Latent Evidence-based Visual Reasoning

Anonymous ACL submission

Abstract

Latent visual reasoning serves as an efficient way for multimodal large language models to preserve fine-grained visual representations. Nevertheless, existing visual latents are typically induced from external priors, generic perceptual targets, or implicit bottlenecks, which may weakly tied to the evidence required by the answer. We propose **L**atent **E**vidence-based **V**isual **R**easoning (**LEVR**), which learns latent cues as answer-supporting visual evidence rather than generic visual thoughts. LEVR introduces *Teacher-Forcing Evidence Discovery*, converting ordinary image-question-answer triples into soft evidence supervision by contrasting attention under answer-conditioned and reference prompts. Concretely, LEVR operates in two cascading stages: *Latent Evidence Induction* aligns cue-to-image attention with the discovered evidence targets, and *Evidence-Block Learning* forces answer generation to route through these intermediates by blocking direct patch attention. Across ten diverse perception-intensive benchmarks, LEVR consistently outperforms state-of-the-art baselines and improves Qwen2.5-VL by 4.7% on average. Extensive analyses further demonstrate that the learned latent cues are both interpretable and effective for visual-centric tasks.

1 Introduction

Multimodal large language models (MLLMs) (Li et al., 2023a; Ahmed et al., 2025; Chen et al., 2024b) have become a central paradigm for vision-language understanding by coupling visual encoders with powerful language backbones. Despite their strong language reasoning ability, many vision-centric tasks often hinge on localized cues, such as counts, spatial relations, depth, or subtle attributes, that are poorly represented by textual descriptions alone (Tong et al., 2024b; Qin et al., 2025; Fu et al., 2024). This motivates a shift toward "thinking with images" (Su et al., 2025; Zhang et al., 2025b; Fan et al., 2026; Shen et al., 2025a), where visual information is preserved as part of the intermediate reasoning process rather than being fully translated into language.

Latent visual reasoning (Li et al., 2025b; Zhang et al., 2025a) offers a compact way to realize this paradigm by introducing visual latent tokens or continuous visual thoughts into the reasoning process, preserving fine-grained visual information beyond textual rationales. As summarized in Figure 1, existing latent visual reasoning methods mainly differ in how visual latents are induced. ***Specialist-prior latents*** inject visual knowledge from external expert models or teacher representations (Li et al., 2025a; Wu et al., 2026), but the resulting latents inherit the objectives and biases of the source models and may not align with the evidence required by a specific question-answer pair. ***Perception-feature latents*** align latent tokens with perceptual targets such as segmentation, depth, edges, or dense visual features (Bigverdi et al., 2025; Qin et al., 2025). Although these targets provide rich visual structure, they are often task-agnostic, and perceptually meaningful regions may not support the current answer. ***Implicitly-induced latents*** encourage latent states to emerge from visual bottlenecks, corrupted inputs, or consistency objectives (Li et al., 2025b; Zhang et al., 2026). While these objectives can make latent states useful for prediction, the resulting latents remain opaque, *i.e.*, their visual semantics are not explicitly localized, and their dependency on the final answer may not correspond to interpretable answer-supporting evidence.

In this work, we propose **L**atent **E**vidence-based **V**isual **R**easoning (**LEVR**), an evidence-aware latent reasoning framework that learns latent cues as *answer-supporting visual evidence* rather than generic visual thoughts. Instead of defining evidence as visually salient regions or regions semantically related to the question (Selvaraju et al., 2017; Petsiuk et al., 2018), LEVR defines evidence from

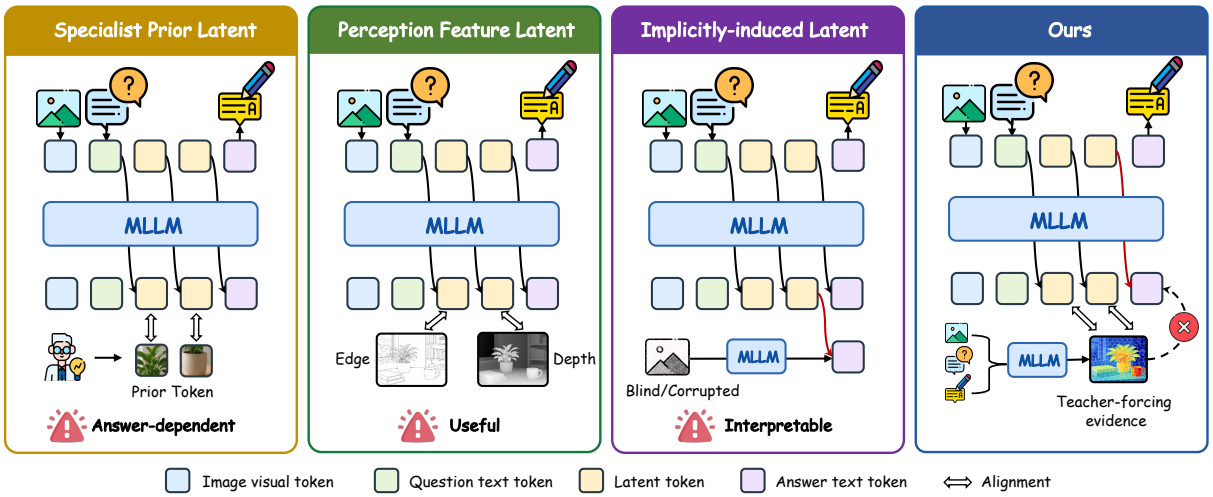

Figure 1: Comparison of current latent-learning paradigms. Prior methods induce visual latents from specialist priors, perceptual features, or implicit objectives, which may lack answer relevance, task-level evidence, or explicit interpretability. Our method discovers answer-supporting evidence via teacher forcing and aligns latent cues with the evidence used for answer generation.

the perspective of *answer reliance*: a visual region is considered evidence if its importance increases when the model is guided toward the correct answer. Under this view, evidence is not simply what the model attends to, but the visual information on *which the answer depends*.

To discover such evidence, LEVR introduces **Teacher-Forcing Evidence Discovery** (TFED), which converts ordinary image-question-answer triples into soft evidence supervision. Unlike methods that rely on step-wise textual rationales or visual reasoning traces (Wei et al., 2022; Shao et al., 2024; Zhang et al., 2025b; Shen et al., 2025a), TFED directly starts from the gold answer and identifies the visual regions that become more influential when the model is guided toward that answer. Specifically, TFED contrasts model attention under an answer-conditioned teacher-forcing prompt and a reference prompt, highlighting regions whose importance increases toward the target answer. The resulting attention contrast is aggregated over SLIC superpixels (Achanta et al., 2012) and converted into soft token-level evidence targets. This process turns standard answer labels into localized visual evidence supervision, without requiring external vision experts, dense perception labels, object boxes, or step-wise rationale annotations.

Based on the discovered evidence, LEVR trains latent evidence cues through a two-stage process. First, **Latent Evidence Induction** inserts compact cue tokens before answer generation and aligns their cue-to-image attention with the discovered evidence targets. It provides explicit localization supervision for the latent cues, enabling them to capture answer-dependent visual regions without relying on step-wise textual rationales (Shao et al., 2024) or reconstructing task-agnostic perceptual targets (Bigverdi et al., 2025; Qin et al., 2025). Then, **Evidence-Block Learning** blocks answer tokens from directly attending to high-prior evidence patches. This prevents the model from bypassing the latent cues and encourages answer generation to use them as intermediate evidence carriers. Together, the two stages make the learned cues both interpretable in where they attend and functional in how they support prediction.

We evaluate LEVR on diverse perception-intensive visual reasoning benchmarks (Fu et al., 2024; Tong et al., 2024a; Wang et al., 2025b; Wu and Xie, 2024). LEVR consistently improves Qwen2.5-VL-7B (Bai et al., 2025) and achieves the best average performance among open-source methods. Extensive quantitative and qualitative analyses further show that the effectiveness of learned cues in localizing fine-grained answer-supporting regions and serving as functional evidence carriers during answer generation.

Our main contributions are threefold: 1) We introduce an answer-reliance view of visual evidence for latent visual reasoning. 2) We propose Teacher-Forcing Evidence Discovery, which derives soft evidence supervision from ordinary image-question-answer triples. 3) We present LEVR, a two-stage evidence-aware latent reasoning framework that trains latent cues to be useful, interpretable, and answer-dependent.

2 Related Work

Latent space reasoning. Latent space reasoning replaces explicit natural-language rationales with

continuous hidden states, soft tokens, or latent trajectories. In language models, prior work has explored hidden chain-of-thought decoding (Liu et al., 2024), latent chain-of-thought vectors (Zhang and Viteri, 2024), continuous latent reasoning as in Coconut (Hao et al., 2024), compressed chains of thought (Cheng and Van Durme, 2024), soft chain-of-thought tokens (Xu et al., 2025), and self-distilled continuous reasoning (Shen et al., 2025b). This paradigm is particularly relevant to multimodal reasoning, where discrete text may be insufficient for preserving fine-grained visual information, *e.g.*, spatial layout and subtle attributes.

Recent MLLM methods extend latent reasoning by inserting visual latent tokens or continuous visual thoughts into the reasoning process. These methods can be roughly grouped by how the latents are induced. Perception-aligned methods supervise latent tokens with dense visual targets or specialist features, including Perception Tokens (Bigverdi et al., 2025) and Chain-of-Visual-Thought (Qin et al., 2025). Prior- or teacher-guided methods use auxiliary supervision, latent sketches, or teacher trajectories, such as Latent Visual Reasoning (Li et al., 2025a), Monet (Wang et al., 2025a), Latent Sketchpad (Zhang et al., 2025a), Sketch-in-Latents (Tong et al., 2025), and LaViT (Wu et al., 2026). Implicitly induced methods encourage visual latents through bottlenecks, corrupted inputs, attention masking, or consistency constraints, such as LIVR (Li et al., 2025b), CrystaL (Zhang et al., 2026), and interleaved latent visual reasoning (Dong et al., 2025). While these approaches make latent states useful for multimodal prediction, their supervision is typically generic, teacher-dependent, or implicit. In contrast, LEVR trains latent tokens as answer-supporting evidence cues, aligning them with the visual regions on which the final answer depends.

Visual evidence for multimodal reasoning. Visual evidence grounding aims to identify the image regions that support a model's prediction. Classical explanation methods localize important regions through gradients or perturbations, such as Grad-CAM (Selvaraju et al., 2017) and RISE (Petsiuk et al., 2018), but they are typically post-hoc and do not ensure that the highlighted regions are functionally used during answer generation. In MLLMs, visual evidence is increasingly incorporated into the reasoning process itself. Search-based methods explicitly locate relevant regions before answering, including guided visual search in V* (Wu and Xie, 2024), adaptive focusing in Chain-of-Focus (Zhang et al., 2025b), and tree-based zooming in Zoom-Eye (Shen et al., 2025a). Trace-based methods construct visual intermediate reasoning processes, including bounding-box-based visual rationales in Visual CoT (Shao et al., 2024) and image-centric reasoning traces in thinking-with-images methods (Su et al., 2025). While these methods make visual evidence more accessible to MLLMs, they often rely on explicit region operations, search procedures, or visual rationales. Moreover, their selected regions may reflect visual saliency or question relevance rather than answer dependence. LEVR instead discovers answer-dependent evidence from ordinary image-question-answer triples and trains latent cues to carry this evidence within standard autoregressive decoding.

3 Approach

We propose LEVR, a latent evidence-based visual reasoning framework that learns latent cues as answer-supporting visual evidence. The key challenge is twofold: the model should identify which visual regions support the answer, and then learn to use the corresponding latent cues during answer generation. To this end, LEVR consists of three components, as illustrated in Figure 2. First, *Teacher-Forcing Evidence Discovery* (§3.1) derives soft evidence targets from image-question-answer triples. Second, *Latent Evidence Induction* (§3.2) aligns latent cue attention with the discovered evidence, making the cues visually grounded. Third, *Evidence-Block Learning* (§3.3) prevents answer tokens from directly accessing evidence patches, encouraging the model to route answer generation through the learned cues.

Problem Setup. For each training example, we are given an image I, a question q, and a gold answer sequence $y = (y_1, \dots, y_T)$. The image is encoded into N image tokens, whose index set is denoted by $\mathcal{V} = \{1, \dots, N\}$. LEVR augments the response with M visual latent cue tokens before answer generation. Specifically, the response sequence $r = (r_1, \dots, r_L)$ follows the format <think> <vis_cue> $\cdots$ <vis_cue> </think> <answer> y </answer>. The <vis_cue> tokens are discrete marker tokens, while their contextual representations $\mathbf{c}_1, \dots, \mathbf{c}_M$ serve as continuous visual latent cues. A multimodal decoder parameterized by θ models the response autoregressively:

$$p_\theta(r \mid I, q) = \prod_{t=1}^{L} p_\theta(r_t \mid I, q, r_{<t}). \qquad (1)$$

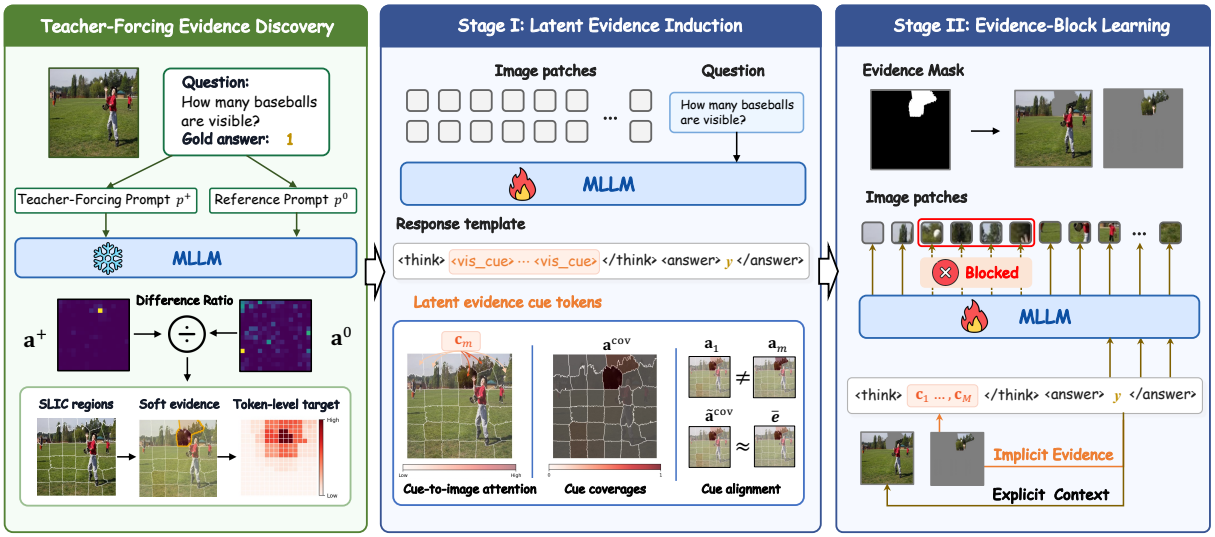

Figure 2: **Overview of LEVR.** *Teacher-Forcing Evidence Discovery* first derives answer-dependent soft evidence targets by contrasting answer-conditioned and reference attention; *Latent Evidence Induction* aligns visual latent cue states with the discovered evidence through cue-to-image attention; and *Evidence-Block Learning* masks direct answer-to-evidence attention during training, encouraging answer generation to use the learned cues as intermediate evidence carriers.

3.1 Teacher-Forcing Evidence Discovery

Instead of relying on human rationales or step-wise annotations, we propose **Teacher-Forcing Evidence Discovery** (TFED), which derives soft evidence targets directly from ordinary image-question-answer triples.

The intuition is that answer-supporting regions should become more important when the model is guided toward the correct answer. However, raw decoder attention often exhibits a generic attention sink, *i.e.*, globally salient regions consistently attract high attention regardless of the target answer (Darcet et al., 2024). To suppress such generic saliency and isolate answer-relevant evidence, TFED contrasts image attention under an answer-conditioned prompt and a reference prompt. The answer-conditioned prompt is p^+ = "Question: q Answer: y", while the reference prompt p^0 asks for a generic image description (Khayatkhoei et al., 2025).

Let $\mathbf{a}^+, \mathbf{a}^0 \in \mathbb{R}^N$ denote the head-averaged image-attention maps from the last text token to all image tokens at a fixed decoder layer $\ell^\star$ under p^+ and p^0, respectively. The relative attention map is:

$$\mathbf{a}^{\text{rel}} = \text{Norm}\left(\mathbf{a}^+ \oslash \max(\mathbf{a}^0, \epsilon)\right), \qquad (2)$$

where $\oslash$ denotes element-wise division, $\max(\cdot, \epsilon)$ is applied element-wise, and $\text{Norm}(\cdot)$ min-max normalizes the map to $[0, 1]$. This ratio suppresses regions that attract generic visual attention and highlights regions whose attention increases specifically under the answer-conditioned prompt.

To obtain spatially coherent evidence targets, we aggregate $\mathbf{a}^{\text{rel}}$ over SLIC superpixels (Achanta et al., 2012). Let R_k denote the k-th superpixel, and let $\rho(i)$ map image token i to its corresponding superpixel. We compute the region-level evidence score as: $s_k = \text{Avg}_{i:\rho(i)=k}\, a_i^{\text{rel}}$. Evidence seed regions are selected by thresholding the region scores: $\mathcal{S}_\tau = \{k : s_k \geq \tau\}$.

We further convert the selected superpixel regions into smooth token-level evidence targets. Let μ_k denote the centroid of region R_k. For each image token i, the evidence target is defined as:

$$e_i = \max_{k \in \mathcal{S}_\tau} \frac{s_k}{\max_{j \in \mathcal{S}_\tau} s_j} \exp\left(-\frac{\|\mu_{\rho(i)} - \mu_k\|_2^2}{2\sigma^2}\right),$$
$$\bar{e}_i = \frac{e_i}{\max_{j \in \mathcal{V}} e_j}. \qquad (3)$$

This Gaussian relaxation reduces the boundary mismatch between superpixels and vision tokens while preserving the relative importance of evidence regions. As a result, TFED converts standard answer supervision into localized soft evidence targets without additional annotations.

3.2 Stage I: Latent Evidence Induction

TFED provides soft evidence targets, but the model still needs to internalize this evidence into latent states. Stage I therefore trains the contextual representations of the visual latent cue tokens to serve as localized evidence carriers by aligning their cue-to-image attention with the evidence target $\bar{\mathbf{e}}$.

Let $\mathcal{D}$ denote the decoder layers used for atten-

tion supervision. For the m-th <vis_cue> position and image token $i \in \mathcal{V}$, let $w_{\ell,h}^{m,i}$ be the attention weight from this cue position to image token i at decoder layer ℓ and head h. We aggregate these weights over selected layers and heads, and renormalize them over image tokens:

$$a_{m,i} = \frac{\sum_{\ell \in \mathcal{D}} \sum_{h=1}^{H} w_{\ell,h}^{m,i}}{\sum_{j \in \mathcal{V}} \sum_{\ell \in \mathcal{D}} \sum_{h=1}^{H} w_{\ell,h}^{m,j}}. \tag{4}$$

When multiple cue tokens are used, we aggregate their attention into a soft coverage map using the noisy-or operator (Wang et al., 2015):

$$a_i^{\text{cov}} = 1 - \prod_{m=1}^{M} (1 - a_{m,i}), \quad \tilde{a}_i^{\text{cov}} = \frac{a_i^{\text{cov}}}{\sum_{j \in \mathcal{V}} a_j^{\text{cov}}}. \tag{5}$$

Here, $\tilde{a}_i^{\text{cov}}$ denotes the normalized coverage assigned by all cue tokens to image token i, allowing different cues to capture complementary regions.

During training, we align this coverage with the TFED evidence target using the cue-attention loss:

$$\mathcal{L}_{\text{cue}} = \left(1 - \sum_{i \in \mathcal{V}} \tilde{a}_i^{\text{cov}} \bar{e}_i\right)^2. \tag{6}$$

This encourages the latent cues to cover answer-supporting evidence regions.

When $M > 1$, we further discourage different cues from attending to the same regions with a diversity loss:

$$\mathcal{L}_{\text{div}} = \frac{1}{M(M-1)} \sum_{m \neq n} \sum_{i \in \mathcal{V}} a_{m,i} a_{n,i}. \tag{7}$$

The final Stage-I objective combines standard language modeling with evidence alignment and diversity regularization:

$$\mathcal{L}_{\text{I}} = \mathcal{L}_{\text{LM}} + \lambda_{\text{cue}} \mathcal{L}_{\text{cue}} + \lambda_{\text{div}} \mathcal{L}_{\text{div}}. \tag{8}$$

Here, $\mathcal{L}_{\text{LM}}$ denotes the autoregressive language modeling loss over the full response sequence r. Through this objective, the visual latent cue states are trained to act as differentiable evidence selectors that localize answer-supporting visual regions.

3.3 Stage II: Evidence-Block Learning

Although Stage I teaches latent cues to localize answer-supporting evidence, localization alone does not ensure that answer generation relies on these cues (Serrano and Smith, 2019). Answer tokens may still directly attend to the same high-evidence image tokens, allowing the model to bypass the latent cue states. Stage II addresses this issue by applying an evidence-blocked attention mask during training. The model retains access to unblocked image tokens as explicit visual context, while the direct path from answer tokens to high-evidence image tokens is removed.

Given the TFED target $\bar{\mathbf{e}}$, we define the blocked evidence tokens as $\mathcal{B}_\delta = \{i \in \mathcal{V} : \bar{e}_i > \delta\}$, where δ is a threshold. Let $\mathcal{P}_\delta$ denote their corresponding positions in the decoder input, and let $\mathcal{A}$ denote the answer-generation query positions inside the <answer> span. For query position t and key position j, the evidence-blocked attention mask is:

$$\Omega_{t,j} = \begin{cases} -\infty, & j > t, \\ -\infty, & t \in \mathcal{A},\ j \in \mathcal{P}_\delta, \\ 0, & \text{otherwise}. \end{cases} \tag{9}$$

The first case preserves standard causal decoding, while the second case prevents answer tokens from directly attending to high-evidence image tokens. All other image tokens remain accessible.

With the evidence-blocked attention mask, Stage II optimizes the autoregressive language modeling loss over the full response sequence:

$$\mathcal{L}_{\text{II}} = -\frac{1}{L} \sum_{t=1}^{L} \log p_\theta(r_t \mid I, q, r_{<t}; \Omega). \tag{10}$$

Although the loss is computed over the full response, the modified mask affects only answer-generation queries in $\mathcal{A}$. Since the latent cues have already learned to attend to the blocked evidence in Stage I, this encourages answer generation to access such evidence through the cue states rather than through direct image-token attention.

3.4 Inference

TFED and evidence-block learning are used only to shape the latent cue states during training. At inference time, LEVR takes only the image and the question as input and follows the same response format with <vis_cue> tokens before the answer span. The contextual states at these cue positions serve as intermediate visual evidence carriers for subsequent answer generation. The final answer is then produced by standard autoregressive decoding, without requiring evidence targets, SLIC superpixels, external vision experts, or attention blocking.

4 Experiments

Implementation details. Following recent latent visual reasoning work (Qin et al., 2025; Zhang

et al., 2026), we use Qwen2.5-VL-7B-Instruct (Bai et al., 2025) as the backbone and fine-tune it with the same LoRA (Hu et al., 2022) setting. To ensure a fair comparison, LEVR is trained on the dataset established by CoVT (Qin et al., 2025), same as Crystal (Zhang et al., 2026). Stage I induces latent evidence cues via cue-to-evidence alignment, and Stage II continues from the resulting checkpoint with evidence-block learning. All experiments are conducted on four NVIDIA H20 GPUs. Our codes and models will be released upon acceptance. More details are provided in Appendix §A.

Evaluation Benchmarks. We evaluate LEVR with VLMEvalKit (Duan et al., 2024) on ten vision-centric multimodal benchmarks. Specifically, we use CVBench from Cambrian-1 (Tong et al., 2024a), reporting CV_{2D} and CV_{3D}, HRBench (Wang et al., 2025b), reporting HR_{4K} and HR_{8K}, BLINK (Fu et al., 2024), RealWorldQA (XAI, 2024), V*Bench (Wu and Xie, 2024), MMVP (Tong et al., 2024b), and MMStar (Chen et al., 2024a). These benchmarks emphasize fine-grained visual understanding, spatial reasoning, high-resolution image comprehension, and vision-dependent multimodal evaluation, covering challenging scenarios such as detailed visual patterns, complex layouts, real-world images, and subtle spatial relations. To further examine robustness against object hallucination, we evaluate on POPE (Li et al., 2023b), which is designed to diagnose hallucinated visual claims in MLLMs.

Baselines. We compare LEVR with the Qwen2.5-VL-7B backbone and representative open-source multimodal reasoning methods. For latent visual reasoning, we include methods with different latent-induction mechanisms: CoVT (Qin et al., 2025) aligns continuous visual tokens with dense perceptual targets; LVR (Li et al., 2025a) and SKILA (Tong et al., 2025) induce visual latents from prior or sketch-based supervision; and LIVR (Li et al., 2025b) and CrystaL (Zhang et al., 2026) obtain visual latents through implicit objectives or bottlenecks. We also compare with Vision-R1 (Huang et al., 2025), a reinforcement-learning-based multimodal reasoning method. We additionally report GPT-4o and Claude-4-Sonnet results as strong closed-source reference models.

4.1 Quantitative Analysis

Comparison with state-of-the-art methods. Table 1 compares LEVR with recent open-source and closed-source MLLMs. LEVR achieves the best open-source average performance, reaching 74.7% and improving the Qwen2.5-VL-7B (Bai et al., 2025) backbone by 4.7%. The gains are especially large on CV_{3D} (+11.3%), HR_{4K} (+4.8%), HR_{8K} (+7.1%), and MMVP (+9.1%), suggesting that latent evidence cues are most beneficial when prediction depends on localized, high-resolution, or fine-grained visual evidence. Compared with recent visual or latent reasoning methods, LEVR remains comparable on their strongest benchmarks, such as CV_{3D}, V*Bench, MMStar, and POPE, while improving the backbone across all ten benchmarks and achieving the highest open-source average. This indicates that answer-dependent evidence supervision provides a transferable signal across heterogeneous perception-intensive tasks, rather than improving only a specific reasoning pattern. More quantitative analyses are left in the Appendix §B.

4.2 Ablation Study

Effect of cue token number. Table 2 studies the effect of the number of latent evidence cue tokens. Introducing a single cue improves the average score from 70.1% to 71.5%, suggesting that one cue is often sufficient to capture the dominant answer-supporting evidence. Using two cues gives a further but modest gain, reaching 71.7%, which indicates that an additional cue can complement the first one when the evidence coverage is incomplete or ambiguous. However, increasing the number to four reduces the average score to 71.3%. This suggests that excessive cue slots may compete for similar evidence regions, making the cue-to-evidence alignment more redundant rather than more informative.

Evidence target construction. We compare different strategies for constructing training evidence targets in Table 3. Random Crop gives the lowest average score, suggesting that supervision from answer-irrelevant regions fails to teach cue tokens where to attend. Vanilla Attention is stronger but may still reflect generic visual saliency, while VisualCoT-BBox (Shao et al., 2024) performs well on V*Bench but relies on external localized rationales that do not consistently transfer across benchmarks. LEVR achieves the best average score of 71.5%, with the strongest results on CV_{2D} and BLINK, indicating that teacher-forcing attention contrast yields a more balanced answer-dependent signal for latent evidence induction.

Table 1: **Comparison with state-of-the-art methods.** Results are accuracy (%) evaluated with VLMEvalKit (Duan et al., 2024). Best and second-best open-source results are marked in **bold** and underline, respectively. ⋆ denotes reproduced results with their official codebases.

Method	CV_{2D}	CV_{3D}	HR_{4K}	HR_{8K}	BLINK	RWQA	V*	MMVP	MMStar	POPE	Avg
Closed-source Methods											
Claude-4-Sonnet	-	-	32.3	22.7	39.6	63.7	15.2	-	-	-	-
GPT-4o	-	-	50.6	46.7	63.0	69.7	42.9	-	-	-	-
Open-source Methods											
Qwen2.5-VL-7B	75.0	73.3	68.6	64.9	55.7	68.6	76.4	70.2	60.3	86.5	70.0
CoVT (Qin et al., 2025)	75.1	**84.7**	71.0	69.4	56.0	**71.6**	78.0	78.3	**65.0**	84.6	73.4
LVR (Li et al., 2025a)	72.0	65.9	69.6	66.1	52.5	67.7	81.7	72.0	61.3	84.2	69.3
Vision-R1 (Huang et al., 2025)	8.0	37.6	64.8	57.0	51.0	9.4	80.1	72.6	60.7	77.8	51.9
LIVR (Li et al., 2025b)	73.9	82.7	72.4	69.0	56.3	69.2	79.1	-	-	**89.5**	-
SKILA (Tong et al., 2025)	**77.3**	80.3	72.0	66.5	56.7	70.5	**84.3**	75.3	64.8	87.3	73.5
Crystal⋆ (Zhang et al., 2026)	72.0	84.3	72.8	69.3	53.0	66.9	79.1	78.3	62.6	84.9	72.3
LEVR	76.3	84.6	**73.4**	**72.0**	**57.8**	70.7	79.1	**79.3**	64.3	88.2	**74.7**
Δ(vs Qwen2.5-VL-7B)	**+1.3**	**+11.3**	**+4.8**	**+7.1**	**+2.1**	**+2.1**	**+2.7**	**+9.1**	**+4.0**	**+1.7**	**+4.7**

Table 2: **Ablation on the number of latent evidence cue tokens.** The 0 token setting removes latent evidence cues and serves as the direct fine-tuning baseline.

Quantity	CV_{2D}	HR_{4K}	BLINK	RWQA	V*	Avg
0 tokens	74.6	72.2	54.8	70.3	78.5	70.1
1 tokens	76.3	73.4	57.8	70.7	79.1	71.5
2 tokens	76.6	71.9	56.9	71.8	81.2	71.7
4 tokens	76.9	72.4	56.8	72.2	78.0	71.3

Table 3: **Ablation on evidence discovery.** Different methods are used to construct training evidence targets: *Random Crop* uses a random local image-token, *Vanilla Attention* uses raw answer-to-image attention, *VisualCoT-BBox* uses external VisualCoT (Shao et al., 2024) bounding boxes.

Method	CV_{2D}	HR_{4K}	BLINK	RWQA	V*	Avg
Random Crop	74.5	72.6	56.0	69.6	78.0	70.1
Vanilla Attention	75.3	74.9	56.5	71.1	78.5	71.3
VisualCoT-BBox	75.3	73.0	56.8	69.0	80.6	71.0
LEVR	76.3	73.4	57.8	70.7	79.1	71.5

Table 4: **Ablation on training stages.** Stage I denotes latent evidence induction, and Stage II is evidence-block learning.

Stage I	Stage II	CV_{2D}	HR_{4K}	BLINK	RWQA	V*	Avg
–	–	74.6	72.2	54.8	70.3	78.5	70.1
✓		75.4	71.6	56.6	71.5	79.6	70.9
	✓	76.0	71.3	57.2	71.5	79.1	71.0
✓	✓	76.3	73.4	57.8	70.7	79.1	71.5

Contribution of training stages. We ablate the two training stages in Table 4 to examine their individual contributions. Stage I alone raises the average score from 70.1% to 70.9%, indicating that cue-to-evidence alignment helps ground the latent cue states. When only Stage II is applied, the average score reaches 71.0%, suggesting that evidence blocking can reduce direct reliance on high-evidence image tokens even without explicit cue alignment. The full two-stage training achieves the best average score of 71.5%, supporting our design of first grounding latent cues in evidence and then encouraging answer generation to use them as an alternative evidence route.

4.3 Qualitative Analysis

Evidence Targets. Figure 4 compares the training evidence targets produced by VisualCoT (Shao et al., 2024) boxes and LEVR on the CoVT dataset (Qin et al., 2025). As seen, VisualCoT boxes often provide coarse object-level supervision: they may cover a broad object region, miss multiple relevant instances, or localize an object without isolating the specific visual cue needed for the answer. For example, in the horse-counting case, box-style supervision may fail to cover all counted instances, while LEVR highlights multiple evidence regions corresponding to the visible horses. In the goggles and boiling examples, LEVR focuses on the answer-critical visual details around the eyes and the pot, rather than assigning supervision to a coarse person or object region. These comparisons show that teacher-forcing attention contrast produces softer and more answer-conditioned evidence maps, enabling cue supervision without human boxes or explicit visual rationales.

Cue Attention. We visualizes the latent cue attention of various visual-centric reasoning examples sampled from BLINK (Fu et al., 2024) in Figure 3. Across diverse task types, LEVR consistently attends to compact regions that are directly involved in answering the question. For instance, in the counting example, the cue can focus on three glasses instead of other objects on the table in the scene. In functional correspondence, it attends to the candidate point corresponding to the reference action. As for relative depth, it highlights the compared point instead of spreading over the background. Similarly, for object localization and jig-

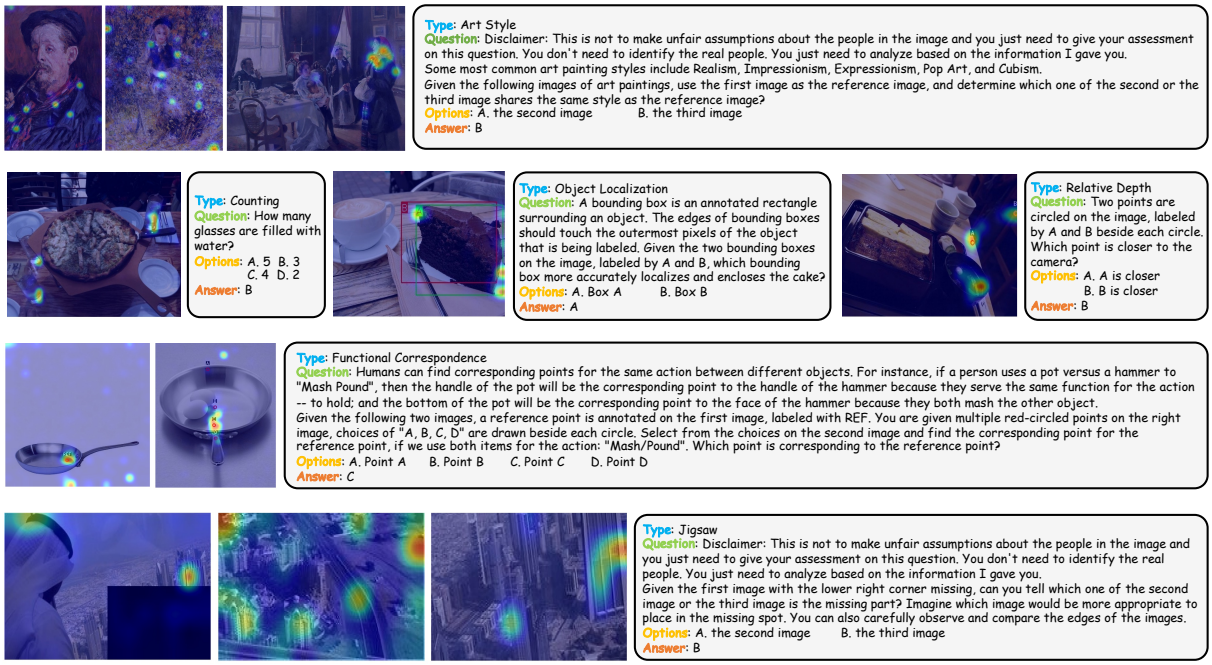

Figure 3: **Visualization of latent cue attention.** These examples are randomly sampled from BLINK (Fu et al., 2024).

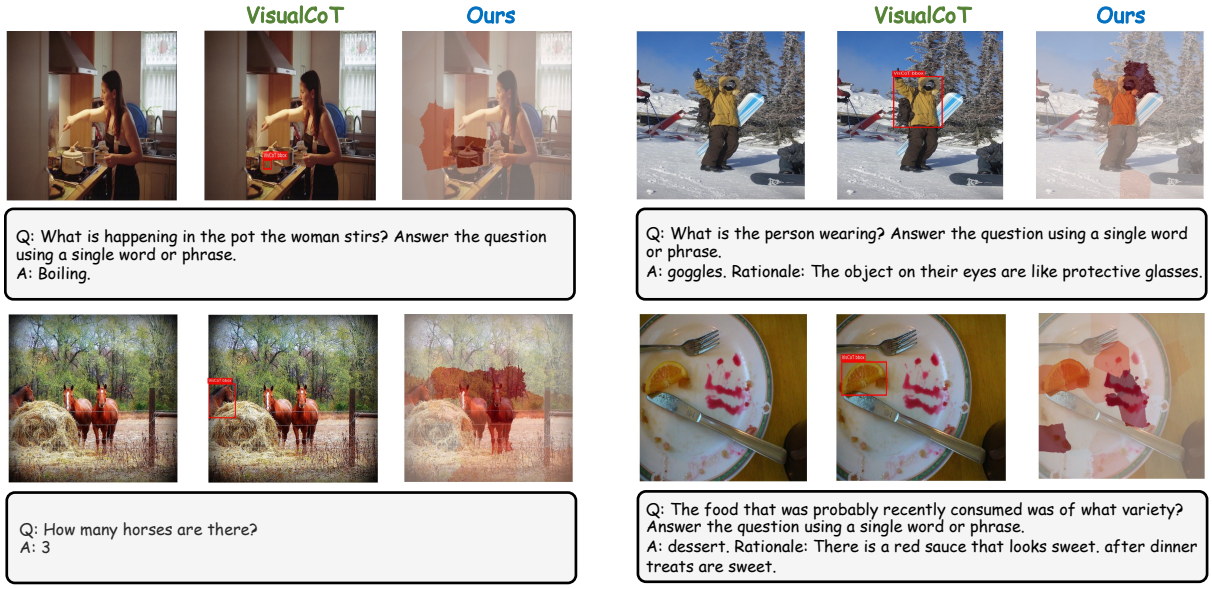

Figure 4: **Visualization of training evidence targets.** We compared the training evidence targets generated by VisualCoT (Shao et al., 2024) and those generated by TFED. The format of the answer is open-ended.

saw reasoning, the attention is concentrated around the relevant object boundary or missing-region context. These examples indicate that the learned cues do not merely capture generic saliency, but act as visual evidence pointers for answer generation. More qualitative results are left in the Appendix §D.

5 Conclusion

We introduced LEVR, an evidence-aware latent visual reasoning framework that learns latent cues as answer-supporting evidence rather than generic visual thoughts. LEVR derives soft evidence supervision from ordinary image-question-answer triples through Teacher-Forcing Evidence Discovery, and uses evidence-block learning to encourage answer generation to rely on the learned cues. Experiments across vision-centric benchmarks show that LEVR consistently outperforms state-of-the-art methods on average performance. Quantitative and qualitative analyses further indicate that the learned cues can localize compact regions sufficient for various visual reasoning, highlighting the importance of answer-dependent visual evidence.

Limitations

LEVR relies on answer-conditioned attention to derive evidence supervision from image-question-answer data. Its effectiveness may therefore be affected by noisy or ambiguous answer annotations and the attention quality of the underlying MLLM. In addition, SLIC-based aggregation may be less precise for very small objects or fine-grained visual structures. Future work may explore more robust evidence discovery mechanisms for such cases.

References

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

Imtiaz Ahmed, Sadman Islam, Partha Protim Datta, Imran Kabir, Md Naseef Ur Rahman Chowdhury, and Ahshanul Haque. 2025. Qwen 2.5: A comprehensive review of the leading resource-efficient llm with potentioal to surpass all competitors.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. 2025. Perception tokens enhance visual reasoning in multimodal language models. In *CVPR*, pages 3836–3845.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024a. Are we on the right way for evaluating large vision-language models? *NeurIPS*, 37:27056–27087.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198.

Jeffrey Cheng and Benjamin Van Durme. 2024. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2024. Vision transformers need registers. In *ICLR*, volume 2024, pages 2632–2652.

Shuai Dong, Siyuan Wang, Xingyu Liu, Chenglin Li, Haowen Hou, and Zhongyu Wei. 2025. Interleaved latent visual reasoning with selective perceptual modeling. *arXiv preprint arXiv:2512.05665*.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *ACM MM*, pages 11198–11201.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Xinze Guan, and Xin Wang. 2026. Grit: Teaching mllms to think with images. *NeurIPS*, 38:116522–116543.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *ECCV*, pages 148–166. Springer.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

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

Mahyar Khayatkhoei, Prateek Chhikara, Filip Ilievski, and 1 others. 2025. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *ICLR*, volume 2025, pages 68194–68213.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. 2025a. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. Llava-onevision: Easy visual task transfer. *TMLR*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742.

Kelvin Li, Chuyi Shang, Leonid Karlinsky, Rogerio Feris, Trevor Darrell, and Roei Herzig. 2025b. Latent implicit visual reasoning. *arXiv preprint arXiv:2512.21218*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. pages 292–305.

Tianqiao Liu, Zui Chen, Zitao Liu, Mi Tian, and Weiqi Luo. 2024. Expediting and elevating large language model reasoning via hidden chain-of-thought decoding. *arXiv preprint arXiv:2409.08561*.

Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*.

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

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? pages 2931–2951.

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

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. 2025a. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. pages 6613–6629.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025b. Codi: Compressing chain-of-thought into continuous space via self-distillation. pages 677–693.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, and 1 others. 2025. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*.

Jintao Tong, Jiaqi Gu, Yujing Lou, Lubin Fan, Yixiong Zou, Yue Wu, Jieping Ye, and Ruixuan Li. 2025. Sketch-in-latents: Eliciting unified reasoning in mllms. *arXiv preprint arXiv:2512.16584*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, and 1 others. 2024a. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *NeurIPS*, 37:87310–87356.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024b. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, pages 9568–9578.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. 2025a. Monet: Reasoning in latent visual space beyond images and language. *arXiv preprint arXiv:2511.21395*.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. 2025b. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *AAAI*, volume 39, pages 7907–7915.

Xinggang Wang, Zhuotun Zhu, Cong Yao, and Xiang Bai. 2015. Relaxed multiple-instance svm with application to object discovery. In *ICCV*, pages 1224–1232.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837.

Linquan Wu, Tianxiang Jiang, Yifei Dong, Haoyu Yang, Fengji Zhang, Shichaang Meng, Ai Xuan, Linqi Song, and Jacky Keung. 2026. Lavit: Aligning latent visual thoughts for multi-modal reasoning. *arXiv preprint arXiv:2601.10129*.

Penghao Wu and Saining Xie. 2024. V?: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, pages 13084–13094.

XAI. 2024. Grok-1.5 vision preview.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025. Softcot: Soft chain-of-thought for efficient reasoning with llms. pages 23336–23351.

Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, and 1 others. 2025a. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in mllms. *arXiv preprint arXiv:2510.24514*.

Jason Zhang and Scott Viteri. 2024. Uncovering latent chain of thought vectors in language models. *arXiv preprint arXiv:2409.14026*.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and 1 others. 2025b. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv e-prints*, pages arXiv–2505.

Yang Zhang, Danyang Li, Yuxuan Li, Xin Zhang, Tianyu Xie, Mingming Cheng, and Xiang Li. 2026. Crystal: Spontaneous emergence of visual latents in mllms. *arXiv preprint arXiv:2602.20980*.

Appendix

This appendix is organized as follows:

- §A provides implementation details, including training setup and hyperparameters.
- §B presents additional quantitative analyses
- §C reports more ablation studies.
- §D provides additional qualitative visualizations.

A Implementation Details

Training setup. Following recent latent visual reasoning work (Qin et al., 2025; Zhang et al., 2026), we fine-tune Qwen2.5-VL-7B-Instruct with LoRA rank 16, and LoRA alpha 32. The language model and vision tower are frozen, while LoRA adapters and the multimodal projector are updated. We use a global batch size of 8 and resize images to 448×448 during training. Stage I is trained for 4,000 steps with learning rates 2×10^{-4} for LoRA adapters. Stage II starts from the Stage-I checkpoint and is trained for several steps (*i.e.*, 150) with learning rates 1×10^{-4} for LoRA adapters.

Hyperparameter Details. TFED (§3.1) extracts answer-conditioned attention from decoder layer $\ell^\star = 22$ and uses 36 SLIC regions, evidence threshold $\tau = 0.1$, and Gaussian smoothing scale $\sigma = 0.05$. For latent evidence induction (§3.2), we use one latent evidence cue token by default ($M = 1$) and set the cue-attention loss weight to $\lambda_{\text{cue}} = 0.25$. The diversity loss is disabled for $M = 1$ with $\lambda_{\text{div}} = 0$; when using multiple cue tokens ($M > 1$), we set $\lambda_{\text{div}} = 0.5$.

B More Quantitative Analysis

Quality of learned latent cues. To examine whether the learned cue regions contain answer-relevant visual evidence, we further evaluate them on V*Bench (Wu and Xie, 2024) in Table 5. Random regions perform poorly under both masking and cropping, while CrystaL cues provide moderate improvements. In contrast, LEVR achieves 76.2% with masked cue regions and 78.8% with cropped cue regions, indicating that its cue-attended regions preserve the visual information needed for answering. This suggests that answer-conditioned evidence supervision learns cue regions that are not only interpretable, but also functionally useful for downstream prediction.

Table 5: **Analysis of the effectiveness of learned latent cues.** We conduct experiments V*Bench (Wu and Xie, 2024). *Mask* preserves only cue-attended regions, while *Crop* directly crops the cue regions for evaluation with Qwen2.5-VL.

Cue Source	Mask	Crop
Random	44.0	37.7
Crystal (Zhang et al., 2026)	57.6	55.0
LEVR	76.2	78.8

Table 6: **Analysis of validity of TFED evidence target.** Counterfactual Mask trains the model on images where TFED evidence regions are masked, under the same setting as LEVR.

Method	CV_{2D}	HR_{4K}	BLINK	RWQA	V*	Avg
Counterfactual Mask	69.9	71.8	54.7	69.9	78.5	69.0
LEVR	76.3	73.4	57.8	70.7	79.1	71.5

Validity of TFED evidence targets. Table 6 examines whether the regions discovered by TFED provide functional evidence for answer prediction. We train a counterfactual variant by masking the TFED evidence regions in the input image while keeping the same training setting as LEVR. Compared with LEVR, Counterfactual Mask reduces the average score from 71.5% to 69.0%, with particularly large drops on CV_{2D} and BLINK. This shows that removing TFED-selected regions weakens the model's ability to learn effective evidence cues, supporting that TFED captures answer-supporting visual regions rather than generic saliency alone.

C More Ablation Study

Effect of cue-attention loss. To evaluate the role of evidence alignment, we vary the cue-attention loss weight λ_{cue} in Table 7. Removing this loss reduces the average score to 70.1%, showing that language modeling alone cannot reliably induce evidence-grounded cue states. A moderate weight of $\lambda_{\text{cue}} = 0.25$ achieves the best average performance of 71.5%, indicating that cue-to-evidence alignment provides useful supervision for latent cue induction. Larger weights degrade performance, with the average score dropping to 70.9% and 70.3% for $\lambda_{\text{cue}} = 0.50$ and 1.00, respectively. This suggests that overly strong attention alignment may over-constrain the cue states and reduce their flexibility for answer generation.

Effect of diversity regularization. We further study the diversity loss under the two-cue setting in Table 8. Without diversity regularization, the model achieves 71.1% on average, suggesting that multiple cues can still provide useful evidence ca-

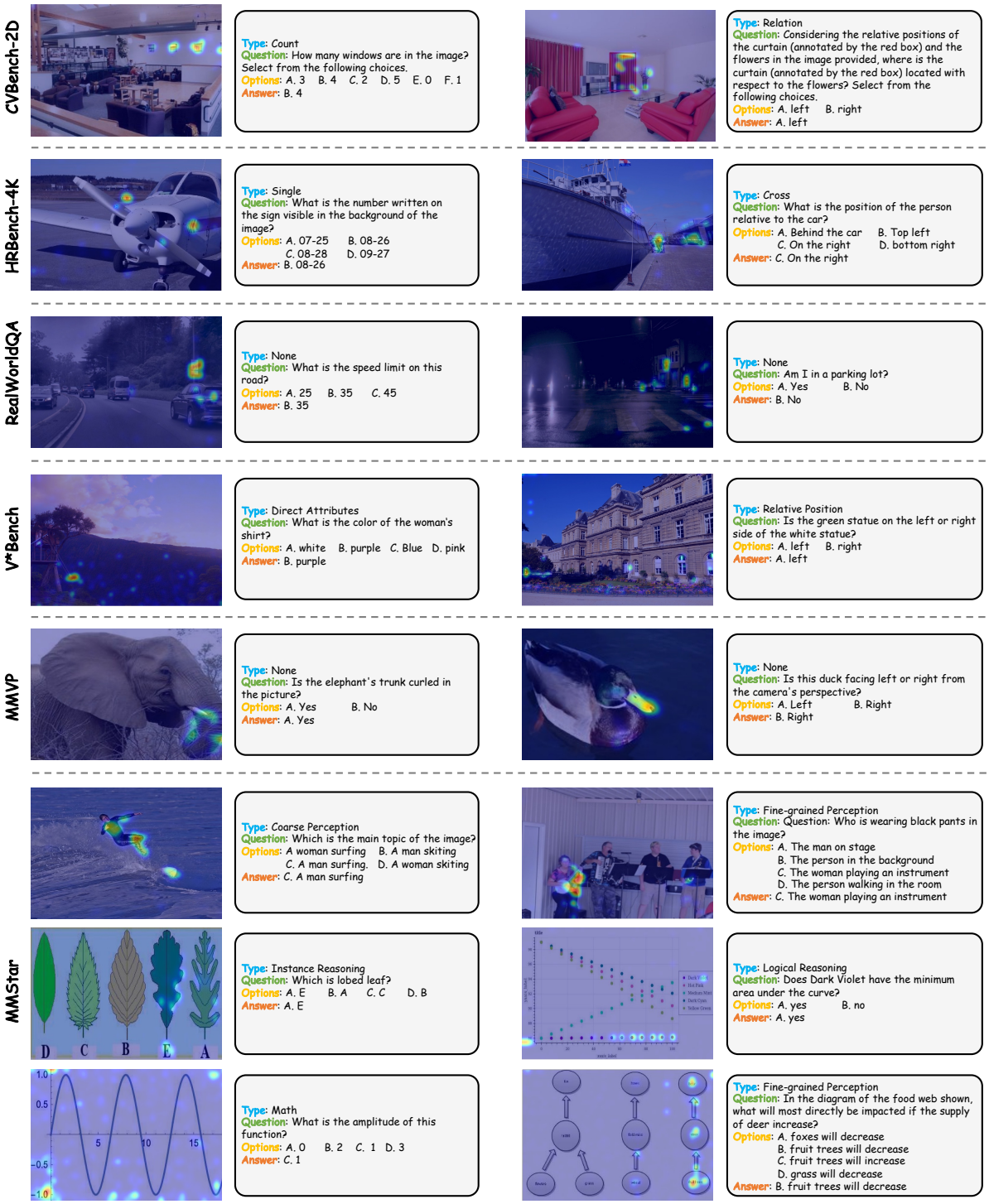

Figure 5: **Visualization of latent cue attention.** These examples are randomly sampled from the left benchmarks.

Table 7: **Ablation on cue-attention loss.**

λ_{cue}	CV_{2D}	HR_{4K}	BLINK	RWQA	V*	Avg
0.00	74.6	72.2	54.8	70.3	78.5	70.1
0.25	76.3	73.4	57.8	70.7	79.1	71.5
0.50	75.9	72.2	56.8	70.5	79.1	70.9
1.00	75.9	72.0	55.7	69.8	78.0	70.3

Table 8: **Ablation on diversity loss.** These experiments all use two latent cue tokens (M=2).

λ_{div}	CV_{2D}	HR_{4K}	BLINK	RWQA	V*	Avg
0.00	75.9	72.6	57.9	70.7	78.5	71.1
0.25	75.8	72.5	57.8	71.1	79.1	71.3
0.50	76.6	71.9	56.9	71.8	81.2	71.7
1.00	75.6	73.5	57.2	69.5	81.7	71.5

pacity. Increasing λ_{div} to 0.50 gives the best average score of 71.7%, indicating that encouraging

Table 9: **Ablation on TFED target region granularity.** Coarse, Default, and Fine use 16, 36, and 64 SLIC regions with $\sigma = 0.08$, 0.05, and 0.03, respectively.

Target Region	CV_{2D}	HR_{4K}	BLINK	RWQA	V*	Avg
Coarse	76.1	72.0	57.7	71.4	78.5	71.1
Default	76.3	73.4	57.8	70.7	79.1	71.5
Fine	76.0	72.0	55.6	70.7	80.6	71.0

Table 10: **Ablation on cue supervision layer.** The layer $\ell^\star$ indicates which decoder attention layer is used to construct the TFED evidence target.

$\ell^\star$	CV_{2D}	HR_{4K}	BLINK	RWQA	V*	Avg
18	76.7	73.6	57.2	71.6	78.0	71.4
22	76.3	73.4	57.8	70.7	79.1	71.5
26	76.4	72.5	56.9	71.5	78.5	71.2

different cues to attend to non-identical regions can improve evidence coverage. However, further increasing the weight to 1.00 slightly reduces the average score to 71.5%. This suggests that diversity is beneficial as a mild regularizer, but excessive separation may force cue tokens away from the most relevant evidence regions.

Effect of TFED target granularity. To analyze the granularity of TFED evidence targets, we vary the number of SLIC regions and the smoothing scale in Table 9. The default setting achieves the best average score of 71.5%, outperforming both coarse and fine target constructions. Coarse targets provide broader evidence regions and perform reasonably well, but may include irrelevant context, leading to a lower average score of 71.1%. Fine targets improve V*Bench but reduce BLINK and the overall average, suggesting that overly fragmented evidence targets can make cue alignment less stable. These results indicate that an intermediate granularity provides a better balance between spatial precision and robust cue supervision.

Effect of TFED attention layer. To examine how the choice of attention layer affects evidence discovery, we vary $\ell^\star$ in Table 10. The results are relatively stable across the tested layers, with average scores ranging from 71.2% to 71.5%. The intermediate layer $\ell^\star = 22$ performs best on average, suggesting that TFED benefits from attention maps that are neither too low-level nor too close to final token prediction. Earlier layers may retain stronger local visual responses, while later layers can become more entangled with language-side generation. In contrast, an intermediate layer provides a more suitable trade-off between spatial localization and answer-conditioned semantics.

D More Qualitative Analysis

Figure 5 provides additional visualizations of latent cue attention across diverse evaluation examples. The learned cues adapt to the type of visual evidence required by each question. For counting and text/number recognition, the attention concentrates on the relevant countable or readable regions, such as the windows in the room, the background sign near the airplane, and the speed-limit sign on the road. For spatial and relational questions, the cues attend to the objects involved in the relation, such as the curtain and flowers, the person relative to the car, the green statue and the white statue, or the duck head used to determine facing direction. For attribute and instance-level reasoning, the attention highlights discriminative local details, including the woman's shirt, the elephant trunk, the surfer, the queried leaf, and the person wearing black pants. The examples involving plots and diagrams further show that cue attention can shift from natural-image objects to abstract visual evidence, such as the amplitude of a curve or the affected node in a food web. Overall, these visualizations suggest that the learned cue states provide task-adaptive evidence localization rather than simply attending to globally salient image regions.