Crystallizing Vision When Reasoning Bungees: Latent Visual Grounding for Hallucination Mitigation in MLRMs

Anonymous ACL submission

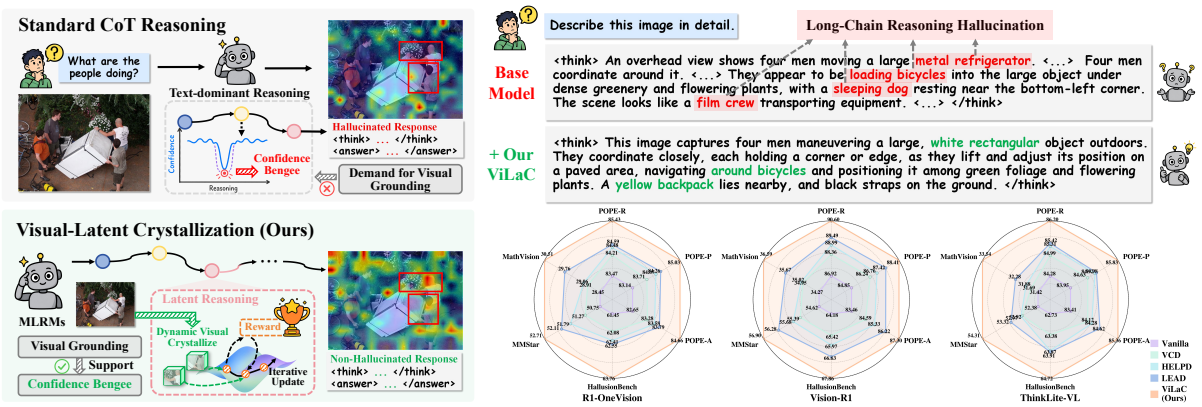


Figure 1: Overview of long-chain reasoning hallucinations and their mitigation via our ViLaC. **Upper Left:** Standard CoT reasoning becomes text-dominant and induces Grounding Bungee. **Upper Right:** ViLaC effectively mitigates long-chain hallucinations. **Bottom Left:** ViLaC crystallizes task-relevant image evidence through latent refinement. **Bottom Right:** ViLaC consistently improves performance across models and benchmarks.

Abstract

Recent multimodal large reasoning models (MLRMs) improve complex visual reasoning with explicit chains, yet still produce hallucinations that are inconsistent with image evidence. Existing hallucination mitigation methods mainly rely on high-cost supplementation of static visual cues to steer an otherwise text-dominant reasoning. Our chain-level analysis reveals a different failure mode, termed **Grounding Bungee**, where hallucinated reasoning trajectories rebound in confidence without corresponding visual support and therefore gradually drift away from image evidence. Based on this observation, we propose **Vi**sual-**La**tent **C**rystallization (**ViLaC**) to ground reasoning in task-relevant image evidence through latent-space exploration, thereby mitigating this bungee-like drift via latent refinement. Specifically, ViLaC introduces optimizable latent think tokens for iterative latent-space exploration, where a confidence-driven reward calibrates the resulting representations for grounded generation. Furthermore, ViLaC dynamically selects visual tokens aligned with evolving latent states, incorporating relevant visual semantics to suppress visually unsupported hallucinations. Extensive experiments show that ViLaC effectively mitigates hallucinations generated by various MLRMs across benchmarks without additional training.

1 Introduction

Multimodal large reasoning models (MLRMs) have advanced complex visual reasoning by generating explicit reasoning chains over image-text inputs (Bai et al., 2025; Qwen Team, 2025; Kimi Team et al., 2025; Yang et al., 2025a). Yet even strong MLRMs can produce linguistically coherent outputs that contradict image evidence, resulting in visual hallucinations (Tong et al., 2024; Guan et al., 2024b; Tian et al., 2026). This problem is especially severe in long-chain reasoning (Tian et al., 2026; Xu et al., 2026a; Lu et al., 2025), where plausible intermediate reasoning can conceal visual errors and make them harder to detect.

Existing attempts to mitigate long-chain hallucination roughly follow two directions. One direction strengthens text-side reasoning through post-training or inference optimization (Yue et al., 2024; Wang et al., 2025b; Havrilla et al., 2024; Shao et al., 2024), while the other supplements visual information during reasoning through external toolkits, visual-token interleaving, and intermediate image generation (Zou et al., 2025; Gao et al., 2025; Zheng et al., 2025; Wang et al., 2025a). These methods turn hallucination mitigation into a costly process of passively supplementing an otherwise text-dominant reasoning chain with external visual cues that are still treated as static inputs. More

importantly, they share an implicit premise: long-chain hallucination stems from persistently insufficient visual information and can therefore be mitigated by adding more external cues.

However, this premise does not fully explain why continuously supplementing external cues still fails to reliably prevent long-chain hallucination. Our analysis reveals a different failure mode, which we term **Grounding Bungee**. Rather than reflecting persistently weak visual support, MLRMs exhibit uneven grounding demand along the reasoning trajectory. Under uncertainty, hallucinated reasoning trajectories often undergo a bungee-like confidence rebound without corresponding visual support, continuing along linguistically plausible yet visually unsupported directions. This explains why passively supplementing static visual cues yields limited returns: once reasoning becomes text-dominant, external visual information no longer serves as an active grounding constraint. The problem is therefore not persistent cue insufficiency, but a bungee-like grounding drift during reasoning.

These observations suggest that mitigating hallucination requires more than supplementing additional visual cues. Recent advances in latent-space reasoning in LLMs offer insight into this shift (Hao et al., 2025; Deng et al., 2024b; Xu et al., 2025b; Li et al., 2025). They show that reasoning can emerge as continuous latent states rather than explicit textual trajectories. Since visual and textual information already reside in a shared semantic space inside MLRMs, task-relevant visual semantics can be reinstated directly as latent evidence.

Motivated by this perspective, we propose **Vi**sual-**La**tent **C**rystallization (**ViLaC**) for hallucination mitigation in MLRMs. Instead of reintroducing visual cues into text-dominant reasoning, ViLaC crystallizes task-relevant image evidence into generation-guiding latent representations to mitigate bungee-like grounding drift. Specifically, ViLaC introduces optimizable latent thought tokens for iterative latent-space exploration, where a confidence-driven reward calibrates latent states before decoding resumes. This enables the model to refine latent states prior to explicit textual generation. ViLaC further dynamically selects visual tokens aligned with the evolving latent state, reinforcing optimized representations with relevant visual semantics and grounding the final generation in image evidence. ViLaC effectively mitigates hallucinations across diverse MLRMs and multimodal benchmarks without additional training.

2 Related Work

Hallucination in MLRMs. Recent MLRMs support long-chain multimodal reasoning, yet their reasoning still largely unfolds as autoregressive text conditioned on an initial visual encoding (Zhang et al., 2024; Lu et al., 2022; Mitra et al., 2024; Xu et al., 2025a; Yang et al., 2025a; Bai et al., 2025; Qwen Team, 2025; Kimi Team et al., 2025). Reasoning-capable multimodal models can still produce responses that sound coherent while departing from image-grounded facts, making hallucination a persistent issue even in long-chain inference and reflective reasoning settings (Tong et al., 2024; Guan et al., 2024b; Wu et al., 2025). Existing remedies typically steer decoding, inject auxiliary visual priors, or retrace evidence after the reasoning trajectory has begun to drift, as in visual contrastive decoding, CLIP-guided decoding, and memory-space retracing (Leng et al., 2024a; Deng et al., 2024a; Zou et al., 2025; Gong et al., 2026). While effective for generic hallucination control, these strategies still rely on externally supplied cues and do not directly address the internal grounding instability of long-chain reasoning.

Latent Space Reasoning. Recent studies increasingly shift from explicit textual reasoning to latent reasoning, where intermediate thought is represented as continuous hidden states rather than fully verbalized tokens or stepwise text sequences (Hao et al., 2025; Deng et al., 2024b; Xu et al., 2025b). This direction aims to preserve richer internal semantics while reducing the overhead of long textual chains and enabling test-time optimization in hidden space (Li et al., 2025; Xu et al., 2025c). In multimodal settings, subsequent work further explores integrating visual information into latent reasoning through hidden-state interleaving, latent token optimization, and uncertainty-aware visual injection, showing that latent-space intervention can support more flexible multimodal reasoning than purely explicit chain generation under complex visual tasks (Pham and Ngo, 2025; Zhang et al., 2026; Gao et al., 2025; Liu et al., 2026).

3 Preliminary and Motivation

3.1 Preliminary

Multimodal large reasoning models (MLRMs) take a visual input $\mathcal{V}$ and a textual query $\mathcal{T}$ to produce a reasoning chain $\mathcal{C}$ and final answer A.

$$\mathcal{F} : (\mathcal{V}, \mathcal{T}) \mapsto (\mathcal{C}, A) \tag{1}$$

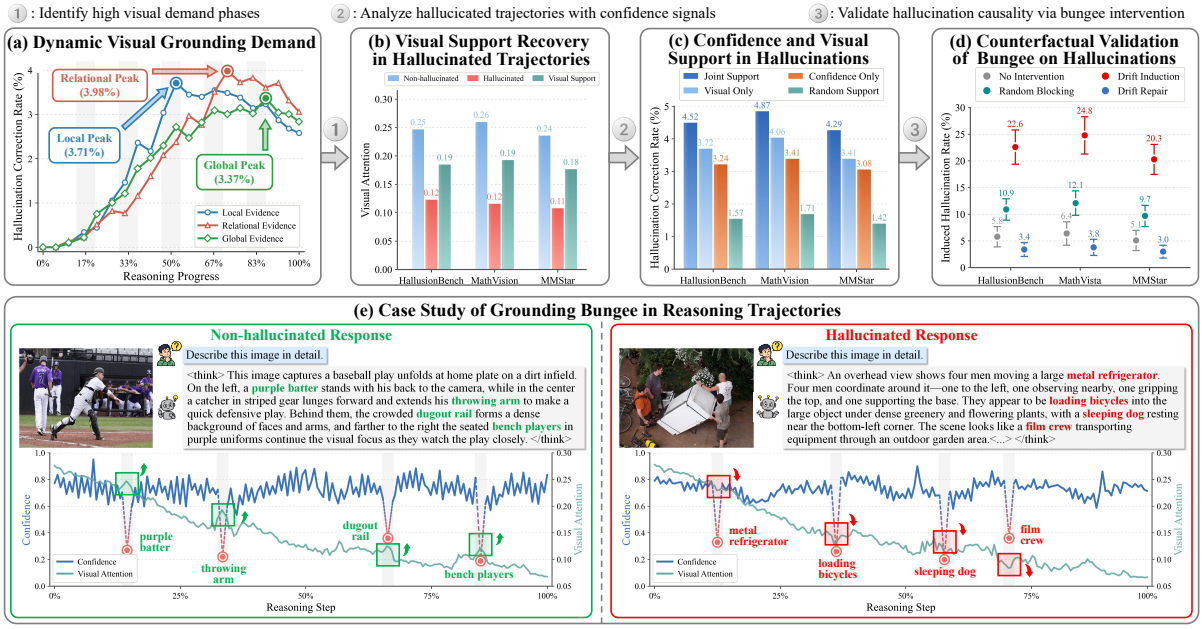


Figure 2: **Motivation analysis of Grounding Bungee. (a)** Visual grounding demand varies across the reasoning trajectory. **(b)** Hallucinated trajectories exhibit weaker attention to visual tokens. **(c)** Joint support combines entropy-based confidence with visual-token attention. **(d)** Counterfactual induction suppresses visual support, while repair restores it. **(e)** Case studies contrast confidence rebound with and without support recovery.

Standard Chain-of-Thought Decoding. Standard MLRMs generate the reasoning chain autoregressively as discretely sampled language tokens:

$$P(\mathcal{C} \mid \mathcal{V}, \mathcal{T}) = \prod_{l} P(c_l \mid c_{<l}, \mathcal{V}, \mathcal{T}) \qquad (2)$$

Each token is predicted conditioned on the multimodal context and prior reasoning, and the final answer is decoded conditioned on $\mathcal{V}, \mathcal{T}$, and $\mathcal{C}$:

$$P(A \mid \mathcal{V}, \mathcal{T}, \mathcal{C}) = \prod_{t} P(a_t \mid a_{<t}, \mathcal{V}, \mathcal{T}, \mathcal{C}) \qquad (3)$$

Accordingly, the final answer is produced after the model has formed an explicit reasoning trajectory $\mathcal{C}$ under the same multimodal inputs.

3.2 Why Long-Chain Reasoning Goes Astray

Recent studies (Tian et al., 2026; Xu et al., 2026a; Lu et al., 2025; Xu et al., 2026b) often attribute multimodal hallucination in long-chain reasoning to a persistent shortage of visual cues, motivating the supplementation of static visual information during text-dominant reasoning. Our chain-level analysis instead reveals a different failure mode, which we term **Grounding Bungee**: grounding demand is uneven along the trajectory, and hallucination emerges when hallucinated reasoning trajectories undergo a bungee-like confidence rebound without corresponding visual support.

Dynamic Demand Leaves Visual Support Unfulfilled. Under a persistent cue-insufficiency account, correction sensitivity should remain roughly uniform along the trajectory. We measure correction sensitivity at normalized position p as:

$$\mathrm{HCR}@p = \frac{N_{\mathrm{corr}}(p)}{N_{\mathrm{hall}}}. \qquad (4)$$

Here, N_{hall} is the number of baseline hallucinated samples, and $N_{\mathrm{corr}}(p)$ the number corrected by intervention at position p. As shown in Figure 2(a), $\mathrm{HCR}@p$ exhibits localized peaks across evidence types, indicating that grounding demand is localized rather than persistent. Figure 2(b) reveals the same mismatch from the support side: non-hallucinated trajectories maintain stronger visual-token attention, while support restoration partially recovers the ratio across benchmarks. Together, these results suggest that hallucination arises from failed support maintenance rather than uniformly insufficient visual information.

A Bungee-like Confidence Rebound Without Visual Support. However, uneven demand alone cannot explain why similar trajectories diverge. Figure 2(e) shows that both responses pass through temporary confidence valleys, indicating that uncertainty itself is not the key anomaly. The divergence emerges afterward: non-hallucinated trajectories recover visual attention together with confidence, whereas hallucinated ones exhibit a bungee-like

confidence rebound while visual support remains weak. Thus, the anomaly is not low confidence itself, but confidence recovery without support restoration. Figure 2(c) supports this view, with Joint Support reaching 4.87% HCR on MathVision (Wang et al., 2024). Hallucination arises when a bungee-like confidence rebounds without support recovery, causing reasoning to drift from image evidence while appearing increasingly certain.

Counterfactual Validation of Grounding Bungee. To move beyond correlation, we conduct counterfactual drift induction and repair tests by modifying visual support under matched conditions. We quantify the effect of intervention policy b using the Induced Hallucination Rate:

$$\mathrm{IHR}(b) = \frac{N_{\mathrm{corr}\to\mathrm{hall}}(b)}{N_{\mathrm{corr}}}. \tag{5}$$

Here, N_{corr} denotes the number of baseline correct samples, and $N_{\mathrm{corr}\to\mathrm{hall}}(b)$ the number that become hallucinated after intervention policy b. As shown in Figure 2(d), drift induction raises IHR to 22.6% on HallusionBench (Guan et al., 2024a), while drift repair reduces it to 3.4%. These results identify Grounding Bungee as a direct driver of hallucination trajectories. Disrupting or restoring visual support changes the final outcome. Additional implementation details are in the appendix.

Takeaway: Grounding Bungee

Hallucinations arise from a bungee-like confidence rebound without visual support under uncertainty as reasoning unfolds.

4 Approach

Motivated by Section 3, we argue that long-chain hallucination does not simply arise from insufficient visual support. Once reasoning becomes text-dominant, external visual cues cease to function as active grounding constraints. Based on this view, ViLaC mitigates bungee-like grounding drift by refining latent think tokens and crystallizing task-relevant visual evidence into latent states.

4.1 Overview

The key challenge is to intervene before unsupported reasoning is committed to explicit text. To this end, ViLaC first performs iterative latent-space exploration over optimizable latent think tokens, then forms grounded latent states from task-relevant visual evidence rather than directly injecting selected visual patches into reasoning.

Latent Think Tokens: We introduce K latent think tokens:

$$\mathcal{H}^{(0)} = [h_1^{(0)}, \ldots, h_K^{(0)}], \quad h_k^{(0)} \in \mathbb{R}^d, \tag{6}$$

where $\mathcal{H}^{(0)}$ denotes their initial latent state. Unlike explicit reasoning tokens $c_l \in \mathcal{C}$, $\mathcal{H}^{(0)}$ is optimized as a continuous latent state before explicit text generation, providing a compact workspace for latent-space exploration before text conversion.

Policy as Latent Space Exploration: We treat latent refinement as local exploration around the current latent state:

$$\tilde{\mathcal{H}}^{(t)} = \mathcal{H}^{(t)} + \epsilon^{(t)}, \quad \epsilon^{(t)} \sim \mathcal{N}(0, \sigma^2 I), \tag{7}$$

which defines a Gaussian exploration policy

$$\pi(\tilde{\mathcal{H}}^{(t)} \mid \mathcal{H}^{(t)}) = \mathcal{N}(\mathcal{H}^{(t)}, \sigma^2 I), \tag{8}$$

This policy probes nearby latent trajectories without committing to explicit text during refinement.

Visual Crystallization: Visual crystallization converts image information into latent visual evidence aligned with evolving latent states. This yields $\mathcal{V}_{\mathrm{lat}}^{(t)}$ from $\mathcal{V}$ conditioned on candidate state $\tilde{\mathcal{H}}^{(t)}$, and forms grounded state $\bar{\mathcal{H}}^{(t)}$ after fusion.

4.2 Confidence-Guided Latent Exploration

Grounding Bungee describes a bungee-like confidence rebound without visual support under uncertainty. Using confidence only after generation cannot regulate this evolving process (Xu et al., 2026b; Zou et al., 2025). We therefore use confidence as an intrinsic optimization signal and refine latent think tokens with a confidence-driven reward.

Confidence-based Reward: Given the grounded latent state $\bar{\mathcal{H}}^{(t)}$, the frozen model produces token distributions $P_i^{(t)}$ for latent positions $i = 1, \ldots, K$. We measure confidence from the concentration of the top-k predictions at each latent position:

$$\mathcal{C}_k\left(P_i^{(t)}\right) = -\sum_{u \in \mathrm{Top}_k(P_i^{(t)})} P_i^{(t)}(u) \log P_i^{(t)}(u), \tag{9}$$

where k denotes the number of top predictions used for confidence estimation. The reward for the current latent state is then defined as

$$\mathcal{R}\left(\bar{\mathcal{H}}^{(t)}\right) = 1 - \frac{1}{K}\sum_{i=1}^{K} \mathcal{C}_k\left(P_i^{(t)}\right), \tag{10}$$

Thus, higher reward corresponds to more concentrated top-k predictions and higher confidence on the grounded state rather than the raw trajectory.

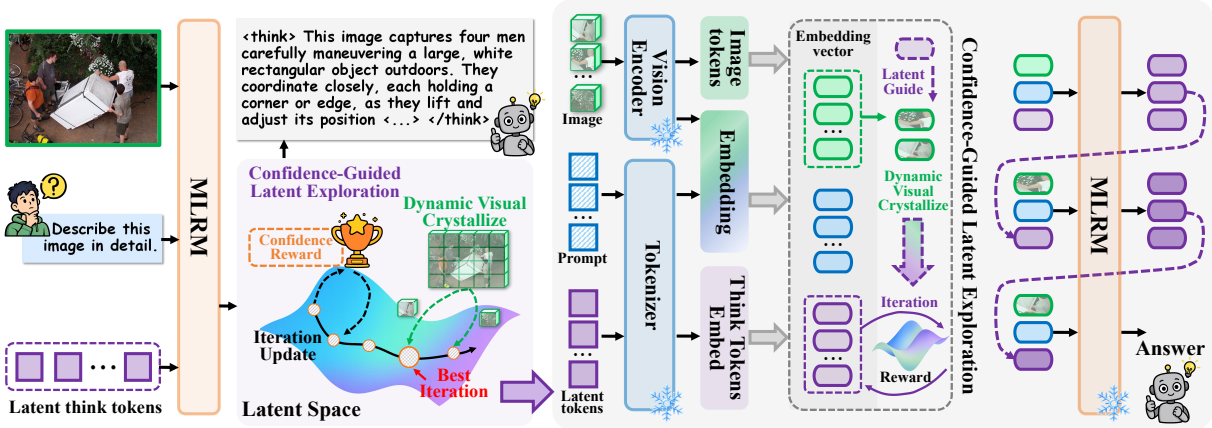


Figure 3: **Overview of our ViLaC method. Left:** Confidence-Guided Latent Exploration iteratively refines latent think tokens, while Dynamic Visual Crystallization fuses candidate latent states with task-relevant visual evidence to form grounded latent states. **Right:** Illustration of the iterative latent-space reasoning process, where latent exploration and visual evidence progressively re-ground the reasoning process toward the final prediction.

Latent Think Update with Policy Gradient: We therefore update the latent think state toward nearby trajectories with higher reward. Formally, we seek a latent state that maximizes the expected reward under the exploration policy:

$$\mathcal{J}(\mathcal{H}) = \mathbb{E}_{\tilde{\mathcal{H}} \sim \pi(\cdot|\mathcal{H})} \left[\mathcal{R}(\bar{\mathcal{H}}) \right], \tag{11}$$

Using REINFORCE (Williams, 1992), we estimate the gradient of this objective with respect to the current latent state

$$\nabla_{\mathcal{H}} \mathcal{J}(\mathcal{H}) = \mathbb{E} \left[\mathcal{R}(\bar{\mathcal{H}}) \, \nabla_{\mathcal{H}} \log \pi(\tilde{\mathcal{H}} \mid \mathcal{H}) \right], \tag{12}$$

Under the Gaussian policy in Equation 8, the gradient term reduces to

$$\nabla_{\mathcal{H}} \log \pi(\tilde{\mathcal{H}}^{(t)} \mid \mathcal{H}^{(t)}) = \frac{\epsilon^{(t)}}{\sigma^2}, \tag{13}$$

which yields the update

$$\mathcal{H}^{(t+1)} = \mathcal{H}^{(t)} + \eta \, \mathcal{R}\left(\bar{\mathcal{H}}^{(t)}\right) \frac{\epsilon^{(t)}}{\sigma^2}, \tag{14}$$

In this way, latent exploration steers the current reasoning state toward nearby latent trajectories that remain more confident after visual grounding.

4.3 Dynamic Visual Crystallization

Confidence-guided exploration alone cannot guarantee grounded reasoning. Instead of directly injecting selected visual patches into reasoning (Zhang et al., 2026; Liu et al., 2026), ViLaC selects visual tokens aligned with the evolving latent state to crystallize latent visual evidence. The resulting grounded latent state $\bar{\mathcal{H}}^{(t)}$ is then used for reward evaluation, enabling confidence assessment conditioned on task-relevant visual evidence.

Algorithm 1: ***Visual-Latent Crystallization***

Input: visual tokens $\mathcal{V}$, text tokens $\mathcal{T}$, latent think token count K, crystallization token count R, optimization steps N_{opt}
Output: final answer A

Encode visual-text context and build visual anchor $\mathcal{V}_{\text{pad}}$
Initialize $\mathcal{H}^{(0)}$ from the current hidden state

Confidence-Guided Latent Exploration

for $t = 0$ to $N_{\text{opt}} - 1$ **do**
  $\epsilon^{(t)} \sim \mathcal{N}(0, \sigma^2 I)$
  $\tilde{\mathcal{H}}^{(t)} \leftarrow \mathcal{H}^{(t)} + \epsilon^{(t)}$ ▷ *latent candidate*

  Dynamic Visual Crystallization

  for $i = 1$ to K **do**
    Compute $s_{ij}^{(t)}$ for all $v_j \in \mathcal{V}$
    $\mathcal{N}_i^{(t)} \leftarrow \text{TopK}_j(s_{ij}^{(t)}, R)$ ▷ *aligned tokens*
  end for
  $\mathcal{V}_{\text{ctx}}^{(t)} \leftarrow \text{SoftAggregate}\left(\{\mathcal{N}_i^{(t)}\}_{i=1}^{K}\right)$
  $\mathcal{V}_{\text{lat}}^{(t)} \leftarrow \mathcal{V}_{\text{pad}} + \mathcal{V}_{\text{ctx}}^{(t)}$ ▷ *latent visual evidence*
  $\bar{\mathcal{H}}^{(t)} \leftarrow \lambda \tilde{\mathcal{H}}^{(t)} + (1-\lambda)\mathcal{V}_{\text{lat}}^{(t)}$

  $r^{(t)} \leftarrow \mathcal{R}(\bar{\mathcal{H}}^{(t)})$ ▷ *reward*
  $\mathcal{H}^{(t+1)} \leftarrow \mathcal{H}^{(t)} + \eta \, r^{(t)} \, \epsilon^{(t)} / \sigma^2$ ▷ *latent update*
end for

Decode from $\mathcal{H}^{(N_{\text{opt}})}$ and output A

Let $\tilde{\mathcal{H}}^{(t)} = \{\tilde{h}_i^{(t)}\}_{i=1}^{K}$ denote the latent think state at iteration t, and $\mathcal{V} = \{v_j\}_{j=1}^{N}$ the image token embeddings. We construct a base visual anchor $\mathcal{V}_{\text{pad}}$ from visual special-token embeddings (e.g., <|vision_start|>). For each latent think token, cosine alignment is computed to select the top-R matched image tokens:

$$s_{ij}^{(t)} = \frac{\left(\tilde{h}_i^{(t)}\right)^{\top} v_j}{\|\tilde{h}_i^{(t)}\| \, \|v_j\|}, \qquad \mathcal{N}_i^{(t)} = \text{TopK}_j\left(s_{ij}^{(t)}, R\right), \tag{15}$$

Category	Method	R1-OneVision (Yang et al., 2025b)				Vision-R1 (Huang et al., 2026)				ThinkLite-VL (Wang et al., 2025c)			
		Acc↑	Δ	F1↑	Δ	Acc↑	Δ	F1↑	Δ	Acc↑	Δ	F1↑	Δ
Random	Vanilla	83.47	+0.00	81.92	+0.00	86.92	+0.00	81.11	+0.00	84.28	+0.00	84.04	+0.00
	+VCD (Leng et al., 2024b)	84.21	+0.74	82.66	+0.74	88.36	+1.44	81.74	+0.63	84.99	+0.71	84.80	+0.76
	+DAMRO (Gong et al., 2024)	84.40	+0.93	82.97	+1.05	88.85	+1.93	82.04	+0.93	85.19	+0.91	85.03	+0.99
	+HELPD (Yuan et al., 2024)	84.48	+1.01	83.05	+1.13	88.99	+2.07	82.21	+1.10	85.31	+1.03	85.18	+1.14
	+EVAS (Zhang et al., 2025)	84.55	+1.08	83.12	+1.20	89.11	+2.19	82.29	+1.18	85.37	+1.09	85.26	+1.22
	+MemVR (Zou et al., 2025)	84.62	+1.15	83.28	+1.36	89.34	+2.42	82.46	+1.35	85.46	+1.18	85.39	+1.35
	+DMLR (Liu et al., 2026)	84.51	+1.04	83.24	+1.32	89.28	+2.36	82.51	+1.40	85.34	+1.06	85.44	+1.40
	+LEAD (Xu et al., 2026b)	84.59	+1.12	83.19	+1.27	89.49	+2.57	82.38	+1.27	85.42	+1.14	85.31	+1.27
	+Ours	**85.43**	**+1.96**	**83.96**	**+2.04**	**90.60**	**+3.68**	**83.21**	**+2.10**	**86.20**	**+1.92**	**86.07**	**+2.03**
Popular	Vanilla	83.14	+0.00	81.66	+0.00	84.85	+0.00	80.63	+0.00	83.95	+0.00	83.72	+0.00
	+VCD (Leng et al., 2024b)	83.71	+0.57	82.44	+0.78	86.24	+1.39	81.29	+0.66	84.63	+0.68	84.39	+0.67
	+DAMRO (Gong et al., 2024)	84.05	+0.91	82.61	+0.95	86.58	+1.73	81.62	+0.99	84.82	+0.87	84.76	+1.04
	+HELPD (Yuan et al., 2024)	84.07	+0.93	82.85	+1.19	86.76	+1.91	81.81	+1.18	84.92	+0.97	84.88	+1.16
	+EVAS (Zhang et al., 2025)	84.13	+0.99	82.92	+1.26	86.88	+2.03	81.79	+1.16	85.02	+1.07	84.86	+1.14
	+MemVR (Zou et al., 2025)	84.20	+1.06	83.05	+1.39	87.09	+2.24	82.03	+1.40	85.11	+1.16	84.97	+1.25
	+DMLR (Liu et al., 2026)	84.27	+1.13	82.83	+1.17	87.11	+2.26	82.07	+1.44	84.97	+1.02	85.11	+1.39
	+LEAD (Xu et al., 2026b)	84.26	+1.12	82.87	+1.21	87.42	+2.57	81.93	+1.30	84.98	+1.03	85.01	+1.29
	+Ours	**85.03**	**+1.89**	**83.73**	**+2.07**	**88.41**	**+3.56**	**82.82**	**+2.19**	**85.83**	**+1.88**	**85.72**	**+2.00**
Adversarial	Vanilla	82.65	+0.00	81.18	+0.00	83.46	+0.00	80.25	+0.00	83.41	+0.00	83.19	+0.00
	+VCD (Leng et al., 2024b)	83.28	+0.63	82.06	+0.88	84.59	+1.13	81.04	+0.79	84.12	+0.71	83.89	+0.70
	+DAMRO (Gong et al., 2024)	83.44	+0.79	82.24	+1.06	85.17	+1.71	81.16	+0.91	84.27	+0.86	84.26	+1.07
	+HELPD (Yuan et al., 2024)	83.58	+0.93	82.19	+1.01	85.33	+1.87	81.29	+1.04	84.28	+0.87	84.36	+1.17
	+EVAS (Zhang et al., 2025)	83.66	+1.01	82.31	+1.13	85.36	+1.90	81.47	+1.22	84.39	+0.98	84.46	+1.27
	+MemVR (Zou et al., 2025)	83.73	+1.08	82.43	+1.25	85.59	+2.13	81.64	+1.39	84.49	+1.08	84.61	+1.42
	+DMLR (Liu et al., 2026)	83.68	+1.03	82.53	+1.35	85.79	+2.33	81.67	+1.42	84.61	+1.20	84.39	+1.20
	+LEAD (Xu et al., 2026b)	83.79	+1.14	82.63	+1.45	86.22	+2.76	81.55	+1.30	84.62	+1.21	84.44	+1.25
	+Ours	**84.66**	**+2.01**	**83.34**	**+2.16**	**87.30**	**+3.84**	**82.49**	**+2.24**	**85.36**	**+1.95**	**85.24**	**+2.05**

Table 1: **Hallucination Mitigation on POPE benchmark.** Accuracy and F1-score (%) comparison under different evaluation settings (i.e., Random, Popular and Adversarial). Δ denotes improvement over vanilla baseline.

where R denotes the number of visual tokens retained for crystallization at each latent position. Soft aggregation of selected tokens yields the visual sequence $\mathcal{V}_{\text{ctx}}^{(t)}$, which combines with visual anchor $\mathcal{V}_{\text{pad}}$ to form crystallized latent evidence:

$$\mathcal{V}_{\text{lat}}^{(t)} = \mathcal{V}_{\text{pad}} + \mathcal{V}_{\text{ctx}}^{(t)}, \tag{16}$$

We then combine the candidate latent state with the crystallized visual evidence:

$$\bar{\mathcal{H}}^{(t)} = \lambda\, \tilde{\mathcal{H}}^{(t)} + (1-\lambda)\mathcal{V}_{\text{lat}}^{(t)}, \tag{17}$$

where λ balances latent reasoning with visual grounding. The crystallized visual evidence is norm-matched to the base visual anchor for stability, and the grounded latent state $\bar{\mathcal{H}}^{(t)}$ is then scored by Equation 10. In this way, relevant visual semantics are crystallized around the evolving latent state and reinforce the optimized representations with task-relevant visual evidence.

5 Experiments

In this section, we evaluate ViLaC through extensive comparisons across multiple MLRMs and benchmarks. We cover both hallucination mitigation and broader understanding settings, and further present ablations on key modules, hyperparameters and inference latency.

5.1 Experimental Settings

Benchmarks: We evaluate ViLaC on two types of benchmarks. For hallucination mitigation, we use POPE (Li et al., 2023) and HallusionBench (Guan et al., 2024a). The former evaluates object-level hallucination under random, popular, and adversarial settings, while the latter measures image-grounded inconsistency. For understanding and reasoning, we use MMStar (Chen et al., 2024) and MathVision (Wang et al., 2024), which cover general and science-specific reasoning tasks.

Models: Our experiments are conducted on three representative MLRMs, covering different post-training paradigms for multimodal reasoning. R1-OneVision (Yang et al., 2025b) and Vision-R1 (Huang et al., 2026) both adopt a two-stage pipeline that first performs supervised fine-tuning and then reinforcement learning, while ThinkLite-VL (Wang et al., 2025c) applies rule-based reinforcement learning directly at scale for reasoning.

Implementation Details: We use 8 latent think tokens and set the number of selected visual candidates to top-k=16 at each iteration. Unless otherwise specified, the number of optimization iterations is 15 with a learning rate of 10^{-3}. To ensure stable exploration in latent space, the perturbation magnitude σ is set to 10%. All experiments are conducted on NVIDIA RTX PRO 6000 GPUs, and further details are provided in the appendix.

Method	R1-OneVision (Yang et al., 2025b)				Vision-R1 (Huang et al., 2026)				ThinkLite-VL (Wang et al., 2025c)			
	Acc↑	Δ	F1↑	Δ	Acc↑	Δ	F1↑	Δ	Acc↑	Δ	F1↑	Δ
Vanilla	61.45	+0.00	58.92	+0.00	64.18	+0.00	61.07	+0.00	62.73	+0.00	60.84	+0.00
+VCD (Leng et al., 2024b)	62.08	+0.63	59.73	+0.81	65.42	+1.24	62.07	+1.00	63.38	+0.65	61.61	+0.77
+DAMRO (Gong et al., 2024)	62.40	+0.95	60.01	+1.09	65.86	+1.68	62.63	+1.56	63.71	+0.98	61.88	+1.04
+HELPD (Yuan et al., 2024)	62.41	+0.96	60.02	+1.10	65.97	+1.79	62.84	+1.77	63.87	+1.14	61.91	+1.07
+EVAS (Zhang et al., 2025)	62.33	+0.88	59.88	+0.96	66.16	+1.98	62.89	+1.82	63.86	+1.13	62.06	+1.22
+MemVR (Zou et al., 2025)	62.51	+1.06	60.22	+1.30	66.39	+2.21	63.18	+2.11	63.98	+1.25	62.24	+1.40
+DMLR (Liu et al., 2026)	62.44	+0.99	60.31	+1.39	66.74	+2.56	63.37	+2.30	64.05	+1.32	62.09	+1.25
+LEAD (Xu et al., 2026b)	62.55	+1.10	60.08	+1.16	66.63	+2.45	63.11	+2.04	63.91	+1.18	62.17	+1.33
+Ours	**63.76**	**+2.31**	**61.34**	**+2.42**	**67.86**	**+3.68**	**63.54**	**+2.47**	**64.72**	**+1.99**	**63.05**	**+2.21**

Table 2: **Hallucination Mitigation on HallusionBench benchmark.** Accuracy and F1-score (%) comparison on HallusionBench across three MLRMs. Δ denotes improvement over the vanilla baseline.

Method	R1-OneVision (Yang et al., 2025b)				Vision-R1 (Huang et al., 2026)				ThinkLite-VL (Wang et al., 2025c)			
	MMStar		MathVision		MMStar		MathVision		MMStar		MathVision	
	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
Vanilla	50.75	47.92	28.45	26.31	54.62	51.44	34.27	31.88	52.38	49.16	31.42	29.07
+VCD (Leng et al., 2024b)	51.27	48.44	28.91	26.84	55.39	52.31	34.95	32.39	52.92	49.61	31.69	29.47
+DAMRO (Gong et al., 2024)	51.68	48.98	29.27	27.15	55.88	52.77	35.56	32.89	53.24	49.94	31.91	29.79
+HELPD (Yuan et al., 2024)	51.79	48.85	29.06	27.27	55.68	52.38	35.02	32.79	52.97	50.16	31.88	29.58
+EVAS (Zhang et al., 2025)	51.59	49.07	29.31	27.24	56.11	52.74	35.49	33.16	53.17	50.27	32.15	29.81
+MemVR (Zou et al., 2025)	51.98	49.41	29.64	27.60	56.34	53.26	35.95	33.28	53.46	50.63	32.43	30.05
+DMLR (Liu et al., 2026)	51.93	49.18	29.95	27.73	56.48	53.36	35.62	33.09	53.29	50.44	32.77	30.38
+LEAD (Xu et al., 2026b)	52.11	49.16	29.76	27.69	56.28	52.88	35.67	33.29	53.32	50.46	32.28	29.90
+Ours	**52.71**	**50.07**	**30.51**	**28.49**	**56.90**	**53.70**	**36.59**	**34.10**	**54.31**	**51.22**	**33.54**	**31.14**
Improvement	**+1.96**	**+2.15**	**+2.06**	**+2.18**	**+2.28**	**+2.26**	**+2.32**	**+2.22**	**+1.93**	**+2.06**	**+2.12**	**+2.07**

Table 3: **Performance of Understanding and Reasoning on MMStar and MathVision.** Results (%) on general and science-specific understanding and reasoning benchmarks across three MLRMs.

5.2 Main Results

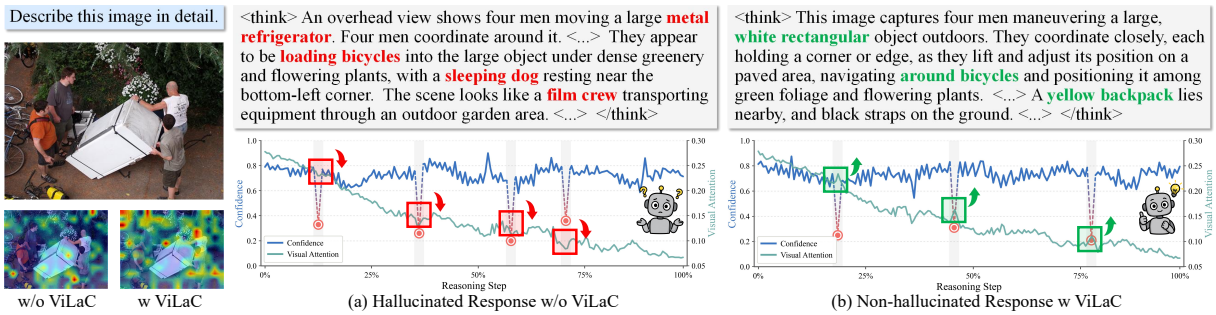
Performance of Hallucination Mitigation: We evaluate ViLaC on two hallucination benchmarks, POPE (Li et al., 2023) and HallusionBench (Guan et al., 2024a). As shown in Table 1, ViLaC consistently achieves the best results across all three backbones and all POPE settings, surpassing LEAD (Xu et al., 2026b) by up to 1.08% in accuracy on Vision-R1 (Huang et al., 2026) under the adversarial setting. Table 2 further shows that ViLaC generalizes to HallusionBench (Guan et al., 2024a), where it also remains consistently better than competing methods, reaching 63.76% accuracy on R1-OneVision (Yang et al., 2025b) and 67.86% on Vision-R1 (Huang et al., 2026). Compared with latent-based DMLR (Liu et al., 2026), ViLaC achieves more consistent gains on hallucination benchmarks. These results indicate that ViLaC effectively mitigates multimodal hallucination across diverse models and settings.

Performance of Understanding and Reasoning: Beyond hallucination-specific benchmarks, we further evaluate ViLaC on MMStar (Chen et al., 2024) and MathVision (Wang et al., 2024) to assess broader understanding and reasoning performance. As shown in Table 3, ViLaC consistently achieves the best results across all three backbones on both benchmarks, with gains of up to 2.28% on Vision-R1 (Huang et al., 2026) for MMStar accuracy and 2.22% for MathVision F1. It also remains stronger than LEAD (Xu et al., 2026b), with margins reaching 1.24% on ThinkLite-VL (Wang et al., 2025c) for MathVision (Wang et al., 2024) F1. It also remains competitive with latent-based DMLR (Liu et al., 2026) on broader reasoning benchmarks.

5.3 Ablation Studies and Analysis

Ablation Study on Key Modules: Table 5 reports the ablation results of the two core modules in ViLaC. Either module alone improves over the vanilla model on both hallucination metrics. On HallusionBench (Guan et al., 2024a), combining both modules improves accuracy from 61.45 to 63.76, showing that confidence-guided latent exploration is most effective when coupled with dynamic visual crystallization rather than used as an isolated confidence-seeking signal.

Ablation Study on Key hyperparameters: Table 4 reports the effect of three key hyperparameters in ViLaC: latent think token count K, retained visual token count R, and fusion weight λ. Across all three backbones, moderate settings consistently achieve the best overall performance. For example, with K=8, Vision-R1 reaches the best HallusionBench (Guan et al., 2024a) accuracy of 67.86, outperforming both smaller and larger settings. This



(a) Hallucinated Response w/o ViLaC (b) Non-hallucinated Response w ViLaC

Figure 4: Visualization of hallucination mitigation. Compared with the hallucinated response without ViLaC, ViLaC yields stronger visual attention near low-confidence steps and more grounded responses.

Model	K					R					λ				
	Val.	POPE-A		HallusionBench		Val.	POPE-A		HallusionBench		Val.	POPE-A		HallusionBench	
		Acc↑	F1↑	Acc↑	F1↑		Acc↑	F1↑	Acc↑	F1↑		Acc↑	F1↑	Acc↑	F1↑
R1-OneVision	6	84.12	82.88	63.18	60.82	12	84.01	82.76	63.05	60.66	0.65	84.09	82.84	63.11	60.72
	8	**84.66**	**83.34**	**63.76**	**61.34**	16	**84.66**	**83.34**	**63.76**	**61.34**	0.70	**84.66**	**83.34**	**63.76**	**61.34**
	10	84.29	83.01	63.41	61.05	20	84.24	82.94	63.29	60.97	0.78	84.18	82.91	63.22	60.88
Vision-R1	6	86.41	81.86	66.98	62.85	12	86.32	81.75	66.82	62.73	0.67	86.37	81.83	66.91	62.81
	8	**87.30**	**82.49**	**67.86**	**63.54**	16	**87.30**	**82.49**	**67.86**	**63.54**	0.73	**87.30**	**82.49**	**67.86**	**63.54**
	10	86.88	82.16	67.31	63.05	20	86.79	82.08	67.22	62.97	0.79	86.73	81.99	67.14	62.89
ThinkLite-VL	6	84.02	83.87	62.97	60.52	12	83.95	83.79	62.84	60.41	0.66	84.00	83.83	62.91	60.47
	8	**85.36**	**84.49**	**64.72**	**62.31**	16	**85.36**	**84.49**	**64.72**	**62.31**	0.75	**85.36**	**84.49**	**64.72**	**62.31**
	10	84.63	84.04	63.58	61.09	20	84.51	83.96	63.44	60.98	0.82	84.46	83.91	63.36	60.91

Table 4: Ablation study on key hyperparameters of ViLaC.

Latent Explore	Visual Crystal	POPE-A		HallusionBench	
		Acc↑	F1↑	Acc↑	F1↑
		82.65	81.18	61.45	58.92
✓		83.74	82.29	62.54	60.06
	✓	83.52	82.04	62.31	59.83
✓	✓	**84.66**	**83.34**	**63.76**	**61.34**

Table 5: Ablation study on the key modules of ViLaC.

Method	R1-OneVision		Vision-R1		ThinkLite-VL	
	Acc↑	Time↓	Acc↑	Time↓	Acc↑	Time↓
Vanilla	61.45	14:54	64.18	16:24	62.73	14:06
+VCD	62.08	32:18	65.42	35:48	63.38	30:36
+DAMRO	62.40	22:32	65.86	23:18	63.71	21:48
+HELPD	62.41	24:11	65.97	25:42	63.87	23:00
+LEAD	62.55	18:53	66.63	19:42	63.91	18:12
+Ours	**63.76**	**17:48**	**67.86**	**18:36**	**64.72**	**17:21**

Table 6: Inference latency analysis of ViLaC.

indicates that ViLaC remains stable within a reasonable hyperparameter range.

Analysis of Inference Latency: Table 6 compares latency and hallucination performance on HallusionBench (Guan et al., 2024a). Across all three backbones, ViLaC achieves the best accuracy while remaining more efficient than prior mitigation methods. This suggests that the gains of ViLaC do not come from disproportionately increasing decoding cost, but from more effective intervention during reasoning. On Vision-R1 (Huang et al., 2026), ViLaC improves accuracy to 67.86% with 18:36 inference time, compared with 19:42 for LEAD (Xu et al., 2026b) and 35:48 for VCD (Leng et al., 2024b), showing a favorable balance between hallucination mitigation and inference efficiency.

Visualization of Hallucination Mitigation: We visualize confidence and visual attention during reasoning in Figure 4. Without ViLaC, hallucinated response introduces unsupported concepts such as *metal refrigerator*, while visual attention continues to decay after confidence drops. In contrast, with ViLaC, non-hallucinated response exhibits stronger visual attention near low-confidence steps, leading to more grounded descriptions. It suggests that confidence recovery is grounded in restored visual support rather than detached from it.

6 Conclusion

In this paper, we study long-chain multimodal hallucination from a chain-level perspective and identify **Grounding Bungee** as a key failure mode in MLRMs. Unlike existing methods that mainly rely on supplementing static visual cues for an otherwise text-dominant reasoning process, ViLaC mitigates hallucination by crystallizing task-relevant image evidence through latent-space exploration. Specifically, ViLaC introduces optimizable latent think tokens for confidence-guided latent refinement and dynamically selects visual tokens aligned with evolving latent states, enabling more grounded generation under uncertainty. Extensive experiments across diverse MLRMs and multiple benchmarks demonstrate that ViLaC effectively reduces hallucination without additional training.

Limitations

Our study focuses on hallucination mitigation in multimodal long-chain reasoning, especially cases where reasoning departs from image evidence under uncertainty. We do not yet evaluate whether the same mechanism transfers equally well to broader multimodal generation scenarios with substantially different interaction patterns, such as open-ended dialogue or video understanding. In addition, our experiments are conducted on three representative MLRMs and several standard benchmarks. Although the results remain consistent across these settings, broader validation across more architectures, reasoning tasks, and multimodal evaluation domains would further establish the generality of Grounding Bungee and the robustness of ViLaC.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yiheng Xu, Jiabo Ye, Xi Zhang, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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

Ailin Deng, Zhirui Chen, and Bryan Hooi. 2024a. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. In *ICLR Workshop*.

Yuntian Deng, Yejin Choi, and Stuart Shieber. 2024b. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. 2025. Interleaved-modal chain-of-thought. In *CVPR*.

Boyang Gong, Yu Zheng, Fanye Kong, Jie Zhou, and Jiwen Lu. 2026. Attention at rest stays at rest: Breaking visual inertia for cognitive hallucination mitigation. *arXiv preprint arXiv:2604.01989*.

Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. 2024. Damro: Dive into the attention mechanism of lvlm to reduce object hallucination. In *EMNLP*, pages 7696–7712.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024a. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, pages 14375–14385.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024b. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, pages 14375–14385.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2025. Training large language models to reason in a continuous latent space. In *COLM*.

Alex Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. 2024. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. 2026. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *ICLR*.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024a. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, pages 13872–13882.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024b. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, pages 13872–13882.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *EMNLP*.

Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. 2026. Reasoning within the mind: Dynamic multimodal interleaving in latent space. In *CVPR findings*.

Haolang Lu, Bolun Chu, WeiYe Fu, Guoshun Nan, Junning Liu, Minghui Pan, Qiankun Li, Yi Yu, Hua Wang, and Kun Wang. 2025. Reallocating attention across layers to reduce multimodal hallucination. *arXiv preprint arXiv:2510.10285*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, pages 2507–2521.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *CVPR*, pages 14420–14431.

Tan-Hanh Pham and Chris Ngo. 2025. Multimodal chain of continuous thought for latent-space reasoning in vision-language models. *arXiv preprint arXiv:2508.12587*.

Qwen Team. 2025. Qwen3-vl: Sharper vision, deeper thought, broader action.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Tu, and Jing Zhang. 2026. More thought, less accuracy? on the dual nature of reasoning in vision-language models. In *ICLR*.

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

Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. 2025a. Reconstructive visual instruction tuning. In *ICLR*, volume 2025, pages 14374–14399.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. 2025b. Vggt: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *NeurIPS*, 37:95095–95169.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025c. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *NeurIPS*, 38:118818–118850.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. 2025. Combating multimodal llm hallucination via bottom-up holistic reasoning. In *AAAI*, pages 8460–8468.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025a. Llava-cot: Let vision language models reason step-by-step. In *ICCV*, pages 2087–2098.

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

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025c. Softcot++: Test-time scaling with soft chain-of-thought reasoning. *arXiv preprint arXiv:2505.11484*.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Zou, Xin Wang, Yuyin Zhou, and Sheng Liu. 2026a. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. In *NeurIPS*, volume 38, pages 82878–82905.

Zhongxing Xu, Zhonghua Wang, Zhe Qian, Dachuan Shi, Feilong Tang, Ming Hu, Shiyan Su, Xiaocheng Zou, Wei Feng, Dwarikanath Mahapatra, and 1 others. 2026b. Thinking in uncertainty: Mitigating hallucinations in mlrms with latent entropy-aware decoding. In *CVPR*.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025a. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, and 1 others. 2025b. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *CVPR*, pages 2376–2385.

Fan Yuan, Chi Qin, Xiaogang Xu, and Piji Li. 2024. Helpd: Mitigating hallucination of lvlms by hierarchical feedback learning with vision-enhanced penalty decoding. In *EMNLP*, pages 1768–1785.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mammoth: Building math generalist models through hybrid instruction tuning. In *ICLR*, volume 2024, pages 40320–40341.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. 2025. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in lvlms. In *EMNLP*, pages 3512–3534.

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*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024.

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

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and 1 others. 2025. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. In *ICML*.

Appendix Overview

This supplementary material provides prompt templates used in our experiments. The contents are organized as follows:

- **Prompt Templates for Datasets (§A):** Prompt formats used for POPE, HallusionBench, MMStar, and MathVision.
- **Details for Motivation Analyses (§B):** Additional definitions, intervention settings, and statistics for the motivation analyses in Section 3.
- **More Experimental Results and Analysis (§C):** Additional analyses of radar-chart trends, heatmap comparison, trajectory visualization, ablation behavior, and latency trade-offs.
- **Case Studies Across Different Datasets (§D):** Qualitative observations across POPE, HallusionBench, MMStar, and MathVision.

A Prompt Templates for Datasets

A.1 Prompt Templates for POPE Dataset

The POPE benchmark (Li et al., 2023) evaluates object-level hallucination through image-grounded binary judgments. In our setting, each POPE sample is formatted as a visual question together with two candidate answers corresponding to *yes* and *no*, so that all models follow a consistent answer space during evaluation. The input format is shown below, where the prompt is highlighted in **green** and the image placeholder is highlighted in **red**.

> A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
>
> **USER: IMAGE**
>
> **Is there a snowboard in the image?**
> **Options: A. yes B. no**
>
> **ASSISTANT:**

A.2 Prompt Templates for HallusionBench Dataset

HallusionBench (Guan et al., 2024a) evaluates whether the model can answer visually grounded

questions without being misled by image-grounded inconsistencies or visual illusions. In our experiments, we preserve the original question for each sample and directly query the model with the associated image. The prompt follows a unified visual question answering format rather than separate binary and multiple-choice subtemplates. An illustrative template is shown below:

> A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
>
> **USER: IMAGE**
>
> **Please answer the question according to the image.**
> **Question: Which object would actually appear at the marked position in the mirror?**
>
> **ASSISTANT:**

A.3 Prompt Templates for MMStar Dataset

MMStar (Chen et al., 2024) is a multimodal reasoning benchmark composed primarily of multiple-choice questions. For this benchmark, we concatenate the original question with its candidate options and ask the model to reason over the image before choosing the final answer. The prompt template is shown below:

> A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
>
> **USER: IMAGE**
>
> **Which option describes the object relationship in the image correctly?**
> **Options: A. The suitcase is on the book.**
> **B. The suitcase is beneath the cat.**
> **C. The suitcase is beneath the bed.**
> **D. The suitcase is beneath the book.**
>
> **Please think step by step and then provide your answer.**
>
> **ASSISTANT:**

A.4 Prompt Templates for MathVision Dataset

MathVision (Wang et al., 2024) evaluates science-oriented multimodal mathematical reasoning. In our experiments, we preserve the original problem statement and append a reasoning-oriented instruction so that the model produces explicit long-chain reasoning before giving the final answer. An illustrative template is shown below:

> A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
>
> **USER: IMAGE**
>
> **Solve the problem and output the answer in the format of \boxed{your answer}.**
> **Question: Bettina chooses five points A, B, C, D and E on a circle and draws the tangent to the circle at point A. She realizes that the five angles marked x are all equally big. How big is the angle $\angle ABD$?**
> **Options: A. $66°$, B. $70.5°$, C. $72°$, D. $75°$, E. $77.5°$**
>
> **Please think step by step and provide your detailed solution.**
>
> **ASSISTANT:**

B Details for Motivation Analyses

This section supplements the motivation analyses in Section 3 and Figure 2. We provide additional details for the correction-based grounding-demand analysis, the visual-support measurements, the counterfactual induction and repair settings, and the statistics behind each subplot in Figure 2.

B.1 Correction Sensitivity and the Definition of HCR@p

To examine whether grounding demand is uniformly distributed along the reasoning chain, we measure how sensitive hallucinated trajectories are to intervention at different positions. For a normalized reasoning position $p \in [0, 1]$, the hallucination correction rate is defined as

$$\mathrm{HCR}@p = \frac{N_{\mathrm{corr}}(p)}{N_{\mathrm{hall}}}, \tag{18}$$

where N_{hall} denotes the number of baseline hallucinated samples and $N_{\text{corr}}(p)$ denotes the number of those samples whose final prediction becomes correct after intervention at position p.

To compare trajectories of different lengths, we first normalize each reasoning chain to the unit interval and divide it into a fixed number of relative positions. For each position bin, we apply the same intervention policy to the corresponding reasoning step and record whether the final answer changes from hallucinated to correct. The resulting $\text{HCR}@p$ curve measures *position-dependent correction sensitivity* rather than average task performance. As shown in Figure 2(a), the curve exhibits clear localized peaks, indicating that the need for grounding is concentrated at a few critical steps rather than spread uniformly across the whole reasoning process.

B.2 Visual-Token Attention Ratio and Joint Support

To analyze how visual grounding differs between successful and hallucinated trajectories, Figure 2(b) compares the amount of attention assigned to visual tokens during reasoning. For a reasoning step t, let A_t^{vis} denote the total attention mass assigned to visual tokens and let A_t^{all} denote the total attention mass over all keys. We define the visual-token attention ratio as

$$\text{VAR}_t = \frac{A_t^{\text{vis}}}{A_t^{\text{all}}}. \tag{19}$$

In practice, we aggregate this ratio over the designated analysis heads and average it over the samples within the same group. Higher values indicate stronger reliance on image evidence. Figure 2(b) shows that non-hallucinated trajectories maintain stronger visual support, whereas hallucinated ones progressively under-attend to visual tokens.

Figure 2(c) further combines support and uncertainty into a single indicator. Confidence alone is insufficient, because hallucinated trajectories can also become confident while drifting away from image evidence. We therefore compute token-level confidence from the entropy of the model prediction and couple it with the visual-token attention ratio: Let

$$\text{Conf}_t = 1 - \frac{\mathcal{H}(P_t)}{\log |\mathcal{Y}|}, \tag{20}$$

where P_t is the next-token distribution at step t, $\mathcal{H}(\cdot)$ denotes entropy, and $|\mathcal{Y}|$ is the output vocabulary size used for normalization. We then define joint support as

$$\text{JS}_t = \text{Conf}_t \cdot \text{VAR}_t. \tag{21}$$

This score favors steps where confidence and visual support are both high. Accordingly, Figure 2(c) ranks intervention locations using both signals rather than confidence alone, showing that the most effective correction points are those where uncertainty and visual dependence co-occur.

B.3 Counterfactual Drift Induction and Repair

The analyses above reveal a strong association between weak support and hallucination, but correlation alone does not show that support failure directly causes the final error. To address this, Figure 2(d) introduces matched counterfactual interventions that explicitly suppress or restore visual support at critical reasoning steps. We consider two complementary settings:

- **Drift induction.** Starting from baseline-correct samples, we suppress the visual support available at a selected critical step and then continue decoding under otherwise identical conditions. We quantify the resulting degradation by the induced hallucination rate

$$\text{IHR}(b) = \frac{N_{\text{corr}\to\text{hall}}(b)}{N_{\text{corr}}}, \tag{22}$$

  where N_{corr} is the number of baseline-correct samples and $N_{\text{corr}\to\text{hall}}(b)$ is the number that become hallucinated under intervention policy b.

- **Drift repair.** Starting from baseline-hallucinated samples, we instead restore or strengthen visual support at the matched critical step and test whether the final prediction can be corrected. This gives a repair-style counterpart to induction and is reported in the main paper as the post-intervention hallucination level. Together, the two settings test whether removing support induces drift and whether restoring support can recover from it.

To keep the comparison controlled, the baseline and intervention runs share the same image, prompt, decoding configuration, and backbone. They differ only in the support manipulation at the target reasoning step. This matched design isolates

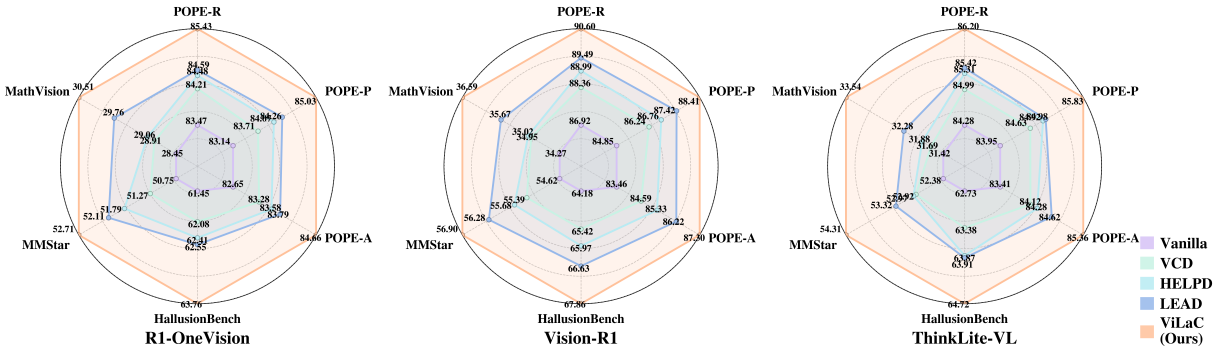


Figure 5: Radar chart comparison of hallucination mitigation methods across multiple benchmarks under R1-OneVision (Yang et al., 2025b), Vision-R1 (Huang et al., 2026), and ThinkLite-VL (Wang et al., 2025c).

the role of support manipulation from unrelated decoding noise. Figure 2(d) shows that disrupting visual support can turn correct trajectories into hallucinated ones, while restoring support can partially recover hallucinated predictions. This provides direct evidence that Grounding Bungee is causally relevant to the final reasoning outcome.

B.4 Statistics for Figure 2

For completeness, we summarize each subplot in Figure 2 below.

- **Figure 2(a).** HCR@p curves obtained by position-wise intervention over baseline hallucinated samples. Each point is the fraction of hallucinated samples corrected by intervention at the corresponding normalized reasoning position.
- **Figure 2(b).** Average visual-token attention ratio of hallucinated, non-hallucinated, and partially repaired trajectories, computed by averaging step-level ratios over samples in each group.
- **Figure 2(c).** Hallucination correction rate under different intervention ranking strategies. Joint support performs best because it prioritizes steps where uncertainty and visual dependence co-occur.
- **Figure 2(d).** Counterfactual induction and repair results under matched intervention settings. The induction bar measures the fraction of originally correct samples that become hallucinated after support suppression, while the repair bar measures the remaining hallucination level after support restoration on originally hallucinated samples.
- **Figure 2(e).** Qualitative case studies showing the evolution of confidence and visual support over reasoning steps. The key contrast is whether confidence recovery is accompanied by support restoration.

C More Experimental Results and Analysis

C.1 Radar Chart Results

Figure 5 presents a compact comparison across POPE (Li et al., 2023), HallusionBench (Guan et al., 2024a), MMStar (Chen et al., 2024), and MathVision (Wang et al., 2024) under three MLRMs, i.e., R1-OneVision (Yang et al., 2025b), Vision-R1 (Huang et al., 2026), and ThinkLite-VL (Wang et al., 2025c). Across all three backbones, ViLaC remains on the outermost contour on nearly all axes, indicating that its gains are not limited to a single benchmark. This advantage is especially clear on the hallucination-oriented axes, including POPE and HallusionBench, where ViLaC consistently outperforms VCD (Leng et al., 2024b), HELPD (Yuan et al., 2024), and LEAD (Xu et al., 2026b). The same trend also extends to MMStar and MathVision, showing that stronger grounding improves hallucination mitigation without sacrificing broader multimodal reasoning performance.

C.2 Heatmap Comparison

Figure 6 provides a qualitative comparison between the base model and ViLaC across several representative examples. Compared with the base model, ViLaC consistently produces more concentrated attention around visually relevant regions, such as key objects, human interactions, and local scene evidence needed for answering the query, whereas the base model tends to spread attention more diffusely over broader but less informative regions. This

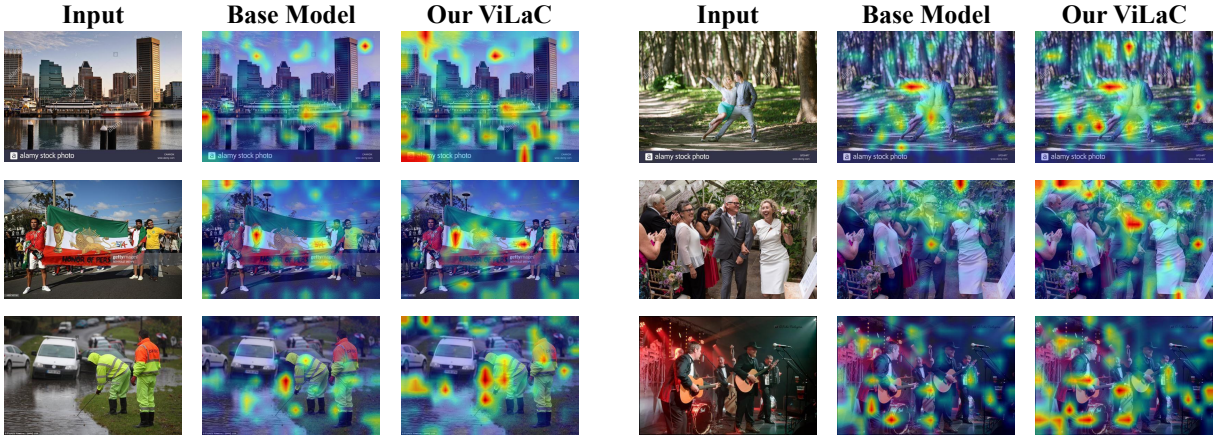


Figure 6: Attention heatmap comparison between the base model and ViLaC. Across diverse scenes, ViLaC produces more concentrated and task-relevant visual attention, while the base model tends to exhibit more diffuse attention over broader but less informative regions.

is especially clear in complex scenes containing multiple people or candidate objects, where ViLaC allocates stronger attention to the local evidence most likely to determine the correct answer. The heatmaps therefore suggest that the gain of ViLaC comes from sharper evidence selection rather than from uniformly amplifying visual responses.

C.3 Visualization Results

Figure 4 highlights a trajectory-level difference that is less obvious from the aggregate metrics alone. In both cases, the model passes through a temporary confidence valley, but the subsequent evolution differs: without ViLaC, the trajectory quickly returns to a confident but weakly grounded state, whereas with ViLaC, the rebound is accompanied by renewed attention to task-relevant regions. This visually complements the Grounding Bungee analysis in Section 3.

C.4 Ablation Results

The ablation results mainly reveal two properties of ViLaC. First, the gains are complementary rather than redundant, since latent exploration and visual crystallization each help on their own but work best when combined on POPE (Li et al., 2023) and HallusionBench (Guan et al., 2024a). Second, the hyperparameter trends are smooth across backbones, with moderate settings consistently performing better than either smaller or larger choices. Together, these results suggest that ViLaC benefits from jointly refining uncertain states and restoring support, while remaining stable to practical parameter variation.

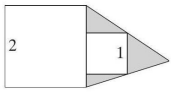
C.5 Latency Analysis

Table 6 suggests that the benefit of ViLaC on HallusionBench (Guan et al., 2024a) comes from more targeted intervention rather than from simply spending substantially more decoding time. Although ViLaC introduces an additional latent refinement stage, its runtime remains much closer to lightweight inference-time methods such as LEAD (Xu et al., 2026b) than to heavier baselines such as VCD (Leng et al., 2024b). This yields a favorable trade-off, especially on backbones where the accuracy gain is accompanied by only a modest latency increase.

D Case Studies Across Different Datasets

To further illustrate the effectiveness of ViLaC, we summarize representative qualitative observations across multiple benchmark settings. Compared with the baseline responses, the outputs generated by ViLaC remain better aligned with the visual content and are less likely to introduce unsupported objects, relations, or intermediate solution steps. This qualitative improvement is consistent with the trends observed on both hallucination-oriented benchmarks such as POPE (Li et al., 2023) and HallusionBench (Guan et al., 2024a), as well as broader reasoning benchmarks such as MM-Star (Chen et al., 2024) and MathVision (Wang et al., 2024). Across these settings, the common pattern is that ViLaC more often redirects the reasoning process back to task-relevant visual evidence before the model settles into a confident but weakly grounded continuation.

The diagram shows two squares: one has a side with a length of 2 and the other (abut on the first square) has a side with a length of 1. What is the area of the shaded zone?



Base Model

<think> The figure contains a square of side length 2 attached to a triangular region, inside which there is a smaller square of side length 1. The shaded region consists of the remaining triangular parts outside the smaller square. Since the large triangle appears symmetric, its total area can be estimated as \(\frac{1}{2}\times 2 \times 2 = 2\). Subtracting the inner square area \(1^2=1\), the shaded area becomes 2 \(- 1 = 1\). However, because the upper and lower shaded triangles appear congruent and there is also a right shaded tip, the total shaded region should instead sum to \frac{3}{2}}.<...></think>

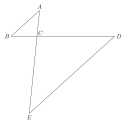
<answer> \boxed{\frac{3}{2}} </answer>

Our ViLaC

<think> The shaded region is the area inside the large triangle-like wedge but outside the small square of side length \(1\). The left side of the wedge spans the full side of the larger square, so its vertical height is \(2\). The wedge narrows linearly to a point on the right. Since the inner square has side length \(1\), its top and bottom corners lie halfway along the wedge height, implying the horizontal distance from the left boundary to the tip is also \(2\). Thus the enclosing triangular region has area \[\frac{1}{2}\times 2 \times 2 = 2.\]
The small square contributes area \[1^2 = 1.\]
Therefore, the shaded area is\[2 - 1 = 1.\]</think>

<answer> \boxed{1}

The area of $\\triangle ABC$ is 6 square centimeters. $\\overline{AB}\\|\\overline{DE}$. $BD=4BC$. What is the number of square centimeters in the area of $\\triangle CDE$?



Base Model

<think> Since \(\overline{AB}\parallel\overline{DE}\), triangles \(ABC\) and \(CDE\) are similar. We are told that \(BD=4BC\). Because \(BD\) includes both \(BC\) and \(CD\), we have\[CD = BD-BC = 4BC-BC = 3BC.\]
Thus the similarity ratio from \(\triangle ABC\) to \(\triangle CDE\) is\[\frac{CD}{BC}=3.\]
Areas scale by the square of the similarity ratio, so \[[\triangle CDE]=6\cdot 3^2=54.\]
However, since point \(C\) lies between the two triangles and the larger triangle appears stretched vertically, an additional scaling factor of 2 should be considered, giving \[54\times 2 = 108.\]</think>

<answer> \boxed{108} </answer>

Our ViLaC

<think> Because \(\overline{AB}\parallel\overline{DE}\), triangles \(ABC\) and \(CDE\) are similar. We are given \[BD=4BC.\]
Since \(BD=BC+CD\), it follows that
\[CD=4BC-BC=3BC.\]
Therefore, the similarity ratio from \(\triangle ABC\) to \(\triangle CDE\) is
\[\frac{CD}{BC}=3.\]
Areas of similar triangles scale by the square of the side ratio, so
\[[\triangle CDE]=[\triangle ABC]\times 3^2=6\times 9=54.\]</think>

<answer> \boxed{54}

Figure 7: Case 1 showing ViLaC mitigating long-chain hallucinations.

These children are standing in a line. Some are facing forwards and others are facing backwards. How many children are holding another child's hand with their right hand?

Base Model

<think> Seven children stand in a line, and we identify which ones hold another child's hand with their right hand based on their facing direction.
- 1st child (forward): holding with the right hand.
- 2nd child (forward): holding with the right hand.
- 3rd child (backward): the hand on the viewer's left is actually the child's right hand, so this child is holding with the right hand.
- 4th child (forward): holding with the right hand.
- 5th child (backward): holding with the right hand.
- 6th child (backward): holding with the right hand.
- 7th child (forward): holding with the right hand.

Thus all seven children are holding another child's hand with their right hand.</think>
<answer> \boxed{7} </answer>

Our ViLaC

<think> Seven children stand in a line, and we identify which ones hold another child's hand with their right hand based on their facing direction.- 1st child (forward): using the right hand.
- 2nd child (forward): using the right hand.
- 3rd child (backward): the visible left side corresponds to the child's actual right hand.
- 4th child (forward): using the left hand, not the right.
- 5th child (backward): using the right hand.
- 6th child (backward): using the right hand.
- 7th child (forward): using the right hand.

So exactly six children are holding another child's hand with their right hand.</think>
<answer> \boxed{6}

Sam paints the 9 small squares in the shape either white, grey or black. What is the minimum number he must paint over so that no two squares sharing a side have the same colour?

Base Model

<think> In the \(3\times3\) grid, there are conflicts between the two adjacent grey squares on the top row and between the connected black squares in the lower middle region.

Repainting the bottom-center black square removes both black conflicts at once, and repainting one grey square fixes the grey conflict, so \(2\) repaintings initially seem sufficient. However, the lower-left white square also appears to create an additional white adjacency pattern, meaning one extra repaint is required. Therefore the minimum number is 3.</think>

<answer> \boxed{3}

Our ViLaC

<think> In the \(3\times3\) grid, conflicts occur between the two adjacent grey squares on the top row and among the connected black squares in the center-bottom region.

Repainting the bottom-center black square removes both black conflicts simultaneously, while repainting one of the grey squares fixes the remaining grey conflict.

After these two repaintings, no adjacent squares sharing a side have the same color. Therefore, the minimum number of repaintings is\[2.\]</think>

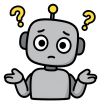
<answer> \boxed{2}

Figure 8: Case 2 showing ViLaC mitigating long-chain hallucinations.

Please describe this image in detail.

Base Model

<think> A person is sitting on a bed playing an electric guitar in a music-themed bedroom filled with posters and recording equipment. <...> The person appears to be practicing casually while surrounded by vintage rock decorations. Several posters of famous band members cover the walls, and there seems to be a recording microphone setup prepared for a studio session nearby. The room also contains what looks like a drum amplifier stack beside the bed, giving the impression of a small home music studio.<...> </think>



Our ViLaC

<think> A person is sitting cross-legged on a bed playing an electric guitar in a bedroom filled with music posters and decorations. <...> The room contains various posters on the walls, patterned curtains, and some audio or electronic equipment near the right side of the image. <...> The person appears focused on playing the guitar in a relaxed indoor setting.<...> </think>

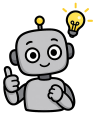
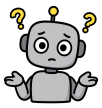


Figure 9: Case 3 showing ViLaC mitigating long-chain hallucinations.

Please describe this image in detail.

Base Model

<think> A person wearing a yellow helmet is kayaking through a fast-moving river in a red kayak. The kayaker appears to be navigating strong currents while holding a double-bladed paddle. Water splashes around the boat as the person moves downstream near dense vegetation. In the distance, there seems to be a waterfall entrance surrounded by forest, suggesting this could be part of an extreme whitewater course. The kayaker also appears to be wearing a professional racing suit and may be participating in a river competition event filmed by outdoor photographers hidden near the riverbank. </think>



Our ViLaC

<think> A person wearing a helmet is paddling a red kayak on a river surrounded by trees and vegetation. <...> The kayaker is using a double-bladed paddle while moving through the water, with small splashes visible around the kayak. <...> Branches and leaves partially appear in the foreground near the riverbank, and the scene takes place outdoors during daytime.<...> </think>

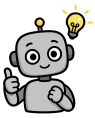
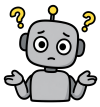


Figure 10: Case 4 showing ViLaC mitigating long-chain hallucinations.

Please describe this image in detail.

Base Model

<think> A messy bedroom is filled with scattered bags, furniture, and unpacked items, suggesting someone has recently moved in or is reorganizing the apartment. The bed is unmade with heavy blankets, while chairs and boxes are spread across the wooden floor. Near the back of the room, there appears to be a large television screen leaning against the wall beside several wooden boards. The open doorway seems to lead into a kitchen area, and the room also contains what looks like musical instrument cases and moving equipment, giving the impression of a musician settling into a new studio apartment.<...> </think>



Our ViLaC

<think> A cluttered bedroom contains an unmade bed, several chairs, bags, cardboard boxes, and miscellaneous household items scattered across the floor. The room has exposed brick walls, wooden flooring, and ceiling lights. An open door leads into a darker adjacent room, while some boards or panels are leaning against the wall near the center of the image. The scene appears to show an apartment during moving, unpacking, or reorganization.<...> </think>

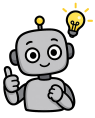
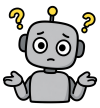


Figure 11: Case 5 showing ViLaC mitigating long-chain hallucinations.

Please describe this image in detail.

Base Model

<think> A balcony patio overlooks a scenic landscape with mountains, trees, and a swimming pool in the background. A small round table is set with breakfast items including cups, bowls, bread, and fruit, suggesting an outdoor morning meal. Sunlight casts shadows across the balcony railing and nearby walls. In the distance, there appears to be a tennis court beside the pool area, and the table also includes what looks like a wine bottle prepared for a luxury brunch. The scene resembles a high-end resort hotel with private recreational facilities and mountain-view accommodations.<...> </think>



Our ViLaC

<think> A balcony with a small outdoor table overlooks a landscaped yard with a swimming pool and mountains in the background. The table is arranged with cups, bowls, plates, and food items, suggesting a breakfast or snack setting. Metal chairs surround the table, and sunlight illuminates the balcony and nearby walls. Trees, grass, and shrubs are visible around the pool area, creating a peaceful outdoor scene.<...> </think>

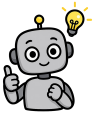
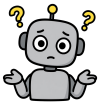


Figure 12: Case 6 showing ViLaC mitigating long-chain hallucinations.

Please describe this image in detail.

Base Model

<think> A waterfront city skyline is reflected in calm harbor water during sunset or early evening. Several modern high-rise buildings stand behind a dock where boats and ferries are moored. A large passenger boat is positioned near the center of the harbor, while smaller vessels are visible along the pier. In the background, there appears to be a sports stadium partially hidden behind the buildings, and the waterfront seems busy with tourist cruise activity preparing for nighttime departures. The tall tower on the right also resembles a luxury hotel skyscraper overlooking the marina district. </think>



Our ViLaC

<think> A city skyline stands along a waterfront with several tall buildings reflected in the calm water. Boats and ferries are docked near the harbor, including a red-and-white passenger vessel near the center of the image. The scene appears to take place during late afternoon or evening under soft sunlight. Piers, posts, and urban buildings line the waterfront area, creating a typical city harbor setting.<...> </think>

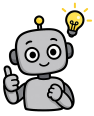
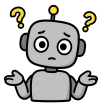


Figure 13: Case 7 showing ViLaC mitigating long-chain hallucinations.

Please describe this image in detail.

Base Model

<think> Two workers wearing high-visibility waterproof clothing are standing beside a flooded roadside during rainy weather. One worker is using a long tool to inspect the water level near the edge of the road, while several vehicles are parked nearby in partially flooded conditions. The scene suggests emergency maintenance or flood monitoring in a residential area. In the background, there appears to be a traffic accident involving multiple stopped vehicles, and the workers may be preparing to place road closure barriers due to rising floodwater. The worker in orange also seems to be communicating with a rescue coordination team further up the road. </think>



Our ViLaC

<think> Two workers wearing bright reflective rain gear are standing beside a flooded roadside on a rainy day. One worker is bending down and using a long tool or stick near the water, while the other stands nearby observing. Several cars are parked or moving along the wet road in the background. The scene appears to involve road maintenance, drainage inspection, or flood monitoring in a residential or suburban area.<...> </think>

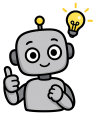
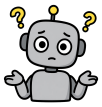


Figure 14: Case 8 showing ViLaC mitigating long-chain hallucinations.

Please describe this image in detail.

Base Model

<think> A joyful wedding ceremony is taking place indoors in a garden-like venue filled with guests and flowers. A smiling bride in a white dress is holding a bouquet while walking arm-in-arm with a man in a gray suit, as attendees around them applaud and celebrate. Several guests are dressed formally and appear to be enjoying the moment. In the background, there seems to be a live wedding band performing for the ceremony, and the couple may be entering a glass-roof reception hall for the evening celebration. The bride also appears to be carrying a bouquet of white lilies and orchids arranged for the formal procession.<...> </think>



Our ViLaC

<think> A wedding ceremony is taking place indoors with guests gathered around and applauding. A smiling bride wearing a white dress is holding a bouquet of flowers while walking beside a man in a gray suit. Other attendees dressed in formal clothing stand nearby in a decorated venue filled with plants and floral arrangements. The atmosphere appears joyful and celebratory.<...> </think>

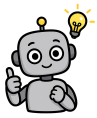


Figure 15: Case 9 showing ViLaC mitigating long-chain hallucinations.