3-VL-8B. Consistent with the layer convention used throughout this appendix, routing is applied *after* the marked layer, so the first affected layers are L16 and L20, respectively.

For this diagnostic, Thinking-Once uses independent top-p routing with $\rho = 0.7$ to select full-resolution core evidence and retains global scene context through background summaries constructed over an 8×8 spatial grid. The compact representation is propagated through the remaining language-model layers without cropping, image re-encoding, or a second visual forward pass. To evaluate post-routing spatial attention on the original grid, attention over the compact representation is projected back in a mass-preserving manner: each retained core token is assigned to its original location, and each background-summary token is redistributed only over its original source set U_g, namely the unselected visual tokens summarized within the corresponding spatial cell, while preserving its total attention mass. Ground-truth boxes are used only to compute Eq. equation 38; they are never used to select routed tokens.

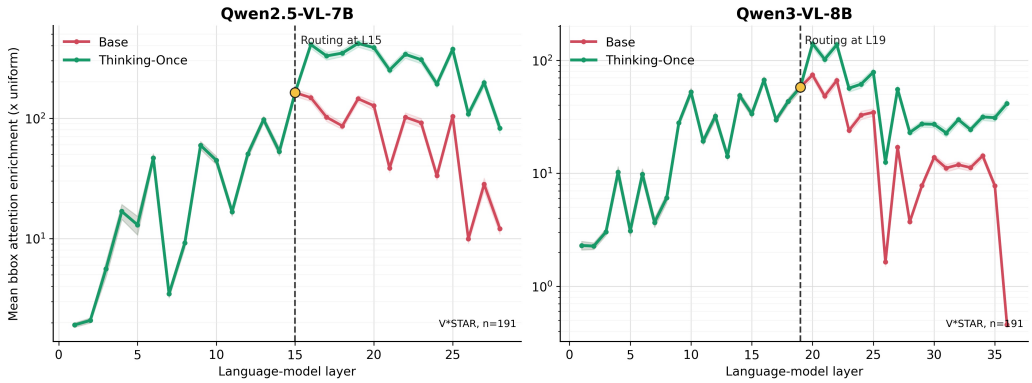


Figure 13: Layer-wise object-attention enrichment on V*Bench. The two panels compare Base and Thinking-Once on all 191 examples with ground-truth object boxes. Enrichment normalizes the fraction of visual attention inside the target box by the fraction of visual tokens covered by that box, so $1\times$ denotes uniform spatial attention and values above $1\times$ denote target-centered concentration. Curves show sample means, shaded regions indicate ± 1 standard error, and the vertical axis is logarithmic. Base and Thinking-Once use identical inputs and routing queries, causing exact pre-routing overlap. Dashed lines and yellow markers indicate routing after L15 for Qwen2.5-VL-7B and after L19 for Qwen3-VL-8B; the first affected layers are therefore L16 and L20. Post-routing attention is projected from the compact core-plus-background representation back to the original visual grid in a mass-preserving manner, with each background summary redistributed only over its original source set U_g. Ground-truth boxes are used only for evaluation, not for routing.

Observed retention. For Qwen2.5-VL-7B, both conditions reach an enrichment of $163.0\times$ at L15. Immediately after routing, Thinking-Once increases the enrichment to $408.3\times$ at L16, compared with $148.1\times$ for Base. The Base model then loses much of its target-centered attention and ends at $12.1\times$ at L28, whereas Thinking-Once retains $82.5\times$, corresponding to a $6.8\times$ final-layer improvement. Qwen3-VL-8B exhibits the same pattern. Both conditions reach $57.9\times$ at L19; at L20, Thinking-Once reaches $140.4\times$, compared with $74.1\times$ for Base. At the final layer, Base falls to $0.46\times$, below the uniform-attention reference, while Thinking-Once maintains $41.5\times$, a $90.5\times$ relative improvement.

Interpretation. Figure 13 provides mechanistic evidence for the motivation behind Thinking-Once. In both model families, substantial bounding-box enrichment has already emerged at intermediate layers, showing that the relevant visual evidence is not necessarily absent or imperceptible. The limitation of the Base model is instead that this object-centered allocation deteriorates and fluctuates strongly during deeper processing. Thinking-Once intervenes precisely when object-specific evidence is accessible, converts the visual sequence into high-confidence core tokens plus compact global context, and routes this representation through the remaining layers. The persistent post-routing separation between the green and red curves indicates that compact evidence routing improves the retention of target-centered visual information across architectures.

Importantly, the curves diverge only after routing. This controlled temporal separation rules out differences in the input image, prompt, query construction, or pre-routing hidden states as explanations for the effect. The figure should therefore be interpreted as evidence for *attention retention and evidence utilization*, rather than as a direct accuracy measurement. Together with the answer-accuracy results and the fixed-token retention diagnostic above, it supports the central claim that Thinking-Once prevents already encoded visual evidence from being diluted or displaced before answer generation.

F.3 Matched-Budget Evidence Coverage Analysis

The main method combines two allocation decisions: how to select the full-resolution core evidence and how to preserve compact contextual support around that core. Because either decision can change the number of routed tokens, an uncontrolled comparison could attribute improved evi-

dence coverage to a larger token budget rather than to a better routing rule. We therefore construct two matched-budget experiments that isolate the effects of query selection and context organization while keeping the corresponding token counts fixed.

Metrics. For each sample, let G_s denote the visual tokens associated with the annotated ground-truth evidence region, and let S_s denote the selected full-resolution core tokens. We measure GT-evidence recall as

$$\mathrm{Recall}_{\mathrm{GT}} = \frac{1}{|\mathcal{D}|} \sum_{s \in \mathcal{D}} \frac{|S_s \cap G_s|}{|G_s|}. \tag{39}$$

To measure whether the retained context remains spatially distributed, we partition the original visual-token map into the same regular 8×8 grid used by Thinking-Once. Spatial-context coverage is the average percentage of grid cells represented by at least one retained or summarized context element. This metric evaluates spatial support rather than task accuracy, and is used only to diagnose how different allocations distribute the same context budget.

Matched core-token budget. For every sample s, Independent first performs minimum-coverage selection separately for each routing query and forms the union of the resulting sets. Its resulting size K_s is then used as the exact per-sample core-token budget for all query-selection variants. Thus, every variant returns exactly K_s core tokens, and differences in GT-evidence recall cannot be explained by retaining more tokens. The variants are defined as follows:

- **Random** uniformly samples K_s visual tokens without using the question.
- **Global** ranks visual tokens using the global question query and retains its top-K_s tokens.
- **Average** averages the normalized evidence distributions of all routing queries and retains the top-K_s tokens under the aggregated distribution.
- **Max** takes the element-wise maximum across the normalized routing-query distributions and retains the top-K_s tokens.
- **Independent** performs minimum-coverage selection for each routing query separately and takes the union, as used by Thinking-Once.

Matched total-token budget. For the context experiment, the core set is held fixed across all variants, with an average of 233.42 core tokens per sample. Each sample uses the same matched context budget across variants, which differ only in how that context budget is allocated:

- **Extra Core** assigns the additional budget to the highest-ranked non-core tokens, without explicitly preserving spatial distribution.
- **Random Context** samples non-core context tokens uniformly at random.
- **Uniform Context** selects non-core tokens at approximately uniform spatial intervals over the visual-token map.
- **Grid Background** partitions the visual-token map into an 8×8 grid and summarizes the unselected tokens within each spatial cell, as used by Thinking-Once.

Table 15 shows that independent routing provides substantially more complete access to the annotated evidence than query-agnostic or query-aggregated alternatives. Random and Global retain only 4.87% and 4.66% of the GT evidence, respectively, indicating that a single generic selection signal is insufficient for questions with multiple semantic demands. Average and Max recover much more evidence, but their aggregation can still suppress tokens that are important to only one entity or one question aspect. Independent reaches 85.26% recall, improving over the stronger aggregated baseline Average by 14.13 percentage points. This result supports the per-query coverage constraint in Eq. equation 15: each routing query should preserve its own evidence before the selected sets are merged.

The context comparison exhibits a complementary pattern. Extra Core covers 80.61% of the spatial cells, showing that assigning more tokens to already salient regions does not guarantee broad contextual support. Random Context and Uniform Context increase coverage to 89.58% and 92.37%, respectively, but can still leave parts of the visual field unsupported. Grid Background reaches

Group	Variant	Coverage (%)
Query routing GT-evidence recall	Random	4.87
	Global	4.66
	Average	71.13
	Max	69.22
	Independent	**85.26**
Context allocation Spatial coverage	Extra Core	80.61
	Random Context	89.58
	Uniform Context	92.37
	Grid Background	**100.00**

Table 15: Matched-budget evidence-coverage analysis. Query-routing variants use the same per-sample core-token count, while context-allocation variants use the same fixed core set and total routed-token budget. Higher values indicate more complete coverage of the corresponding evidence structure.

100.00% coverage under the same total budget, exceeding Uniform Context by 7.63 percentage points. The gain therefore comes from organizing context according to the image's spatial structure, rather than from increasing the number of routed tokens.

Together, the two controlled comparisons isolate the roles of the method's two main components. Independent minimum-coverage routing protects evidence associated with distinct semantic queries, while structured grid summaries preserve compact support across the full visual field. Their advantages under matched budgets show that Thinking-Once benefits from better evidence allocation, not merely from retaining more tokens.

F.4 Pixel-Level Evidence Restoration Translates into Accuracy

The coverage analysis above measures which tokens are retained, and the enrichment diagnostic measures where attention is allocated, but neither directly ties the visual content of the target region to the final answer. We therefore conduct a pixel-level restoration experiment that asks a complementary question at the input level: with the full image and every pixel outside the annotated target boxes kept intact, does gradually restoring the pixels inside the target region increase the probability of answering correctly? This experiment tests the role of target-region visual evidence; it is not a causal experiment linking predicted boxes or IoU to accuracy.

Setup. The experiment covers all 191 V*Bench questions (115 attribute and 76 relative-position) with Qwen2.5-VL-7B-Instruct. For each image, the pixels inside the ground-truth boxes are first replaced by the median RGB color computed over the pixels outside the boxes; the original pixels are then progressively restored from the center of each box. We evaluate five restoration levels (0%, 25%, 50%, 75%, and 100%). When a question involves two boxes, both boxes are processed at the same level. The 0% level masks only the target regions rather than the whole image, and the 100% level is pixel-identical to the decoded original image. Ground-truth coordinates are used only to construct the image variants and are never provided to the model as hints or coordinate inputs. All five levels share the same questions, options, and inference configuration (bf16, FlashAttention-2, greedy decoding, answers constrained to valid option letters), and only full-image inference results are counted, without cropping or routing.

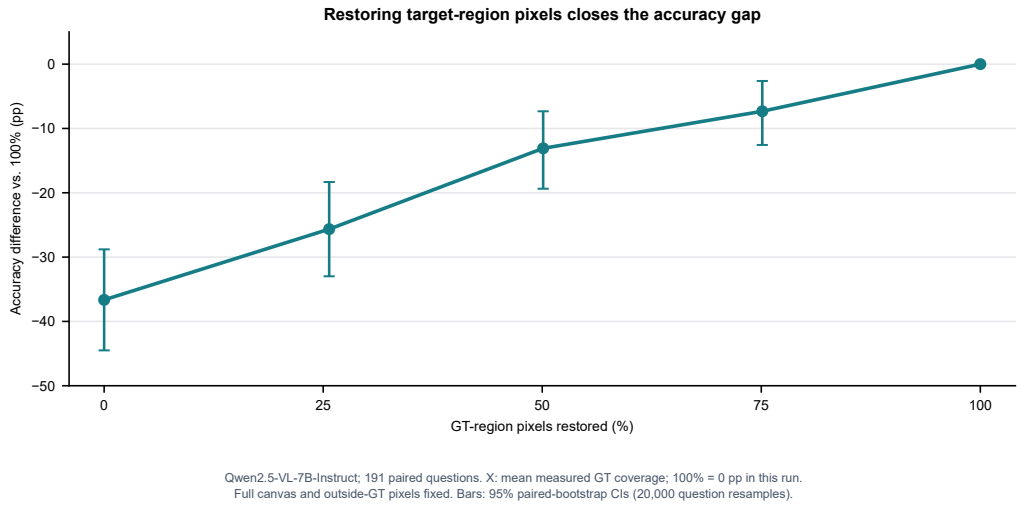


Figure 14: Accuracy on V*Bench as a function of the fraction of restored ground-truth-region pixels. The horizontal axis denotes the fraction of target-region pixels restored from the masked variants, and the vertical axis denotes the per-question paired accuracy difference relative to the 100% restoration condition of the same run. Error bars are computed from 20,000 question-level bootstrap resamples over the 191 questions.

Results. Figure 14 reports the per-question paired accuracy difference relative to the 100% condition of the same run. Accuracy increases from 80/191 correct answers when the target regions are fully masked to 150/191 when they are fully restored, an overall improvement of 36.6 percentage points, with the intermediate restoration levels shown in the figure.

Interpretation. This experiment connects evidence availability to answer accuracy at the input level and provides the pixel-level counterpart of the Remove-Oracle necessity signal in the main paper: when the evidence-bearing region is masked while the surrounding context is fully preserved, the remaining image is not sufficient to answer, and accuracy recovers as the region content returns. Two caveats qualify this reading. First, the median-color fill leaves a silhouette of the masked region, which may hint at the target location; second, restoration proceeds from the box center outward, which may expose the most informative features earlier than a uniform restoration order would. The curve should therefore be read as evidence that target-region content carries answer-determining information, rather than as a causal estimate relating predicted-box IoU to accuracy.

F.5 Complementary Evidence that Useful Information Already Survives Encoding

The main text uses LongCat-Next (Team, 2026) to show that lightweight decoders can reconstruct image structure, object boundaries, and spatial layout from frozen vision-encoder outputs. This does not prove that every high-resolution question is answerable from the encoded representation. Instead, it establishes a narrower but necessary condition for single-visual-pass routing: local visual information can survive visual encoding and remain available to later multimodal layers.

Evidence from interleaved visual reasoning and vision tool-use training supports the same distinction from another direction. Yang et al. (Yang et al., 2026) show that removing interleaved images from several visual CoT systems often causes only small performance changes, suggesting that intermediate images are not always the dominant source of the observed gain. DeepEyes further shows that text-only chain-of-thought training can already improve high-resolution VQA over the base model (Zheng et al., 2026), indicating that part of the improvement may come from reasoning alignment or task adaptation rather than newly acquired visual observations alone. Similarly, Ma et al. (Ma et al., 2026) decompose crop-and-zoom tool-use reinforcement learning and find that a large part of the training progress can be attributed to tool-free intrinsic improvement.

These findings do not imply that crops, zoomed views, or visual tools are unnecessary. They instead show that performance gains from such systems should not be attributed solely to additional visual evidence. In many cases, improved performance may also come from better use of information already present in the original visual encoding or in intermediate hidden states. This is consistent with recent evidence that large models may internally encode relevant evidence even when it is not effectively surfaced in the final prediction (Liu et al., 2025b). Thinking-Once is designed for this utilization-limited regime: the required evidence is at least partially encoded, but it is diluted, misallocated, or discarded before answer generation.

F.6 From Oracle Interventions to an Evidence-Routing Window

The layer-wise oracle interventions in the main paper separate three factors that are otherwise entangled: token budget, token identity, and intervention depth. *Random Tokens* preserve the same number of tokens as the entity-region tokens, *Oracle Core Tokens* preserve only tokens inside the ground-truth box, and *Remove Oracle Tokens* preserves all tokens except those inside the box. This design tests whether performance depends merely on using fewer tokens, or on preserving the causal identity of the evidence.

The comparisons provide two complementary signals. The advantage of Oracle Core Tokens over Random Tokens shows a sufficiency signal: a compact entity-centered subset can support strong reasoning when it contains the relevant evidence. The failure of Remove Oracle Tokens gives a necessity signal: abundant non-target information cannot reliably replace evidence removed from the entity region. Because Random Tokens use the same budget as Oracle Core Tokens, and Remove Oracle Tokens keeps substantially more tokens, the result cannot be explained by sequence length alone. What matters is which tokens are routed and when their influence is redirected.

The layer dependence of these interventions further defines an *evidence-routing window*. Early layers may preserve fine visual detail but lack stable question-conditioned correspondence. Late layers contain richer multimodal states, but too little computation remains for a routing operation to change the reasoning trajectory. Intermediate layers provide the useful overlap: question-relevant evidence has become identifiable, while enough downstream computation remains for the routed evidence to affect the answer. The earliest oracle optimum should therefore not be interpreted as the practical routing layer, because ground-truth boxes provide perfect localization that is unavailable at inference time. A usable routing layer must jointly provide reliable question-conditioned localization and sufficient remaining depth.

F.7 Why the Evidence Set Must Preserve Context

Entity tokens form the evidence core, but the core alone is not always sufficient. The background-masking analysis in the main paper shows a non-monotonic effect: moderate background removal can improve accuracy, while aggressive masking degrades it. This indicates that background tokens contain a mixture of interference and support. Some background tokens are redundant or harmful, but others encode spatial references, relational cues, or attribute-comparison evidence needed to interpret the target entity.

This observation explains why object-only cropping can be brittle in HR-VQA. A question may refer to a single entity while asking about its relation to another object, its relative position, or an attribute comparison whose evidence lies outside the entity box. Removing all surrounding tokens preserves the referent but can destroy part of the structure required to answer the question. Thus, the routing objective should not be to keep the smallest possible entity-only crop. It should preserve the smallest evidence structure that still supports the question.

Attention-sink interventions reveal a related form of context. Although attention-sink tokens can distort attention-based localization, suppressing them reduces accuracy across models and benchmarks in the main analysis. This shows that a token that is spatially diffuse or visually non-discriminative can still carry useful hidden-state information, such as aggregation or information transfer signals. Together, background masking and sink suppression indicate that effective HR-VQA evidence is structured rather than merely sparse: it consists of a dense entity-centered core plus a compact support system that preserves explicit and implicit context.

F.8 Applicability Boundaries and Falsifiability

The evidence-allocation account is intended for a utilization-limited regime, not for all HR-VQA failures. Routing cannot recover details that the vision encoder has irreversibly removed, nor can it correct an initial encoding that never represents the relevant region. It may also provide limited benefit when the base model already allocates evidence reliably, or when the question requires new visual observations outside the original input. In such cases, cropping, zooming, or multi-round visual search remains appropriate. Thinking-Once is therefore complementary to evidence reacquisition methods: it addresses failures caused by inefficient preservation and allocation of available evidence, rather than failures caused by absent evidence.

This account makes several testable predictions. First, routing should help more on questions involving relations, comparisons, or cross-instance context than on questions solvable from an isolated attribute, because such questions require both a precise entity anchor and supporting context. Second, gains should be larger for crowded high-resolution inputs, where relevant evidence competes with many irrelevant tokens and is more vulnerable to dilution. Third, interventions should be most effective at intermediate layers, after question-conditioned correspondence has emerged but before downstream reasoning has become too close to the output. Fourth, removing all context or suppressing sink states should hurt performance when those tokens carry useful support signals.

The account would be weakened if these patterns consistently failed to appear. For example, if routing mainly improved isolated attribute questions, if gains were unrelated to visual clutter, if query-agnostic pruning at arbitrary layers matched the routing gains, or if context removal never hurt performance, then the proposed evidence-allocation explanation would be insufficient. These criteria make the analysis not only a motivation for Thinking-Once, but also a falsifiable description of when and why single-visual-pass routing should work.