

Decompose, Look, and Reason: Reinforced Latent Reasoning for VLMs

Mengdan Zhu[♣], Senhao Cheng[♣], Liang Zhao^{♣*}

♣Emory University ♠University of Michigan, Ann Arbor
{mengdan.zhu, liang.zhao}@emory.edu
senhaoc@umich.edu

Abstract

Vision-Language Models often struggle with complex visual reasoning due to the visual information loss in textual CoT. Existing methods either add the cost of tool calls or rely on localized patch-based embeddings that are insufficient to extract semantics in multi-step reasoning. We propose "*Decompose, Look, and Reason*" (DLR), a reinforced latent reasoning framework that dynamically decomposes queries into textual premises, extracts premise-conditioned continuous visual latents, and deduces answers through grounded rationales. We introduce a three-stage training pipeline and propose a novel *Spherical Gaussian Latent Policy*, to enable effective exploration in the latent space. Extensive experiments on vision-centric benchmarks show that DLR consistently outperforms strong baselines, including text-only, interleaved multimodal CoT, and latent reasoning methods, while providing superior stepwise interpretability.

1 Introduction

Vision-Language Models (VLMs) have made remarkable progress on visual question answering (Goyal et al., 2017; Antol et al., 2015), vision-language understanding (Li et al., 2023; Zhu et al., 2025b), vision-language generation (Zhang et al., 2021; Zhu et al., 2025a), and vision-language reasoning (Wang et al., 2024; Yu et al., 2026). The success of integrating Chain-of-Thought (CoT) into large language models (LLMs) for enhancing reasoning on complex tasks has inspired its extension to multimodal reasoning. Early approaches relied on text-only multimodal CoT (MCoT) (Zhang et al., 2023), which translates visual inputs into textual descriptions. However, this textual abstraction inevitably loses some important visual information. Therefore, subsequent work evolved toward interleaved MCoT, which explicitly incorporates

localized visual signals like cropped patches or bounding boxes (Shao et al., 2024a; Gao et al., 2025) into the reasoning steps. This line of work later developed into "thinking with images", where the model actively manipulates or edits images during reasoning. Such edits mainly include zooming in (Shen et al., 2025; Zhang et al., 2025; Su et al., 2026), drawing auxiliary lines and sketches (Hu et al., 2024), and highlighting boxing regions for understanding (Fu et al., 2025). Despite being visually grounded, these approaches incur extra cost due to external tool calls or program-based image manipulations, and are inherently limited by the set of available external tools. Alternatively, latent space reasoning models can be more efficient, as they avoid external tool calling by directly projecting intermediate visual information into an internally continuous embedding space. Existing latent visual reasoning methods typically learn continuous representations by reconstructing query-relevant region-of-interest (ROI) visual embeddings (Li et al., 2026) or the embeddings of auxiliary images (Wang et al., 2026b).

Patch-based methods such as Interleaved MCoT and "thinking with images" approaches, and prior latent reasoning methods have the following limitations: Such methods remain tied to explicit localized visual regions but fail to semantically isolate the desired elements. (1) An ROI or patch-based method may *over-include* information by bundling together all visual content within a selected region, including irrelevant context that are not needed for the current reasoning step. (2) At the same time, it may *under-include* when the required evidence is inherently non-local, such as a global layout, more abstract concepts, or a cross-patch relation. For example, Fig. 2(b) requires global and cross-patch visual information to determine the dominant color. Moreover, prior latent visual reasoning methods often insert the latent only once. For instance, Fig. 2(a) presents a complex logi-

*Corresponding author.

cal reasoning problem, where each reasoning step requires attending to different regions and their logical relationship, which cannot be captured by a single ROI.

To address these limitations, we introduce **Decompose, Look, and Reason: Reinforced Latent Reasoning (DLR)** framework for VLMs. Our framework mimics the human cognitive process of "Decompose $\rightarrow$ Look $\rightarrow$ Reason": (1) **Decompose the premise dynamically**: The model dynamically generates a textual premise or subquestion, determining *what* specific details need to be verified in the image and *when* to look for them. (2) **Look premise-conditioned latents**: A visual grounder attends to the image and conditioned on the hidden state of the textual premise, extracting continuous latent embeddings that capture *where* to look. Unlike patch-based latent reasoning methods, these latent tokens enable more efficient representation of image-related visual thoughts, covering both localized visual information and non-local latent semantics. (3) **Reason the latent-grounded rationale**: Conditioned on the injected visual latents, the VLM generates a textual rationale to explain in detail and eventually deduce the final answer.

To unleash the potential of this DLR framework, we propose a progressive three-stage training pipeline. In stage I: Pretraining, we establish a foundational cross-modal alignment and ensure that the continuous latents can accurately extract visual semantics corresponding to specific textual conditions. Next, in Stage II: Supervised Fine-tuning, we teach the VLM with the decomposition capability to internalize the structured DLR format. However, SFT strictly relies on teacher-forced log-likelihood, inherently bounding the visual grounder’s capacity to deterministic feature extraction without active exploration. To break this bottleneck, Stage III employs reinforcement learning to unlock true latent exploration. By adapting GRPO with our proposed novel latent policy optimization SGLP, we enable the model to actively explore the continuous visual manifold.

In summary, our main contributions are:

- We propose "Decompose, Look, and Reason", a reinforced latent reasoning framework, which dynamically decomposes the query into premises requiring visual verification, while simultaneously extracts premise-conditioned visual latents. The two components are mutually reinforcing, enabling the framework to progressively improve

both the VLM text policy and latent visual policy to find an optimal reasoning trajectory.

- We introduce a progressive three-stage training pipeline with a novel latent policy optimization. The proposed Spherical Gaussian Latent Policy (SGLP) intrinsically aligns with the hyperspherical manifold of vision-language representations and enables direct latent exploration without magnitude collapse. This effectively bridges the gap in multimodal latent RL and breaks the deterministic limitations of SFT.
- We conduct extensive experiments on multiple vision-language benchmarks spanning visual perception, mathematical reasoning, and vision-language understanding. DLR consistently outperforms strong baselines, including text-only, interleaved MCoT, and latent reasoning methods. Further ablations and case studies verify the contributions of each component, while also demonstrating that DLR yields more interpretable stepwise visual reasoning through premise-conditioned latents.

2 Related Work

2.1 Multimodal Chain-of-Thought Reasoning

Multimodal Chain-of-Thought (MCoT) reasoning extends textual CoT to settings where the input, intermediate reasoning process, or output may involve non-linguistic modalities such as images (Wang et al., 2025). Early MCoT methods generate textual intermediate rationales conditioned on visual inputs. Representative works (Zhang et al., 2023; Lu et al., 2022) show that inserting natural-language reasoning steps substantially improves visual question answering and science reasoning. But their intermediate reasoning process remains text-centric. Subsequent work explores interleaved MCoT, where visual content is explicitly inserted into the reasoning chain. For instance, Visual CoT introduces intermediate supervision by annotating a key bounding box (Shao et al., 2024a) while ICoT constructs interleaved reasoning by selecting image regions according to the model’s attention. (Gao et al., 2025). These methods improve visual grounding, but still rely on patch-based visual embeddings.

2.2 Think with Images

Recent work has evolved from interleaved MCoT reasoning to "thinking with images" (Su et al., 2025b). In this paradigm, the model actively ed-

its, manipulates, or augments visual evidence during intermediate reasoning steps. Representative methods equip vision-language models with drawing lines, or boxes during reasoning (Hu et al., 2024), and perform image edits such as highlighting, boxing, and masking to enhance structured visual understanding (Fu et al., 2025). More recent methods further generalize this paradigm by learning to invoke external vision tools adaptively during reasoning (Su et al., 2026, 2025a). However, they typically rely on external tools or executable programs to manipulate images. As a result, they often incur additional computational overhead and remain constrained by the coverage and capability of the predefined toolset (Su et al., 2025b). In contrast, our method explores directly internal latent visual reasoning rather than external tool calls.

2.3 Latent Space Reasoning

An alternative line of work seeks to perform visual reasoning directly in continuous latent space. Representative methods differ in how these latents are defined and supervised. Some studies reconstruct visual embeddings by query-relevant ROI (Li et al., 2026; Yang et al., 2026) and auxiliary images (Wang et al., 2026b; Zhu et al., 2024). However, the explicit ROI supervision limits the expressiveness and scalability of the visual latents. The intermediate reasoning process in these methods is typically interleaved by a single visual embedding. More recently, IVT-LR (Chen et al., 2026) can perform interleaved vision-text reasoning in latent space by combining latent textual states with selected visual embeddings. In contrast, our method could provide multi-step visual verifications without ROI supervision. It first establishes textual decomposition to iteratively decide *what* and *when* to inspect, then ground the specific premise with visual latents, and produce a more accurate rationale based on the retrieved evidence.

3 Reinforced Latent Reasoning

3.1 Method Overview

We formulate the multimodal reasoning task as a multi-step reasoning process, following a "Decompose, Look, and Reason" paradigm. Given an input image I and a user question q , our framework dynamically interleaves the visual grounding subquestion or premise p with latent visual evidence $\mathbf{z}$ and its corresponding rationale r . This process generates an iterative reasoning trajectory $\tau = \{(p^{(t)}, \mathbf{z}^{(t)}, r^{(t)})\}_{t=1}^N$ that ultimately deduces the final answer a .

The architecture consists of two primary learnable components parameterized by $\Theta = \{\theta, \phi\}$: a VLM policy P_θ governing discrete text generation, and a latent visual grounder policy P_ϕ responsible for continuous latent embeddings. The generation process unfolds in three steps:

- ◊ **Decompose:** The VLM P_θ learns to parse the current reasoning trajectory to mark a textual premise p that needs further visual verification, enclosing it within the special tokens `<premise>` and `</premise>`.
- ◊ **Look:** The latent visual grounder P_ϕ takes the image features and the hidden state of `</premise>`, s_t , to actively search for visual evidence with respect to p , yielding continuous latent embeddings $\mathbf{z} \in \mathbb{R}^{l \times d}$ inside the `<vis_thought>` and `</vis_thought>` tokens.
- ◊ **Reason:** Conditioned on the injected visual evidence $\mathbf{z}$, the VLM P_θ generates a textual rationale r enclosing with the `<rationale>` and `</rationale>` and deduces the final answer a .

As shown in Fig. 1, our framework couples textual reasoning and latent visual grounding through a joint factorization of the reasoning trajectory:

$$P_\Theta(\mathbf{y}, \mathbf{z} | I, q) = P_\theta(\mathbf{y} | I, q, \mathbf{z}) P_\phi(\mathbf{z} | I, q), \quad (1)$$

which can be further expanded in Eq. 4. This formulation makes the two components jointly optimized: *better textual decomposition guides more precise visual grounding, and richer visual latents ground stronger rationales*. As a result, jointly optimizing the two components allows them to mutually reinforce each other, making it more likely to discover an optimal reasoning trajectory.

3.2 Three-stage Training Pipeline

To effectively train this dynamic and interleaved latent reasoning framework, we propose a three-stage training pipeline.

3.2.1 Stage I: Pretraining

At the first stage, we warm up the latent visual grounder to establish basic cross-modal alignment between the continuous visual space and the discrete textual representation space of the VLM.

Visual Grounder We freeze the pretrained VLM backbone and optimize a lightweight visual grounder. The frozen visual encoder extracts image features: $V = \Phi_{\text{vis}}(I) \in \mathbb{R}^{m \times d}$, where m is the number of visual tokens after the vision encoder and d is the hidden dimension.

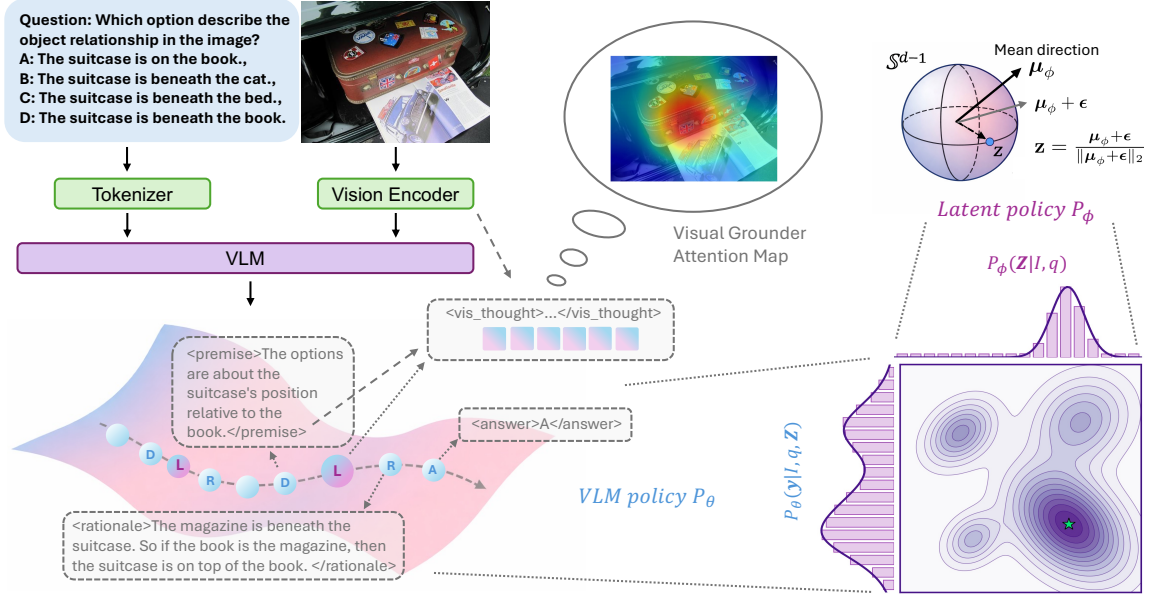


Figure 1: Overview of the *Decompose–Look–Reason* (DLR) framework and its reinforcement finetuning training objectives. The bottom-right contour illustrates the joint optimization landscape induced by the VLM text policy p_θ and the latent visual policy p_ϕ .

On the language side, we feed the question q into the frozen language model and take the hidden state of its last valid token as the textual condition: $h_q = \Phi_{\text{txt}}(q) \in \mathbb{R}^d$.

Similarly, we encode the answer a with the same frozen language model and use the hidden state of its last valid token as the target text embedding: $h_a = \Phi_{\text{txt}}(a) \in \mathbb{R}^d$.

We introduce a set of L learnable latent queries $\mathbf{z}_0 \in \mathbb{R}^{L \times d}$, which serve as trainable slots for extracting question-relevant visual evidence.

The visual grounder contains two cross-attention layers followed by a feed-forward network. First, the latent queries are conditioned on h_q . Next, the question-conditioned latent queries attend over image features V to extract relevant evidence. Finally, it passes a feed-forward network to obtain the latent visual representation: $\mathbf{z} \in \mathbb{R}^{L \times d}$.

Contrastive Alignment To ensure the latent embeddings $\mathbf{z}$ capture the precise visual semantics necessary to deduce the answer, we employ an InfoNCE contrastive loss. We mean-pool the latent $\bar{\mathbf{z}} = \frac{1}{L} \sum_{k=1}^L \mathbf{z}_k$ and L_2 -normalize it alongside the target answer embedding:

$$\hat{\mathbf{z}} = \frac{\bar{\mathbf{z}}}{\|\bar{\mathbf{z}}\|_2}, \quad \hat{\mathbf{h}}_a = \frac{\mathbf{h}_a}{\|\mathbf{h}_a\|_2}. \quad (2)$$

The bidirectional contrastive loss guides the pooled visual evidence to align with the correct answer's

semantic embedding in the latent space:

$$\mathcal{L}_{v \rightarrow t} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\hat{\mathbf{z}}_i^\top \hat{\mathbf{h}}_{a,i}/\tau)}{\sum_{j=1}^N \exp(\hat{\mathbf{z}}_i^\top \hat{\mathbf{h}}_{a,j}/\tau)},$$

$$\mathcal{L}_{t \rightarrow v} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\hat{\mathbf{h}}_{a,i}^\top \hat{\mathbf{z}}_i/\tau)}{\sum_{j=1}^N \exp(\hat{\mathbf{h}}_{a,j}^\top \hat{\mathbf{z}}_i/\tau)},$$

$$\mathcal{L}_{\text{pretrain}} = \frac{1}{2}(\mathcal{L}_{v \rightarrow t} + \mathcal{L}_{t \rightarrow v}). \quad (3)$$

3.2.2 Stage II: Supervised Finetuning (SFT)

In the second stage, we teach the model to follow our "Decompose, Look, and Reason" paradigm through supervised finetuning. We construct an annotated visual reasoning SFT dataset based on Vision-R1-cold dataset (Huang et al., 2026) using an LLM (e.g., GPT-5-mini) with the prompt in Appendix A to add our DLR special tokens. During SFT, both the VLM π_θ and the visual grounder π_ϕ are jointly optimized. The VLM learns to break down the reasoning trajectory into a sequence of image related sub-questions or premises p . Simultaneously, the visual grounder is trained as a deterministic feature extractor to output latent embeddings $\mathbf{z}^{(t)}$ that maximize the likelihood of the ground-truth rationales and the final answer. The joint probability distribution for generating this trajectory can be factorized into the product of the

VLM policy π_θ and the latent visual policy π_ϕ :

$$P_\Theta(\mathbf{y}, \mathbf{z}|I, q) = \prod_{t=1}^N \left[\underbrace{P_\theta(p^{(t)}|I, q, \mathcal{H}_{<t})}_{\text{Decompose}} \cdot \underbrace{P_\phi(\mathbf{z}^{(t)}|I, s_t)}_{\text{Look}} \cdot \underbrace{P_\theta(r^{(t)}|I, q, \mathcal{H}_{<t}, p^{(t)}, \mathbf{z}^{(t)})}_{\text{Reason}} \right] \cdot P_\theta(a|I, q, \mathcal{H}_{\leq N}), \quad (4)$$

where $\mathcal{H}_{<t}$ denotes the cumulative discrete context history up to step t .

In practice, optimizing the continuous latent variables $\mathbf{z}^{(t)}$ with a discrete cross-entropy objective is infeasible. Therefore, we represent the latents with a sequence of L discrete placeholder tokens (e.g., <VIS0> to <VIS31>) and implement a two-pass forward mechanism. Specifically, the VLM first extracts the hidden state s_t at the </premise> token, which the grounder π_ϕ then uses to generate the continuous visual latents $\mathbf{z}^{(t)}$. Next, we replace the placeholder tokens with $\mathbf{z}^{(t)}$ and continue generating the subsequent rationale $r^{(t)}$ and answer a . When computing the training objective, we mask the labels of these placeholder tokens. The optimization objective is thus formulated as the autoregressive Cross-Entropy (CE) loss over the text tokens:

$$\mathcal{L}_{SFT}(\Theta) = -\mathbb{E}_{\mathbf{y} \sim \mathcal{D}} \left[\sum_{t=1}^N \left(\log \pi_\theta(p^{(t)}|I, q, \mathcal{H}_{<t}) + \log \pi_\theta(r^{(t)}|I, q, \mathcal{H}_{<t}, p^{(t)}, \mathbf{z}^{(t)}) \right) + \log \pi_\theta(a|I, q, \mathcal{H}_{\leq N}) \right]. \quad (5)$$

By minimizing this CE loss, the visual grounder π_ϕ is implicitly optimized via gradients backpropagated from the subsequent rationales $r^{(t)}$ and the final answer a . While SFT successfully teaches the model to internalize the structured DLR format, it relies strictly on teacher-forced log-likelihood. This inherently bounds the visual grounder’s capacity to deterministic feature extraction without active exploration, which motivates our latent policy in Stage III.

3.2.3 Stage III: Reinforcement Finetuning for Continuous Latent Embeddings

Limitations of GRPO and Geometric Inductive Bias on Latent Embeddings. Prior work on latent visual reasoning (Li et al., 2026; Yang et al., 2026) perform GRPO (Shao et al., 2024b) following SFT. But standard GRPO can only be applied to discrete text tokens and can not optimize continuous latent embeddings due to the lack of a defined token distribution. Thus, previous work treats visual latents as deterministic variables, confining the RL

exploration to only the LLM’s text generation. The visual latents are frozen or updated indirectly via gradients backpropagated from the textual RL loss. This lack of a visual latent direct exploration limits the latent representation learning, easily falling into the suboptimal bounded by its SFT.

To address this limitation, we propose a stochastic latent policy for the visual grounder. Crucially, this policy should be aligned with the geometry of the vision-language feature space: it is inherently constrained to a hyperspherical manifold, where semantic information (e.g., cosine similarity in CLIP) is encoded primarily in the *direction* rather than the *magnitude* of feature vectors. Formally, let the semantic manifold be the unit hypersphere $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$. A standard Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$ defined in Euclidean space $\mathbb{R}^d$ is geometrically mismatched for this manifold, as it conflates magnitude variance with semantic variance (Wang et al., 2026b). To resolve this, we propose a Spherical Gaussian Latent Policy (SGLP), which intrinsically aligns with the geometric inductive bias of the vision-language feature space. Instead of applying unconstrained Euclidean noise, our visual grounder predicts a L_2 -normalized mean direction $\boldsymbol{\mu}_\phi \in \mathbb{S}^{d-1}$. We then inject isotropic exploration noise and explicitly re-project the sampled vector back onto the unit hypersphere:

$$\mathbf{z} = \frac{\boldsymbol{\mu}_\phi + \boldsymbol{\epsilon}}{\|\boldsymbol{\mu}_\phi + \boldsymbol{\epsilon}\|_2}, \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}). \quad (6)$$

This formulation ensures that the continuous exploration operates on the angular space, since its norm is fixed as $\|\mathbf{z}\|_2 = 1$. By mathematically decoupling the semantic direction from the vector magnitude, our approach eliminates the risk of magnitude collapse.

Latent Policy Optimization. We adapt the Dr. GRPO (Liu et al., 2025) for our joint policy optimization. For each question q , we sample a group of G trajectories $\{\tau_1, \dots, \tau_G\}$ from the current joint policy $\pi_{\Theta_{old}}$, where $\Theta = \{\theta, \phi\}$. The advantage A_i for the i -th trajectory is estimated by subtracting the group mean:

$$A_i = r^{(i)} - \mu(\{r^{(j)}\}_{j=1}^G). \quad (7)$$

The total objective $\mathcal{J}(\Theta)$ is the sum of the text policy objective and the latent policy objective:

$$\mathcal{J}(\Theta) = \mathcal{J}_{text}(\theta) + \mathcal{J}_{latent}(\phi). \quad (8)$$

While $\mathcal{J}_{text}(\theta)$ follows Dr. GRPO formulation for discrete text tokens, we derive $\mathcal{J}_{latent}(\phi)$ for our latent policy. For a latent $\mathbf{z}_{old}$ generated by the old policy mean $\mu_{\phi_{old}}$, the optimization objective is:

$$\mathcal{J}_{latent}(\phi) = \mathbb{E} \left[\frac{1}{G} \sum_{i=1}^G \min(\rho_i A_i, \text{clip}(\rho_i, 1 - \epsilon, 1 + \epsilon) A_i) \right], \quad (9)$$

where the importance sampling ratio ρ_i is computed using the spherical Gaussian probability densities:

$$\begin{aligned} \rho_i &= \frac{\pi_{\phi}(\mathbf{z}_{old}^{(i)})}{\pi_{\phi_{old}}(\mathbf{z}_{old}^{(i)})} \\ &= \exp \left(\frac{\|\mathbf{z}_{old}^{(i)} - \mu_{\phi_{old}}^{(i)}\|_2^2 - \|\mathbf{z}_{old}^{(i)} - \mu_{\phi}^{(i)}\|_2^2}{2\sigma^2} \right). \end{aligned} \quad (10)$$

Theorem 3.1 (Latent Policy Effect). *By maximizing $\mathcal{J}_{latent}(\phi)$, the expected update of the visual grounder increases the alignment between mean direction μ_{ϕ} and latent samples $\mathbf{z}_{old}$ that receive high advantage. Equivalently, the optimization encourages the latent policy to move toward semantically task-relevant directions on the hypersphere.*

Proof. The formal proof is in Appendix B. $\square$

Reward Design. We design a dense reward function r to guide the exploration. The total reward for the i -th trajectory is defined as:

$$r^{(i)} = R_{outcome}^{(i)} + \beta \cdot \mathbb{I}(R_{outcome}^{(i)} > 0) \cdot R_{focus}^{(i)}. \quad (11)$$

(1) Outcome Reward ($R_{outcome}$): This is a binary sparse reward indicating the correctness of the model-generated answer $\hat{a}$ against the ground-truth answer a .

$$R_{outcome}^{(i)} = \begin{cases} 1 & \text{if ExactMatch}(\hat{a}^{(i)}, a^{(i)}) \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

(2) Visual Grounder’s Focus Reward (R_{focus}):

To encourage visual grounding, we align the visual grounder’s attention map $\mathcal{A}_{latent}$ with an oracle attention map $\mathcal{A}_{oracle}$ derived from a frozen, strong vision-language model (e.g., SigLIP-SO400M) conditioned on the premise p . We define the focus reward as the exponential negative Kullback-Leibler (KL) divergence:

$$R_{focus}^{(i)} = \exp \left(-\lambda \cdot D_{KL}(\mathcal{A}_{oracle}^{(i)} \parallel \mathcal{A}_{latent}^{(i)}) \right). \quad (13)$$

The indicator function $\mathbb{I}(\cdot)$ in the total reward acts as a filter: we postulate that visual alignment is

only meaningful when it contributes to a correct inference. If the model answers incorrectly, the focus reward is suppressed to zero to avoid reinforcing "hallucinated" attention patterns that do not yield factual correctness.

4 Experiments

4.1 Experiment Setup

Evaluation Benchmarks To assess our DLR framework, we conduct comprehensive evaluations across four vision-centric benchmarks that collectively measure diverse capabilities in visual reasoning and understanding. For visual detail understanding, we evaluate on V* Bench (Wu and Xie, 2024), which assesses VLM’s capability to perform fine-grained attribute recognition and relative spatial reasoning. For mathematical visual reasoning, we evaluate on MathVista (Lu et al., 2024), a benchmark designed to assess mathematical reasoning in visual contexts. For broad multidisciplinary multimodal reasoning, we evaluate on MMMU-Pro (Yue et al., 2025). Finally, for general multimodal capability, we evaluate on MM-Star (Chen et al., 2024), a vision-indispensable benchmark that covers six core capabilities and eighteen fine-grained axes. We report the Pass@1 accuracy using greedy decoding. We set a maximum length of 40960.

Baselines We compare DLR against several representative vision-language reasoning baselines. These include our model backbone Qwen3-VL-8B-Thinking (Bai et al., 2025) which is a strong text-only baseline; ICoT (Gao et al., 2025), an interleaved multimodal CoT baseline that constructs paired visual-textual intermediate rationales by selecting image patches based on attention maps; PixelReasoner (Su et al., 2026), a "thinking with images" approach that performs intermediate reasoning through pixel-space visual operations; and LVR (Li et al., 2026), a latent visual reasoning method that conducts reasoning in continuous visual embedding space. All open-source baselines are adapted on top of the same Qwen3-VL-8B-Thinking backbone as DLR to ensure fair comparison.

Implementation Details We adopt Qwen3-VL-8B-Thinking as our base VLM. The visual grounder is initialized with $L = 32$ learnable continuous queries. The focus reward parameter $\beta = 0.1$ and $\lambda = 1$. Experiments are conducted on 4 NVIDIA H200 GPUs. More details in Appendix C.

Model	V*			MathVista	MMMU-Pro	MMStar
	Overall	Attribute	Spatial	Testmini	Overall	Overall
<i>Proprietary Model</i>						
GPT-4o	67.5*	72.2*	60.5*	60.0*	51.9*	-
<i>Open-Source Model based on Qwen3-VL-8B-Thinking</i>						
Qwen3-VL-8B-Thinking	79.6	81.7	76.3	79.2	60.9	72.3
ICoT	79.6	81.7	76.3	77.1	61.2	67.0
LVR	82.2	82.6	81.6	77.3	60.5	72.7
PixelReasoner	80.1	82.6	76.3	74.2	56.1	68.7
DLR (ours)	83.8	84.3	82.9	82.7	63.5	76.2
Δ over Backbone	+4.2	+2.6	+6.6	+3.5	+2.6	+3.9

Table 1: Performance on real-world perception, math, and reasoning benchmarks. The best results are highlighted in **bold**. Results marked with “*” are reported by other papers (Wang et al., 2026a,b).

4.2 Main Results

Tab. 1 presents the main results of DLR on four vision-centric benchmarks. Overall, DLR achieves the best performance among all baselines across all evaluated benchmarks, even surpassing the proprietary model GPT-4o with approximately 200B parameters. On V* Bench, DLR achieves 83.8 overall accuracy, outperforming the backbone by 4.2%, LVR by 1.6%, and PixelReasoner by 3.7%. The gains are consistent across both attribute and spatial subcategories, confirming that DLR improves fine-grained visual detail understanding in both attribute recognition and relative position reasoning. We adapted baselines to the same Qwen3-VL-8B-Thinking backbone as ours for a fair comparison. ICoT inserts its interleaved visual evidence based on “\n” tokens. On short reasoning tasks like V*, the model would fail to trigger newlines, causing ICoT to degenerate into the backbone performance. LVR’s design is inherently compatible with Qwen3-VL-Thinking’s <think> block: its latent visual reasoning tokens (<|lvr_start|>, <|lvr_end|>) are inserted inside the <think> block, and its SFT data preserves the CoT structure. As a result, LVR benefits from the stronger Qwen3-VL backbone. In contrast, PixelReasoner’s tool calls are placed outside of the <think> block in Qwen3-VL’s implementations by design, so PixelReasoner’s SFT data cannot be wrapped into the thinking mode. When PixelReasoner’s SFT is applied with this non-thinking data format for 1–2 epochs, the model tends to prioritize learning the tool-calling pattern at the expense of forgetting the original thinking capability that the backbone was trained on, which prevents it from building upon the backbone’s existing rea-

soning abilities. On MathVista, DLR obtains 82.7, exceeding Qwen3-VL baseline by 3.5% and surpassing the strongest open-source baseline LVR by 5.4%. This result suggests that the proposed DLR mechanism is particularly beneficial for mathematical visual reasoning, where the model must progressively inspect diagrams or figures and accumulate evidence over multiple steps rather than relying on a single-pass latent representation like LVR. On MMMU-Pro, DLR achieves improves over Qwen3-VL baseline by 2.6%. This shows that DLR generalizes beyond localized perception tasks to challenging multidisciplinary reasoning scenarios, where better premise decomposition and premise-conditioned visual verification both contribute to improved answer accuracy. Results on MMStar show that DLR improves not only improves the complex reasoning tasks but also broader multimodal understanding.

In summary, we have several key findings: (i) **the text-only baseline is consistently weaker than methods that explicitly incorporate visual evidence in intermediate steps**, confirming the importance of grounded visual verification. In complex tasks, Qwen3-VL baseline sometimes generates very long outputs (over 30k+ tokens) but repeats uncertain self-verification without sufficient visual grounding. (ii) **“think with images” methods such as PixelReasoner improve over the text-only baseline, but stills trails latent visual reasoning methods like LVR**. It suggests that internal latent grounding can be more efficient than external visual-editing tools. (iii) **among latent reasoning methods, DLR consistently outperforms LVR**, indicating **dynamically interleaving multi-step premise-conditioned latent visual grounding is**

Model	V*	MathVista	MMMU-Pro	MMStar
DLR-SFT	80.5	65.1	54.6	64.6
w/o pretraining	79.5	63.0	53.6	63.2
DLR	83.8	82.7	63.5	76.2
w/o R_{focus}	82.6	66.5	55.1	64.8
w/o $\mathcal{J}_{latent}$	81.5	57.1	54.9	63.3

Table 2: Ablations of the components in DLR.

more effective than constraining the reasoning process to a single, coarse latent embedding.

4.3 Ablation Study

Effect of the Pretraining, SFT, and RL. In Tab. 2, we observe that removing pretraining consistently hurts performance on all benchmarks. This shows that the pretraining stage provides a better initialization for the visual grounder by establishing foundational cross-modal alignment before structured reasoning is learned during SFT. Without this warm-up stage, the latent visual grounder is less capable of extracting premise-relevant visual evidence. During SFT, it internalizes the structured DLR format. Compared with DLR-SFT, the full DLR achieves a 3.3% improvement on V*. This shows that the RL stage can break the deterministic limitations of SFT, enabling the model to actively explore visual evidence and to facilitate finding potentially correct reasoning paths.

Effect of the Latent Policy Optimization. The most critical component of stage III is the latent policy optimization. Removing $\mathcal{J}_{latent}$ in Tab. 2 leads to a drastic performance degradation, particularly on MathVista, where the score drops from 82.7% to 57.1%. This confirms that our proposed SGLP is essential for guiding effective exploration in continuous visual space.

Effect of the Focus Reward. Removing R_{focus} results in a decline in V* performance by 1.2%, suggesting that aligning the visual grounder’s attention with an oracle grounding signal provides some guidance for stable latent exploration.

4.4 Quantitative Faithfulness Analysis

Beyond the qualitative attention visualizations, we further quantitatively evaluate whether the visual grounder’s attention faithfully identifies image regions that are important to the model’s predictions. Following the comprehensiveness framework (DeYoung et al., 2020), we mask the top image regions identified by the visual grounder’s attention map and measure the resulting accuracy degradation on V* Bench. As a control, we randomly mask the same number of image regions.

Masking Strategy	V* Accuracy	Δ
No mask (DLR)	83.8	–
Random mask	77.7	-6.1
DLR attention mask	36.5	-47.3

Table 3: Quantitative faithfulness analysis on V* Bench.

As shown in Tab. 3, random masking reduces accuracy by only 6.1 percentage points, whereas masking DLR attention regions causes a 47.3-point drop. This large degradation confirms that the visual grounder’s attention faithfully captures the image regions critical to the model’s predictions.

4.5 Case Study

Both cases in Fig. 2 highlight a **common challenge for text-only multimodal reasoning: The model often generates long but weakly visually grounded reasoning chains**. In contrast, DLR explicitly decomposes the problem into premise-level subproblems, retrieves premise-conditioned visual evidence, and then produces grounded rationales before answering. In the first example, the baseline severely **overthinks** the problem. Although it repeatedly waits to check possible rules for the circular diagrams again and again, it fails to reliably ground these hypotheses in the image and ultimately predicts the wrong answer after generating **15,177 tokens**. The baseline keeps revising speculative rules such as counting green segments, checking angular relations, and testing row-wise patterns, but still converges to an incorrect option. This example reveals that a text-only reasoning process can become verbose and unstable when the task fundamentally depends on precise visual comparison. By contrast, DLR solves the same problem through a much more structured reasoning process. It first generates a premise that identifies the key visual subproblem—namely, that the green segments are arranged in consistent relative positions across the grid—and then retrieves visual evidence conditioned on this premise. The subsequent rationale is directly grounded in the retrieved evidence, allowing the model to verify the pattern step by step. Rather than forcing the model to solve the entire problem in a single opaque trajectory, DLR follows an explicit *decompose* $\rightarrow$ *look* $\rightarrow$ *reason* loop that is better aligned with how humans solve difficult visual tasks.

The second example further demonstrates the interpretability of our framework. The text-only

- space. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 39117–39129.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4443–4458.
- Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. 2025. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*.
- Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. 2025. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19520–19529.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.
- Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379.
- Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Zhe Xu, Yao Hu, Shaohui Lin, et al. 2026. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pages 63794–63812.
- Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. 2026. Latent visual reasoning. In *International Conference on Learning Representations*, volume 2026, pages 148076–148090.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pages 23439–23554.
- Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in neural information processing systems*, 35:2507–2521.
- Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024a. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024b. Deepseek-math: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. 2025. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6613–6629.
- Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhui Chen. 2026. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. *Advances in Neural Information Processing Systems*, 38:8222–8251.
- Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. 2025a. Open-thinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*.
- Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. 2025b. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*.
- Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. 2026a. VI-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *Advances in Neural Information Processing Systems*, 38:30865–30891.

- Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. 2026b. Monet: Reasoning in latent visual space beyond image and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12030–12040.
- Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.
- Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.
- Penghao Wu and Saining Xie. 2024. V*: Guided visual search as a core mechanism in multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13084–13094. IEEE.
- Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. 2026. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 33510–33520.
- Xinlei Yu, Zhangquan Chen, Yongbo He, Tianyu Fu, Guanting Dong, Cheng Yang, Chengming Xu, Yue Ma, Xiaobin Hu, Zhe Cao, et al. 2026. The latent space: Foundation, evolution, mechanism, ability, and outlook. *arXiv preprint arXiv:2604.02029*.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186.
- Han Zhang, Weichong Yin, Yewei Fang, Lanxin Li, Boqiang Duan, Zhihua Wu, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vilg: Unified generative pre-training for bidirectional vision-language generation. *arXiv preprint arXiv:2112.15283*.
- Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. 2025. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv e-prints*, pages arXiv–2505.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.
- Mengdan Zhu, Senhao Cheng, Guangji Bai, Yifei Zhang, and Liang Zhao. 2025a. Cross-modal rag: Sub-dimensional retrieval-augmented text-to-image generation. *arXiv e-prints*, pages arXiv–2505.
- Mengdan Zhu, Raasikh Kanjiani, Jiahui Lu, Andrew Choi, Qirui Ye, and Liang Zhao. 2025b. Latent-explainer: explaining latent representations in deep generative models with multimodal large language models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4529–4538.
- Mengdan Zhu, Zhenke Liu, Bo Pan, Abhinav Angirekula, and Liang Zhao. 2024. Explaining latent representations of generative models with large multimodal models. *arXiv preprint arXiv:2402.01858*.

A Prompt for Constructing SFT Data

<image>\nThe question is “{question}” and the reasoning process is “{reasoning}”. Within the reasoning process, focus on the content between <think> and </think> and perform the following annotation: Identify every major argument that requires looking at the image and is immediately followed by sentence(s) that describe the specific details obtained from observing the image, which serve as the crucial rationale for deriving the answer.

Wrap each such major argument with the tags <premise> and </premise>. Only tag the arguments that require looking at the image to get the following details. Wrap the corresponding supporting details with <rationale> and </rationale>. For example: <premise>The image shows white suitcases stacked.</premise><rationale>On the top suitcase in the foreground, there’s a wide-brimmed hat with a black band and a red fabric draped over it.</rationale> or <premise>There are two parallel lines, m and n, cut by a transversal.</premise><rationale>Angle 1 is above line m on the right side of the transversal, and angle 2 is below line n on the left side of the transversal.</rationale>

The final output should be the original reasoning with <premise> and <rationale> tags inserted in place. Keep all original text unchanged except for inserting the tags.

Figure 3: The Prompt Template for Constructing the Reasoning Process for SFT.

B Gradient Derivation for Latent Policy Optimization

We formally derive the gradient update rule for the latent visual grounder, parameterized by ϕ , to demonstrate how the Spherical Gaussian Latent Policy (SGLP) explores the semantic space.

Let $\mathbf{z}^{(t)} \in \mathbb{S}^{d-1}$ be the continuous latent embedding generated at step t conditioned on state s_t . From the Policy Gradient Theorem, the gradient of the expected return objective $\mathcal{J}_{latent}(\phi)$ with respect to the visual grounder parameters ϕ is given

by:

$$\nabla_{\phi} \mathcal{J}_{latent}(\phi) = \mathbb{E} \left[\sum_t A \nabla_{\phi} \log \pi_{\phi}(\mathbf{z}^{(t)} | s_t) \right],$$

For our SGLP, both the sampled latent $\mathbf{z}^{(t)}$ and the predicted mean direction $\boldsymbol{\mu}_{\phi}(s_t)$ are strictly L_2 -normalized. As a result, the squared Euclidean distance mathematically simplifies into a scaled cosine similarity. The log-probability of the policy is therefore proportional to the dot product:

$$\log \pi_{\phi}(\mathbf{z}^{(t)} | s_t) \approx \frac{1}{\sigma^2} \mathbf{z}^{(t)\top} \boldsymbol{\mu}_{\phi}(s_t) + C.$$

Taking the derivative of this log-probability with respect to ϕ yields:

$$\begin{aligned} \nabla_{\phi} \log \pi_{\phi}(\mathbf{z}^{(t)} | s_t) &\approx \frac{1}{\sigma^2} \nabla_{\phi} \left(\mathbf{z}^{(t)\top} \boldsymbol{\mu}_{\phi}(s_t) \right) \\ &= \frac{1}{\sigma^2} \mathbf{z}^{(t)\top} \nabla_{\phi} \boldsymbol{\mu}_{\phi}(s_t). \end{aligned}$$

Substituting this back into the policy gradient formulation, the parameter update rule governed by the GRPO advantage A becomes:

$$\phi_{new} \leftarrow \phi_{old} + \alpha \mathbb{E} \left[\sum_t \frac{A}{\sigma^2} \mathbf{z}^{(t)\top} \nabla_{\phi} \boldsymbol{\mu}_{\phi}(s_t) \right].$$

Some Takeaways on the Manifold:

- **Directional Alignment** ($\mathbf{z}^{(t)\top} \nabla_{\phi} \boldsymbol{\mu}_{\phi}$): Because both vectors reside on the unit hypersphere, the gradient guides the predicted mean vector $\boldsymbol{\mu}_{\phi}$ along the surface of the manifold without altering its magnitude.
- **Advantage Weighting** (A): The group-relative advantage A acts as a directional scalar. If $A > 0$ (the better latent $\mathbf{z}^{(t)}$ focused on the right visual region and represents better extracted premise-aligned representation, which led to a correct rationale and answer), the gradient pulls the mean direction $\boldsymbol{\mu}_{\phi}$ towards the explored $\mathbf{z}^{(t)}$. Conversely, if $A < 0$, it pushes $\boldsymbol{\mu}_{\phi}$ away.

C Implementation Details

All trainings use the AdamW optimizer with a weight decay of 0.01 and a cosine learning rate scheduler with a 10% linear warmup.

Stage I: Pretraining. In the first stage, the base VLM is completely frozen. We train only the visual grounder on 443,757 visual question-answering

pairs in the training set of VQAv2 dataset (Goyal et al., 2017) using the symmetric InfoNCE loss. We set the contrastive temperature $\tau = 0.07$ and train 3 epochs with a learning rate of 1×10^{-4} and batch size of 16.

Stage II: SFT. We train the model on our custom annotated Vision-R1-cold dataset (200k) (Huang et al., 2026) to internalize the DLR format. During SFT, the visual encoder and the base LLM weights remain frozen. We apply LoRA fine-tuning with rank $r = 64$, $\alpha = 128$, and a dropout rate of 0.05. We train for 2 epochs with a learning rate of 5×10^{-5} and an effective batch size of 16.

Stage III: Latent Policy Optimization. For the reinforcement finetuning stage, we implement our SGLP framework on the ViRL dataset (39k) (Wang et al., 2026a). The policy samples a group of $G = 4$ trajectories per query. We train for 1 epoch with a VLM learning rate is 1×10^{-6} with sampling temperature set to 1.0, and an effective batch size of 8. The visual grounder learning rate is 1×10^{-5} .

D Computational Overhead.

DLR introduces only a lightweight visual grounder of approximately 7M parameters, corresponding to less than 0.09% of the 8B VLM backbone. During Stages II and III, the grounder is jointly optimized with the VLM and therefore introduces only a small additional parameter overhead. The only additional training stage compared with a standard SFT + RL pipeline is Stage I pretraining, where the entire VLM backbone is frozen and only the visual grounder is optimized. Moreover, the relative parameter overhead decreases with backbone scale, accounting for less than 0.01% when scaling to 70B+ models.

E LLM Usage Disclosure

We use large language models to correct the grammar and improve the clarity of writing in this paper.