LATENT NOISE MASK ELIMINATES VISUAL REDUNDANCY OF MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) often fail in fine-grained visual reasoning, as question-relevant visual cues are diluted by dense and redundant image tokens. Recent multimodal reasoning methods usually extend chain-of-thought from language models into visual or latent spaces, seeking to add intermediate reasoning states while overlooking the negative impact of redundant visual tokens. We propose **L**at**E**nt **N**oise ma**S**k (LENS), a question-conditioned visual evidence purification framework that empowers MLLMs to reason with cleaner visual cues in latent space. LENS introduces a lightweight *Lens Evidence Token* (LET) to score which visual tokens support the current question and preserve them during decoding. Guided by the LET scores, it injects adaptive latent noise into low-relevance tokens, softly suppressing distractors without changing the model backbone or token sequence. With only one temporary learnable control token and a lightweight noise generator, LENS adds minimal overhead while improving the base MLLM by 2.4–6.4 points on most VQA datasets and by 4.1–6.4 points on grounding tasks. These results show that multimodal reasoning can benefit more directly from cleaner question-relevant visual evidence than from simply extending the reasoning trace.

1 INTRODUCTION

Multimodal large language models (MLLMs) have become a central interface for visual understanding and multimodal reasoning (Bai et al., 2025; Li et al., 2024; Wang et al., 2025b). By mapping images into token sequences that can be processed by large language models, they inherit strong language reasoning ability and support tasks such as visual question answering, spatial reasoning, chart understanding, and visual grounding. However, reliable visual reasoning *often depends on a few local cues rather than the whole image*. Accordingly, many failures arise when small objects, fine-grained attributes, or spatial relations are diluted by background regions, nearby distractors, or language priors (Wang et al., 2025c; Zhang et al., 2024). **The key bottleneck is not only weak reasoning but also the difficulty of isolating question-relevant evidence from dense visual tokens.**

Recent works address this gap by making MLLMs reason longer or reason in richer spaces. Textual chain-of-thought prompting encourages step-by-step inference and has been extended to multimodal tasks (Wei et al., 2022; Zhang et al., 2023). Other methods introduce visual cues through image crops, bounding boxes, visual tools, or image generation (Shao et al., 2024; Zheng et al., 2025; Su et al., 2025). Latent-space methods further move intermediate reasoning into continuous representations, e.g., Mirage interleaves latent visual tokens with text, DMLR refines latent think tokens and injects relevant patches, and VisMem stores latent vision memories during generation (Yang et al., 2025; Liu et al., 2025a; Yu et al., 2025b). **Most recent multimodal reasoning methods improve reasoning by adding more textual, visual, or latent context.** These studies show that visual and latent spaces can support reasoning beyond pure text, but they fail to directly reduce the influence of irrelevant image tokens.

However, this additive view leaves a simpler question underexplored. Images are already encoded as dense visual token sequences, and many tokens are unrelated to a specific question. For instance, when answering the color of a small object, background tokens and unrelated objects may dominate the context. When grounding a target object, visually similar neighbors may provide strong but misleading cues. **When the needed evidence is already present, the main challenge becomes**

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

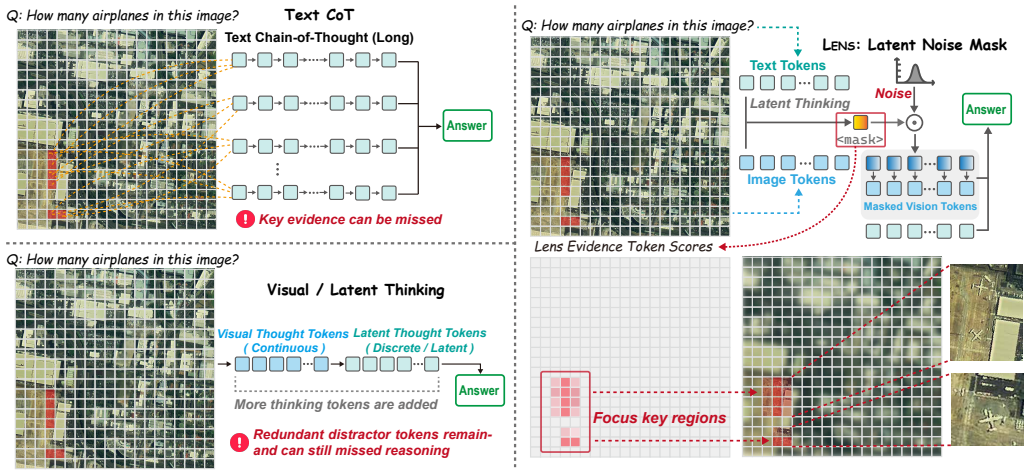


Figure 1: Motivation of LENS. Existing multimodal reasoning methods often add visual or latent reasoning states, yet redundant image tokens can still distract the model. LENS learns *Lens Evidence Token* (LET) to score visual evidence and softly suppress irrelevant visual tokens in latent space.

visual selectivity rather than visual availability. In such cases, adding more reasoning context not only leaves distractors untouched, but can also make decoding harder by increasing the amount of irrelevant context that the model must sift through.

In this paper, we study multimodal reasoning from the perspective of visual evidence purification. **LENS learns which visual tokens support the current question and suppresses the remaining tokens before they interfere with decoding.** This differs from token pruning (Yu et al., 2025a; Bigverdi et al., 2025) or hard region selection (Shao et al., 2024; Fu et al., 2025). Removing tokens can be brittle when the relevance prediction is imperfect, and it may change the token layout expected by the backbone. Instead, we use a soft latent-space intervention that keeps the original sequence structure while reducing the influence of distractor tokens. As shown in Fig. 1, the goal is not to generate longer textual thoughts or add more visual thoughts, but to clean the visual evidence that the model already receives (An et al., 2026; Shao & Li, 2025).

To this end, we propose Latent Noise Mask, denoted as LENS, a question-conditioned visual evidence suppression framework for MLLMs. **LENS provides a lightweight question-conditioned intervention that purifies visual evidence without external tools, costly rationales, or backbone modification.** It introduces a temporary *Lens Evidence Token* (LET) that estimates the relevance of each visual token to the current question. Instead of using expensive chain-of-thought traces, LENS uses object-level annotations as evidence supervision by mapping question-relevant boxes to visual patches. Guided by the LET scores, it preserves evidence tokens and injects adaptive latent noise into low-relevance tokens (Li, 2024). The backbone and token sequence structure remain unchanged, making the framework easy to attach to existing MLLMs.

Our experiments show that this shift from visual addition to visual suppression is effective. Across VQA and grounding benchmarks, LENS improves the base MLLM by 2.4–6.4 points on most VQA datasets and by 4.1–6.4 points on grounding tasks. Qualitative results further show that the predicted LET scores align with question-relevant regions and reduce attention to distractors. **The results show that cleaner question-relevant visual evidence can improve multimodal reasoning more directly than simply extending the reasoning trace.** Our contributions are summarized as follows:

- We reveal that question irrelevant visual distractors in dense visual tokens can compete with key evidence and hinder fine grained visual reasoning, which motivates us to explore for suppressing visual interference than simply adding more tokens with visual information.
- We propose a lightweight visual evidence purification framework that learns question conditioned token relevance and softly suppresses low relevance visual tokens in latent space.
- Experiments show the advantage in both accuracy and efficiency over other methods, demonstrating the effectiveness and generality of reasoning with cleaner visual evidence.

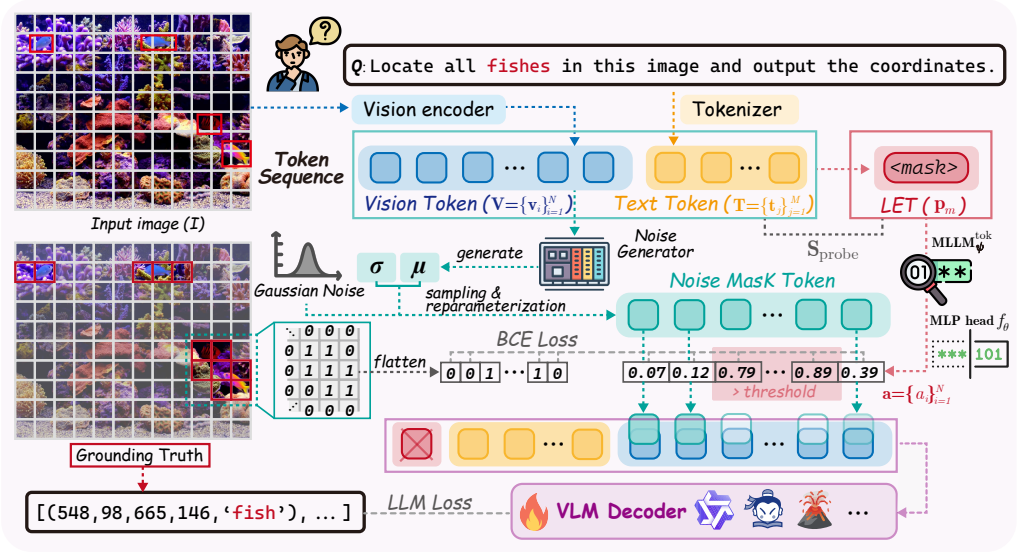


Figure 2: Overview of LENS. Evidence probing appends a temporary ⟨mask⟩ token, whose hidden state is decoded by a simple MLP head f_θ into the LET scores. A threshold τ converts the scores into a suppression gate, and a small noise generator predicts token-specific latent noise for low-gate visual tokens. The control token is removed before final decoding, so the MLLM receives the purified sequence $\widetilde{\mathbf{V}}$ with the original token layout.

2 RELATED WORK

Visual and multimodal reasoning Chain-of-thought prompting improves language reasoning by exposing intermediate rationales (Wei et al., 2022), and has been adapted to multimodal tasks that require visual evidence (Zhang et al., 2023; Mondal et al., 2024). Recent MLLM methods add visual evidence through grounded rationales, selected regions, structured prompts, visual sketching, image editing, or tool use (Shao et al., 2024; Gao et al., 2025; Mitra et al., 2024; Hu et al., 2024; Zheng et al., 2025; Su et al., 2025; Fu et al., 2025; Li et al., 2025a; Zhao et al., 2025; Wang et al., 2025a; Fan et al., 2025). These methods show that explicit visual context can improve perception and reasoning, but they usually expand the reasoning trace or modify the visual input. **LENS addresses a complementary problem. It keeps the original image sequence and suppresses question-irrelevant visual tokens before decoding.**

Latent perception and token-level intervention Latent-space reasoning moves intermediate computation into continuous states, reducing the cost of generating long textual chains (Hao et al., 2024; Li et al., 2025b). Multimodal latent methods further interleave, optimize, predict, or retrieve latent visual context for reasoning (Yang et al., 2025; Liu et al., 2025a; Qin et al., 2025; Yu et al., 2025b). Perception-aware and token-level methods also improve visual reasoning through reinforcement learning, perception objectives, selected visual tokens, or visual prompts (Liu et al., 2025b; Shen et al., 2025; Huang et al., 2025; Wang et al., 2025c; Chen et al., 2025; Lei et al., 2025; Yu et al., 2025a; Bigverdi et al., 2025). LENS is closest to this line, but differs in the intervention target. Rather than adding memory, retrieving patches, or pruning tokens, **it introduces a question-conditioned *Lens Evidence Token* and applies latent noise to tokens with low LET scores while preserving the original token layout.**

3 METHODOLOGY

3.1 PROBLEM SETUP AND METHOD OVERVIEW

Problem Setup. We consider an MLLM that receives an image–question pair (I, Q). The image encoder maps I into visual tokens $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^{N}$, and the tokenizer maps Q into text tokens $\mathbf{T} =$

$\{\mathbf{t}_j\}_{j=1}^{M}$. A standard MLLM decodes the answer sequence $Y = \{y_l\}_{l=1}^{L}$ from

$$P_\psi(Y \mid \mathbf{V}, \mathbf{T}) = \prod_{l=1}^{L} P_\psi(y_l \mid \mathbf{V}, \mathbf{T}, y_{<l}). \tag{1}$$

This formulation gives all visual tokens the same access to the decoder. In many visual reasoning tasks, however, only a small subset of tokens provides direct evidence for the question. The remaining tokens often describe background regions, unrelated objects, or similar distractors. These tokens can weaken the useful evidence before answer generation.

Method Overview. As illustrated in Fig. 1 and 2, LENS aims to purify the visual context before decoding through two coupled stages, i.e., **evidence probing** and **latent suppression**. **Evidence probing** appends a temporary control token $\langle mask \rangle$ to the joint input $[\mathbf{V}, \mathbf{T}]$. The hidden state of this token is decoded into a question-conditioned LET score vector $\mathbf{a} \in [0,1]^N$, where $a_i \in [0,1]$ measures the reliability of visual token $\mathbf{v}_i$ for the current question; the control token is removed afterwards. **Latent suppression** converts the LET scores into suppression gates g_i, which guide the construction of a purified visual sequence $\widetilde{\mathbf{V}} = \{\widetilde{\mathbf{v}}_i\}_{i=1}^{N}$ by forming $\widetilde{\mathbf{v}}_i = \mathbf{v}_i + g_i \mathbf{r}_i$, where $\mathbf{r}_i$ denotes adaptive latent noise. The final answer is then decoded as $P_\psi(Y \mid \widetilde{\mathbf{V}}, \mathbf{T})$. In this way, LENS achieves visual selectivity while preserving the original visual token count and backbone interface, without cropping images, pruning tokens, generating intermediate images, or adding a persistent reasoning trace.

3.2 EVIDENCE PROBING

LENS estimates token relevance through a short probing path. Given $\mathbf{V}$ and $\mathbf{T}$, we append the temporary *Lens Evidence Token* $\langle mask \rangle$ after the text tokens and form

$$\mathbf{S}_{\text{probe}} = [\mathbf{v}_1, \ldots, \mathbf{v}_N, \mathbf{t}_1, \ldots, \mathbf{t}_M, \mathbf{p}_m], \tag{2}$$

where $\mathbf{p}_m$ represents the temporary $\langle mask \rangle$ control token in the joint MLLM input space. This token is appended only to the probing sequence and is not part of the original visual-token or text-token spaces. Then, deocoder $\text{MLLM}_\psi^{\text{tok}}(\cdot)$ outputs a sequence of hidden states based on $\mathbf{S}_{\text{probe}}$,

$$\mathbf{H}_{\text{probe}} = \text{MLLM}_\psi^{\text{tok}}(\mathbf{S}_{\text{probe}}), \tag{3}$$

$$\mathbf{h}_m = [\mathbf{H}_{\text{probe}}]_m. \tag{4}$$

Since $\mathbf{h}_m$ attends to both visual and textual tokens, it provides a compact question-aware summary for evidence prediction. The LET only opens this probing path and is not used during final answer decoding.

A single-layer MLP head f_θ maps the LET hidden state to a dense evidence score vector over visual tokens

$$\mathbf{a} = \sigma(f_\theta(\mathbf{h}_m)), \quad \mathbf{a} \in [0,1]^N. \tag{5}$$

Each score a_i indicates how likely $\mathbf{v}_i$ supports the answer. This LET score vector is conditioned on the question, so the same image region may be evidence for one question and a distractor for another.

We train the *Lens Evidence Token* with object-level evidence supervision instead of chain-of-thought traces. Let $\mathcal{B}_Q$ denote the question-relevant boxes, and let Ω_i denote the image patch covered by visual token $\mathbf{v}_i$. The token label is

$$z_i = \begin{cases} 1, & \exists B \in \mathcal{B}_Q,\ \Omega_i \cap B \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

A visual token is labeled positive if its image patch overlaps at least one question-relevant box, and is labeled negative otherwise. The LET supervision loss is

$$\mathcal{L}_{\text{LET}} = -\frac{1}{N} \sum_{i=1}^{N} \left[z_i \log a_i + (1 - z_i) \log(1 - a_i) \right]. \tag{7}$$

This supervision teaches the probe where visual evidence is located, while avoiding expensive rationale annotations.

Table 1: Main comparison on 10 benchmarks for visual understanding, reasoning, and grounding. VQA scores evaluate answer quality, F1@0.5 evaluates grounding quality, and averages summarize VQA, grounding, and overall performance. The **best** and second best values are marked.

Methods	General VQA Tasks							Grounding Tasks F1@0.5					Avg.
	CUB	GQA	OpenImg	SROIE	VSR	MSVQA	Avg.	COCO	Obj365	RUOD	Visdrone	Avg.	
Vanilla	68.60	64.06	50.00	88.71	66.54	50.41	64.72	12.88	9.12	9.42	20.44	12.97	44.02
SFT	86.45	72.65	82.08	92.49	76.58	63.96	79.04	46.50	40.02	63.63	39.60	47.44	66.40
Visual-RFT	88.21	68.95	73.74	94.28	72.28	62.14	76.60	55.89	36.81	50.24	46.32	47.32	64.89
VLM-R1	87.84	74.69	84.81	94.46	78.71	65.66	81.03	50.48	41.62	66.59	43.07	50.44	68.79
PAPO	86.96	70.97	74.86	93.64	75.99	60.55	77.16	50.43	38.45	55.26	42.05	46.55	64.92
VPT	91.06	76.37	85.04	94.36	80.18	65.71	82.12	31.54	25.29	47.99	23.32	32.04	62.09
LVR	90.28	72.61	78.68	90.45	76.93	63.50	78.74	49.77	39.89	64.78	42.34	49.20	66.92
DMLR	90.26	75.56	85.13	88.41	78.45	64.28	80.35	44.39	38.67	58.79	35.28	44.28	65.92
Vismem	89.77	71.44	80.58	90.67	75.45	62.78	78.45	45.29	39.25	62.16	37.67	46.09	65.51
LENS-SFT	**91.50**	79.01	86.89	95.20	80.69	66.39	83.28	52.00	44.10	70.05	44.09	52.56	70.99
LENS-GRPO	91.24	**83.95**	**88.10**	**95.92**	**82.43**	**67.55**	**84.87**	**60.93**	**51.32**	**74.51**	**58.41**	**61.29**	**75.44**

3.3 LATENT SUPPRESSION

The LET scores become effective only when they change how visual tokens influence decoding. LENS therefore converts $\mathbf{a}$ into a latent reliability gate. The gate preserves tokens with high evidence scores and perturbs tokens with low evidence scores in the same latent space used by the MLLM.

For each visual token $\mathbf{v}_i \in \mathbb{R}^d$, a lightweight noise generator G_ϕ predicts the mean and positive scale of a token-specific noise distribution:

$$(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) = G_\phi(\mathbf{v}_i), \quad \boldsymbol{\sigma}_i > 0. \tag{8}$$

We then sample $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and obtain latent noise by reparameterization

$$\mathbf{r}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}_i. \tag{9}$$

This makes the perturbation feature-aware rather than fixed random noise.

Given a threshold τ, the LET-guided gate is $g_i = \frac{1}{\tau}\mathrm{ReLU}(\tau - a_i)$, where $0 \le g_i \le 1$. When $a_i \ge \tau$, the gate is zero and the token remains unchanged. When $a_i < \tau$, the gate increases as the evidence score decreases. The purified token is

$$\widetilde{\mathbf{v}}_i = \mathbf{v}_i + g_i \mathbf{r}_i. \tag{10}$$

The resulting sequence $\widetilde{\mathbf{V}} = \{\widetilde{\mathbf{v}}_i\}_{i=1}^N$ keeps the original token layout. Thus the decoder still receives a full visual sequence, but low-reliability tokens become less stable as evidence for answer generation.

3.4 OPTIMIZATION AND INFERENCE

Training. LENS is first trained with supervised fine-tuning. The answer loss is computed after latent suppression, so the model learns to answer from $\widetilde{\mathbf{V}}$. The LET loss grounds the probe with token-level labels.

$$\mathcal{L}_{\mathrm{SFT}} = \mathcal{L}_{\mathrm{ans}} + \beta \mathcal{L}_{\mathrm{LET}}. \tag{11}$$

Here, β balances answer learning and LET supervision. After supervised fine-tuning, we refine the evidence policy with reinforcement fine-tuning. The supervised LET loss trains token scores independently, while latent suppression depends on the selected evidence set. We model the LET score vector as a Bernoulli policy during training

$$b_i \sim \mathrm{Bernoulli}(a_i), \quad \mathbf{b} = \{b_i\}_{i=1}^N. \tag{12}$$

The binary action $\mathbf{b}$ indicates selected evidence tokens. Given the label $\mathbf{z}$, the set-level reward is the F1 score

$$R(\mathbf{b}, \mathbf{z}) = \frac{2 \cdot \sum_{i=1}^N b_i z_i}{\sum_{i=1}^N b_i + \sum_{i=1}^N z_i + \varepsilon}. \tag{13}$$

Table 2: Generalization results across 9 base models from Qwen3-VL (Bai et al., 2025), InternVL3.5-VL (Wang et al., 2025b), and Qwen3.5 (Qwen Team, 2026). Green ↑ and red ↓ values denote absolute changes from the corresponding base model, showing how the same visual evidence purification strategy transfers across model families and scales.

Base Model	CUB	GQA	OpenImg	SROIE	VSR	MSVQA	COCO	Obj365	RUOD	Visdrone	Avg.
Qwen3-VL-2B	83.5	70.5	82.8	93.8	74.3	62.8	44.0	36.9	62.5	32.4	64.3
+ LENS-SFT	88.1 ↑4.6	78.4 ↑7.9	86.7 ↑3.8	94.7 ↑0.9	78.7 ↑4.5	65.0 ↑2.2	47.8 ↑3.8	39.5 ↑2.7	65.9 ↑3.4	35.0 ↑2.5	68.0 ↑3.6
+ LENS-GRPO	89.6 ↑6.2	78.7 ↑8.2	87.5 ↑4.6	94.3 ↑0.5	79.7 ↑5.5	68.0 ↑5.2	52.5 ↑8.5	42.0 ↑5.2	66.9 ↑4.4	44.3 ↑11.8	70.3 ↑6.0
Qwen3-VL-4B	86.5	72.7	82.1	92.5	76.6	64.0	46.5	40.0	63.6	39.6	66.4
+ LENS-SFT	91.5 ↑5.1	79.0 ↑6.4	86.9 ↑4.8	95.2 ↑2.7	80.7 ↑4.1	66.4 ↑2.4	52.0 ↑5.5	44.1 ↑4.1	70.1 ↑6.4	44.1 ↑4.5	71.0 ↑4.6
+ LENS-GRPO	91.2 ↑4.8	84.0 ↑11.3	88.1 ↑6.0	**95.9** ↑3.4	**82.4** ↑5.9	67.6 ↑3.6	**60.9** ↑14.4	51.3 ↑11.3	74.5 ↑10.9	58.4 ↑18.8	75.4 ↑9.0
Qwen3-VL-8B	89.5	72.2	85.2	94.7	76.2	65.1	49.9	41.2	68.2	42.3	68.4
+ LENS-SFT	**92.3** ↑2.8	79.6 ↑7.4	87.4 ↑2.2	95.7 ↑1.0	79.2 ↑3.0	70.2 ↑5.1	52.2 ↑2.3	43.3 ↑2.1	72.9 ↑4.7	45.7 ↑3.5	71.8 ↑3.4
+ LENS-GRPO	91.7 ↑2.2	**82.4** ↑10.2	**88.3** ↑3.0	95.5 ↑0.8	81.8 ↑5.6	**70.9** ↑5.9	60.0 ↑10.1	**52.7** ↑11.4	**75.6** ↑7.4	**60.2** ↑17.9	**75.9** ↑7.4
InternVL3.5-2B	86.9	73.9	84.1	92.1	76.5	66.0	33.0	22.7	38.5	17.2	59.1
+ LENS-SFT	90.5 ↑3.6	83.7 ↑9.8	86.8 ↑2.7	94.1 ↑2.0	78.3 ↑1.8	68.2 ↑2.2	49.7 ↑16.7	36.3 ↑13.6	56.5 ↑18.0	29.3 ↑12.1	67.3 ↑8.2
InternVL3.5-4B	89.7	74.7	84.8	93.3	77.0	66.7	47.6	37.9	53.6	28.5	65.4
+ LENS-SFT	90.9 ↑1.2	87.5 ↑12.8	86.3 ↑1.5	95.3 ↑2.0	**83.0** ↑6.0	70.5 ↑3.8	53.7 ↑6.1	**45.1** ↑7.2	64.7 ↑11.1	37.2 ↑8.7	71.4 ↑6.0
InternVL3.5-8B	91.3	77.6	86.2	93.8	80.7	68.9	48.4	39.5	55.3	29.9	67.1
+ LENS-SFT	**93.1** ↑1.8	**90.1** ↑12.5	**87.3** ↑1.1	**96.0** ↑2.2	82.5 ↑1.8	**71.5** ↑2.6	**54.0** ↑5.6	44.6 ↑5.1	**66.9** ↑11.6	**37.4** ↑7.5	**72.3** ↑6.9
Qwen3.5-2B	86.2	72.7	84.5	92.7	76.7	65.9	38.4	34.9	59.0	36.4	64.7
+ LENS-SFT	89.1 ↑2.9	77.6 ↑4.9	88.3 ↑3.8	93.7 ↑1.0	78.1 ↑1.4	69.0 ↑3.1	43.2 ↑4.8	38.8 ↑3.9	63.6 ↑4.6	37.3 ↑0.9	67.8 ↑3.1
Qwen3.5-4B	89.8	79.9	87.0	94.7	80.2	67.5	41.3	38.9	64.5	39.1	68.3
+ LENS-SFT	91.3 ↑1.6	89.2 ↑9.3	**90.4** ↑3.4	94.5 ↓0.2	81.4 ↑1.2	69.4 ↑1.8	48.2 ↑6.8	42.3 ↑3.4	65.9 ↑1.5	40.6 ↑1.6	71.3 ↑3.0
Qwen3.5-9B	91.3	79.6	84.5	**95.0**	81.7	69.7	44.5	42.2	66.4	42.3	69.7
+ LENS-SFT	**93.3** ↑2.0	**90.9** ↑11.3	88.3 ↑3.8	94.7 ↓0.3	**82.4** ↑0.7	**71.0** ↑1.3	**52.7** ↑8.2	**45.4** ↑3.2	**68.4** ↑2.0	**44.5** ↑2.2	**73.1** ↑3.4

The small constant ε avoids numerical instability. We maximize the expected reward while regularizing the policy toward the supervised checkpoint

$$\mathcal{J}_{\mathrm{RFT}}(\theta) = \mathbb{E}_{\mathbf{b}\sim\pi_\theta(\cdot|I,Q)}\left[R(\mathbf{b},\mathbf{z})\right] - \lambda D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid I,Q)\|\pi_{\mathrm{SFT}}(\cdot \mid I,Q)\right). \tag{14}$$

Inference. During inference, LENS first inserts the temporary $\langle mask\rangle$ and predicts LET scores $\mathbf{a}$. It then removes the control token, applies the threshold gate g_i and forms $\widetilde{\mathbf{v}}_i = \mathbf{v}_i + g_i\mathbf{r}_i$ for each visual token, yielding the purified sequence $\widetilde{\mathbf{V}}$ with LFT-guided latent suppression. The generation sequence is

$$\mathbf{S}_{\mathrm{gen}} = [\widetilde{\mathbf{v}}_1, \ldots, \widetilde{\mathbf{v}}_N, \mathbf{t}_1, \ldots, \mathbf{t}_M], \tag{15}$$

and the answer is sampled from

$$Y \sim P_\psi(\cdot \mid \mathbf{S}_{\mathrm{gen}}). \tag{16}$$

The only extra cost is the probing pass and the latent suppression operation.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Benchmarks. We evaluate LENS on 10 datasets covering VQA and grounding. For VQA, we use CUB (Wah et al., 2011), GQA (Hudson & Manning, 2019), OpenImages (Kuznetsova et al., 2020), SROIE (Huang et al., 2019), VSR (Liu et al., 2023), and MSVQA (Jiang et al., 2026). For grounding, we use COCO2017 (Lin et al., 2014), Objects365 (Shao et al., 2019), RUOD (Fu et al., 2023), and VisDrone (Zhu et al., 2021). We report VQA scores following the Visual CoT benchmark (Shao et al., 2024). For grounding, we report F1@0.5 following object detection metrics. More dataset details are provided in Appendix A.

Baselines. We compare LENS with Vanilla and SFT, as well as 7 representative reasoning and perception baselines. These include Visual-RFT (Liu et al., 2025b), VLM-R1 (Shen et al., 2025), PAPO (Wang et al., 2025c), VPT (Yu et al., 2025a), LVR, DMLR (Liu et al., 2025a), and VisMem (Yu et al., 2025b).



Figure 3: Visualization of LENS on VQA tasks. For each example, the left image shows the question-conditioned LET scores mapped to visual patches, and the right image shows the visual effect after applying token-level latent noise to low-relevance visual tokens. LENS preserves answer-supporting evidence such as attributes, objects, relations, and OCR fields, while suppressing irrelevant background or distracting regions. This illustrates its advantage in purifying visual evidence for fine-grained answer generation.

Implementation Details. All experiments except Table 2 use Qwen3-VL-4B and 8 NVIDIA H100 80G GPUs. We set β to 0.2 to balance the two loss terms and set τ to 0.5 corresponding to the binary cross-entropy loss. More implementation details are provided in Appendix B.

4.2 MAIN RESULTS

The main comparison in Table 1 shows that LENS improves MLLM performance by cleaning visual evidence rather than adding more reasoning context. Compared with SFT, LENS-SFT raises the overall average from 66.40 to 70.99, and LENS-GRPO further increases it to 75.44. Against the strongest non-LENS baseline, LENS-GRPO improves the overall average by 6.65 points. It improves the VQA average from 79.04 to 84.87 and the grounding average from 47.44 to 61.29 over SFT, showing that evidence purification benefits both answer prediction and spatial localization.

The VQA results indicate that suppressing redundant visual tokens improves fine-grained answer generation while preserving broad multimodal reasoning ability. LENS-SFT improves all six VQA datasets over SFT, with gains from 2.43 points on MSVQA to 6.36 points on GQA. The reinforcement refinement further raises the VQA average to 84.87 and obtains the best results on five of the six VQA datasets. The largest improvement appears on GQA, where LENS-GRPO reaches 83.95 compared with 72.65 for SFT. This pattern suggests that the learned LET scores help the model focus on evidence needed for compositional and fine-grained reasoning.

The grounding results provide stronger evidence that visual selectivity is the main source of improvement. Grounding requires the model to locate question-relevant regions, so distractor suppression should directly improve this metric. LENS-SFT increases the grounding average from 47.44 to 52.56 over SFT, while LENS-GRPO raises it to 61.29. The gains of LENS-GRPO are especially large on VisDrone, COCO, Object365, and RUOD, reaching 18.81, 14.43, 11.30, and 10.88 points over SFT. Together with the component ablations in Table 3, these improvements support the claim that LET-guided latent suppression reduces the influence of irrelevant visual tokens.

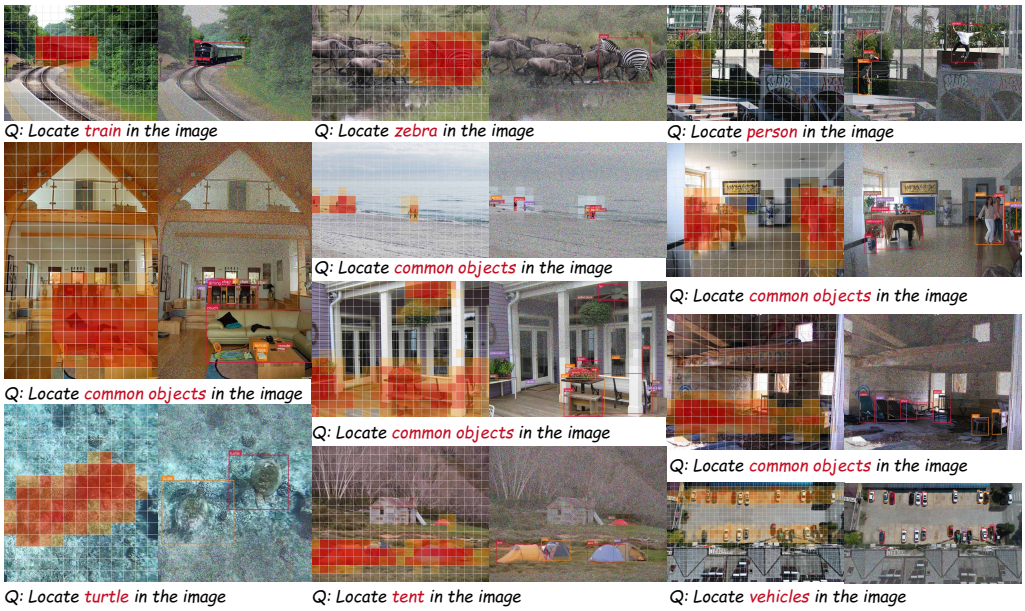


Figure 4: Visualization of LENS on grounding tasks. For each example, the left image shows the LET scores over visual patches, and the right image shows the visual effect after applying token-level latent noise to low-relevance visual tokens. LENS keeps target objects and multiple queried instances visually stable, while perturbing surrounding clutter and non-target distractors. This illustrates its advantage in improving question-conditioned spatial localization.

Table 2 shows that LENS is a transferable visual evidence purification strategy across model families and scales. With supervised training alone, LENS-SFT improves the overall average on all nine base models by 3.0–8.2 points. The gains hold for Qwen3-VL, InternVL3.5-VL, and Qwen3.5 models, indicating that the intervention is not tied to one backbone. On Qwen3-VL models, reinforcement refinement further enlarges the overall gains to 6.0, 9.0, and 7.4 points for 2B, 4B, and 8B models. Although Qwen3.5 shows two small drops on SROIE, the overall averages remain consistently higher, suggesting that LENS improves visual selectivity with limited task-specific tradeoffs.

4.3 ABLATION STUDY AND VISUALIZATIONS

4.3.1 ABLATION STUDY

The ablation results in Table 3 show that LENS works because evidence prediction and latent suppression act together. Evidence probing alone improves the average score from 66.40 to 68.04, which indicates that the temporary $\langle mask \rangle$ token learns a useful question-conditioned LET scores. However, this gain is limited because the LFT only identifies relevant regions and does not yet reduce the influence of redundant tokens during decoding. This supports the main claim that visual selectivity must be connected to an explicit latent intervention.

Different masking types confirm that visual suppression must be guided by the learned LET scores and soft noise. EP+RM drops the average score to 60.95, and the degradation is especially clear on tasks that require precise evidence, e.g. SROIE and GQA. It shows that simply perturbing visual tokens can destroy answer-supporting cues. In addition, EP+ZM/AM replaces the low-score token with zero padding or average visual token, which also drop the average score to 58.66/59.67, respectively. In contrast, EP+LS reaches 70.99 and improves all benchmarks over the baseline, showing that latent suppression with soft noise is effective when it preserves high-score tokens and weakens low-score distractors.

LET-guided visual suppression significantly outperforms applying noise based on attention scores. *Attention mask* drops the average score to 65.33. We suspect that the attention of a certain layer does not always focus on direct visual evidence, so applying visual suppression based on the attention scores of that layer is inadvisable. Moreover, this approach requires obtaining the

Table 3: Ablation on Qwen3-VL-4B evaluating the role of evidence probing, random masking, latent suppression, and reinforcement refinement in visual evidence purification. EP denotes *Evidence probing*, RM denotes *Random mask*, Attention mask directly uses the attention scores of the last layer of the decorder to add noise to the visual token sequence, ZM replaces low-score tokens with zero padding, AM replaces them with average visual tokens, and LS denotes *Latent suppression*.

Methods	CUB	GQA	OpenImg	SROIE	VSR	MSVQA	COCO	Obj365	RUOD	Visdrone	Avg.
Baselines	86.45	72.65	82.08	92.49	76.58	63.96	46.50	40.02	63.63	39.60	66.40
EP	87.62	74.89	83.65	93.15	77.67	64.36	48.98	41.85	66.52	41.67	68.04
EP+RM	85.60	68.34	84.13	69.90	70.67	60.48	41.47	34.52	60.73	33.70	60.95
EP+ZM	87.23	63.99	82.38	56.43	69.55	61.41	43.02	36.84	56.28	29.44	58.66
EP+AM	86.82	65.47	83.23	59.88	68.19	62.68	44.91	36.62	58.97	29.92	59.67
Attention mask	87.63	69.31	83.97	92.42	72.28	64.90	46.48	39.07	58.96	38.31	65.33
EP+LS	**91.50**	**79.01**	**86.89**	**95.20**	**80.69**	**66.39**	**52.00**	**44.10**	**70.05**	**44.09**	**70.99**
SFT+GRPO	90.85	76.69	85.03	94.62	78.71	67.25	52.75	46.96	65.21	48.36	70.64
EP+LS+GRPO	91.72	81.65	87.20	95.13	79.35	67.32	56.34	48.46	71.96	56.58	73.57
LENS-GRPO	**91.24**	**83.95**	**88.10**	**95.92**	**82.43**	**67.55**	**60.93**	**51.32**	**74.51**	**58.41**	**75.44**

complete attention weights, making it impossible to use acceleration algorithms (such as flash attention), which significantly reduces training and inference efficiency. This indicates that the proposed method has a clear advantage over applying visual suppression based on model attention.

The reinforcement results further show that set-level evidence quality is crucial for the final visual context. SFT+GRPO raises the average score from 66.40 to 70.64, showing that policy optimization improves the base model. When the same optimization is applied after LENS-SFT, the average rises to 73.57, which shows that a cleaner LET scoring signal provides a stronger starting point. LENS-GRPO obtains the best average score of 75.44, indicating that optimizing the evidence policy with a set-level reward better matches the goal of selecting complete question-relevant evidence. **These ablations also clarify why LENS improves both VQA and grounding.** The LET scores alone brings moderate gains because it exposes question-relevant regions, while latent suppression turns these scores into a cleaner decoding context. The large drop of EP+RM shows that suppression is useful only when it follows evidence structure. The stronger result of LENS-GRPO further suggests that complete evidence sets matter more than isolated high-score patches.

4.3.2 VISUALIZATIONS

Fig. 3 and Fig. 4 visualize the effect of LENS on VQA and grounding tasks. In each pair, the left image maps the predicted LET scores to visual patches, while the right image visualizes the image-space effect of the latent-space intervention on low-relevance visual tokens. For VQA tasks (Fig. 3), LENS highlights answer-supporting cues such as attributes, objects, relations, and OCR fields, while the latent-noise effect mainly appears on backgrounds or unrelated regions. This shows its advantage in purifying fine-grained visual evidence for answer generation. For grounding tasks (Fig. 4), LENS preserves target objects and multiple queried instances, while perturbing surrounding clutter and non-target distractors. This shows its advantage in improving question-conditioned spatial localization. Together, the visualizations show that the LET scores align with question-relevant regions and that latent suppression concentrates the decoding context around those regions.

5 CONCLUSION

This paper presents LENS, a visual evidence purification framework that improves multimodal reasoning by suppressing question-irrelevant visual tokens in latent space. Instead of adding longer textual traces, extra visual inputs, or persistent latent memories, LENS introduces a question-conditioned *Lens Evidence Token* supervised by object-level annotations and uses adaptive latent noise to weaken low-relevance tokens while preserving the original backbone and token sequence. Experiments across VQA and grounding benchmarks show consistent gains over strong training, token-level, and latent reasoning baselines. Ablations and visualizations further confirm that the improvement comes from coupling evidence probing with LET-guided latent suppression. These results suggest that cleaner visual evidence is a direct and effective path toward more reliable fine-grained vision-language reasoning.

AI USE STATEMENT

In this work, generative AI tools were used only for English-language polishing. They were not used for research ideation or hypothesis formulation, theoretical or mathematical development, methodology or experiment design, method implementation, translation, data generation or processing, qualitative analysis, or result interpretation.

REPRODUCIBILITY STATEMENT

Reproducibility details are provided in the appendix: Appendix A documents datasets, task formats, splits, and metrics; Appendix B specifies implementation and inference settings; Appendix C describes evidence-supervision construction; Appendix D gives the reinforcement fine-tuning setup and rewards; Appendix E records baseline implementations and the reproduction protocol; and Appendix G reports computational costs. Together with the method definitions and optimization objectives in the main text, these details specify the training, inference, and evaluation procedures used in this work.

REFERENCES

Hongjun An, Wenhan Hu, Sida Huang, Siqi Huang, Ruanjun Li, Yuanzhi Liang, Jiawei Shao, Yiliang Song, Zihan Wang, Cheng Yuan, et al. Ai flow: Perspectives, scenarios, and approaches. *Vicinagearth*, 3(1):1, 2026.

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

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3836–3845, 2025.

Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. *arXiv preprint arXiv:2506.05331*, 2025.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images. *arXiv preprint arXiv:2505.15879*, 2025.

Chenping Fu, Risheng Liu, Xin Fan, Puyang Chen, Hao Fu, Wanqi Yuan, Ming Zhu, and Zhongxuan Luo. Rethinking general underwater object detection: Datasets, challenges, and solutions. *Neurocomputing*, 517:243–256, 2023. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2022.10.039.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*, 2025.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19520–19529, 2025.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1516–1520. IEEE, 2019.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Kai Jiang, Siqi Huang, Xiangyu Chen, Jiawei Shao, Hongyuan Zhang, Ping Luo, and Xuelong Li. Multimodal continual learning with mllms from multi-scenario perspectives, 2026. URL https://arxiv.org/abs/2511.18507.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 2886–2903, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025a.

Hengli Li, Chenxi Li, Tong Wu, Xuekai Zhu, Yuxuan Wang, Zhaoxin Yu, Eric Hanchen Jiang, Song-Chun Zhu, Zixia Jia, Ying Nian Wu, et al. Seek in the dark: Reasoning via test-time instance-level policy gradient in latent space. *arXiv preprint arXiv:2505.13308*, 2025b.

Xuelong Li. Positive-incentive noise. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):8708–8714, 2024. doi: 10.1109/TNNLS.2022.3224577.

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 *European conference on computer vision*, pp. 740–755. Springer, 2014.

Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. Reasoning within the mind: Dynamic multimodal interleaving in latent space. *arXiv preprint arXiv:2512.12623*, 2025a.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2034–2044, 2025b.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14420–14431, 2024.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 18798–18806, 2024.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

Jiawei Shao and Xuelong Li. Ai flow at the network edge. *IEEE Network*, 40(1):330–336, 2025.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8430–8439, 2019.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *Technical Report CNS-TR-2011-001*, 2011.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 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.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*, 2025c.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025.

Runpeng Yu, Xinyin Ma, and Xinchao Wang. Introducing visual perception token into multimodal large language model. *arXiv preprint arXiv:2502.17425*, 2025a.

Xinlei Yu, Chengming Xu, Guibin Zhang, Zhangquan Chen, Yudong Zhang, Yongbo He, Peng-Tao Jiang, Jiangning Zhang, Xiaobin Hu, and Shuicheng Yan. Vismem: Latent vision memory unlocks potential of vision-language models. *arXiv preprint arXiv:2511.11007*, 2025b.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Ming Li, Qilong Wu, Kaipeng Zhang, and Chen Wei. Pyvision: Agentic vision with dynamic tooling. *arXiv preprint arXiv:2507.07998*, 2025.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2021.

APPENDIX

A DATASET AND EVALUATION DETAILS

The evaluation suite is designed to test whether LENS can select fine-grained visual evidence across answer generation and spatial localization. Table A.1 summarizes the datasets, task forms, split sizes, input formats, and evaluation metrics. The six VQA datasets cover attribute recognition, compositional reasoning, OCR, spatial relation understanding, and multi-choice activity understanding. The four grounding datasets cover common objects, long-tail objects, underwater objects, and dense aerial scenes. This combination is important because visual evidence purification should improve both the answer produced from selected evidence and the localization of that evidence.

For VQA evaluation, each prediction is compared with the reference answer by Qwen3.5-Flash. The judge receives the question, the reference answer, and the model prediction, then returns a scalar score in $[0, 1]$. We average the sample scores and report the normalized percentage score. This protocol gives partial credit when a prediction is semantically close to the reference, which is useful for open-ended VQA and OCR-style answers.

For grounding evaluation, we first extract all predicted boxes from the model output. Each predicted category is mapped to the closest category in the dataset vocabulary by Qwen3.5-Flash. Predictions without a valid category match are removed. We then match predicted boxes and ground-truth boxes of the same mapped category at IoU threshold 0.5 and compute the dataset-level F1 score. This evaluation penalizes both missed objects and unsupported detections, so it directly reflects whether the model localizes the complete evidence set.

The two benchmark groups provide complementary evidence for the central claim. VQA measures whether selected visual evidence supports answer generation, while grounding measures whether the selected evidence is spatially correct. Consistent gains on both groups indicate that LENS improves question-conditioned visual selectivity across answer-generation and spatial-localization formats.

B IMPLEMENTATION DETAILS

The implementation keeps the MLLM backbone and the visual token sequence interface unchanged. Unless otherwise stated, experiments use Qwen3-VL-4B as the base model and train on 8 NVIDIA H100 80G GPUs. Input images are resized under a maximum pixel budget of 1000000 and a minimum pixel budget of 3136. Spatial sizes are aligned to the visual patch stride of 32. Thus the number of visual tokens is bounded by the resized patch grid and remains below about 1000 tokens, while very small images still keep at least three visual tokens.

The temporary $\langle mask \rangle$ token is appended to the end of the text input in the probing pass, after the image tokens and question tokens have been formed as in Eq. 2. The hidden state at this temporary position is used only to predict the LET scores. The token is removed before answer decoding, so it does not become a persistent reasoning token and does not alter the generation interface.

The probe head is a lightweight one-layer MLP implemented as a linear projection followed by a sigmoid activation. Given the hidden state $\mathbf{h}_m$ at the temporary token, the head produces N logits and converts them into the visual evidence prior $\mathbf{a} \in [0, 1]^N$. Here N is the number of visual tokens for the current image. The head therefore adds only a small number of parameters and its output is directly aligned with the visual token grid.

The noise generator is a two-branch MLP that predicts feature-aware perturbation parameters for each visual token. Each branch contains a linear layer, a ReLU activation, and a linear output layer. One branch predicts $\boldsymbol{\mu}_i$ and the other predicts the raw scale parameter for $\boldsymbol{\sigma}_i$. The final latent perturbation is sampled by reparameterization,

$$\mathbf{r}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{B.1}$$

Latent suppression is applied after the visual encoder has produced visual token embeddings and before these embeddings are merged with the text tokens for decoding. This position lets LENS suppress visual distractors without changing image preprocessing, tokenizer behavior, language decoding, or the output format.

Table A.1: Dataset and evaluation details. Train and validation sizes are counted from our annotation files. Judge score denotes VQA scoring by Qwen3.5-Flash, and F1@0.5 denotes class-normalized grounding F1.

Dataset	Task	Train	Val	Input	Metric
VQA					
CUB	Attribute	10,056	492	Image question	Judge score
GQA	Compositional	98,149	978	Image question	Judge score
OpenImages	Open domain	43,053	945	Image question	Judge score
SROIE	OCR	2,486	686	Document query	Judge score
VSR	Spatial	3,376	404	Image statement	Judge score
MSVQA	Multi choice	13,603	500	Image question	Judge score
Grounding					
COCO2017	Common objects	118,287	1,000	Localization prompt	F1@0.5
Objects365	Long tail objects	200,000	1,000	Localization prompt	F1@0.5
RUOD	Underwater objects	13,744	249	Localization prompt	F1@0.5
VisDrone	Aerial objects	6,004	249	Localization prompt	F1@0.5

We train for 1 epoch with batch size 128, learning rate 4×10^{-5}, AdamW optimizer, warmup ratio 0.03, weight decay 0, and bf16 mixed precision. The loss weight for evidence supervision is $\beta = 0.2$, and the deterministic inference threshold is $\tau = 0.5$. The same preprocessing and decoding settings are used across baselines unless a baseline requires its own official setting.

The only additional forward computation is the short probing pass used to estimate $\mathbf{a}$. After that, the model performs standard autoregressive decoding with the same token sequence length, using purified visual embeddings $\widetilde{\mathbf{V}}$ instead of the original visual embeddings $\mathbf{V}$.

C Evidence Supervision Construction

LENS uses object-level boxes to supervise the question-conditioned *Lens Evidence Token* without requiring chain-of-thought rationales or pixel-level masks. Each training sample provides an image, a question or localization prompt, an answer, and a set of question-relevant boxes. We first resize the image with the same preprocessing used by the MLLM. If the original image has width W and height H, and the resized image has width $\widetilde{W}$ and height $\widetilde{H}$, every box $B = (x_1, y_1, x_2, y_2)$ is mapped to

$$\widetilde{B} = \left(\left\lfloor x_1 \frac{\widetilde{W}}{W} \right\rfloor, \left\lfloor y_1 \frac{\widetilde{H}}{H} \right\rfloor, \left\lfloor x_2 \frac{\widetilde{W}}{W} \right\rfloor, \left\lfloor y_2 \frac{\widetilde{H}}{H} \right\rfloor \right). \tag{C.1}$$

The resized image is divided into a 32 by 32 patch grid. Let patch Ω_i be the spatial area of visual token $\mathbf{v}_i$. The binary evidence label is

$$z_i = \mathbb{I}\left[\max_{\widetilde{B} \in \mathcal{B}_Q} \operatorname{area}(\Omega_i \cap \widetilde{B}) > 0 \right]. \tag{C.2}$$

The labels are flattened from left to right and top to bottom, which gives a token-level target $\mathbf{z} \in \{0, 1\}^N$ aligned with the visual token order.

Multiple question-relevant objects are handled by taking the union over boxes. If boxes overlap, a patch is still labeled once, so overlapping annotations do not overweight a region. Small objects are retained as long as they intersect at least one patch. This rule is important for CUB attributes, OCR fields, VisDrone objects, and RUOD objects, where the evidence can occupy a small part of the image.

For OCR-oriented samples such as SROIE, the evidence boxes correspond to text regions needed to answer the question. Patches intersecting those text boxes are labeled as positive, while unrelated document regions remain negative. For common object grounding datasets such as COCO2017, Objects365, RUOD, and VisDrone, the prompt asks the model to locate common or dataset-defined objects, so all annotated target boxes in the sample are treated as the relevant evidence set.

We remove invalid boxes with non-positive width or height after resizing. If a sample has no valid box or produces no positive patch after preprocessing, its LET loss is masked out for that sample. The answer loss can still be used when the answer annotation is valid. This filtering avoids training the probe with empty or contradictory evidence labels.

This construction turns inexpensive box supervision into token-level LET supervision. Box-derived supervision provides the distinction required by latent suppression between likely evidence tokens and broad distractors. The supervision also remains question-conditioned because the positive boxes are selected according to the current question or localization prompt rather than image saliency alone.

D Reinforcement Fine-Tuning Details

Reinforcement fine-tuning refines the LET scores as a set-level evidence-selection policy. After supervised training, each LET score a_i is interpreted as the Bernoulli parameter for token selection,

$$\pi_\theta(\mathbf{b} \mid I, Q) = \prod_{i=1}^{N} a_i^{b_i}(1-a_i)^{1-b_i}, \quad b_i \in \{0,1\}. \tag{D.1}$$

During training, $\mathbf{b}$ is sampled stochastically. This sampling exposes the model to boundary tokens whose LET scores are uncertain, which acts as data augmentation for evidence selection. During inference, we do not sample. We instead use the deterministic threshold $\tau = 0.5$ and preserve tokens with $a_i \geq \tau$.

We optimize the evidence policy with GRPO. For each image-question pair, we sample $K = 4$ evidence masks from the current Bernoulli policy. The evidence reward follows the set F1 score between the sampled mask and the target label,

$$R_{\text{ev}}(\mathbf{b}^{(k)}, \mathbf{z}) = \frac{2\sum_{i=1}^{N} b_i^{(k)} z_i}{\sum_{i=1}^{N} b_i^{(k)} + \sum_{i=1}^{N} z_i + \varepsilon}. \tag{D.2}$$

The group baseline is the mean reward of the K samples, and the normalized advantage is

$$\widehat{A}^{(k)} = \frac{R_{\text{ev}}(\mathbf{b}^{(k)}, \mathbf{z}) - \frac{1}{K}\sum_{\ell=1}^{K} R_{\text{ev}}(\mathbf{b}^{(\ell)}, \mathbf{z})}{\text{std}_\ell\left(R_{\text{ev}}(\mathbf{b}^{(\ell)}, \mathbf{z})\right) + \varepsilon}. \tag{D.3}$$

The policy objective is

$$\mathcal{L}_{\text{GRPO}} = -\frac{1}{K}\sum_{k=1}^{K} \widehat{A}^{(k)} \log \pi_\theta(\mathbf{b}^{(k)} \mid I, Q) + \lambda D_{\text{KL}}\left(\pi_\theta(\cdot \mid I, Q) \| \pi_{\text{SFT}}(\cdot \mid I, Q)\right), \tag{D.4}$$

where $\lambda = 0.01$. The KL term keeps the policy close to the supervised LET policy and prevents reward optimization from collapsing to overly sparse or overly dense masks.

For answer-level reinforcement baselines, we use the same task reward implementation for all RL methods. The reward dispatches by dataset type. General VQA uses normalized text or choice accuracy, SROIE uses OCR text similarity, MSVQA uses yes-no matching, count accuracy, list F1, or choice accuracy, and grounding uses box F1 at IoU 0.5 after category normalization. All answer-level RL baselines share this task reward; LENS additionally uses the evidence reward above to refine its token-selection policy.

The sampled mask is used only during training. Given $\mathbf{b}^{(k)}$, selected tokens are preserved and unselected tokens receive stronger latent suppression. At inference, the deterministic mask from τ is used to construct $\widetilde{\mathbf{V}}$. This difference avoids test-time randomness while preserving the training benefit of stochastic exploration around uncertain evidence boundaries.

The RFT stage therefore improves the completeness of the selected evidence set. Supervised learning labels each token independently, whereas GRPO rewards the whole selected set. This is better aligned with VQA and grounding, where the model often needs several supporting patches or multiple object instances to answer correctly.

E Full Baseline and Reproduction Protocol

All baselines are reproduced under the same backbone, data, and evaluation protocol. For each trainable method, we replace the original backbone with Qwen3-VL-4B and use the same training split described in Appendix A. We keep the default hyperparameters in the released code unless the backbone replacement requires a dimension or tokenizer adjustment. This protocol isolates the method design from changes in data scale, model capacity, and evaluation scripts.

Vanilla **Vanilla measures the zero-training capability of the Qwen3-VL-4B backbone.** We directly run inference with the shared preprocessing, prompting, decoding, and evaluation scripts. This baseline shows how much visual understanding and grounding ability is already present before any task-specific adaptation.

SFT **SFT measures the effect of full-parameter supervised adaptation under the same data budget as LENS.** We fine-tune Qwen3-VL-4B on the same training data and use the implementation settings in Appendix B. This baseline separates the gain from ordinary supervised training from the gain brought by question-conditioned evidence probing and latent suppression.

Visual-RFT **Visual-RFT is reproduced as a reinforcement fine-tuning baseline with visual verifiable rewards.** We use the official Visual-RFT implementation (Liu et al., 2025b), replace the base MLLM with Qwen3-VL-4B, and keep its default GRPO hyperparameters. The method samples multiple responses with reasoning tokens and final answers, then updates the policy with task rewards such as visual classification correctness or localization quality.

VLM-R1 **VLM-R1 evaluates an R1-style reinforcement learning pipeline for vision-language tasks.** We use the official VLM-R1 implementation (Shen et al., 2025) with Qwen3-VL-4B and the same training data. Its policy update follows the released GRPO setting and optimizes visual understanding or grounding outputs through verifiable rewards.

PAPO **PAPO tests whether perception-aware policy optimization improves multimodal reasoning under the same data setting.** We use the official PAPO implementation (Wang et al., 2025c), replace the backbone with Qwen3-VL-4B, and keep the default PAPO hyperparameters. The method adds an implicit perception loss to RLVR by comparing rollout likelihoods under clean and corrupted visual inputs, encouraging the policy to rely on informative visual evidence while reasoning.

VPT **VPT is reproduced without external visual encoders to keep the comparison fair.** We use the official VPT implementation (Yu et al., 2025a) with Qwen3-VL-4B and the same training data. VPT introduces visual perception tokens that can trigger extra region perception or visual re-encoding. In our reproduction, we only use the CLIP style mode with the model native visual encoder and do not add DINO, SAM, or other external encoders, because those encoders would provide additional visual knowledge unavailable to the other baselines.

LVR **LVR is kept as a two-stage latent visual reasoning baseline.** We use the official LVR implementation, replace the base MLLM with Qwen3-VL-4B, and preserve both stages in the original recipe. The first stage performs supervised fine-tuning that jointly learns latent visual reasoning and text generation, while the second stage applies reinforcement learning to refine the latent reasoning process with response-level rewards.

DMLR **DMLR is evaluated as a training-free test-time latent reasoning method.** Since DMLR (Liu et al., 2025a) does not require additional training, we apply the official inference procedure on the SFT checkpoint. It refines latent think tokens at test time through confidence-guided latent updates and dynamically injects selected visual features into the latent reasoning process.

VisMem **VisMem is reproduced with its two-stage latent memory training pipeline.** We use the official VisMem implementation (Yu et al., 2025b), replace the base MLLM with Qwen3-VL-4B, and keep the original memory learning procedure. The method first learns short-term and long-

Table F.1: Additional ablations on the evidence threshold τ and the LET loss weight β. VQA Avg. averages the six VQA datasets, Grounding Avg. averages the four grounding datasets, and Avg. averages all ten datasets.

Setting	CUB	GQA	OpenImg	SROIE	VSR	MSVQA	**VQA Avg.**	COCO	Obj365	RUOD	Visdrone	**Grounding Avg.**	**Avg.**
Threshold τ													
$\tau = 0.3$	88.87	78.45	85.84	93.16	77.23	68.10	81.94	53.20	43.10	69.57	46.29	53.04	70.38
$\tau = 0.5$	91.50	79.01	86.89	95.20	80.69	66.39	**83.28**	52.00	44.10	70.05	44.09	52.56	70.99
$\tau = 0.7$	90.69	78.75	87.05	92.73	78.47	70.54	83.04	51.79	44.49	71.34	45.02	**53.16**	**71.09**
LET loss weight β													
$\beta = 0.1$	89.75	78.62	85.86	94.05	78.65	65.89	82.14	50.77	43.65	69.68	44.27	52.09	70.12
$\beta = 0.2$	91.50	79.01	86.89	95.20	80.69	66.39	**83.28**	52.00	44.10	70.05	44.09	52.56	**70.99**
$\beta = 0.4$	90.59	77.84	84.95	92.15	77.45	67.18	81.69	53.88	44.87	71.36	44.92	**53.76**	70.52

term latent vision memories, then learns how to invoke these memories during inference to support perceptual retention and semantic consolidation.

All trainable baselines use the same data and base model size, while method-specific training stages are retained when they are part of the official recipe. As a result, the comparison mainly reflects whether a method adds textual reasoning, visual prompts, latent reasoning, latent memories, test-time latent updates, or question-conditioned visual evidence suppression. LENS keeps the backbone interface unchanged and learns to predict and suppress question-irrelevant visual tokens.

F Additional Ablation Studies

We evaluate sensitivity to the inference threshold and LET loss weight. We vary the inference threshold τ in $\{0.3, 0.5, 0.7\}$ and the LET loss weight β in $\{0.1, 0.2, 0.4\}$. The backbone, training data, decoding setting, and evaluation scripts are kept unchanged. Table F.1 reports all benchmark scores together with VQA, grounding, and overall averages.

The threshold τ mainly controls the balance between preserving weak evidence and removing distractors. At inference, tokens with $a_i \geq \tau$ are preserved and tokens with lower LET scores receive latent suppression. The preserved token ratio and the mean suppression strength are

$$\rho_\tau = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[a_i \geq \tau\right], \tag{F.1}$$

$$\bar{g}_\tau = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\tau} \mathrm{ReLU}\left(\tau - a_i\right). \tag{F.2}$$

When $\tau = 0.3$, the intervention is conservative and keeps more marginal patches active. This protects possible evidence, but it also leaves distractors in the context, giving an overall average of 70.38. Increasing τ to 0.5 raises the VQA average to 83.28 and gives the best results on CUB, GQA, SROIE, and VSR. This shows that moderate suppression helps fine-grained answer generation. With $\tau = 0.7$, the grounding average reaches 53.16 and the overall average reaches 71.09, but SROIE and VSR drop compared with $\tau = 0.5$. This pattern indicates that aggressive suppression can help localization while perturbing small OCR or spatial cues.

The loss weight β shows a similar tradeoff between LET supervision and answer flexibility. The supervised objective is

$$\mathcal{L}_{\mathrm{SFT}}(\beta) = \mathcal{L}_{\mathrm{ans}} + \beta \mathcal{L}_{\mathrm{LET}}, \tag{F.3}$$

$$\nabla \mathcal{L}_{\mathrm{SFT}}(\beta) = \nabla \mathcal{L}_{\mathrm{ans}} + \beta \nabla \mathcal{L}_{\mathrm{LET}}. \tag{F.4}$$

When $\beta = 0.1$, the LET score vector is weakly supervised and the overall average is 70.12. The default $\beta = 0.2$ obtains the best overall average of 70.99 and the best VQA average of 83.28, which suggests that the LET scores are strong enough to guide suppression without dominating answer learning. Increasing β to 0.4 improves the grounding average to 53.76, including the best scores on COCO, Object365, RUOD, and VisDrone. However, the VQA average decreases to 81.69. This means stronger box driven supervision improves spatial selectivity, but can reduce flexibility for open ended answer generation.

Table G.1: Inference cost comparison on the VQA test split. TTFT denotes Time To First Token, TPOT denotes Time Per Output Token, and Avg. Length denotes the generated answer length. Timing is measured on one H100 GPU with five warm-up rounds and ten measured rounds.

Method	TTFT ms			TPOT ms			Avg. Length
	bs=1	bs=8	bs=16	bs=1	bs=8	bs=16	
GT	–	–	–	–	–	–	14.46
SFT	71.484±6.722	41.730±0.688	40.840±0.583	35.010±3.057	6.131±0.398	3.574±0.217	14.29
Visual-RFT	188.409±7.687	188.077±2.807	198.696±2.680	35.206±2.985	7.018±0.617	4.537±0.424	20.89
VLM-R1	72.165±6.655	41.863±0.501	39.271±0.620	34.015±2.039	6.258±0.895	4.755±0.260	17.08
LVR	88.497±4.099	74.970±0.749	77.040±0.367	31.937±1.785	6.512±0.269	4.089±0.144	28.44
PAPO	73.218±7.097	41.087±0.735	39.268±0.599	35.417±3.625	6.894±0.735	4.659±0.975	14.91
VPT	172.081±22.503	128.372±7.544	119.615±3.237	105.011±6.952	29.409±1.635	17.312±0.791	34.88
VisMem	151.129±24.425	97.919±7.532	89.136±5.133	64.379±5.983	38.455±3.589	23.512±2.333	22.41
LENS	122.163±11.163	66.886±0.995	65.412±0.832	34.156±4.239	6.161±0.539	3.494±0.268	15.98

These results support the default setting used in the main experiments. The best overall threshold is only 0.10 points above $\tau = 0.5$, while $\tau = 0.5$ gives the strongest VQA average and a balanced grounding score. For β, the default value gives the best overall score and avoids the VQA degradation observed at $\beta = 0.4$. Across the tested settings, the overall average varies by 0.71 points over τ and by 0.87 points over β. These results reflect the intended mechanism of LENS, question-conditioned LET scores preserve answer-supporting patches and latent suppression weakens visually redundant regions.

G COMPUTATIONAL COST ANALYSIS

LENS adds a localized probing cost while keeping the decoding interface unchanged. Let N be the number of visual tokens, M be the number of text tokens, L be the output length, and d be the hidden dimension. A standard MLLM mainly pays the autoregressive decoding cost $C_{\mathrm{decode}}(N, M, L)$. LENS adds one evidence probing pass and one token-wise latent suppression operation,

$$C_{\mathrm{LENS}} \approx C_{\mathrm{probe}}(N, M) + C_{\mathrm{sup}}(N, d) + C_{\mathrm{decode}}(N, M, L). \tag{G.1}$$

The purified visual sequence keeps the same length as the original visual sequence,

$$\left|\widetilde{\mathbf{V}}\right| = |\mathbf{V}| = N. \tag{G.2}$$

Thus, the decoder-visible visual sequence remains at length N, and the added computation consists of one evidence-probing pass and one token-wise latent-suppression operation.

The timing data show that the extra cost of LENS is paid before decoding rather than during each generated token. Using SFT as the closest backbone matched baseline, the TTFT overhead is

$$\Delta_{\mathrm{TTFT}}(b) = \mathrm{TTFT}_{\mathrm{LENS}}(b) - \mathrm{TTFT}_{\mathrm{SFT}}(b), \tag{G.3}$$

which gives 50.68 ms, 25.16 ms, and 24.57 ms for bs=1, bs=8, and bs=16. In contrast, the per token change is

$$\Delta_{\mathrm{TPOT}}(b) = \mathrm{TPOT}_{\mathrm{LENS}}(b) - \mathrm{TPOT}_{\mathrm{SFT}}(b), \tag{G.4}$$

which gives -0.85 ms, 0.03 ms, and -0.08 ms for the same batch sizes. The near zero TPOT change confirms that LENS does not make autoregressive decoding slower after the visual evidence has been purified.

The comparison with reasoning and memory baselines further supports the efficiency claim. Visual-RFT has much larger TTFT and a longer average output length of 20.89. LVR and VisMem also increase answer length because they rely on latent reasoning or memory usage. VPT has the largest TPOT cost, reaching 105.011 ms at bs=1, because additional visual perception changes the decoding workload. By contrast, LENS combines its accuracy gains with an average response length of 15.98 tokens, close to SFT (14.29) and GT (14.46).

The cost profile matches the design goal of visual evidence purification. The probing pass estimates $\mathbf{a}$ once, and the suppression operation applies

$$\widetilde{\mathbf{v}}_i = \mathbf{v}_i + \frac{1}{\tau}\mathrm{ReLU}\left(\tau - a_i\right)\mathbf{r}_i, \tag{G.5}$$

before standard decoding. Since this operation is linear in the number of visual tokens and creates no new token sequence, the remaining overhead is predictable.

H Additional Qualitative Visualizations

The additional visualizations extend the main qualitative evidence to harder perception settings. The main text already shows that LENS can focus on answer-supporting regions in VQA and target objects in grounding. This section adds three groups of cases that are more likely to expose visual redundancy. The purpose is to show that the same question-conditioned LET scores remain useful when the evidence is small, visually ambiguous, or text based.

- **Small or distant targets test whether fine local evidence can be preserved.** These cases examine objects that occupy only a few visual patches, where background tokens can easily dilute the useful signal.
- **Occluded, hidden, and dense scenes test whether distractors can be suppressed.** These cases examine cluttered images where similar textures, nearby objects, and partial visibility make visual evidence ambiguous.
- **OCR fields test whether the LET scores are truly question-conditioned.** These cases examine invoices where the correct region changes with the requested field, e.g., address, date, or company name.

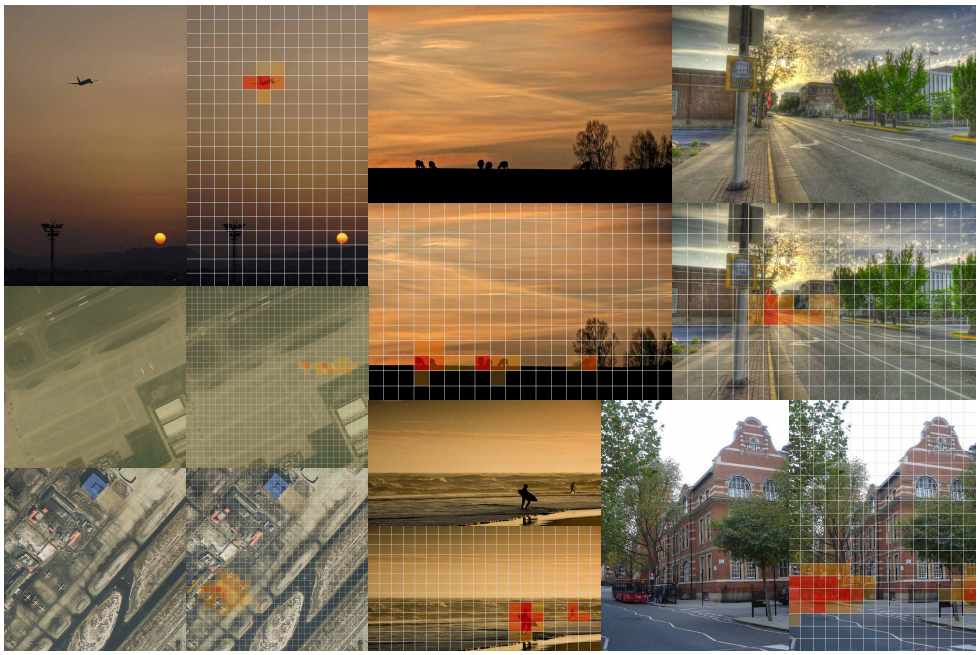


Figure H.1: Additional small object visualizations. The predicted LET scores focus on small targets such as distant aircraft, far people, road signs, small aerial objects, pedestrians, and vehicles.

The small object cases show that LENS can preserve fine evidence even when the target occupies only a few visual patches. In Fig. H.1, the highlighted regions are concentrated around tiny objects rather than spread over the sky, road, sea, or building background. This behavior is important for the proposed latent suppression mechanism. The LET assigns higher scores to compact target regions, so latent noise mainly weakens irrelevant context while preserving the local evidence used for the answer.

The cluttered scene cases show why visual evidence purification is more suitable than simply adding more reasoning context. In Fig. H.2, the first underwater example contains a turtle whose texture and color are close to the surrounding reef, making it difficult even for human observers to separate the animal from the background. The learned LET scores still give strong responses around the turtle and other queried regions, while suppressing large areas of water and coral. Similar behavior appears in scenes with riders, flags, building facades, indoor objects, and multiple people. These examples show that LET responses remain concentrated around queried regions in cluttered scenes, consistent with reduced distractor interference.

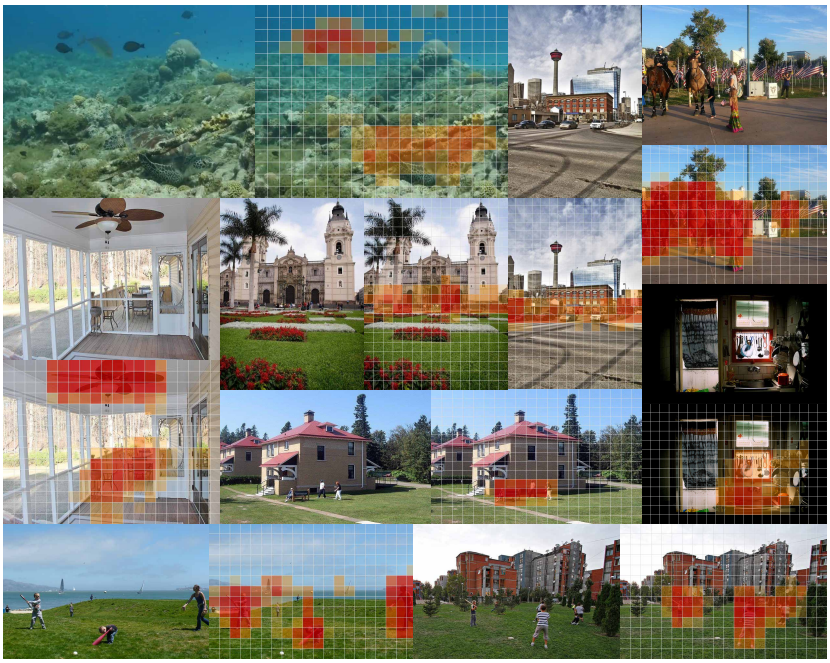


Figure H.2: Additional visualizations for occluded, hidden, and dense scenes. The examples include underwater targets, cluttered street scenes, indoor occlusion, building scenes, and crowded outdoor scenes.

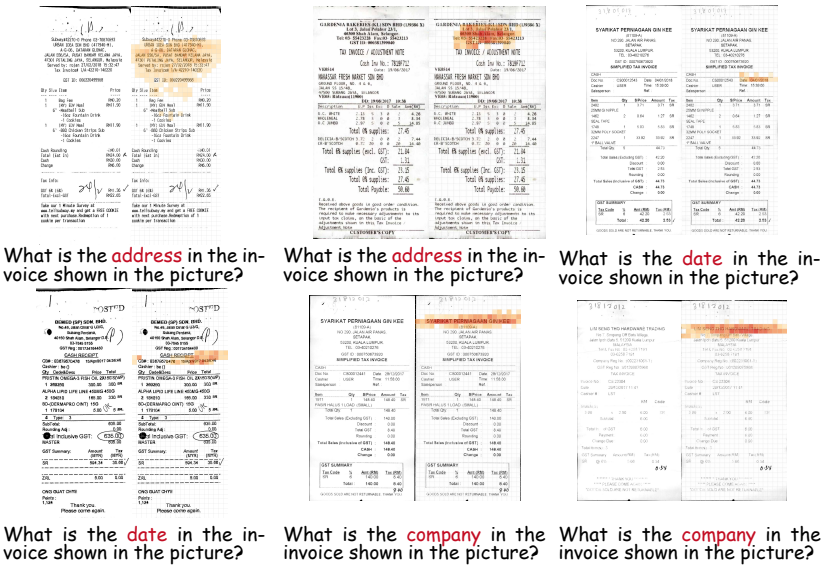


Figure H.3: Additional OCR visualizations. The predicted LET scores move to the text field required by the question, including address, date, and company information.

The OCR cases demonstrate that the LET scores are conditioned on the question rather than fixed document saliency. Invoices contain many visually similar text lines, and a generic document attention pattern may focus on headers, totals, or dense item rows regardless of the question. In Fig. H.3, the highlighted patches shift according to the requested field. Address questions activate address blocks, date questions activate date lines, and company questions activate merchant names. This behavior matches the objective of LENS, which uses the question to select evidence through the LET scores before latent suppression. As a result, redundant text regions are weakened and the decoder receives a cleaner document representation for OCR style reasoning.