LOOK TWICE: TRAINING-FREE EVIDENCE HIGHLIGHTING FOR KNOWLEDGE-BASED VISUAL QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge-based Visual Question Answering (KB-VQA) requires Multimodal Large Language Models (MLLMs) to identify and combine fine-grained visual cues with retrieved textual evidence. However, retrieval often introduces noisy and partially relevant content, while images contain distracting visual regions, causing pretrained MLLMs to overlook the evidence that actually supports the answer. To address this, we introduce Look Twice (LoT), a training-free inference-time framework that turns the model's own internal attention into an explicit multimodal evidence-selection mechanism. LoT first leverages the model's internal attention patterns to identify query-relevant image regions and textual sentences, filters attention sinks and distracting content, and reformulates the input to explicitly highlight the selected evidence before answer generation. The method requires no parameter updates, auxiliary models, or architectural modifications. Across four KB-VQA benchmarks and ten off-the-shelf MLLMs ranging from 2B to 38B parameters, LoT improves every evaluated backbone, with average gains of up to $+12.5$ accuracy points. It also provides further gains when combined with established context-refinement strategies, yielding additional improvements over already refined inputs. These results establish LoT as a general and effective mechanism for enabling pretrained MLLMs to exploit available multimodal evidence more accurately.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) (Caffagni et al., 2024a; Liu et al., 2024a) integrate LLMs (Zhao et al., 2023) with visual encoders (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025) to jointly process images and text through a unified generative interface, enabling applications such as visual dialogue and open-ended question answering.

A particularly challenging scenario arises when answering a question that requires information that is not directly observable in the image. Many real-world queries depend on external knowledge, domain-specific facts, or contextual information that can not be reliably stored in the model parameters. Knowledge-based Visual Question Answering (KB-VQA) addresses this challenge by augmenting the model input with passages retrieved from external knowledge sources (Mensink et al., 2023; Chen et al., 2023). In this setting, the model must reason jointly over the visual input and the retrieved textual elements to generate the final answer.

The key challenge, however, is not merely retrieving more information, but identifying and effectively exploiting the evidence that supports the answer. Retrieved contexts often contain noisy, redundant, or only partially relevant passages (Caffagni et al., 2024b), which can distract the model when directly appended to the prompt. Existing approaches address this issue through increasingly sophisticated retrieval-augmented pipelines, auxiliary reasoning modules (Cocchi et al., 2025; Compagnoni et al., 2026), passage filtering (Hong et al., 2025; Ye et al., 2026), or re-ranking mechanisms (Yan & Xie, 2024; Yang et al., 2025). While effective, these approaches typically rely on additional training, multi-stage processing, or task-specific components, increasing computational overhead and system complexity.

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.

Evidence selection is equally challenging in the visual modality. Many KB-VQA questions rely on evidence localized in specific image regions, yet MLLMs often rely on coarse visual cues rather than precise spatial grounding (Bi et al., 2025; Deng et al., 2025; Wang et al., 2025a). This issue can be further exacerbated by *attention sinks* in large Transformer architectures (Gu et al., 2025; Xiao et al., 2024), where semantically uninformative tokens or patches receive disproportionate attention, thereby weakening evidence localization. These failures reveal a common bottleneck: irrelevant passages and distracting visual regions can dominate generation despite rich multimodal inputs. KB-VQA therefore requires not only better retrieval, but also inference-time mechanisms that prioritize relevant evidence.

In this work, we investigate whether the internal attention dynamics of pretrained MLLMs can be leveraged to identify query-relevant multimodal evidence. Building on this, we introduce **Look Twice (LoT)**, a *training-free* inference-time framework that improves multimodal evidence selection by explicitly highlighting query-relevant cues in both retrieved text and the input image. The key idea is to let the model *look twice* at the input. LoT first leverages the model's internal attention patterns to estimate the relevance of visual regions and textual elements to the query. These signals are then used to reformulate the multimodal input by highlighting the selected evidence while filtering out irrelevant or distracting content before final answer generation (Fig. 1).

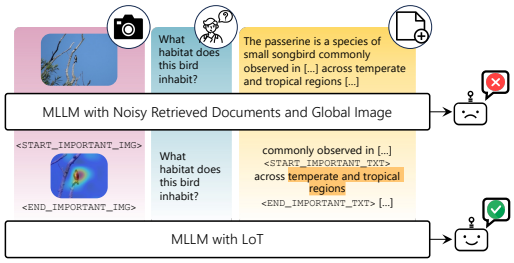


Figure 1: Given an image, question, and retrieved context, an MLLM may miss relevant multimodal evidence. LoT uses internal attention to identify and highlight relevant visual and textual evidence, improving answer generation.

Specifically, for visual input, LoT identifies the target entity in the question and aggregates the corresponding attention over visual tokens to obtain a query-specific relevance map. A filtering step further reduces spurious activations associated with attention sinks. For textual input, LoT uses attention signals to identify the most informative sentences in the retrieved context. The resulting visual and textual evidence is highlighted, while irrelevant content is discarded, yielding a refined multimodal input for answer generation. Unlike existing methods that refine either modality through external modules, predefined regions, bounding boxes, or separately trained selectors, LoT jointly identifies and highlights evidence from both modalities using only the MLLM's internal attention.

We evaluate LoT across four KB-VQA benchmarks (Chen et al., 2023; Mensink et al., 2023; Lerner et al., 2022; Hu et al., 2023) and ten off-the-shelf MLLMs ranging from 2B to 38B parameters. LoT improves every evaluated backbone without any additional training, with average gains from $+3.1$ to $+12.5$ accuracy points. Its adaptable design also enables straightforward integration into existing KB-VQA inference pipelines (Wei et al., 2022; Li et al., 2026; Compagnoni et al., 2026; Yan & Xie, 2024), where it provides further gains without modifying or retraining their components. Together, these results establish LoT as an effective and broadly applicable mechanism for improving multimodal evidence utilization across models, datasets, and inference settings.

Contributions. Our contributions are summarized as follows:

- We introduce Look Twice (LoT), a *training-free* inference-time framework for KB-VQA that improves how pretrained MLLMs exploit multimodal evidence.
- We show that internal attention signals can be leveraged to jointly identify query-relevant visual and textual evidence, enabling LoT to highlight useful cues and suppress irrelevant content before answer generation.
- We conduct an extensive evaluation across four KB-VQA benchmarks and ten MLLMs spanning 2B–38B parameters. LoT improves every evaluated backbone, achieving average gains up to $+12.5$, and delivers further improvements when integrated into diverse inference pipelines.

2 Related Work

Knowledge-Based VQA. Knowledge-Based Visual Question Answering (KB-VQA) requires models to combine visual understanding with knowledge beyond what is directly observable in the image. While early benchmarks (Marino et al., 2019; Schwenk et al., 2022) introduced this challenge,

recent datasets such as Encyclopedic-VQA (Mensink et al., 2023) and InfoSeek (Chen et al., 2023) require fine-grained reasoning over large-scale knowledge sources. To address these, recent approaches rely on sophisticated retrieval-augmented pipelines, additional supervision, or task-specific training to better integrate external evidence (Caffagni et al., 2024b; Yan & Xie, 2024; Yuan et al., 2026; Hong et al., 2025; Cocchi et al., 2025; Compagnoni et al., 2026; Yang et al., 2025). In contrast, we investigate how pretrained MLLMs exploit retrieved textual context and visual evidence directly at inference time. Our training-free evidence highlighting strategy improves multimodal evidence selection without additional supervision, parameter updates, or architectural modifications.

Attention Dynamics in Vision-Language Models. Prior work has extensively analyzed how Transformer models distribute attention, revealing structured patterns involving sparsity, head specialization, and redundancy (Clark et al., 2019; Voita et al., 2019; Kobayashi et al., 2020; Michel et al., 2019; Cordonnier et al., 2020). More recent work highlights the phenomenon of *attention sinks*, where a small subset of tokens or image regions captures a disproportionate amount of attention while other potentially relevant elements are ignored. In LLMs, such sinks often correspond to special or punctuation tokens and can hinder effective long-context reasoning (Gu et al., 2025; Xiao et al., 2024; Yu et al., 2024). Similar behaviors have been observed in vision-language models, where background or visually salient patches dominate attention despite contributing little to the final prediction (Woo et al., 2025; Kang et al., 2025b). Together, these analyses suggest that attention is influenced not only by the input but also by systematic biases. In KB-VQA, this can lead models to focus on salient yet irrelevant tokens or regions instead of query-relevant evidence.

Inference-Level Attention Analysis. A growing line of work modifies attention maps at inference time to improve grounding and reduce hallucinations in large models. In LLMs, attention scores identify relevant textual evidence (Liu et al., 2025). In multimodal settings, some approaches identify dominant visual tokens and redistribute attention toward informative regions (Kang et al., 2025b). Others leverage internal attention dynamics to enhance semantically central visual regions during generation (Lyu et al., 2026), suppress disruptive instruction tokens that distort cross-modal alignment (Chen et al., 2025), or reinforce high-focus attention patterns across heads to strengthen visual conditioning (Zhang et al., 2024). Other approaches improve fine-grained visual perception by reprocessing localized regions, deriving crops from the model attention or gradient signals (Zhang et al., 2025), constructing object-relevance maps from cached token representations (Zhong et al., 2025), or performing confidence-guided hierarchical exploration (Shen et al., 2025).

In contrast, LoT jointly refines visual and retrieved textual evidence for knowledge-intensive VQA. It uses internal attention as a relevance signal to highlight query-relevant content and suppress distractions before generation, without modifying model parameters or architecture.

3 Proposed Method

We introduce Look Twice (LoT), a *training-free* inference-time framework that improves multimodal evidence selection in MLLMs by *highlighting* query-relevant cues in both the retrieved context and the input image. Our key idea is to treat the model internal attention distributions as an implicit relevance signal. Specifically, we (i) estimate which context tokens and visual regions the model considers relevant to the query, (ii) filter sink-driven attention patterns in the visual modality to improve evidence localization, (iii) suppress irrelevant or distracting content, and (iv) highlight the resulting textual evidence and image region through lightweight prompt-level markers during final answer generation. This approach requires no parameter updates and no architectural modifications. An overview of LoT can be seen in Fig. 2.

3.1 Preliminaries

Task Definition. We consider a pretrained MLLM that processes an image together with a textual prompt and generates answers autoregressively. Given an input image I and a textual question $\mathbf{X}_T = [x_1, \ldots, x_{N_T}]$, the model first converts the input image into a sequence of N_V visual tokens, denoted as $\mathbf{X}_V$, using its vision encoder and multimodal projector. In the KB-VQA setting, the model additionally receives retrieved textual evidence. A retrieval module selects the top-n textual documents from a knowledge base (*e.g.*, composed of Wikipedia articles), which are concatenated into a context sequence $\mathbf{X}_C = [c_1, \ldots, c_{N_C}]$. The model processes a single multimodal sequence obtained by concatenating visual tokens, question tokens, and retrieved context,

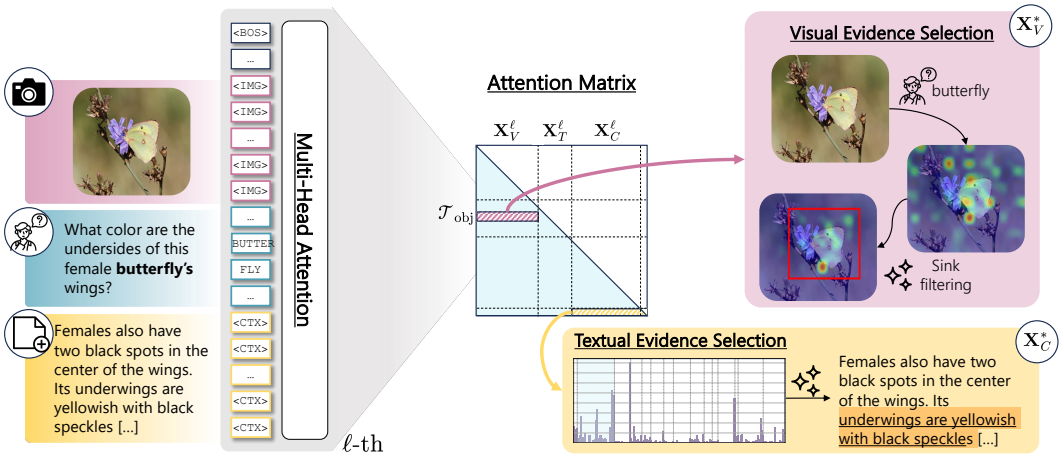


Figure 2: Overview of the proposed Look Twice (LoT). LoT inspects the MLLM internal attention to identify relevant visual regions and textual evidence, discards irrelevant content, and highlights the selected cues for final answer generation.

$\mathbf{X} = [\mathbf{X}_V; \mathbf{X}_T; \mathbf{X}_C]$ with total length $S = N_V + N_T + N_C$. During inference, the model generates an answer autoregressively conditioned on this sequence.

Attention Notation. To integrate information across the question, visual inputs, and retrieved knowledge, MLLMs rely on stacked causal self-attention layers that model interactions among tokens in the multimodal sequence. The input sequence is processed by a Transformer decoder with L layers and K attention heads. At layer ℓ, the hidden states are denoted as $\mathbf{H}^\ell = [\mathbf{h}_1^\ell, \mathbf{h}_2^\ell, \ldots, \mathbf{h}_S^\ell] \in \mathbb{R}^{S\times d}$, where $\mathbf{h}_i^\ell \in \mathbb{R}^d$ denotes the representation of token i at layer ℓ, and d is the hidden dimension. The initial representation $\mathbf{H}^0$ corresponds to the embeddings of the concatenated multimodal input sequence $\mathbf{X}$. Each layer applies causal multi-head self-attention (Vaswani et al., 2017) to propagate information across tokens. We denote by $\mathbf{A}^{\ell,k} \in \mathbb{R}^{S\times S}$ the attention matrix at layer ℓ and head $k \in \{1, \ldots, K\}$, where $\mathbf{A}^{\ell,k}[i, j]$ represents how strongly token i attends to token j. Our method leverages attention patterns between tokens corresponding to the question, visual inputs, and retrieved context to estimate multimodal evidence relevance with respect to the query.

3.2 Training-Free Multimodal Evidence Highlighting

During an initial analysis stage, LoT generates a single token and uses the internal attention dynamics of the model to identify query-relevant visual regions and textual evidence. These cues are then explicitly exposed to the model during full answer generation. Our approach operates by extracting attention signals from the pretrained MLLM, filtering spurious activations caused by attention sinks while discarding irrelevant information, and highlighting the resulting evidence through lightweight prompt markers at inference time.

Self-Guided Visual Evidence Selection. On the visual side, our objective is to identify the image regions that are truly relevant for answering the query by exploiting the model internal attention between the question tokens $\mathbf{X}_T$ and the visual tokens $\mathbf{X}_V$ as an *implicit relevance signal*.

Given the attention matrix $\mathbf{A}^{\ell,k} \in \mathbb{R}^{S\times S}$, we focus on the interactions between the target object in the question and the visual tokens. To this end, we use an off-the-shelf NLP library to identify the target-object tokens and extract the corresponding *object-to-visual* attention submatrix:

$$\mathbf{A}^{\ell,k}_{\text{obj}\to\text{vis}} = \mathbf{A}^{\ell,k}[\mathcal{T}_{\text{obj}}, \mathcal{V}] \in \mathbb{R}^{|\mathcal{T}_{\text{obj}}|\times N_V}, \tag{1}$$

where $\mathcal{V} = \{1, \ldots, N_V\}$ denotes the indices of visual tokens in the concatenated sequence and $\mathcal{T}_{\text{obj}} \subseteq \{N_V + 1, \ldots, N_V + N_T\}$ the indices of textual tokens referring to the target object. For example, given the question "*What color are the undersides of this female butterfly's wings?*", $\mathcal{T}_{\text{obj}}$ corresponds to the token indices in the attention map associated with "*female butterfly*". By restricting attention to these tokens, we explicitly capture how the object mentioned in the question interacts with the visual input, providing a focused measure of object-specific visual grounding.

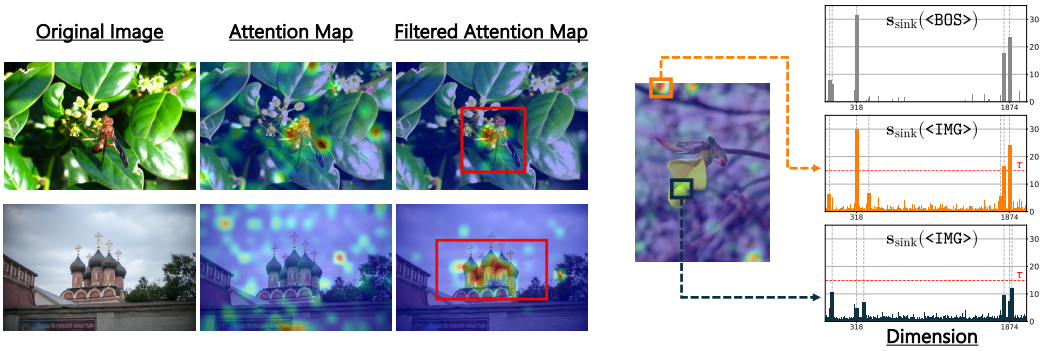


Figure 3: Qualitative examples of our visual evidence selection pipeline (left). Sink-score plots (right) show that irrelevant visual tokens (orange) exhibit high activations in sink dimensions, similarly to the BOS token, a known attention sink in LLMs, while relevant tokens (blue) remain moderate. LoT filters tokens whose sink score exceeds τ.

Since cross-modal interactions are not uniformly localized but emerge across multiple layers and heads (Shi et al., 2025; Yin et al., 2025), and the target object may span multiple textual tokens, we aggregate object-to-visual attention signals across the object tokens, a subset of layers L_{vis}, and all attention heads. The resulting vector $\mathbf{a}_{\text{vis}}$ provides a single relevance score per visual token and serves as the initial spatial relevance signal:

$$\mathbf{a}_{\text{vis}} = \frac{1}{|\mathcal{T}_{\text{obj}}| \cdot |L_{\text{vis}}| \cdot K} \sum_{i \in \mathcal{T}_{\text{obj}}} \sum_{\ell \in L_{\text{vis}}} \sum_{k=1}^{K} \mathbf{A}^{\ell,k}_{\text{obj}\to\text{vis}}[i] \in \mathbb{R}^{N_V}. \tag{2}$$

Multi-Layer Attention Sink Filtering. While $\mathbf{a}_{\text{vis}}$ captures object-conditioned visual relevance, Transformer models are known to exhibit *attention sinks* (Gu et al., 2025; Xiao et al., 2024): tokens that attract disproportionate attention mass regardless of their semantic contribution. This can produce spurious attention concentration and weaken grounding. In multimodal settings, certain visual tokens similarly receive high attention across layers despite being spatially or semantically uninformative (Kang et al., 2025b), as shown in Fig. 3 (right, middle row).

To distinguish sink-driven responses from genuine object relevance, we identify the hidden dimensions most strongly associated with attention sinks. Inspired by Kang et al. (2025b), we analyze the hidden representations of visual tokens across a selected subset of layers L_{sink}. For each layer $\ell \in L_{\text{sink}}$, let $\mathbf{H}^{\ell} \in \mathbb{R}^{S\times d}$ denote the hidden states and $\mathbf{H}^{\ell}_{\mathcal{V}} \in \mathbb{R}^{N_V \times d}$ the slice corresponding to visual tokens. We identify the hidden dimensions that consistently exhibit disproportionately high normalized activations for the BOS token in the base LLM, which is known to act as a prototypical attention sink (Xiao et al., 2024). The resulting set of sink dimensions, denoted as $\mathcal{D}_{\text{sink}} \subseteq \{0, \ldots, d-1\}$, corresponds to a fixed subset of hidden dimensions determined by the LLM backbone. While the specific dimensions depend on the model architecture, they remain constant across input instances (Fig. 3, right).

We define the attention sink score vector $\mathbf{s}_{\text{sink}}$ as the maximum absolute activation across $\mathcal{D}_{\text{sink}}$, normalized along the hidden dimension d, and averaged across the layers in L_{sink}:

$$\mathbf{s}_{\text{sink}} = \frac{1}{|L_{\text{sink}}|} \sum_{\ell \in L_{\text{sink}}} \frac{\max_{m \in \mathcal{D}_{\text{sink}}} \left|\mathbf{H}^{\ell}_{\mathcal{V}}[:, m]\right|}{\|\mathbf{H}^{\ell}_{\mathcal{V}}\|_{\text{row}}} \in \mathbb{R}^{N_V}, \tag{3}$$

where $\|\mathbf{H}^{\ell}_{\mathcal{V}}\|_{\text{row}} = \sqrt{\frac{1}{d}\sum_{m=0}^{d-1} \mathbf{H}^{\ell}_{\mathcal{V}}[:, m]^2}$ is the row-wise RMS norm of $\mathbf{H}^{\ell}_{\mathcal{V}}$. Intuitively, $\mathbf{s}_{\text{sink}}$ quantifies, for each visual token, the activation along the sink dimensions relative to the overall magnitude of its hidden representation. Tokens with unusually high values in $\mathbf{s}_{\text{sink}}$ are therefore likely to be attention sinks.

To suppress these artifacts, we filter visual tokens whose sink score exceeds a threshold τ. Specifically, tokens where $\mathbf{s}_{\text{sink}} > \tau$ are marked as sinks and their corresponding entries in $\mathbf{a}_{\text{vis}}$ are set to zero. The resulting filtered visual relevance vector is reshaped to match the dimensions of the visual feature map, producing a 2D attention map $\mathbf{M}_{\text{vis}} \in \mathbb{R}^{H\times W}$ that reflects the model implicit grounding of the target object over image regions. Importantly, this filtering is applied only during

the analysis of the first generated token; the attention maps used in the subsequent full generation step remain unchanged. Fig. 3 (left, third column) shows the obtained attention maps.

Bounding Box Extraction. We now use the filtered attention map to localize the query entity and extract the corresponding image crop. To convert the attention map into an explicit spatial region, we interpret $\mathbf{M}_{\text{vis}}$ as a spatial probability distribution by normalizing it. We then compute the weighted centroid (c_x, c_y) of the attention map to locate the center of the highlighted visual region:

$$c_x = \sum_{h,w} w\, \mathbf{M}_{\text{vis}}(h,w), \quad c_y = \sum_{h,w} h\, \mathbf{M}_{\text{vis}}(h,w). \tag{4}$$

To quantify the spatial spread of the evidence, we compute the weighted standard deviations along each axis:

$$\sigma_x = \sqrt{\sum_{h,w} (w - c_x)^2 \mathbf{M}_{\text{vis}}(h,w)}, \quad \sigma_y = \sqrt{\sum_{h,w} (h - c_y)^2 \mathbf{M}_{\text{vis}}(h,w)}. \tag{5}$$

The bounding box is finally defined as $(x_1, y_1, x_2, y_2) = (c_x - \beta\sigma_x,\ c_y - \beta\sigma_y, c_x + \beta\sigma_x,\ c_y + \beta\sigma_y)$, which captures the region containing the most relevant visual evidence. The parameter β controls the spatial extent of the bounding box, acting as a scaling factor that adjusts how much surrounding visual context is included.

Self-Guided Textual Evidence Selection. Complementary to visual evidence localization, LoT extends the same self-guided selection strategy to the retrieved context. In fact, only a subset of the retrieved context $\mathbf{X}_C$ is relevant to the query (Cocchi et al., 2025; Yan & Xie, 2024; Yang et al., 2025), yet standard MLLMs often attend broadly across all context tokens, introducing noise and distracting reasoning (Caffagni et al., 2024b). Building on this observation, LoT leverages the model internal self-attention as an implicit relevance signal to distinguish informative from distracting sentences in $\mathbf{X}_C$. The resulting relevance estimates are then used to highlight the most informative evidence while discarding sentences deemed irrelevant before answer generation.

Intuitively, we measure how strongly the model focuses on each token in the retrieved context when generating the answer. Specifically, we extract the *last-to-context* attention submatrix, corresponding to the attention from the last input token at position t to all context tokens:

$$\mathbf{A}^{\ell,k}_{\text{last}\to\text{ctx}} = \mathbf{A}^{\ell,k}[t, \mathcal{C}] \in \mathbb{R}^{1\times N_C}, \tag{6}$$

where $\mathcal{C} \subseteq \{N_V + N_T + 1, \ldots, N_V + N_T + N_C\}$ denotes the indices of the tokens corresponding to the retrieved context in the concatenated multimodal sequence.

As in the visual modality, to obtain a robust textual relevance score $\mathbf{a}_{\text{txt}}$, we aggregate these attention signals across multiple layers L_{txt} and heads K, capturing diverse reasoning patterns learned at different depths of the model:

$$\mathbf{a}_{\text{txt}} = \frac{1}{|L_{\text{txt}}| \cdot K} \sum_{\ell \in L_{\text{txt}}} \sum_{k=1}^{K} \mathbf{A}^{\ell,k}_{\text{last}\to\text{ctx}} \in \mathbb{R}^{N_C}. \tag{7}$$

The token-level scores are then averaged within each sentence to obtain $\hat{\mathbf{a}}_{\text{txt}}$. Sentences whose relevance score exceeds a threshold α are selected as relevant textual evidence and explicitly highlighted. Conversely, sentences whose score falls below a lower threshold α_{drop} are considered irrelevant and removed from the retrieved context before answer generation. Sentences with scores between the two thresholds are retained without highlighting to preserve supporting context.

Inference with Evidence Refinement. The final step of LoT refines the multimodal input by highlighting the identified relevant textual and visual evidence while discarding content considered irrelevant. This guides the MLLM toward the most informative cues and reduces the influence of potentially distracting information during answer generation. To perform evidence highlighting efficiently, we adopt a prompt-augmentation strategy employing prompt-level markers. For visual evidence, we crop the image using the predicted bounding box and provide only the resulting region to the model enclosed by the markers `<START_IMPORTANT_IMG>` and `<END_IMPORTANT_IMG>`. For textual evidence, the selected spans in $\mathbf{X}_C$ are wrapped with the markers `<START_IMPORTANT_TXT>`

and <END_IMPORTANT_TXT>. Sentences identified as irrelevant are removed from the retrieved context, producing the refined textual input $\mathbf{X}_C^*$. We update the task instructions to explicitly inform the model that the markers denote key evidence and should not appear in the output. The model then generates the final response from the original question $\mathbf{X}_T$ and the refined multimodal input $\mathbf{X}_V^*, \mathbf{X}_C^*$, using the updated prompt template.

This inference-time mechanism leaves the model parameters unchanged and requires minimal prompt modifications. Algorithm 1 in Appendix A details the complete method, while the complete prompt templates are provided in Appendix G.

4 Experiments

4.1 Datasets

We evaluate LoT on four KB-VQA benchmarks: Encyclopedic-VQA (E-VQA) (Mensink et al., 2023), InfoSeek (Chen et al., 2023), ViQuAE (Lerner et al., 2022), and OVEN (Hu et al., 2023). E-VQA contains 221k question-answer pairs over 16.7k entities, including both single-hop and multi-hop questions, and is supported by a knowledge base of roughly 2M Wikipedia pages. InfoSeek comprises approximately 1.3M image-question-answer triplets spanning 11k entities, including previously unseen ones; its official setup uses 100k pages from a 6M-page Wikipedia knowledge base. We additionally report results on the ViQuAE test set and the OVEN validation set. Following prior work (Cocchi et al., 2025; Compagnoni et al., 2026; Wu et al., 2025), retrieval for ViQuAE and OVEN uses the E-VQA and InfoSeek knowledge bases, respectively. Together, these benchmarks span diverse scales, entity distributions, and reasoning challenges.

4.2 Implementation Details

Models and Baselines. We evaluate our method on a diverse set of MLLMs with different architectures and parameter scales. Specifically, we consider Qwen2-VL (Wang et al., 2024) (2B, 7B), Qwen2.5-VL (Bai et al., 2025b) (3B, 7B, 32B), Qwen3-VL (Bai et al., 2025a) (4B, 8B), and InternVL3.5 (Wang et al., 2025b) (4B, 8B, 38B).

Retrieval Pipeline. We adopt a cross-modal entity retrieval pipeline to retrieve relevant knowledge for each input image. Following Yang et al. (2025), each Wikipedia entity in the knowledge base is represented by a concise textual summary. We encode all entity summaries using the EVA-CLIP text encoder (Sun et al., 2024). Given an input image, we obtain its representation with the EVA-CLIP visual encoder and perform image-to-text retrieval through inner-product similarity search with FAISS (Johnson et al., 2021). The top-n retrieved documents are concatenated to form the textual context provided to the MLLM, with $n = 3$ in our experiments. When re-ranking is enabled, we retrieve the top-3 documents and retain the top-5 sections for answer generation.

Evidence Selection. During generation of the first output token, we analyze attention patterns across different decoder layer ranges, selecting modality-specific subsets corresponding to where MLLMs typically process each modality. Prior work indicates that textual evidence is primarily consolidated in deeper layers, whereas visual grounding emerges in intermediate layers (Kang et al., 2025a; Liu et al., 2025; Jiang et al., 2025). Attention sinks, instead, may arise throughout the network. Accordingly, we estimate textual relevance from the second half of the decoder layers, denoted by L_{txt}. Following Compagnoni et al. (2026), we use spaCy for lightweight target-object extraction. Visual relevance is computed over the middle half of the decoder, spanning one-quarter to three-quarters of the total depth and denoted by L_{vis}. We perform sink detection over the same intermediate range, setting $L_{\text{sink}} = L_{\text{vis}}$. D_{sink} is found following Kang et al. (2025b).

During sink filtering, the threshold τ is set to the 25th percentile of the sink scores computed over all visual tokens. For textual evidence refinement, we highlight sentences whose attention score exceeds half of the maximum sentence-level score, setting $\alpha = 0.5 \cdot \max(\hat{\mathbf{a}}_{\text{txt}})$. We discard sentences whose score is lower than one third of the highlighting threshold, using $\alpha_{\text{drop}} = \alpha/3$. Sentences whose scores lie between α_{drop} and α are retained without highlighting. For InternVL3.5 backbones, we use a more conservative filtering criterion and halve the default dropping threshold. Finally, in the bounding box extraction step we set $\beta = 2$. Further details on design choices, sink dimensions and selection of hyperparameter values are provided in Appendices C and D.

Table 1: Performance on KB-VQA benchmarks. For each backbone, we compare the standard retrieval-augmented model with the same model augmented with LoT. Results are reported on the E-VQA test set, InfoSeek validation set, OVEN validation set, and ViQuAE test set. Avg denotes the average score across the four "All" columns; improvements over baseline are in green.

		E-VQA		InfoSeek			OVEN	ViQuAE		
		Single	All	Unseen-Q	Unseen-E	All	All	All	Avg	
Small	▼ Qwen2-VL-2B	17.5	16.0	5.3	5.6	5.4	1.2	15.0	9.4	
	+ LoT (Ours)	**20.5**	**18.5**	**21.8**	**21.4**	**21.6**	**14.5**	**29.1**	**20.9**	**Δ+11.5**
	▼ Qwen2.5-VL-3B	30.3	28.0	22.6	22.2	22.4	11.6	22.9	21.2	
	+ LoT (Ours)	**32.6**	**30.2**	**30.3**	**29.8**	**30.1**	**29.5**	**34.7**	**31.1**	**Δ+9.9**
	▼ Qwen3-VL-4B	35.0	32.7	28.0	28.6	28.3	23.5	34.7	29.8	
	+ LoT (Ours)	**36.9**	**35.1**	**29.9**	**30.3**	**30.1**	**25.7**	**40.5**	**32.9**	**Δ+3.1**
	▼ InternVL3.5-4B	29.4	26.4	28.8	29.1	29.0	7.9	36.4	24.9	
	+ LoT (Ours)	**33.7**	**31.0**	**29.9**	**29.0**	**29.5**	**16.3**	**41.9**	**29.7**	**Δ+4.8**
Medium	▼ Qwen2-VL-7B	25.6	22.9	24.2	24.7	24.4	11.1	33.0	22.9	
	+ LoT (Ours)	**29.6**	**26.6**	**33.9**	**33.2**	**33.6**	**30.5**	**50.6**	**35.3**	**Δ+12.5**
	▼ Qwen2.5-VL-7B	32.1	30.2	23.9	25.1	24.5	20.2	36.4	27.8	
	+ LoT (Ours)	**35.0**	**32.2**	**29.3**	**30.9**	**30.1**	**30.0**	**50.5**	**35.7**	**Δ+7.9**
	▼ Qwen3-VL-8B	36.5	34.8	29.1	30.4	29.7	18.1	43.7	31.6	
	+ LoT (Ours)	**38.0**	**36.3**	**32.7**	**32.2**	**32.5**	**19.6**	**54.0**	**35.6**	**Δ+4.0**
	▼ InternVL3.5-8B	31.3	28.8	29.4	29.8	29.6	20.8	44.5	30.9	
	+ LoT (Ours)	**33.7**	**31.4**	**31.6**	**31.1**	**31.3**	**26.4**	**55.4**	**36.1**	**Δ+5.2**
Large	▼ Qwen2.5-VL-32B	35.3	33.7	26.7	26.1	26.4	13.3	37.9	27.8	
	+ LoT (Ours)	**37.8**	**35.5**	**29.6**	**28.5**	**29.0**	**15.6**	**49.0**	**32.3**	**Δ+4.5**
	▼ InternVL3.5-38B	33.8	31.6	33.1	33.0	33.1	20.2	51.5	34.1	
	+ LoT (Ours)	**35.0**	**33.4**	**33.6**	**33.2**	**33.4**	**24.8**	**57.6**	**37.3**	**Δ+3.2**

4.3 Experimental Results

Main Results. Table 1 reports the performance of LoT on E-VQA, InfoSeek, OVEN, and ViQuAE. Across all model families, scales, and benchmarks, LoT consistently improves performance. The largest average gains are observed for Qwen2-VL-7B and Qwen2-VL-2B, which improve by $+12.5$ and $+11.5$ points, respectively. LoT also yields substantial improvements for Qwen2.5-VL-3B ($+9.9$) and Qwen2.5-VL-7B ($+7.9$), demonstrating that the benefits of multimodal evidence refinement are not restricted to a specific architecture. The improvements are particularly pronounced on benchmarks requiring effective reasoning over retrieved knowledge. For instance, Qwen2-VL-7B improves from 11.1 to 30.5 on OVEN and from 33.0 to 50.6 on ViQuAE. Gains also remain consistent for larger models, with InternVL3.5-38B increasing from 34.1 to 37.3 on average.

Overall, the average improvement ranges from $+3.1$ to $+12.5$ points across the evaluated models, showing that attention-guided evidence highlighting is a general and robust mechanism, yielding consistent and substantial accuracy gains regardless of backbone capacity.

Table 2: Comparison of inference-time strategies combined with LoT. Gray rows denote strategies that employ no additional trained components.

	E-VQA		InfoSeek		
	Base	+LoT	Base	+LoT	
▼ Qwen2.5-VL-3B	28.0	**30.2**	22.4	**30.1**	**Δ+4.95**
+ CoT (Wei et al., 2022)	28.1	**30.1**	29.5	**30.1**	**Δ+1.30**
+ Critic (Compagnoni et al., 2026)	30.1	**31.3**	30.6	**31.1**	**Δ+0.85**
+ Trained Critic (Compagnoni et al., 2026)	33.3	**34.4**	33.5	**34.4**	**Δ+1.00**
+ EchoSight Re-ranker (Yan & Xie, 2024)	29.8	**31.1**	30.7	**32.0**	**Δ+1.30**
+ Qwen3-VL-Reranker-8B (Li et al., 2026)	30.5	**31.6**	**33.6**	**33.6**	**Δ+0.55**
▼ Qwen2.5-VL-7B	30.2	**32.2**	24.5	**30.1**	**Δ+3.80**
+ CoT (Wei et al., 2022)	28.6	**32.1**	28.5	**30.1**	**Δ+2.55**
+ Critic (Compagnoni et al., 2026)	30.1	**30.6**	29.5	**29.8**	**Δ+0.40**
+ Trained Critic (Compagnoni et al., 2026)	35.5	**36.3**	34.0	**35.2**	**Δ+1.00**
+ EchoSight Re-ranker (Yan & Xie, 2024)	32.9	**34.3**	30.4	**32.8**	**Δ+1.90**
+ Qwen3-VL-Reranker-8B (Li et al., 2026)	33.6	**34.0**	32.6	**34.0**	**Δ+0.90**
▼ Qwen2.5-VL-32B	33.7	**35.5**	26.4	**29.0**	**Δ+2.20**
+ CoT (Wei et al., 2022)	32.8	**35.6**	**29.7**	29.3	**Δ+1.20**
+ Critic (Compagnoni et al., 2026)	35.7	**37.0**	29.8	**30.9**	**Δ+1.20**
+ Trained Critic (Compagnoni et al., 2026)	37.9	**40.7**	34.4	**35.2**	**Δ+1.80**
+ EchoSight Re-ranker (Yan & Xie, 2024)	36.9	**38.4**	28.8	**30.0**	**Δ+1.35**
+ Qwen3-VL-Reranker-8B (Li et al., 2026)	36.8	**38.6**	31.0	**31.3**	**Δ+1.05**

Integration with Existing Inference Strategies. Table 2 evaluates the compatibility of LoT with a diverse set of KB-VQA inference strategies. We consider training-free methods, including chain-of-thought prompting (CoT) (Wei et al., 2022) and a prompt-based critic model following Compagnoni et al. (2026), where the same zero-shot MLLM used for answer generation is prompted to retain or discard candidate passages. We evaluate the task-specific critic model released by Compagnoni et al. (2026), which was fine-tuned for passage filtering, and specialized re-ranking models (Li et al., 2026; Yan & Xie, 2024)[1]. Across these settings, LoT provides complementary gains, reaching up to $+2.55$ points when combined with

[1]All configurations use $n = 3$, except the trained Critic, which follows its original $n = 20$ setting.

existing inference strategies and up to $+4.95$ points for the standard retrieval-augmented baseline. Thus, LoT remains effective even when reasoning, passage filtering, or re-ranking is already applied.

Computational Analysis. LoT limits memory overhead by monkey-patching the forward pass to compute only the attention scores required for evidence selection, while retaining FlashAttention-2 for all remaining computations. On E-VQA with Qwen2.5-VL-3B, its evidence selection reduces the input length of the second pass by 71.93%: -124.0 visual tokens (-54.24%) and $-2,139.3$ textual tokens (-72.90%). Considering GPU-side model execution, LoT overall adds 14.1% latency overhead over the baseline and $+42.8\%$ FLOPs. Further details can be found in Appendix E.

4.4 Ablation Studies

Effectiveness of Multimodal Highlighting. Table 3 isolates the contribution of visual and textual highlighting on the 3B and 7B versions of Qwen2.5-VL. Textual highlighting provides the largest overall gains, which is consistent with the retrieval-heavy nature of KB-VQA: retrieved documents introduce substantially more textual content into the multimodal input, and the external knowledge required to answer the question is primarily conveyed through this modality. Nevertheless, visual highlighting remains beneficial, improving over the corresponding baseline in most settings and providing additional gains when combined with textual refinement. In particular, the full LoT configuration achieves the strongest results across all benchmarks, reaching 32.2 on E-VQA, 30.1 on InfoSeek, 30.0 on OVEN, and 50.5 on ViQuAE for the 7B backbone. Appendix B covers matched-budget baselines, alternative attention-based highlighting strategies, and visual-only results on standard MLLM benchmarks; Appendix D analyzes evidence-selection quality beyond answer accuracy.

Evidence Availability Analysis. To assess how LoT exploits retrieved evidence, we evaluate E-VQA separately when the retriever is able to find the answer-bearing document (*retrieval hit*) and when it does not (*retrieval miss*). As shown in Fig. 4 (left), LoT achieves substantial gains of $+5.6$ and $+6.8$ on retrieval hits, demonstrating its ability to identify and emphasize the relevant evidence when available. On retrieval misses, gains are smaller but remain positive ($+0.9$ and $+0.4$): jointly highlighting the most informative visual cues and removing the least relevant sentences, LoT reduces noise and helps the model rely better on the image.

Table 3: Ablation evaluating visual and textual highlighting independently and jointly.

	Highlight		E-VQA		InfoSeek			OVEN	ViQuAE
	Visual	Textual	Single	All	U-Q	U-E	All	All	All
▼ Qwen2.5-VL-3B	-	-	30.3	28.0	22.6	22.2	22.4	11.6	22.9
	✓	-	31.6	29.4	24.1	23.7	23.9	16.6	23.6
	-	✓	31.7	29.1	29.8	29.2	29.5	28.0	34.6
+ LoT (Ours)	✓	✓	**32.6**	**30.2**	**30.3**	**29.8**	**30.1**	**29.5**	**34.7**
▼ Qwen2.5-VL-7B	-	-	32.1	30.2	23.9	25.1	24.5	20.2	36.4
	✓	-	32.4	30.2	24.4	25.4	24.9	21.0	34.3
	-	✓	34.7	32.1	28.8	29.9	29.4	28.7	49.6
+ LoT (Ours)	✓	✓	**35.0**	**32.2**	**29.3**	**30.9**	**30.1**	**30.0**	**50.5**

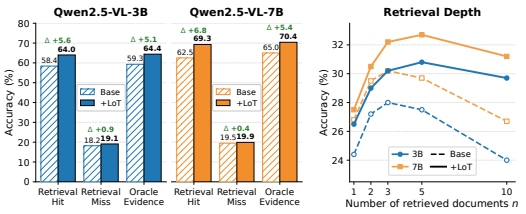


Figure 4: Performance on E-VQA (All) under different evidence-availability (left) and retrieval-depth settings varying n (right).

Effectiveness with Oracle Evidence. To isolate the effect of evidence utilization from retrieval quality, in Fig. 4 (left) we also evaluate the oracle setting on E-VQA in which each sample is provided only with its ground-truth Wikipedia page. LoT improves accuracy from 59.3 to 64.4 for the 3B model and from 65.0 to 70.4 for the 7B model. Notably, even when provided with only the correct page and minimal distracting context, LoT still delivers substantial gains.

Effect of Retrieval Depth. In Fig. 4 (right), we vary the number of retrieved documents from 1 to 10 on E-VQA. LoT consistently outperforms the baselines. Notably, increasing n from 3 to 5 benefits LoT, while the baselines degrade under additional retrieval noise.

5 Conclusion

We introduced LoT, a training-free inference-time framework that uses internal attention signals to refine visual and textual evidence before answer generation. Across four KB-VQA benchmarks and ten off-the-shelf MLLMs, LoT consistently improves the model performance and complements existing refinement strategies, demonstrating that model-internal attention can enhance multimodal evidence utilization without additional training, auxiliary models, or architectural changes.

AI USE STATEMENT

In this work, we used generative AI tools to draft, revise, and rewrite portions of the manuscript based on content and instructions provided by the authors, with the goal of improving clarity, readability, grammar, and style. The scientific ideas, methodology, experiments, results, and conclusions were developed by the authors. All AI-assisted text was carefully reviewed and edited by the authors for technical correctness, originality, and consistency with the intended meaning. We take responsibility for the final content of this work, including all text produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the complete source code for LoT in the supplementary material, including a README and the scripts required to run our methodology. The main paper reports the experimental setup, retrieval pipeline, implementation details, and hyperparameter choices, while the appendix provides additional implementation details, prompt templates, hyperparameter analyses, and further experimental results. Publicly available models, datasets, and external resources used in our experiments can be obtained from their original sources.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, 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, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025b.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. Unveiling Visual Perception in Language Models: An Attention Head Analysis Approach. In *CVPR*, 2025.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The Revolution of Multimodal Large Language Models: A Survey. In *ACL Findings*, 2024a.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs. In *CVPR Workshops*, 2024b.

Beitao Chen, Xinyu Lyu, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Attention Hijackers: Detect and Disentangle Attention Hijacking in LVLMs for Hallucination Mitigation. *arXiv preprint arXiv:2503.08216*, 2025.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? In *EMNLP*, 2023.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What Does BERT Look at? An Analysis of BERT's Attention. In *ACL*, 2019.

Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visual Question Answering. In *CVPR*, 2025.

Alberto Compagnoni, Marco Morini, Sara Sarto, Federico Cocchi, Davide Caffagni, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. ReAG: Reasoning-Augmented Generation for Knowledge-based Visual Question Answering. In *CVPR*, 2026.

Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the Relationship between Self-Attention and Convolutional Layers. In *ICLR*, 2020.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or Vision: Do Vision-Language Models Have Blind Faith in Text? In *CVPR*, 2025.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When Attention Sink Emerges in Language Models: An Empirical View. In *ICLR*, 2025.

Yuyang Hong, Jiaqi Gu, Qi Yang, Lubin Fan, Yue Wu, Ying Wang, Kun Ding, Shiming Xiang, and Jieping Ye. Knowledge-based Visual Question Answer with Multimodal Processing, Retrieval and Filtering. In *NeurIPS*, 2025.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain Visual Entity Recognition: Towards Recognizing Millions of Wikipedia Entities. In *CVPR*, 2023.

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 *CVPR*, 2025.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-Scale Similarity Search with GPUs. *IEEE Trans. on Big Data*, 7(3):535–547, 2021.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your Large Vision-Language Model Only Needs A Few Attention Heads For Visual Grounding. In *CVPR*, 2025a.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See What You Are Told: Visual Attention Sink in Large Multimodal Models. In *ICLR*, 2025b.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. In *EMNLP*, 2020.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. ViQuAE, a Dataset for Knowledge-based Visual Question Answering about Named Entities. In *ACM SIGIR*, 2022.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *CVPR*, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In *ECCV*, 2024b.

Zhining Liu, Rana Ali Amjad, Ravinarayana Adkathimar, Tianxin Wei, and Hanghang Tong. SelfElicit: Your Language Model Secretly Knows Where is the Relevant Evidence. In *ACL*, 2025.

Guangtao Lyu, Qi Liu, Chenghao Xu, Jiexi Yan, Muli Yang, Xueting Li, Fen Fang, and Cheng Deng. Revealing and Enhancing Core Visual Regions: Harnessing Internal Attention Dynamics for Hallucination Mitigation in LVLMs. In *ACL Findings*, 2026.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *CVPR*, 2019.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *ACL Findings*, 2022.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic VQA: Visual Questions About Detailed Properties of Fine-Grained Categories. In *ICCV*, 2023.

Paul Michel, Omer Levy, and Graham Neubig. Are Sixteen Heads Really Better than One? In *NeurIPS*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge. In *ECCV*, 2022.

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 *EMNLP*, 2025.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision Function Layer in Multimodal LLMs. In *NeurIPS*, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA Models That Can Read. In *CVPR*, 2019.

Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. EVA-CLIP-18B: Scaling CLIP to 18 Billion Parameters. *arXiv preprint arXiv:2402.04252*, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *NeurIPS*, 2017.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting. In *ACL*, 2019.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Sudong Wang, Yunjian Zhang, Yao Zhu, Jianing Li, Zizhe Wang, et al. Towards Understanding How Knowledge Evolves in Large Vision-Language Models. In *CVPR*, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*, 2022.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't Miss the Forest for the Trees: Attentional Vision Calibration for Large Vision Language Models. In *ACL Findings*, 2025.

Penghao Wu and Saining Xie. V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs. In *CVPR*, 2024.

Wenyi Wu, Zixuan Song, Kun Zhou, Yifei Shao, Zhiting Hu, and Biwei Huang. Towards General Continuous Memory for Vision-Language Models. In *NeurIPS*, 2025.

xAI. Grok-1.5 Vision Preview, 2024. URL https://x.ai/blog/grok-1.5v.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient Streaming Language Models with Attention Sinks. In *ICLR*, 2024.

Yibin Yan and Weidi Xie. EchoSight: Advancing Visual-Language Models with Wiki Knowledge. In *EMNLP Findings*, 2024.

Wei Yang, Jingjing Fu, Rui Wang, Jinyu Wang, Lei Song, and Jiang Bian. OMGM: Orchestrate Multiple Granularities and Modalities for Efficient Multimodal Retrieval. In *ACL*, 2025.

Wei Ye, Yixin Su, Yueguo Chen, Longxiang Gao, Jianjun Li, Ruixuan Li, and Rui Zhang. QKVQA: Question-Focused Filtering for Knowledge-based VQA. *arXiv preprint arXiv:2601.13856*, 2026.

Hao Yin, Guangzong Si, and Zilei Wang. Lifting the Veil on Visual Information Flow in MLLMs: Unlocking Pathways to Faster Inference. In *CVPR*, 2025.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and Harnessing Hidden Attention Sinks: Enhancing Large Language Models without Training through Attention Calibration. In *ICML*, 2024.

Xu Yuan, Liangbo Ning, Wenqi Fan, and Qing Li. mKG-RAG: Leveraging Multimodal Knowledge Graphs in Retrieval-Augmented Generation for Knowledge-intensive VQA. In *ACM SIGIR*, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *ICCV*, 2023.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs Know Where to Look: Training-free Perception of Small Visual Details with Multimodal LLMs. In *ICLR*, 2025.

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. Seeing Clearly by Layer Two: Enhancing Attention Heads to Alleviate Hallucination in LVLMs. *arXiv preprint arXiv:2411.09968*, 2024.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223*, 2023.

Liangyu Zhong, Fabio Philipp Rosenthal, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, and Leo Schwinn. FOCUS: Internal MLLM Representations for Efficient Fine-Grained Visual Question Answering. In *NeurIPS*, 2025.

A LoT Algorithm

Algorithm 1 summarizes the LoT pipeline. Given an image, a question, and retrieved context, LoT first generates a single output token to extract object-to-visual and last-to-context attention scores across decoder layers and heads. The aggregated visual attention is filtered using the detected sink dimensions and converted into a query-specific bounding box. For text, token-level attention is averaged within each sentence to estimate its relevance: highly relevant sentences are selected for highlighting, low-relevance sentences are discarded, and intermediate-scoring sentences are retained unchanged. The selected crop and textual evidence are then used to reconstruct a multimodal input with dedicated markers highlighting the most informative content. Finally, this refined input is used to generate the final answer.

Algorithm 1 Look Twice (LoT): Training-Free Multimodal Evidence Highlighting

Input: image $\mathbf{X}_V$, question $\mathbf{X}_T$, retrieved context $\mathbf{X}_C$
$\mathbf{A} \leftarrow \text{MLLM}([\mathbf{X}_V; \mathbf{X}_T; \mathbf{X}_C])$ ▷ first forward pass to obtain attention matrix
for each (ℓ, k) **do**
  $\mathbf{A}^{\ell,k}_{\text{obj}\to\text{vis}} \leftarrow \mathbf{A}^{\ell,k}[\mathcal{T}_{\text{obj}}, \mathcal{V}]$ ▷ object-to-visual attention
  $\mathbf{A}^{\ell,k}_{\text{last}\to\text{ctx}} \leftarrow \mathbf{A}^{\ell,k}[t, \mathcal{C}]$ ▷ last-to-context attention
end for
$\mathbf{a}_{\text{vis}} \leftarrow$ average over $k, \ell, \mathcal{T}_{\text{obj}}$ ▷ Eq. 2
$\mathbf{a}_{\text{txt}} \leftarrow$ average over k, ℓ ▷ Eq. 7
Visual Evidence Selection
$\mathbf{s}_{\text{sink}} \leftarrow$ compute sink score vector using $\mathcal{D}_{\text{sink}}$ ▷ Eq. 3
$\mathbf{a}_{\text{vis}} \leftarrow \mathbf{a}_{\text{vis}}$ where $\mathbf{s}_{\text{sink}} < \tau$ else 0 ▷ sink filtering
$\mathbf{M}_{\text{vis}} \leftarrow$ reshape $\mathbf{a}_{\text{vis}}$ in 2D attention map
$(c_x, c_y), (\sigma_x, \sigma_y) \leftarrow \mathbf{M}_{\text{vis}}$ ▷ centroid and spread (Eq. 4, 5)
$(x_1, y_1, x_2, y_2) \leftarrow (c_x - \beta\sigma_x,\ c_y - \beta\sigma_y,\ c_x + \beta\sigma_x,\ c_y + \beta\sigma_y)$ ▷ evidence bbox
Textual Evidence Selection
$\hat{\mathbf{a}}_{\text{txt}} \leftarrow \text{avg}(\text{NLP}(\mathbf{a}_{\text{txt}}))$ ▷ average over sentence
$\mathbf{X}_C \leftarrow \mathbf{X}_C[\hat{\mathbf{a}}_{\text{txt}} > \alpha_{drop}]$ ▷ discard irrelevant sentences
Identify sentences where $\hat{\mathbf{a}}_{\text{txt}} > \alpha$ ▷ textual evidence
Multimodal Evidence Highlighting
$\mathbf{X}^*_V, \mathbf{X}^*_C \leftarrow$ add evidence markers
answer $\leftarrow \text{MLLM}([\mathbf{X}^*_V; \mathbf{X}_T; \mathbf{X}^*_C])$ ▷ generate answer on highlighted evidence

B Additional Experimental Results

B.1 Additional Ablation Studies

Comparison with Non-Attention-Based Selection Strategies. To determine whether the gains of LoT arise simply from shortening the textual context or cropping the image, Table 4 compares LoT against several non-attention-based selection strategies operating under the same input budget. For all textual controls, the visual input is left unchanged and the selector targets a token budget corresponding to 30% of the valid retrieved context. *Random* constructs a deterministic, sample-specific random ordering of the candidate sentences and retains sentences from this ordering until the token budget is reached. *Leading* instead preserves the original context order and retains the earliest sentences that fit within the budget. *BM25* ranks sentences by their lexical relevance to the question, using term-frequency saturation, inverse document frequency, and length normalization. The BM25 statistics are computed over the sentences in the retrieved context of each example, and the highest-ranked sentences are retained. *Retrieval Round-Robin* follows the original ranking of the retrieved passages and interleaves their sentences: it considers the first sentence from each passage in retrieval order, followed by the second sentence from each passage, and so forth, until the token budget is reached. Finally, *Prompted (Two-Pass)* asks the same MLLM, conditioned on the image and question, to return a ranked list of the sentence IDs that are most useful for answering the question. The model then performs a second generation pass using only the selected textual evidence and the original image. For vision, *Center Crop* selects a central region covering 46% of

Table 4: Comparison of LoT with budget-matched non-attention-based selection strategies on E-VQA and InfoSeek. Textual methods retain 30% of the context, while visual methods retain 46% of the image area.

	E-VQA		InfoSeek		
	Single-Hop	All	Unseen-Q	Unseen-E	All
▼ Qwen2.5-VL-3B	30.3	28.0	22.6	22.2	22.4
Random	22.2	21.2	22.2	22.1	22.1
Leading	24.1	23.0	28.2	28.6	28.4
BM25	30.6	28.2	28.5	28.0	28.3
Retrieval Round-Robin	22.6	21.7	24.2	25.2	24.7
Prompted (Two-Pass)	30.5	28.2	28.7	28.2	28.4
LoT (Textual)	**31.7**	**29.1**	**29.8**	**29.2**	**29.5**
▼ Qwen2.5-VL-7B	32.1	30.2	23.9	25.1	24.5
Random	23.9	23.3	22.6	23.5	23.1
Leading	26.9	25.8	27.3	29.6	28.4
BM25	34.5	31.9	27.5	28.9	28.2
Retrieval Round-Robin	25.5	24.4	24.0	25.9	24.9
Prompted (Two-Pass)	33.6	31.4	27.4	28.8	28.1
LoT (Textual)	**34.7**	**32.1**	**28.8**	**29.9**	**29.4**
▼ Qwen2.5-VL-3B	30.3	28.0	22.6	22.2	22.4
Center Crop	31.2	29.2	23.9	23.2	23.5
Random Crop	31.1	29.1	22.1	21.6	21.9
LoT (Visual)	**31.6**	**29.4**	**24.1**	**23.7**	**23.9**
▼ Qwen2.5-VL-7B	32.1	**30.2**	23.9	25.1	24.5
Center Crop	31.8	29.9	23.3	24.3	23.8
Random Crop	31.6	29.5	23.3	24.3	23.8
LoT (Visual)	**32.4**	**30.2**	**24.4**	**25.4**	**24.9**

the image area, whereas *Random Crop* selects an equally sized region at a random location. These visual controls operate on the original textual context.

For text, random and retrieval-based selectors often substantially reduce performance, while stronger lexical and prompted selectors yield more competitive results. Nevertheless, LoT achieves the best overall accuracy on both datasets and model sizes. This indicates that its improvements can not be explained solely by retaining fewer tokens, but depend on selecting evidence according to the query-conditioned relevance signals of the model. A similar pattern emerges for visual selection. Generic center and random crops provide inconsistent benefits and can reduce performance, particularly for the 7B model. Overall, these comparisons show that targeted multimodal evidence selection is more effective than generic input compression at an equivalent evidence budget.

Comparison with Different Visual Highlighting Strategies. Table 5 compares alternative strategies for presenting the selected visual evidence. Beyond the baseline, which uses only the original input image, we consider drawing the estimated bounding box on the original image (*w/ bbox on input image*), providing only the selected crop (*w/ only selected evidence*), and using both the original image and the crop either without markers (*w/ input image and evidence (no highlight)*) or with highlighting markers around both images (*w/ highlight on both input image and evidence*). The complete LoT strategy instead provides only the selected crop enclosed by highlighting markers (*highlight on evidence only*).

Drawing the bounding box on the original image provides no consistent benefit, while using only the selected crop yields at most limited improvements. Providing both the original image and the crop without markers generally degrades results, suggesting that duplicated visual content introduces additional noise. Highlighting both images is more effective, but remains inconsistent across models and datasets. Overall, LoT achieves the best results in all settings, indicating that explicitly highlighting the localized evidence is more effective than preserving or duplicating the full image. This design also avoids the additional visual tokens required to process the original image.

Comparison with Different Textual Highlighting Strategies. Table 6 compares different uses of the textual relevance scores produced by LoT. Beyond the unmodified baseline, we consider highlighting the complete retrieved context (*w/ highlight on all retrieved content*) or the full passages

Table 5: Ablation study on visual highlighting strategies.

	E-VQA		InfoSeek		
	Single	All	U-Q	U-E	All
▼ Qwen2.5-VL-3B (only input image)	30.3	28.0	22.6	22.2	22.4
w/ bbox on input image	29.6	27.4	22.8	22.2	22.5
w/ only selected evidence	30.7	28.3	22.7	22.3	22.5
w/ input image and evidence (no highlight)	30.0	27.8	21.4	20.8	21.2
w/ highlight on both input image and evidence	**31.6**	29.3	23.0	22.4	22.7
LoT (Visual) – highlight on evidence only	**31.6**	**29.4**	**24.1**	**23.7**	**23.9**
▼ Qwen2.5-VL-7B (only input image)	32.1	30.2	23.9	25.1	24.5
w/ bbox on input image	32.0	29.9	23.5	24.9	24.2
w/ only selected evidence	32.1	29.9	23.9	25.1	24.5
w/ input image and evidence (no highlight)	30.7	28.4	23.2	24.5	23.8
w/ highlight on both input image and evidence	31.7	29.6	22.5	25.4	24.0
LoT (Visual) – highlight on evidence only	**32.4**	**30.2**	**24.4**	**25.4**	**24.9**

Table 6: Ablation study on textual highlighting strategies.

	E-VQA		InfoSeek		
	Single	All	U-Q	U-E	All
▼ Qwen2.5-VL-3B	30.3	28.0	22.6	22.2	22.4
w/ highlight on all retrieved content	30.6	28.2	24.1	23.5	23.8
w/ highlight on entire passage	31.2	28.8	25.0	23.9	24.4
w/ highlight on evidence only (no discard)	**31.7**	**29.1**	24.5	24.0	24.2
w/ keep only evidence	30.5	28.5	27.0	26.3	26.7
LoT (Textual)	**31.7**	**29.1**	**29.8**	**29.2**	**29.5**
▼ Qwen2.5-VL-7B	32.1	30.2	23.9	25.1	24.5
w/ highlight on all retrieved content	32.0	30.1	24.1	25.0	24.5
w/ highlight on entire passage	32.1	30.0	26.0	25.9	26.0
w/ highlight on evidence only (no discard)	33.6	31.4	24.9	25.4	25.1
w/ keep only evidence	33.9	31.6	26.3	26.4	26.4
LoT (Textual)	**34.7**	**32.1**	**28.8**	**29.9**	**29.4**

containing the selected evidence (*w/ highlight on entire passage*). At sentence level, *w/ highlight on evidence only (no discard)* marks sentences whose scores exceed α while retaining the remaining context, whereas *w/ keep only evidence* removes all non-selected sentences, equivalently setting $\alpha_{\text{drop}} = \alpha$. The complete LoT strategy highlights high-relevance sentences, discards low-relevance ones, and retains intermediate-scoring sentences unchanged.

Highlighting the complete retrieved context provides only marginal benefits, whereas more fine-grained sentence-level strategies are generally more effective than passage-level highlighting. Retaining only the selected evidence substantially improves InfoSeek, but can remove supporting context that remains useful, particularly on E-VQA. The complete LoT strategy achieves the best performance across all model and dataset configurations, showing that explicit highlighting and selective filtering play complementary roles.

B.2 Results on Standard MLLM Benchmarks

To further assess the robustness of our visual-highlighting strategy, we evaluate LoT on standard MLLM benchmarks beyond KB-VQA. Specifically, we report results on RealWorldQA (xAI, 2024), V-Star (Wu & Xie, 2024), TextVQA (Singh et al., 2019), and ChartQA (Masry et al., 2022), covering real-world visual reasoning, scene-text understanding, and chart reasoning. Since these benchmarks do not provide retrieved textual context, we apply LoT using visual highlighting only. Moreover, as several questions require global scene understanding or relationships between multiple objects, we retain the full input image and append the selected crop enclosed by the visual highlighting markers. This preserves the complete visual context while explicitly directing the model toward the most query-relevant region.

As shown in Table 7, LoT improves the average performance of all five evaluated backbones, with gains ranging from $+1.0$ to $+4.2$ points. These results show that highlighting query-relevant image

Table 7: Performance comparison on standard MLLM benchmarks. For each backbone, we compare the original model with the same model augmented with LoT using only visual highlighting. Results are reported on RealWorldQA, V-Star, TextVQA, and ChartQA. Avg denotes the average score across the four benchmarks; improvements over the baseline are shown in green.

	RealWorldQA	V-Star	TextVQA	ChartQA	**Avg**	
▼ Qwen2.5-VL-3B	59.1	66.5	62.5	79.1	66.8	
+ LoT (Ours)	**61.7**	**71.2**	**66.4**	**79.5**	**69.7**	**△+2.9**
▼ Qwen3-VL-4B	66.7	70.2	74.7	80.7	73.1	
+ LoT (Ours)	**71.6**	**78.5**	**76.5**	**82.3**	**77.2**	**△+4.2**
▼ Qwen2.5-VL-7B	65.0	71.2	75.7	77.0	72.2	
+ LoT (Ours)	**67.5**	**73.3**	**77.9**	**79.4**	**74.5**	**△+2.3**
▼ Qwen3-VL-8B	66.8	73.3	76.9	82.0	74.8	
+ LoT (Ours)	**69.7**	**77.0**	**77.2**	**82.8**	**76.7**	**△+1.9**
▼ Qwen2.5-VL-32B	65.6	**73.8**	72.4	79.7	72.9	
+ LoT (Ours)	**67.1**	72.3	**74.0**	**82.2**	**73.9**	**△+1.0**

regions improves visual evidence utilization even without retrieved textual context, extending the applicability of LoT beyond KB-VQA.

C Additional Implementation Details

C.1 Pre-Processing and Model-Specific Details

Target Object Identification. As described in the main paper, visual evidence highlighting requires identifying the target object referred to in the question. We denote by $\mathcal{T}_{\text{obj}}$ the set of indices of the input tokens corresponding to the queried object. These indices are then used to select the row indices of the submatrix $\mathbf{A}^{\ell,k}_{\text{obj}\to\text{vis}}$. To extract $\mathcal{T}_{\text{obj}}$, we employ the spaCy toolkit[2], using dependency parsing and part-of-speech annotations through a lightweight, linguistically grounded procedure.

Given an input question, we analyze its syntactic structure to identify the noun phrase that captures the focus of the question. In particular, we consider nouns associated with interrogative constructions (*e.g.*, *what*, *which*, *who*) as well as nouns involved in relevant syntactic relations, such as verbal or prepositional dependencies. Once a candidate head noun is identified, we expand it to the corresponding full noun phrase by including syntactically attached modifiers, such as compound nouns and adjectival modifiers, obtaining a semantically coherent span that represents the queried object [3]. If this procedure does not identify a valid target-object span, we fall back to using the last token in input (*i.e.*, the same token used for text highlighting) as $\mathcal{T}_{\text{obj}}$. This fallback ensures that visual-evidence extraction remains well-defined even for questions whose syntactic structure does not contain an explicit or reliably detectable target object[4].

Table 8: Coverage of the dataset-specific entity-selection procedure.

Dataset	**#Q**	**Found**	**Fallback**
E-VQA	5,750	99.93%	0.07%
InfoSeek	71,335	100.00%	0.00%
OVEN	3,291	91.01%	8.99%
ViQuAE	1,257	98.01%	1.99%
Overall	81,633	99.60%	0.40%

To validate this step, we measure how often the entity-selection procedure identifies a valid textual target for extracting the visual attention map. The results in Table 8 reflect the different linguistic characteristics of the four datasets. E-VQA and InfoSeek mostly contain templated questions in which the visual entity is explicitly introduced by demonstrative expressions such as *this*, *that*, *these*, or *those*. In these cases, the nominal expression associated with the demonstrative provides a reliable attention target. Consequently, the procedure identifies an entity in 99.93% of E-VQA questions and all InfoSeek questions. OVEN contains less regular, often short or underspecified questions. Nevertheless, a valid noun phrase is identified in 2,995 of the 3,291 questions (91.01%). The remaining 296 questions

[2]Specifically, we use the `en_core_web_sm` model available at `https://spacy.io/models/en`.

[3]For example, in the sentence "*John James Audubon became famous for painting what?*", *John James Audubon* is selected as the visual subject.

[4]For example, "*What is green?*" contains no noun phrase denoting the object that should be localized.

often describe an attribute without explicitly naming the corresponding object. ViQuAE instead contains longer, more compositional questions that may mention multiple entities. We therefore use named-entity recognition and select the last named entity detected in the question, which generally corresponds to the entity closest to the requested information. This strategy identifies a valid target in 1,232 of the 1,257 questions (98.01%), with the last-token fallback used for the remaining 25.

Context Sentence Splitting. To obtain sentence-level relevance scores (*e.g.*, $\hat{\mathbf{a}}_{\mathrm{txt}}$) from the token-level attention vector $\mathbf{a}_{\mathrm{txt}}$, we first segment the retrieved context into sentences using the English spaCy pipeline, which is also employed for target-object extraction. Specifically, given a context $\mathbf{X}_C$, we obtain its sentences as $[\mathcal{S}(\mathbf{X}_C) = s_1, \ldots, s_M]$. For each sentence s_i, we identify the corresponding context tokens and compute its relevance score by averaging their token-level attention values. The resulting sentence-level scores are then used to determine which sentences should be highlighted, retained as supporting context, or discarded.

Model Sink Dimensions. Table 9 reports the hidden dimensions identified as attention sinks and filtered for each evaluated model. To determine them, we process 100 randomly sampled examples from the E-VQA validation split and rank the dimensions of the `BOS` hidden representation according to their average activation magnitude. We then select the dimensions that consistently exhibit the highest values across the analyzed samples and keep them fixed in all subsequent experiments.

Table 9: Attention-sink dimensions filtered for each MLLM.

Model	Sink dimensions
Qwen2-VL-2B-Instruct	1073, 534, 940
Qwen2-VL-7B-Instruct	2570, 458
Qwen2.5-VL-3B-Instruct	318, 1874, 1819
Qwen2.5-VL-7B-Instruct	458, 2570
Qwen2.5-VL-32B-Instruct	4675, 3094
Qwen3-VL-4B-Instruct	0
Qwen3-VL-8B-Instruct	1838
InternVL3.5-4B	4, 396, 0
InternVL3.5-8B	2276, 233
InternVL3.5-38B	731

Under standard causal self-attention, the `BOS` token cannot attend to any subsequent token. Therefore, as long as it always occupies the first sequence position and is encoded identically, its hidden representation is expected to be largely input-independent, making the associated sink dimensions stable across examples. Evaluating multiple samples serves primarily to verify this stability empirically and to reduce the influence of model-specific numerical effects or input-formatting artifacts on the selected dimensions.

C.2 Additional Baseline Implementation Details

We evaluate the compatibility of LoT with reasoning, passage-filtering, and re-ranking strategies. In each case, the underlying method first processes the retrieved context according to its original design, after which LoT refines the resulting visual and textual evidence before answer generation.

Chain-of-Thought Prompting (Wei et al., 2022). We augment the standard generation prompt by instructing the MLLM to reason step by step over the image and retrieved context before producing the final answer. The backbone, retrieved documents, and decoding settings remain unchanged, requiring neither training nor additional models.

Prompt-based Critic (Compagnoni et al., 2026). Each passage from the top-3 retrieved documents is independently evaluated by the same zero-shot MLLM used for answer generation, which is prompted to classify it as useful or irrelevant given the image and question following the original prompt used in (Compagnoni et al., 2026). Only passages classified as useful are retained and provided to the final generation step.

Trained Critic (Compagnoni et al., 2026). We also employ the official Critic model proposed in (Compagnoni et al., 2026), based on Qwen2.5-VL-3B and fine-tuned for passage-relevance classification. Following its original configuration, the Critic model filters passages from the top-20 retrieved documents before they are provided to the generator.

EchoSight Re-ranker (Yan & Xie, 2024). We divide the top-3 retrieved documents into sections and adapt the trained EchoSight Q-Former re-ranker to score each candidate section according to its relevance to the image-question pair. Unlike the complete EchoSight pipeline, we retain our original retrieval results and use only its multimodal re-ranking component. The sections are then ordered according to their relevance scores, and the top-5 are provided to the answer generator. This

Table 10: Selection of α on the E-VQA validation split using Qwen2.5-VL-3B.

α	Acc.	Prec.	Rec.	Avg. sent.
$0.25\ \max(\hat{\mathbf{a}}_{\text{txt}})$	32.14	57.72	93.89	5.86
$\mathbf{0.50}\ \max(\hat{\mathbf{a}}_{\text{txt}})$	**32.61**	62.07	77.62	2.03
$0.75\ \max(\hat{\mathbf{a}}_{\text{txt}})$	31.36	68.97	60.44	1.04

Table 11: Selection of α_{drop} on the E-VQA validation split using Qwen2.5-VL-3B.

α_{drop}	Acc.	Ctx. red.
$\alpha/2.0$	31.59	74.47%
$\alpha/\mathbf{3.0}$	**32.61**	70.45%
$\alpha/4.0$	31.92	66.71%

configuration keeps the initial retrieval pipeline and the final generator fixed, isolating the effect of the re-ranking component.

Qwen3-VL-Reranker-8B (Li et al., 2026). Following the same section-level setting, we use the released Qwen3-VL-Reranker-8B to estimate the relevance of each candidate section to the multimodal image-question query. The sections are ordered according to the resulting relevance scores, and the top-5 are provided to the answer generator. The initial retrieval pipeline and the final generator remain unchanged.

C.3 Computing Infrastructure

Experiments were conducted on compute nodes equipped with four NVIDIA A100 GPUs with 64 GiB of HBM2e memory each, one 32-core Intel Xeon Platinum 8358 CPU at 2.6 GHz, and 512 GiB of DDR4 system memory. All model-inference jobs were restricted to a single compute node. The standard allocation comprised one GPU, 8 CPU cores, and 128 GiB of system memory. Depending on the memory requirements of the backbone, individual jobs used between one and four GPUs. The largest model configurations used all four GPUs, corresponding to 256 GiB capacity of GPU memory. Dataset shards were evaluated independently using Slurm array jobs; no multi-node model parallelism was employed.

The experiments were run with Python 3.10.19, PyTorch 2.6.0 (CUDA 12.6 build), Transformers 4.57.1, Accelerate 1.12.0, FlashAttention-2 2.8.3, and FAISS-GPU 1.7.2. Image processing used Pillow 10.3.0 and OpenCV 4.11.0, while spaCy 3.7.5 was used for the target-object extraction procedure. Since LoT is training-free, the reported computation concerns retrieval, inference, and evaluation only; no model parameters were optimized.

D Hyperparameter Analysis and Discussion

D.1 Textual Evidence Selection

Selection of α. We select α on the E-VQA validation split. Sentence precision measures the fraction of highlighted sentences found in the oracle document, whereas document recall measures the fraction of retrieval-hit samples containing at least one such sentence. As shown in Table 10, setting α to 0.5 of the maximum sentence-level attention score provides the best trade-off: it preserves substantially higher recall than the 0.75 setting, while improving precision and eliciting considerably fewer sentences than the 0.25 setting.

Selection of α_{drop}. We select the sentence-dropping threshold on the E-VQA validation split, by evaluating $\alpha_{\text{drop}} \in \{\alpha/2, \alpha/3, \alpha/4\}$. As shown in Table 11, the most aggressive setting reduces the context by 74.47%, but yields lower accuracy, while retaining more context with $\alpha_{\text{drop}} = \alpha/4$ does not improve performance. We therefore adopt $\alpha_{\text{drop}} = \alpha/3$, which achieves the highest validation accuracy while reducing the retrieved context by 70.45%. For InternVL3.5, we note that sentence-level attention distributions are more sharply peaked, causing the default threshold $\alpha/3$ to discard approximately 84% of the context. We therefore use $\alpha/6$, reducing the discarded fraction to approximately 70% and aligning its filtering behavior with the other model families.

D.2 Visual Evidence Selection

Attention Map Filtering. To obtain spatial attention maps, we project the attention scores of visual tokens back onto their corresponding image regions, accounting for each visual tokenization scheme of the model. For the Qwen2-VL, Qwen2.5-VL, and Qwen3-VL checkpoints used in our experiments, the vision encoder merges each 2×2 spatial group of neighboring patches into a single

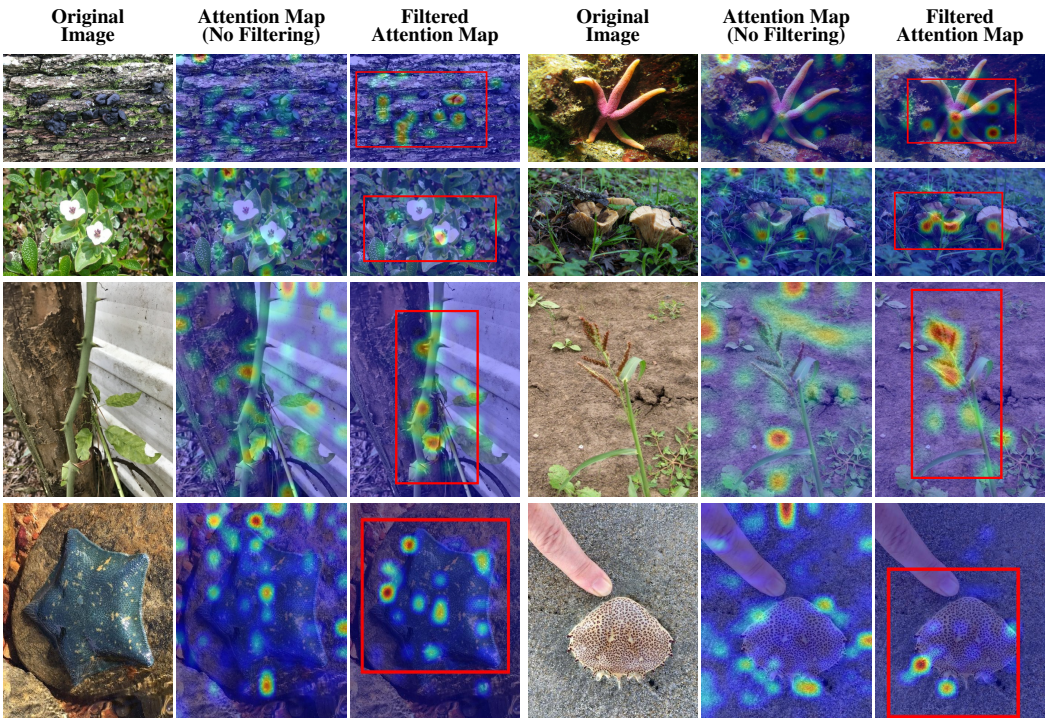


Figure 5: Qualitative examples of attention-map filtering.

visual token, with a spatial merge size of 2. We therefore reconstruct the post-merging visual grid, whose height and width are those of the original patch grid divided by 2.

For InternVL3.5, we exclude the final set of visual tokens corresponding to the global image thumbnail. The remaining patch representations are arranged into their original two-dimensional grid and processed by the pixel-shuffle operation of the model with a downsampling ratio of 0.5. Consequently, each visual token combines the features of a spatial 2×2 group of vision-encoder patches. When dynamic image tiling is used, we reconstruct the attention map independently for each image tile and map it back to the corresponding region of the original image.

Fig. 5 shows some qualitative examples of the effect of attention-sink filtering. The raw attention maps contain scattered activations, often far from the target object, which can lead to inaccurate bounding-box estimation. After filtering, the attention becomes more concentrated around the queried object, producing the more accurate localizations shown by the red bounding boxes.

Bounding Box Extraction Method. To extract visual evidence from the model, we experiment with several strategies to aggregate filtered attention maps and derive bounding boxes that localize the relevant visual regions. We consider three main bounding box extraction methods:

- **Min-Max Coordinates:** This approach considers all non-zero values in the attention map and computes the bounding box by taking the minimum and maximum coordinates along each axis.
- **Morphological Filtering:** The attention map is first thresholded over 0.1 to create a binary map. Morphological closing operations (dilation followed by erosion) are then applied to obtain a connected region, and the bounding box is computed around it.
- **Weighted Centroid:** This method computes the centroid of the attention map weighted by the attention values themselves. The bounding box is then defined around the centroid using the standard deviation of the attention distribution along each axis, scaled by a multiplier to control box size.

For this analysis, we focus on E-VQA. Since the dataset does not provide bounding-box annotations for the queried objects, we use predictions from GroundingDINO (Liu et al., 2024b) (Base), an open-vocabulary object detector, as proxy ground truth. Table 12 reports results in terms of Intersection over Union (IoU), Coverage, Precision, and Center Distance. As shown, the weighted centroid

Table 12: Comparison of bounding box extraction methods for target object localization.

Method	IoU ↑	Coverage ↑	Precision ↑	Center Distance ↓
Min-Max Coordinates	0.352	0.891	0.500	0.090
Morphological Filtering	0.399	0.582	0.716	0.084
Weighted Centroid	0.487	0.814	0.626	0.071

Table 13: Effect of attention-sink filtering on visual evidence localization.

Configuration	IoU	Acc@0.5	Coverage	Precision
w/o sink filtering	0.388	0.083	**0.999**	0.388
w/ sink filtering	**0.489**	**0.441**	0.869	**0.554**

method consistently achieves the best trade-off between coverage and IoU scores, indicating that it captures the most relevant regions of the attention map while remaining tightly localized around the target object. The min-max method produces overly large and sparse bounding boxes, leading to lower precision, while the morphological approach generates compact regions but occasionally fails to cover all relevant attention peaks. Overall, the weighted centroid method provides the best trade-off between accuracy and compactness, making it the most effective strategy for bounding box extraction.

Bounding Box Extraction Hyperparameters. On E-VQA, GroundingDINO detections cover approximately 48% of the image area, with the median bounding-box ratio falling within the 40–50% range. In particular, about 53% of the detected regions occupy less than half of the image. We therefore focus the hyperparameter analysis on relatively compact entities, for which localization quality is especially critical: tighter boxes enable the resulting crop to isolate the relevant visual evidence while excluding more of the surrounding background.

Furthermore, Table 14 reports a grid search on the COCO (Lin et al., 2014) validation split. We restrict the analysis to objects whose ground-truth bounding boxes cover between 25% and 50% of the image area, as these relatively compact entities provide a more informative test of whether the attention-based procedure can isolate the target from the surrounding scene. Since COCO is an object-detection benchmark and does not provide question-answer pairs, we generate a synthetic query for each evaluated object category that explicitly refers to the target entity. The question and image are then provided to the MLLM, and the bounding box extracted from its attention map is evaluated against the corresponding COCO annotation. Across the tested configurations, a standard-deviation multiplier of $\beta = 2$ consistently provides the strongest localization performance. Combined with a sink threshold at the 25th percentile, it achieves the highest mean IoU and Acc@IoU≥ 0.5, and is therefore used in all subsequent experiments.

Effect of Attention-Sink Filtering. To isolate the contribution of attention-sink filtering, we apply the same weighted-centroid localization procedure with and without removing sink-associated activations. Following the hyperparameter analysis, we evaluate both variants on the COCO validation subset containing objects that occupy 25–50% of the image area, while keeping all other localization settings fixed. As shown in Table 13, filtering produces tighter and more accurate regions, substantially improving IoU, Acc@IoU≥ 0.5, and precision, with only a moderate reduction in coverage. Without filtering, coverage is nearly 100% because attention sinks often occur near the image boundaries, as also shown in Fig. 5. Removing them therefore yields a better overall localization trade-off.

Importantly, the relatively aggressive threshold τ only excludes sink tokens from estimating the target location, not from the final crop. Because the crop is defined around the weighted centroid, filtered patches may still be included within the resulting bounding box.

This design is particularly suitable for KB-VQA, where the queried entity is often small relative to the full image. Restricting localization to the strongest object-conditioned activations reduces the influence of residual attention on background regions and yields a more precise centroid. The subsequent box construction, controlled by $\beta = 2$, then restores an appropriate spatial margin around this centroid, compensating for the reduced set of contributing tokens and covering the target more completely. Together, these properties explain why the selected sink threshold improves, rather than harms, localization performance in Table 14.

Table 14: Analysis of bounding-box extraction hyperparameters applied to the filtered attention maps. We report mean IoU and localization accuracy at IoU ≥ 0.5. P_q denotes the q-th percentile of the sink-score distribution.

τ	β	IoU ↑	Acc@IoU≥0.5 ↑
P_{10}	1	0.267	0.023
P_{10}	2	0.461	0.360
P_{10}	3	0.405	0.170
P_{25}	1	0.317	0.045
P_{25}	2	**0.489**	**0.441**
P_{25}	3	0.388	0.090
P_{50}	1	0.342	0.087
P_{50}	2	0.484	0.408
P_{50}	3	0.374	0.029
P_{75}	1	0.356	0.103
P_{75}	2	0.481	0.399
P_{75}	3	0.368	0.013

Table 15: Layer-range ablation on the E-VQA validation split using Qwen2.5-VL-3B. When varying the textual range, visual highlighting is fixed to the middle half; when varying the visual range, textual highlighting is fixed to the last half.

Layer range	Single-Hop	All
Text Highlighting		
First half	21.6	21.3
Middle half	30.1	28.4
Last half	**35.3**	**32.6**
Visual Highlighting		
First half	35.2	**32.6**
Middle half	**35.3**	**32.6**
Last half	34.1	31.9

D.3 Layer-Range Selection

Prior work indicates that textual evidence is primarily consolidated in deeper decoder layers, whereas visual grounding tends to emerge in intermediate layers (Kang et al., 2025a; Liu et al., 2025; Jiang et al., 2025). Guided by these findings, we set L_{txt} to the last half of the decoder and L_{vis} to its middle half. Table 15 empirically validates these literature-guided choices. When varying one modality-specific range, the other is kept fixed to its selected configuration. The final half yields the best performance for textual highlighting, while the middle half achieves the highest Single-Hop accuracy and ties the best overall result for visual highlighting.

E Additional Details on Computational Analysis

LoT requires only a single additional token generation step, followed by lightweight post-processing of the extracted attention maps. The entire procedure requires neither additional training nor model fine-tuning.

Our implementation is designed to avoid large memory footprint by monkey-patching the forward pass. We retain standard FlashAttention (version 2) for all attention computations while restricting the attention matrix computation to a small subset of QK^T slices, resulting in $(|\mathcal{T}_{\text{obj}}|+1)\times S$ instead of $S \times S$, where $(|\mathcal{T}_{\text{obj}}| + 1) \ll S$ with long contexts since $|\mathcal{T}_{\text{obj}}|$ is on average $\sim$2.

Token Reduction. To quantify the impact of these operations on E-VQA test split with Qwen2.5-VL-3B, we measure the average number of input and output tokens. LoT substantially reduces the multimodal input used for final answer generation by an average of 2.3k tokens per sample, corresponding to a 71.93% reduction. Specifically, it removes an average of 124.0 visual tokens (54.24%) and 2.1k text tokens (72.90%) per sample. Thus, LoT produces a substantially more compact multimodal input, primarily by filtering irrelevant textual context while also reducing the visual token count.

Latency Breakdown. We further measure the average per-sample latency of the two stages on E-VQA test split using Qwen2.5-VL-3B. Baseline answer generation requires 1.057 seconds per sample. LoT spends 0.341 seconds on multimodal evidence highlighting and 0.866 seconds on answer generation from the refined input. Therefore, LoT requires 1.207 seconds per sample, corresponding to an absolute increase of 0.150 seconds and a 14.1% latency overhead. Notably, generation from the refined input is faster than baseline generation, decreasing from 1.057 to 0.866 seconds per sample, an 18.09% reduction.

Computational Cost. We also estimate the floating-point operations required by each stage on E-VQA test split using Qwen2.5-VL-3B. Baseline answer generation requires 19.995 TFLOPs per sample, whereas generation from the LoT-refined input requires only 3.783 TFLOPs, corresponding to an 81.08% reduction. The evidence-selection stage requires 24.767 TFLOPs, resulting in a total

cost of 28.551 TFLOPs per sample for the complete pipeline. This represents a 42.79% increase relative to baseline answer generation.

Run-to-Run Variability. All evaluations use greedy decoding with temperature set to 0 and sampling disabled. Thus, no stochastic token sampling is involved, removing the main source of run-to-run variation in generated answers. We therefore report a single evaluation run for each model–dataset configuration rather than averages and standard deviations over repeated runs. Repeating the complete evaluation across all large-scale MLLMs and benchmarks would incur substantial computational cost while providing limited additional information about sampling variability.

F Additional Details on KB-VQA Benchmarks

In KB-VQA, models are required to answer questions by retrieving and leveraging information from an external knowledge base, typically composed of structured or unstructured resources such as Wikipedia. We evaluate our approach on several representative KB-VQA datasets, all of which link images and questions to Wikipedia entities and provide corresponding knowledge bases.

Encyclopedic-VQA (Mensink et al., 2023). This dataset contains 221k question-answer pairs, each associated with up to five images and covering 16.7k fine-grained entities. Questions are divided into *single-hop* and *two-hop* types: single-hop questions can be answered using information from a single Wikipedia page, whereas two-hop questions require sequential retrieval across multiple pages. The dataset is divided into training, validation, and test splits comprising 1M, 13.6k, and 5.8k samples. Main results are reported on the test split, while hyperparameters are selected on the validation split. The dataset also provides an external knowledge base derived from Wikipedia, comprising about 2M pages, each including the article title, textual content, and associated images. In our experiments, we employ the original 2M-page knowledge base provided with the dataset.

InfoSeek (Chen et al., 2023). It contains 1.3M image-question-answer triplets associated with about 11k unique Wikipedia entities. The dataset is divided into training, validation, and test splits, comprising roughly 934k, 73k, and 348k samples, respectively. Both validation and test sets include questions about unseen entities. InfoSeek also provides an external knowledge base covering approximately 6M Wikipedia entities. Following previous works (Caffagni et al., 2024b; Cocchi et al., 2025), experiments are conducted using a knowledge base of 100k pages and employing 3-shot examples in the prompt to align the response with the dataset format.

OVEN (Hu et al., 2023). This dataset is designed to link an image and a textual query to the corresponding Wikipedia entity, covering approximately 6M entities. OVEN includes nearly 5M training examples, along with validation and test splits, as well as a human-verified evaluation set. Importantly, the dataset explicitly distinguishes between *seen* entities (available during training) and *unseen* entities, enabling evaluation of the model ability to generalize to novel concepts. In our experiments, we employ the same knowledge base used in InfoSeek and report results on the query split of the dataset.

ViQuAE (Lerner et al., 2022). It is a knowledge-based VQA dataset focused on entity-centric questions grounded in Wikipedia. It contains approximately 3.7k question-answer pairs associated with images and entity-centric queries. Each example links an image to a corresponding Wikipedia entity and requires retrieving relevant encyclopedic knowledge to answer the question. In our experiments, we employ the same knowledge base used in Encyclopedic-VQA. In the main results, we report the exact-match metric of the dataset.

G Multimodal Inference Prompts with Markers

After the relevant textual sentences and visual regions have been identified and extracted, we explicitly guide the model to attend to these elements during its response generation process by marking them with special tokens. For the textual highlighting, sentences selected as relevant are inserted between the tokens: `<START_IMPORTANT_TXT>` and `<END_IMPORTANT_TXT>`. While in the visual highlighting, the original image is cropped employing the extracted bounding box and inserted between the markers: `<START_IMPORTANT_IMG>` and `<END_IMPORTANT_IMG>`. These markers do not represent single tokens, they will simply be tokenized into multiple tokens by the tokenizer. The prompt-box below illustrates the final structure of the user template used at inference time.

User Template

```
<START_IMPORTANT_IMG>
[image_cropped]
<END_IMPORTANT_IMG>
[question]
The following paragraphs may contain useful information to help
answer the question correctly:
[sentence]
...
<START_IMPORTANT_TXT>
[evidence_sentence]
<END_IMPORTANT_TXT>
...
[sentence]
```

Finally, the prompt box below presents the system prompt. For each modality, the prompt explicitly instructs the model to prioritize and ground its reasoning in the content enclosed within the corresponding special markers.

System Prompt

```
Answer the encyclopedic question about the given image.  Don't
mention the visual content of the image in your output.  Directly
output the answer of the question according to the context.
If the paragraphs do not contain the information required to answer
the question, you should answer the question using your knowledge.
<START_IMPORTANT_IMG> and <END_IMPORTANT_IMG> are used to mark the
important visual evidence.  Do not output the markers.
<START_IMPORTANT_TXT> and <END_IMPORTANT_TXT> are used to mark the
important textual evidence.  Do not output the markers.
```

H Additional Qualitative Results

We also report additional qualitative results in Fig. 6. The examples show how the model identifies the main subject of the question and localizes it in the image (red bbox), allowing it to focus on the most informative visual regions. On the textual side, LoT highlights the sentences within the retrieved documents that are most relevant for answering the question.

I Limitations

LoT relies on a two-stage inference procedure. Although the refined input substantially reduces the cost of final answer generation, the model must still perform an additional forward pass and repeat the prefill stage, increasing overall latency and computational cost. Moreover, LoT requires access to internal attention maps and hidden representations, and is therefore not directly applicable to black-box models exposed only through APIs. For syntactically atypical questions, the lightweight spaCy parser may fail to identify the target object, in which case we fall back to the final input token; nevertheless, this remains substantially cheaper than invoking a separate billion-parameter LLM. The implementation also introduces non-negligible CPU-side preprocessing for token-to-context alignment and sentence-span construction, although these operations could be further optimized. Finally, LoT treats internal attention as a heuristic relevance signal rather than a faithful explanation of the model prediction. Consequently, inaccurate relevance estimates may discard useful evidence, particularly when it is diffuse across the image or distributed across multiple textual passages. Performance is also sensitive to the filtering strength, as shown in Table 10, highlighting the trade-off between removing distracting context and preserving potentially useful evidence.

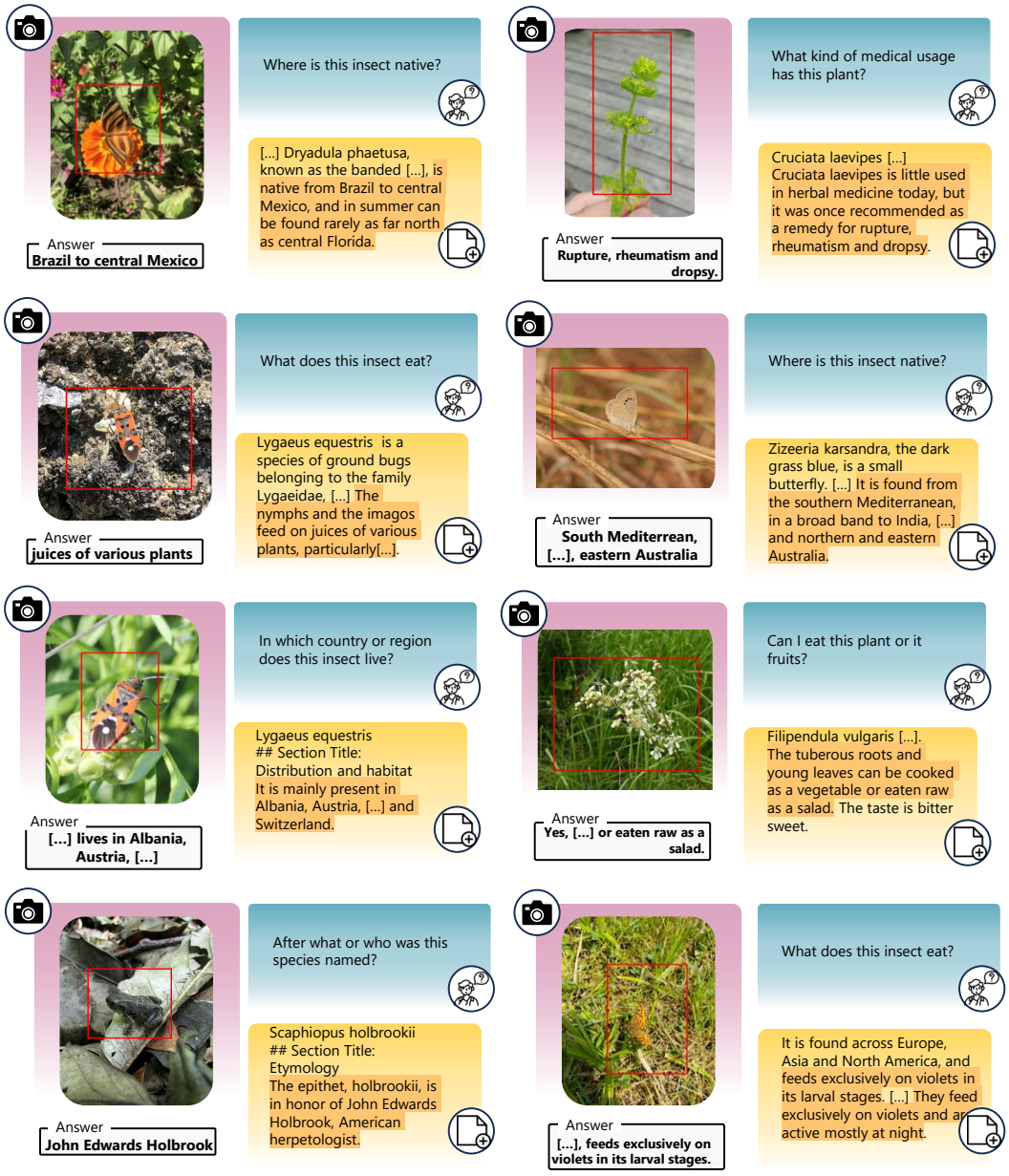


Figure 6: Qualitative examples of LoT highlighting query-relevant visual regions and textual evidence, enabling the model to generate the correct answer on E-VQA.