GELVR: Geometry-Consistent Latent Visual Reasoning in Multimodal LLMs

Jiaqi Wang[1]* **Yi Feng**[2]* **Xian Wu**[1]†
[1]Tencent Jarvis Lab, China [2]Beijing Jiaotong University, China
jackjqqwang@tencent.com yifeng@bjtu.edu.cn kevinxwu@tencent.com

Abstract

Multimodal Chain-of-Thought (CoT) reasoning has significantly advanced large models, yet forcing continuous visual evidence into discrete text or fixed tool calls inevitably sacrifices fine-grained detail. Recent paradigms attempt to reason directly within a continuous latent space; however, supervising latent slots in isolation fails to account for their collective geometry, resulting in a fragmented latent space that lacks structural guidance. To ensure structural integrity across training stages, we propose GELVR (Geometry-Consistent Latent Visual Reasoning), a framework that establishes geometric consistency as a governing principle: the relational structure is explicitly *aligned* during SFT and actively *preserved* throughout RL, anchoring both stages to a common geometric scaffold. To operationalize this paradigm during the reward-driven phase, we further introduce GEPO (Geometry-preserving Policy Optimization), a policy-optimization algorithm tailored for the hybrid discrete-continuous action spaces of latent reasoning. GEPO employs a spherical von Mises-Fisher (vMF) policy to respect the decoder's inherent geometry and integrates the geometry-preserving regularizer directly into the RL objective. This ensures that reward optimization and structural integrity advance jointly rather than in tension. Extensive experiments across multiple visual reasoning benchmarks demonstrate that GELVR consistently outperforms state-of-the-art baselines, particularly in high-resolution and fine-grained perception tasks. Comprehensive ablations further validate the necessity of each geometry-consistent component in establishing stable and coherent latent reasoning. The code repository is available at `https://github.com/JackeyWang96/GeLVR`.

1 Introduction

Multimodal large language models (MLLMs) have made rapid progress on tasks that require joint understanding of visual and textual content. Chain-of-thought (CoT) reasoning [Wei et al., 2022] is central to this progress, and recent multimodal CoT systems extend it to visual queries [Zhang et al., 2024, Xu et al., 2024, Meng et al., 2025, Huang et al., 2025]. Yet the way these systems handle visual evidence remains fundamentally text-mediated: either the decoder verbalizes what it sees before reasoning over it (*"thinking about images"*, Figure 1(a)) [Zhang et al., 2024, Xu et al., 2024, Huang et al., 2025], or it invokes external tools (e.g. crop, zoom) to re-perceive the image mid-reasoning (*"thinking with images"*, Figure 1(b)) [Wu and Xie, 2024, Hu et al., 2024, Zheng et al., 2025]. Both paradigms share one limitation: fine-grained visual information is eventually squeezed through a discrete text channel or a predefined tool vocabulary, bottlenecking the bandwidth of visual reasoning.

To preserve fine-grained visual evidence throughout the reasoning process, a more recent line of work, exemplified by pioneering methods such as Mirage [Yang et al., 2025], LVR [Li et al.,

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

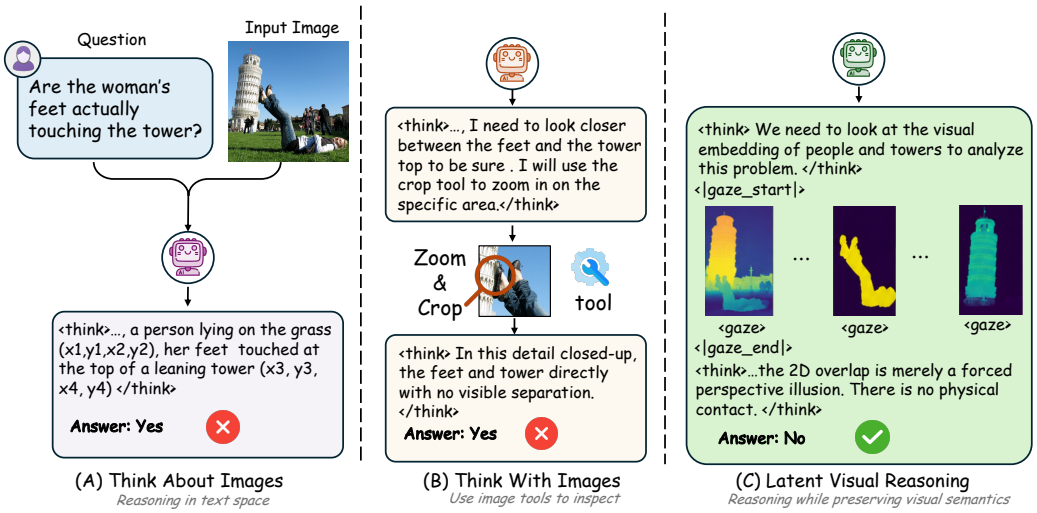

Figure 1: **Comparison of visual reasoning paradigms. (a) Thinking about images**: suffers from a semantic bottleneck, as discretizing continuous visual scenes discards fine-grained evidence. **(b) Thinking with images**: relies on external tools (e.g., crop), introducing latency and brittle invocations while fragmenting the holistic spatial context. **(c) Latent visual reasoning (Ours)**: performs reasoning natively within the continuous visual latent space, seamlessly unified with text.

2025b], SkiLa [Tong et al., 2025], Monet [Wang et al., 2025a], and HyLaR [Cheng et al., 2026], moves reasoning directly into the decoder's continuous latent space, delivering visual *thoughts* without ever rendering pixels (Figure 1(c)). While promising, these methods share a common limitation: they treat the latent segment as a bag of independent points, completely ignoring the inherent **relational geometry** among the latent slots. Worse still, this shortcoming leads to a severe geometric inconsistency across training stages. In standard pipelines, any relational structure implicitly formed during Supervised Fine-Tuning (SFT) is silently eroded by scalar rewards during Reinforcement Learning (RL). This motivates a **geometry-consistent paradigm**: one that explicitly enforces *geometry alignment* during SFT, and ensures strict *geometry preservation* under RL.

Following this principle, we propose **GELVR** (**Ge**ometry-Consistent **L**atent **V**isual **R**easoning), a framework that instantiates geometry consistency as two complementary, stage-level objectives operating on a single relational quantity. At the SFT stage, GELVR realizes *geometry alignment* by explicitly supervising, beyond the usual point-wise reconstruction, the pairwise relational structure of the decoder's latent trajectory against that of an external visual target manifold—so the latent segment acquires a well-defined geometric shape rather than a loose cluster of points. At the RL stage, the same relational quantity is reused with a different reference: the trajectory produced by the frozen SFT policy. This enables *geometry preservation*, which anchors the online latent trajectory against the learned structure, allowing the reward to optimize individual latents without disrupting their joint geometry. The two stages therefore share one relational object and differ only in which reference they point to: what SFT installs is exactly what RL protects.

To instantiate the RL phase of our framework, we introduce **GEPO** (**Ge**ometry-preserving **P**olicy **O**ptimization), a group-based policy optimization algorithm tailored to the hybrid discrete-continuous action space of our decoder. A key challenge in continuous-latent RL is that standard Gaussian policies operate in Euclidean space. This fundamentally mismatches the near-spherical geometry of RMSNorm'ed [Zhang and Sennrich, 2019] hidden states, inevitably pulling trajectories off the manifold established during SFT. GEPO elegantly resolves this by modeling continuous actions with a spherical von Mises–Fisher (vMF) policy that natively respects the decoder's geometry. Furthermore, it integrates the SFT-anchored regularizer directly into the RL objective, ensuring that reward-driven optimization and structural preservation are pursued *jointly* at every update.

Our main contributions are summarized as follows.

- **A geometry-consistent paradigm for latent visual reasoning.** We propose GELVR, a framework that establishes geometric consistency as a governing principle, unifying

geometry alignment in SFT and *geometry preservation* in RL by anchoring both stages to a common geometric scaffold.

- **GEPO, a geometry-preserving RL algorithm for hybrid action spaces.** We introduce GEPO, a GRPO-style algorithm tailored for hybrid discrete-continuous reasoning. GEPO resolves the latent-space mismatch via a spherical vMF policy, decouples discrete-continuous trust regions, and integrates an SFT-anchored regularizer to ensure reward optimization and structural preservation advance jointly rather than in tension.
- **Superiority in complex and high-resolution visual reasoning.** Extensive evaluations across multiple benchmarks demonstrate that GELVR significantly outperforms state-of-the-art latent-reasoning baselines, especially on tasks requiring fine-grained perception and high-resolution understanding. Rigorous ablations further confirm that our geometry-consistent components are indispensable for preventing semantic collapse and ensuring stable policy optimization.

2 Related Work

2.1 Explicit Visual Reasoning in Multimodal LLMs

Work on explicit visual reasoning in multimodal LLMs can be broadly divided into *thinking about images*, which performs chain-of-thought [Wei et al., 2022] entirely in text space, and *thinking with images*, which grants the model discrete visual actions during reasoning. Text-space methods verbalize step-by-step reasoning after a single visual encoding: early work fine-tunes on curated visual CoT data [Zhang et al., 2024, Xu et al., 2024], while more recent efforts reward outcome correctness or grounded behaviors via rule-based RL [Meng et al., 2025, Peng et al., 2025, Huang et al., 2025, Liu et al., 2025]. Because the image is consumed only once, however, details that are small or off-axis at encoding time cannot be re-inspected mid-chain. Tool-based methods close this loop by exposing zoom/crop utilities or iterative visual search [Wu and Xie, 2024, Hu et al., 2024, Zheng et al., 2025], and generative variants such as Janus-Pro [Chen et al., 2025] let a unified decoder render auxiliary images as intermediate thoughts. Both threads pay a similar price: tool-based methods are bounded by a fixed action interface and incur latency that grows with reasoning depth, while generative variants further tie the model's visual understanding to its rendering capacity. Yet human visual reasoning itself does not require re-rendering the image or verbalizing every perceptual cue, and many of the operations these routes exteriorize are arguably already implicit in a modern MLLM's attention stack.

2.2 Latent Visual Reasoning

Latent reasoning performs the chain of thought directly in the decoder's hidden space, bypassing token-level discretization and external rendering. In language-only LLMs, Coconut [Hao et al., 2024] feeds the last hidden state back as the next input, and CCoT [Cheng and Durme, 2024] compresses long reasoning chains into a small set of dense representations. Multimodal extensions insert dedicated latent slots into the decoder and supervise them against a visual target: Mirage [Yang et al., 2025] distills latent visual tokens from mean-pooled embeddings of auxiliary images, enabling "machine mental imagery" without rendering pixels; LVR [Li et al., 2025b] instead aligns each slot, via MSE, to the next image-patch hidden state from the same backbone ViT; SkiLa [Tong et al., 2025] interleaves text thoughts with sketch-in-latent segments MSE-aligned to a projected frozen-SigLIP2 encoding of the reference sketch, discarding the auxiliary encoder at inference; and Monet [Wang et al., 2025a] adopts a self-distillation route in which a frozen copy of the same backbone supplies hidden-state targets, followed by a Visual-Latent Policy Optimization (VLPO) RL stage.

In concurrent work, HyLaR [Cheng et al., 2026] introduces Decoupled Policy Optimization (DePO), an RL algorithm that models continuous latent actions with a von Mises–Fisher (vMF) policy and applies separate clipping ranges to text and latent updates. These latent visual reasoning methods differ in their visual targets and optimization schemes, but do not explicitly match pairwise relations among latent slots. Like HyLaR, we adopt a vMF policy to model continuous latent actions. Our framework, GELVR, further incorporates explicit relational regularization into both training stages: during SFT, it aligns the pairwise structure of predicted latents with that of visual targets; during RL, it regularizes deviations from the relational structure of trajectories produced by a frozen SFT model.

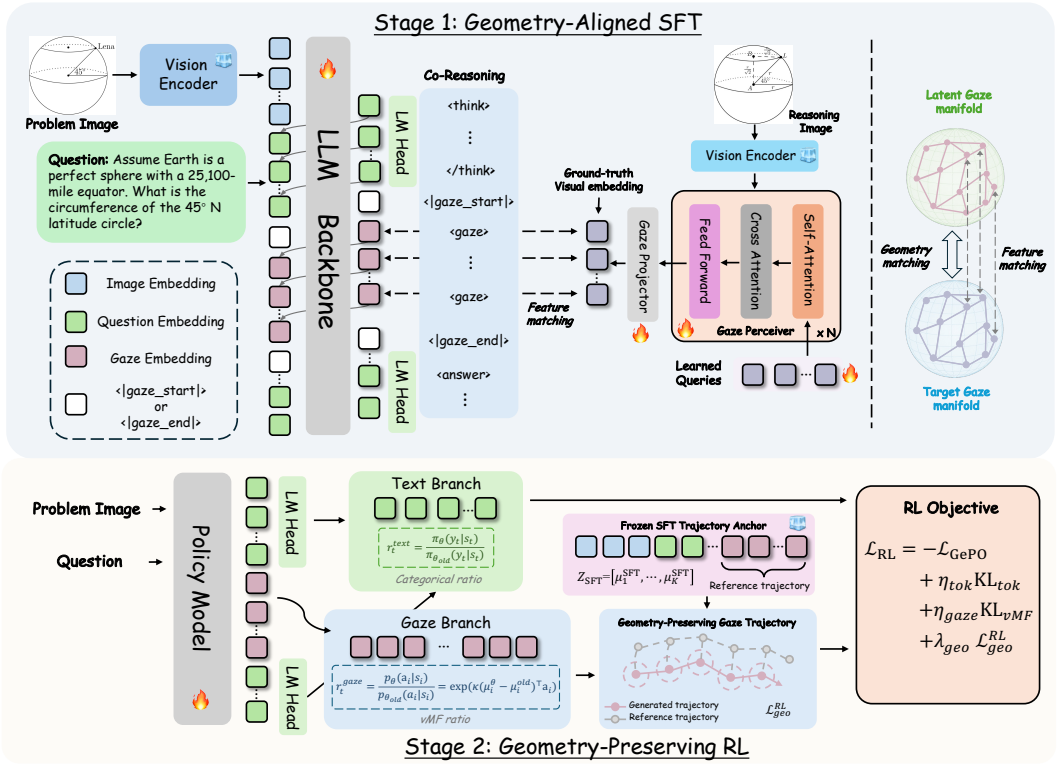

Figure 2: **Two-stage training pipeline of GELVR. (Top) Geometry-Aligned SFT.** The model interleaves discrete text and continuous latent gaze tokens. Generated gaze manifolds are aligned to target visual embeddings via a dual objective: point-wise MSE for features and Gram–Frobenius loss for structural geometry. **(Bottom) Geometry-Preserving RL (GEPO).** Policy optimization is decoupled into categorical (text) and spherical vMF (gaze) branches. To prevent representation drift, GEPO geometrically anchors the generated gaze trajectory to a frozen SFT reference, preserving pairwise relational structures during RL.

3 Method

In this section, we present **GELVR** (**Ge**ometry-Consistent **L**atent **V**isual **R**easoning), a novel multimodal Chain-of-Thought (CoT) method that enables MLLMs to combine discrete textual reasoning with a continuous latent visual reasoning trace within a single generation process. The central design principle of GELVR is to maintain a consistent geometric structure within the visual latent space. We first provide an overview of the GELVR method and its discrete–continuous decoding paradigm (§3.1). Subsequently, we detail our two-stage training methodology: §3.2 describes the SFT phase, which establishes this latent geometry through a targeted alignment objective; finally, §3.3 presents **GEPO** (**Ge**ometry-preserving **P**olicy **O**ptimization), an RL algorithm that further refines reasoning capabilities via task rewards while strictly preventing geometric drift.

3.1 Overview

As shown in Figure 2 (top), GELVR inherits the standard MLLM architecture—a vision encoder with a modality projector, and an LLM decoder. In a conventional MLLM, the decoder autoregressively emits discrete tokens through the language-modeling head, so even when the model needs to reason about a spatial relation or a fine-grained visual cue, its intermediate thoughts must first be serialized into words; the output space is constrained to the discrete vocabulary, limiting the model's ability to reason directly in the visual domain. To remove this constraint, we introduce two control tokens, `<|gaze_start|>` and `<|gaze_end|>`, which delimit a sequence of latent visual tokens (i.e. a gaze sgement). Upon emitting `<|gaze_start|>`, the model enters a gaze mode in which textual generation is suspended and a sequence of continuous hidden states—which we call *gaze tokens*—is

produced as internal visual reasoning signals; these hidden states are *not* mapped through the LM head and are never materialized as pixels, so gaze tokens remain purely latent. Once <|gaze_end|> is emitted, ordinary text decoding resumes. A full reasoning trace therefore interleaves discrete textual steps with continuous gaze segments, producing what can be viewed as a computational analogue of human visual thinking: the model alternates between verbalizing what it knows and silently "looking" at what it still needs to inspect, all within a single autoregressive pass and without requiring external reasoning images as inputs or performing explicit pixel-level generation.

Merely emitting visual tokens, however, does not guarantee coherent visual reasoning. Existing training recipes supervise each visual token only against its own target, treating the visual segment as a bag of independent points rather than a structured whole. The *relational geometry* among the tokens—which visual tokens should be close on the manifold and which should be far—never becomes an explicit training signal. During the RL phase, this vulnerability is exacerbated: scalar task reward pulls individual tokens in arbitrary directions that maximize the reward, leaving the off-diagonal relational structure free to drift and, in the worst case, to collapse. For the gaze segment to serve as a coherent visual thinking trace, its relational structure must be explicitly installed during supervision and actively protected under reward. These two requirements motivate the two-stage training pipeline of GELVR (Figure 2). A geometry-aligned SFT phase installs the relational structure of the gaze trace, supervised by an auxiliary visual branch that enforces a dual objective—point-wise feature matching and structural geometry matching. This auxiliary branch is safely discarded at inference(§3.2). Subsequently, the RL phase anchors this structure against reward-driven drift via our proposed algorithm, **Ge**ometry-preserving **P**olicy **O**ptimization (GEPO) (§3.3). Together, they realize the central promise of GELVR: the relational geometry of the gaze trace, once established, remains geometrically consistent across the entire training trajectory.

3.2 Geometry-Aligned Supervised Fine-Tuning

The supervised fine-tuning (SFT) stage equips the decoder with the ability to enter gaze mode on demand and to populate each gaze segment with semantically meaningful latent tokens. Because the gaze tokens never pass through the LM head, they cannot be supervised by a next-token likelihood; instead, we supervise them against visual target embeddings extracted from a region of interest relevant to the query. The model is then jointly trained to predict the next text token and to reconstruct these target embeddings at the gaze positions, so that the two modalities are learned under a single autoregressive objective.

Ground-truth target extraction for gaze tokens. To produce ground-truth visual embeddings for the K gaze tokens, we introduce an auxiliary module, the *Gaze Perceiver*, that is active only during SFT training (Figure 2, top-right). Inspired by Q-Former [Li et al., 2023], the Gaze Perceiver distills an image into a small set of query-relevant visual embeddings through a group of learnable queries into the decoder. Concretely, a frozen SigLIP2 encoder [Tschannen et al., 2025] processes each ground-truth reasoning image into $P{=}729$ patch tokens of dimension D_v. We aggregate these using K learnable queries ($\in \mathbb{R}^{K\times D_v}$) processed through $N{=}2$ Gaze Perceiver blocks. By sequentially applying inter-query self-attention, patch-to-query cross-attention, and a feed-forward layer, the Perceiver extracts localized visual evidence, producing K query-aggregated embeddings ($\in \mathbb{R}^{K\times D_v}$). A linear *Gaze Projector* maps these embeddings to the LLM's hidden dimension D, yielding the target sequence $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_K] \in \mathbb{R}^{K\times D}$. Within the gaze segment, the decoder autoregressively generates K <gaze> tokens, and their corresponding hidden states $\mathbf{z}_i$ are supervised directly against $\mathbf{t}_i$. Crucially, $\mathbf{T}$ is computed end-to-end without stopping gradients; this allows the gaze loss to backpropagate into both the decoder and the visual modules, actively pulling the visual targets into the LLM's native semantic space.

Training objective. Let $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_K] \in \mathbb{R}^{K\times D}$ denote the gaze-position hidden states produced by the decoder, which live in the same decoder hidden space as the ground-truth targets $\mathbf{T}$ above. The model is jointly trained to predict the next discrete text token and to reconstruct the continuous ground-truth embeddings at the gaze positions. We apply the standard autoregressive cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$ over the text tokens of the reasoning trace and answer. Simultaneously, a *feature-matching* term pulls each decoder gaze state $\mathbf{z}_i$ toward its corresponding target embedding $\mathbf{t}_i$,

$$\mathcal{L}_{\mathrm{feat}} \;=\; \frac{1}{K}\sum_{i=1}^{K} \big\|\mathbf{z}_i - \mathbf{t}_i\big\|_2^2. \tag{1}$$

Prior latent visual reasoning methods such as LVR [Li et al., 2025b], SkiLa [Tong et al., 2025], and HyLaR [Cheng et al., 2026] rely on the MSE feature-matching objective $\mathcal{L}_{\text{feat}}$ alone, which supervises each visual token independently and leaves the pairwise relational geometry among the K tokens free to drift—meaning a model can achieve a seemingly low $\mathcal{L}_{\text{feat}}$ error while still suffering from severe structural distortions in the latent space. We address this gap by adding an explicit *geometry-matching* term that directly aligns the pairwise relational structure of the gaze sequence with that of the target sequence. Denoting the pairwise cosine matrix of a unit sequence $\mathbf{X}$ as $\text{Gram}_{ij}(\mathbf{X}) = \langle \bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j \rangle$ where $\bar{\mathbf{x}}_i = \mathbf{x}_i / \|\mathbf{x}_i\|$, the term is the Frobenius distance between the two Gram matrices,

$$\mathcal{L}_{\text{geo}}^{\text{SFT}} \;=\; \frac{1}{K(K-1)} \sum_{i \neq j} \big(\text{Gram}_{ij}(\mathbf{Z}) - \text{Gram}_{ij}(\mathbf{T})\big)^2. \tag{2}$$

For feature matching, both the predicted and target embeddings participate in backpropagation. For geometry matching, we detach the target relation matrix and use it as a structural reference. The overall SFT objective combines the three components with scalar weights: $\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{CE}} + \alpha \mathcal{L}_{\text{feat}} + \beta \mathcal{L}_{\text{geo}}^{\text{SFT}}$. While feature matching anchors each token to its absolute visual semantics, geometry matching preserves the structural topology of the target manifold. Together, these dual signals align the gaze representations not just as isolated points, but as a geometrically coherent whole.

3.3 Geometry-Preserving Reinforcement Learning

While the SFT stage establishes a geometrically structured gaze manifold, RL task rewards can easily distort it by pulling individual tokens in unconstrained directions, ignoring their relational geometry. Prior methods fail to address this geometry–reward conflict. LVR [Li et al., 2025b] applies policy gradients exclusively to discrete text tokens, leaving the continuous latents to rely on weak, indirect updates through downstream text likelihoods. Conversely, VLPO in Monet [Wang et al., 2025a] optimizes latents directly via a Gaussian policy in $\mathbb{R}^D$. In concurrent work, HyLaR [Cheng et al., 2026] introduces Decoupled Policy Optimization (DePO), combining vMF-inspired latent policy optimization with separate clipping ranges for text and latent actions. However, these objectives do not explicitly constrain changes in the pairwise relational structure among latent slots, motivating an explicit geometry-preservation objective during RL.

Building on DePO [Cheng et al., 2026], our **GePO** (**Ge**ometry-preserving **P**olicy **O**ptimization) treats text tokens and continuous gaze latents as two action-space projections of a *single* policy π_θ: a categorical softmax head on the vocabulary $\mathcal{V}$, and a spherical von Mises–Fisher (vMF) head on the unit hypersphere S^{D-1}. The unified log-probability for this theoretical model is stated in Appendix A.1. Crucially, to prevent reward-driven drift of the latent trajectory, we further introduce an *SFT-anchored trajectory Gram regularizer* that preserves its pairwise cosine structure. As illustrated in Figure 2 (bottom), GePO applies the same clipped PPO surrogate to both action spaces with separate clip ranges, and adds the Gram anchor on top.

Hybrid action space: vMF gaze policy. We first describe the gaze branch under a normalized vMF model. For gaze position i, let s_i denote the state and $\boldsymbol{\mu}_i^\theta, \boldsymbol{\mu}_i^{\text{old}} \in S^{D-1}$ be the current and cached old vMF mean directions. The gaze head predicts $\boldsymbol{\mu}_i^\theta = L_2\text{-normalize}(\text{RMSNorm}(\mathbf{h}_i))$ with a fixed shared concentration κ. Under this model, we evaluate both policies at the cached rollout direction $\mathbf{a}_i \equiv \boldsymbol{\mu}_i^{\text{old}}$; hence the shared vMF normalizer cancels and the log-ratio becomes

$$\log \frac{\pi_\theta(\mathbf{a}_i \mid s_i)}{\pi_{\text{old}}(\mathbf{a}_i \mid s_i)} = \kappa\big(\boldsymbol{\mu}_i^\theta - \boldsymbol{\mu}_i^{\text{old}}\big)^\top \mathbf{a}_i = \kappa\big(\cos(\boldsymbol{\mu}_i^\theta, \boldsymbol{\mu}_i^{\text{old}}) - 1\big). \tag{3}$$

This theoretical mode-referenced ratio measures only the angular deviation from the old rollout direction; see Appendix A.2 for the derivation. The K gaze slots are open-loop, each consuming the same placeholder `<gaze>`, so all directions are produced in one deterministic forward pass. In practice, we retain the magnitudes of latent representations and compute gaze ratios from differences of scaled inner-product scores under the same stored rollout context. These scores incorporate both directional and magnitude information. We use separate clipping ranges for text and gaze actions, with a tighter interval for the gaze branch, to limit surrogate gains from large ratio changes. Clipping is applied to these practical score ratios; implementation details are provided in Appendix A.4.

SFT-anchored trajectory Gram regularizer. To make geometry preservation operational, after SFT we cache once, for every RL prompt, a deterministic mode rollout $\mathbf{Z}_{\text{SFT}} = [\boldsymbol{\mu}_1^{\text{SFT}}, \ldots, \boldsymbol{\mu}_K^{\text{SFT}}] \in$

$(S^{D-1})^K$ under the frozen SFT policy together with its Gram matrix $\mathrm{Gram}(\mathbf{Z}_{\mathrm{SFT}})$; this *trajectory anchor* takes no gradient. Per-step clipping acts on individual action-score ratios, but does not explicitly constrain the $\mathcal{O}(K^2)$ pairwise relations encoded by the off-diagonal entries of the Gram matrix. Let $\mathbf{M}_\theta = [\boldsymbol{\mu}_1^\theta, \ldots, \boldsymbol{\mu}_K^\theta]$ be the online trajectory of mean directions on the same prompt. We close the gap with a trajectory-level regularizer

$$\mathcal{L}_{\mathrm{geo}}^{\mathrm{RL}} \;=\; \frac{1}{K(K-1)} \sum_{i \neq j} \Big((\boldsymbol{\mu}_i^\theta)^\top \boldsymbol{\mu}_j^\theta \;-\; (\boldsymbol{\mu}_i^{\mathrm{SFT}})^\top \boldsymbol{\mu}_j^{\mathrm{SFT}} \Big)^2, \tag{4}$$

which reuses the Frobenius–Gram form of (2) with the external target $\mathbf{T}$ replaced by $\mathbf{Z}_{\mathrm{SFT}}$, and is a deterministic, differentiable function of θ through $\boldsymbol{\mu}_i^\theta$. SFT thus performs a *two-manifold alignment*, while RL performs a *single-manifold preservation*: the trajectory is free to move on the sphere but its pairwise cosine structure is anchored to the shape SFT had learned.

GEPO objective. We use standard GRPO group-relative advantage estimation [Shao et al., 2024] broadcast to every action position of each rollout. Let the index sets $\mathcal{T}$ and $\mathcal{S}$ denote, respectively, the text and gaze positions of a rollout (distinct from the bold-upright SFT target sequence $\mathbf{T}$ in (2)), and let π_{ref} be the frozen SFT policy. Both $\mathcal{L}_{\mathrm{tok}}$ and $\mathcal{L}_{\mathrm{gaze}}$ are clipped PPO surrogates of the standard form $\mathbb{E}\big[\min\big(r_t\,A_t,\ \mathrm{clip}(r_t,\,1-\epsilon,\,1+\epsilon)\,A_t\big)\big]$, where r_t is a position-specific ratio between the current and old policies, A_t is the rollout's single group-relative advantage broadcast uniformly to every action position of that rollout and ϵ is a per-branch clipping range. The two branches differ only in *which r_t and which ϵ they use*. On text positions $\mathcal{T}$, $\mathcal{L}_{\mathrm{tok}}$ uses the usual categorical ratio $r_t^{\mathrm{tok}} = \pi_\theta(y_t \mid s_t)/\pi_{\mathrm{old}}(y_t \mid s_t)$ with clip ϵ_{tok}. On gaze positions $\mathcal{S}$, $\mathcal{L}_{\mathrm{gaze}}$ uses the practical inner-product surrogate ratio r_i^{gaze} described above, with a tighter clip ϵ_{gaze}; its computation is detailed in Appendix A.4. The two branches are combined into a single decoupled surrogate,

$$\mathcal{L}_{\mathrm{GePO}}(\theta) \;=\; \mathcal{L}_{\mathrm{tok}}\big(\theta \mid \mathcal{T}; \epsilon_{\mathrm{tok}}\big) \;+\; \alpha\, \mathcal{L}_{\mathrm{gaze}}\big(\theta \mid \mathcal{S}; \epsilon_{\mathrm{gaze}}\big), \tag{5}$$

and the overall GEPO loss adds per-branch KL regularization and our trajectory Gram anchor,

$$\mathcal{L}_{\mathrm{RL}}(\theta) \;=\; -\mathcal{L}_{\mathrm{GePO}}(\theta) \;+\; \eta_{\mathrm{tok}}\,\mathrm{KL}_{\mathrm{tok}} \;+\; \eta_{\mathrm{gaze}}\,\mathrm{KL}_{\mathrm{vMF}} \;+\; \lambda_{\mathrm{geo}}\,\mathcal{L}_{\mathrm{geo}}^{\mathrm{RL}}. \tag{6}$$

$\mathrm{KL}_{\mathrm{tok}}$ is the standard sample-based categorical KL to π_{ref}, averaged over $\mathcal{T}$. Writing $\boldsymbol{\mu}_i^{\mathrm{ref}}$ for the mean direction of π_{ref} re-evaluated at the *current* state s_i—note that this is distinct from the cached offline anchor $\boldsymbol{\mu}_i^{\mathrm{SFT}}$ used in $\mathcal{L}_{\mathrm{geo}}^{\mathrm{RL}}$, which is produced once under the SFT prompt before RL starts—the vMF KL between current and reference policies admits, under shared fixed κ, a closed-form scaled cosine distance $W_\kappa\,(1 - (\boldsymbol{\mu}_i^\theta)^\top \boldsymbol{\mu}_i^{\mathrm{ref}})$, where the scalar W_κ is a constant derived from κ and D; in (6) we average this per-position term over $\mathcal{S}$, so that η_{gaze} and W_κ jointly control its effective strength. The full derivation of W_κ and all implementation details are given in Appendix A.4. The last term $\lambda_{\mathrm{geo}}\mathcal{L}_{\mathrm{geo}}^{\mathrm{RL}}$ is the single geometry-preserving ingredient of GEPO.

4 Experiments

4.1 Experimental Settings

Training Data. Following Tong et al. [2025], we filter the Zebra-CoT dataset [Li et al., 2025a] for the SFT stage, yielding 101K SFT samples across science, 2D reasoning, and visual logic. For RL, we aggregate prompts from DeepEyes [Zheng et al., 2025] and Thyme [Zhang et al., 2025b] to focus on high-resolution search and fine-grained perception.

Evaluation Benchmarks. We evaluate on four primary benchmarks targeting fine-grained perception and reasoning: V*Bench [Wu and Xie, 2024] (Visual Search), HR-Bench-4K/8K [Wang et al., 2025b] (High-Resolution), and MME-RealWorld-Lite [Zhang et al., 2025a] (Perception and Reasoning). We additionally evaluate generalization on MMVP [Tong et al., 2024], MMStar [Chen et al., 2024], and BLINK [Fu et al., 2024], and hallucination robustness on HallusionBench [Guan et al., 2024], with results in Table 2.

Baselines. We benchmark GELVR against four categories of representative baselines: **(1) Proprietary frontier MLLMs**—GPT-4o [OpenAI, 2024]; **(2) Open-source general MLLMs**—Qwen2.5-VL-7B [Bai et al., 2025], and InternVL3-8B [Zhu et al., 2025]; **(3) Think-with-images models**—DeepEyes [Zheng et al., 2025], Thyme [Zhang et al., 2025b], and Mini-o3 [Lai et al., 2025]; **(4) latent visual reasoning models**—LVR [Li et al., 2025b], SkiLa [Tong et al., 2025], Monet [Wang et al., 2025a], and HyLaR [Cheng et al., 2026].

Table 1: **Main results on real-world perception and reasoning benchmarks.** The best value in each column is highlighted in **bold**, including ties. "$*$" denotes paper-reported results; "†" denotes our evaluations of released weights under identical settings. Mini-o3 uses its official multi-turn setting (Avg@32 on V* and Avg@8 on HR-Bench). "–" denotes unavailable results on the listed benchmark splits; Mini-o3 does not specify the Lite split for its MME-RealWorld result.

Model	V*Bench			HR-Bench-4K			HR-Bench-8K			MME-RealWorld-Lite		
	Overall	Attribute	Spatial	Overall	FSP	FCP	Overall	FSP	FCP	Overall	Reasoning	Perception
Proprietary Model												
GPT-4o* [OpenAI, 2024]	67.5	72.2	60.5	59.0	70.0	48.0	55.5	62.0	49.0	52.0	48.3	54.4
Open-Source General MLLMs												
InternVL3-8B* [Zhu et al., 2025]	70.20	67.80	73.70	70.00	78.80	61.30	69.30	78.80	**59.80**	48.60	44.80	51.00
Qwen2.5-VL-7B* [Bai et al., 2025]	76.44	77.39	75.00	68.00	80.25	55.75	63.75	73.75	53.75	45.75	39.73	49.62
Think-with-Images Models												
Thyme* [Zhang et al., 2025b]	82.20	83.50	80.30	77.00	91.00	63.00	72.00	86.50	57.50	55.20	49.10	59.10
DeepEyes* [Zheng et al., 2025]	83.25	84.35	81.58	71.25	83.75	58.75	65.13	77.00	53.25	54.28	50.53	56.53
Mini-o3* [Lai et al., 2025]	**88.20**	–	–	**77.50**	–	–	**73.30**	–	–	–	–	–
Latent Visual Reasoning Models												
LVR* [Li et al., 2025b]	81.70	–	–	69.60	–	–	66.10	–	–	49.10	43.50	52.70
SkiLa* [Tong et al., 2025]	84.30	–	–	72.00	–	–	66.50	–	–	54.10	50.20	56.60
SkiLa† [Tong et al., 2025]	79.05	80.00	77.63	71.88	77.50	**66.25**	66.25	76.25	56.25	53.67	50.00	56.03
Monet* [Wang et al., 2025a]	83.25	83.48	82.89	71.00	85.25	56.75	68.00	79.75	56.25	55.50	**51.07**	58.34
Monet† [Wang et al., 2025a]	81.68	82.61	80.26	69.00	81.25	56.75	64.37	74.25	54.49	55.34	50.80	58.25
HyLaR* [Cheng et al., 2026]	83.77	82.61	85.53	75.00	**93.75**	56.25	70.50	88.25	52.75	–	–	–
GELVR-SFT (Ours)	82.72	84.34	80.26	73.63	93.00	54.25	67.25	82.25	52.25	52.63	49.33	54.75
GELVR-SFT+GEPO (Ours)	85.86	**85.22**	**86.84**	77.13	**93.75**	60.50	70.63	**88.50**	52.75	**55.81**	50.67	**59.11**
Δ *vs. Qwen2.5-VL-7B*	+9.42	+7.83	+11.84	+9.13	+13.50	+4.75	+6.88	+14.75	-1.00	+10.06	+10.94	+9.49

4.2 Main Results

The benchmarks detailed in Table 1 feature ultra-high-resolution, visually cluttered environments where target regions constitute a minuscule fraction of the image. Whereas general MLLMs are notoriously susceptible to severe hallucinations when isolating these elusive targets, GELVR decisively overcomes this perceptual bottleneck. As demonstrated across all four benchmarks, our model delivers consistent performance gains; remarkably, on several sub-metrics, it matches or even outperforms heavy agentic pipelines that strictly rely on external cropping tools. Quantitatively, GELVR substantially elevates the baseline capabilities of its Qwen2.5-VL-7B initialization [Bai et al., 2025], posting robust increases of **+9.42** on V* [Wu and Xie, 2024] and **+9.13** on HR-Bench-4K [Wang et al., 2025b], with a further **+10.06** on MME-RealWorld-Lite [Zhang et al., 2025a]; the gains concentrate on spatial sub-metrics—V* Spatial (+11.84), HR-Bench FSP (+13.50 at 4K, +14.75 at 8K), and MME-RealWorld Reasoning (+10.94)—exactly where the geometry-alignment objective in §3.2 is expected to act. Conversely, we attribute the marginal -1.00 drop on HR-Bench-8K FCP to an acceptable "alignment tax": hyper-focusing on local regions for extreme single-target localization (driving the +14.75 FSP surge) inevitably sacrifices minor global context needed for macro-level multi-object relations. Among the compared latent visual reasoning models, GELVR achieves the highest overall scores on the four benchmarks. To rule out pipeline artefacts, we re-evaluate two open-source latent baselines (SkiLa†, Monet†) from their released weights under our identical protocol. On V*, GELVR (85.86%) outperforms SkiLa† (79.05%) [Tong et al., 2025] and Monet† (81.68%) [Wang et al., 2025a] by margins of 6.81% and 4.18%, respectively. This superiority extends to high-resolution evaluations: on HR-Bench-4K it exceeds SkiLa† and Monet† by 5.25% and 8.13%; on HR-Bench-8K by 4.38% and 6.26%; and on MME-RealWorld-Lite by 2.14 and 0.47 points, respectively. The gap is stable across visual-search, ultra-high-resolution, and real-world-reasoning splits, suggesting that the improvement stems from the geometry-consistent mechanism rather than from any single benchmark. Compared with the paper-reported HyLaR results [Cheng et al., 2026], GELVR scores 2.09, 2.13, and 0.13 points higher on V*, HR-Bench-4K, and HR-Bench-8K, respectively. Mini-o3's official multi-turn results [Lai et al., 2025] are higher on these three benchmarks; these paper-reported comparisons use different inference and evaluation settings.

Generalization across benchmarks. Table 2 reports results on MMVP, MMStar, and BLINK. Both SFT and GePO improve all three scores over the preceding stage. Relative to Qwen2.5-VL-7B, the full pipeline gains 8.33, 2.63, and 3.21 points, respectively, achieving the highest MMVP score among the compared models. Thyme and HyLaR obtain the highest MMStar and BLINK scores, respectively. On HallusionBench, the score improves from 56.57 for Qwen2.5-VL-7B to 60.47 after SFT and 63.46 after GePO, a total gain of 6.89 points (Table 2).

Table 2: **Generalization and hallucination robustness.** We report both training stages of GELVR on Qwen2.5-VL-7B. The highest reported value in each column is in **bold**. HyLaR* and the HallusionBench baseline scores are taken from Cheng et al. [2026]; these paper-reported results may use different evaluation settings. "–" denotes an unreported result.

Model	MMVP	MMStar	BLINK	HallusionBench
Open-Source General MLLMs				
Qwen2.5-VL-7B [Bai et al., 2025]	65.67	59.70	53.60	56.57
Think-with-Images Models				
Thyme [Zhang et al., 2025b]	–	**65.90**	56.10	55.60
DeepEyes [Zheng et al., 2025]	70.00	58.73	51.08	62.57
Latent Visual Reasoning Models				
LVR [Li et al., 2025b]	64.00	57.93	53.60	**65.19**
Monet [Wang et al., 2025a]	68.00	60.33	50.71	56.36
HyLaR* [Cheng et al., 2026]	73.67	62.00	**57.14**	63.68
GELVR-SFT (Ours)	71.67	60.60	55.08	60.47
GELVR-SFT+GEPO (Ours)	**74.00**	62.33	56.81	63.46

Table 3: **Ablation on geometry-consistent components and RL algorithms.** Evaluates the effects of $\mathcal{L}_{\text{geo}}^{\text{SFT}}$, RL policy class, and $\mathcal{L}_{\text{geo}}^{\text{RL}}$. All RL runs share the same SFT initialization.

SFT Phase		RL Phase			V*	HR-4K	HR-8K	Avg.
$\mathcal{L}_{\text{feat}}$	$\mathcal{L}_{\text{geo}}^{\text{SFT}}$ (β)	Algorithm	Policy Type	$\mathcal{L}_{\text{geo}}^{\text{RL}}$ (λ)				
Step 1: Geometry alignment in SFT (no RL).								
✓	×	–	–	–	80.63	71.75	66.86	73.08
✓	✓	–	–	–	**82.72**	**73.63**	**67.25**	**74.53**
Step 2: Choice of RL algorithm (initialized from optimal SFT).								
✓	✓	GRPO [Shao et al., 2024]	Categorical	×	80.10	73.63	67.13	73.62
✓	✓	VLPO [Wang et al., 2025a]	Gaussian	×	82.19	74.97	68.06	75.07
Step 3: Geometry preservation in RL (our full pipeline).								
✓	✓	GEPO (ours)	Spherical vMF	×	84.82	76.25	70.50	77.19
✓	✓	**GEPO (ours)**	**Spherical vMF**	✓	**85.86**	**77.13**	**70.63**	**77.87**

4.3 Ablation Studies

Ablation on Geometry-Consistent Components and RL Algorithms. Table 3 decomposes our pipeline. **(1) SFT Geometry:** Adding $\mathcal{L}_{\text{geo}}^{\text{SFT}}$ to $\mathcal{L}_{\text{feat}}$ increases the average score by $+1.45$. Gains are largest on V* and HR-Bench-4K, confirming that supervising pairwise relations directly benefits fine-grained grounding. **(2) Policy class:** Starting from identical SFT checkpoints, plain GRPO [Shao et al., 2024] (Euclidean categorical) regresses below SFT, while VLPO [Wang et al., 2025a] offers marginal gains ($+0.54$). In contrast, GEPO's spherical vMF policy surges to 77.19 ($+2.12$ over VLPO). **(3) RL Preservation:** Enabling $\mathcal{L}_{\text{geo}}^{\text{RL}}$ adds a final $+0.68$ average gain. The three geometric pieces ($\mathcal{L}_{\text{geo}}^{\text{SFT}}$, vMF, $\mathcal{L}_{\text{geo}}^{\text{RL}}$) are additive rather than redundant, collectively resolving complementary issues of representation drift and reward-driven distortion.

Ablation on Latent Budget. We study how the latent budget interacts with training by varying $K_{\text{train}} \in \{8, 16, 24, 32\}$ and sweeping $K_{\text{test}} \in \{8, \ldots, 64\}$; Figure 3 reports V* Overall after SFT and HR-Bench-4K Overall after GEPO RL. First, latent thinking consistently improves over the no-latent Qwen2.5-VL-7B baseline: even latent-8 reaches 79.58 on V* at K_{test}=8, above the 76.44 baseline. Second, SFT models over-specialize to their training horizon: performance peaks near $K_{\text{test}} \approx K_{\text{train}}$ and degrades under extrapolation, e.g., latent-16 drops from 82.72 to 78.53 when K_{test} increases to 64. Third, GEPO with $\mathcal{L}_{\text{geo}}^{\text{RL}}$ substantially reduces this decay: after RL, performance remains stable across budgets, and longer-horizon models perform best. For example, latent-32 retains 76.25 at K_{test}=64 after reaching 77.13 at K_{test}=32, while its SFT counterpart drops by more than 5 points over the same interval. Together with Table 3, these results support our mechanism: $\mathcal{L}_{\text{geo}}^{\text{SFT}}$ establishes stable latent geometry, $\mathcal{L}_{\text{geo}}^{\text{RL}}$ preserves it under reward optimization, and the spherical vMF policy makes this preservation effective in the decoder-facing latent manifold.

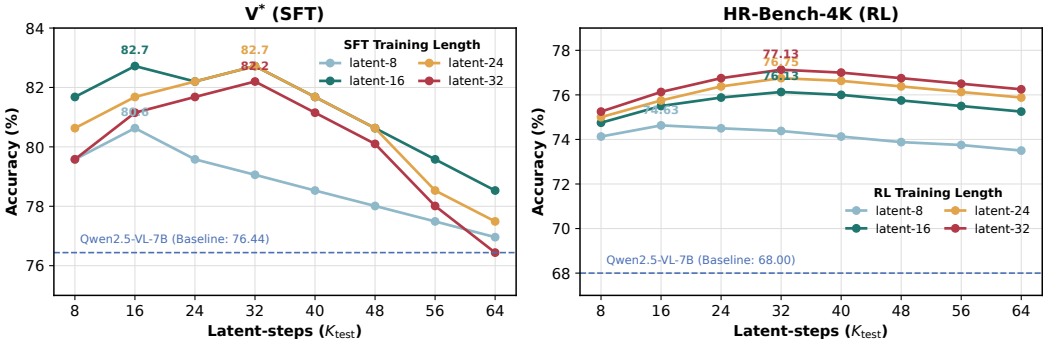

Figure 3: **Latent-step extrapolation.** Models trained with fixed latent budgets (K_{train}) are evaluated across varying inference lengths (K_{test}). **Left (V*, SFT):** models overfit to their training horizon, with sharp decay when $K_{\text{test}} \gg K_{\text{train}}$. **Right (HR-Bench-4K, RL):** reinforcement learning resolves this boundary limitation, enabling robust test-time compute scaling.

5 Conclusion

We presented **GELVR**, a framework that establishes geometric consistency as a governing principle for latent visual reasoning. Addressing the structural fragmentation in prior isolated supervision methods, GELVR utilizes a common geometric scaffold to unify *geometry alignment* in SFT with *geometry preservation* in RL. This paradigm is operationalized via **GEPO**, a policy optimization algorithm tailored for hybrid action spaces that employs a spherical vMF policy and decouples trust regions to advance reward optimization and structural integrity jointly. Extensive experiments and ablations across multiple benchmarks validate GELVR's superiority in high-resolution perception and confirm the necessity of each geometry-consistent component. Looking ahead, we plan to investigate more expressive non-Euclidean geometric constraints to better capture the intricate topological structures of high-dimensional visual reasoning manifolds.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *CoRR*, abs/2502.13923, 2025. doi: 10.48550/arXiv.2502.13923.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *CoRR*, abs/2501.17811, 2025. doi: 10.48550/arXiv.2501.17811.

Jeffrey Cheng and Benjamin Van Durme. Compressed chain of thought: Efficient reasoning through dense representations. *CoRR*, abs/2412.13171, 2024. doi: 10.48550/arXiv.2412.13171.

Tao Cheng, Shi-Zhe Chen, Hao Zhang, Yixin Qin, Jinwen Luo, and Zheng Wei. HyLaR: Hybrid latent reasoning with decoupled policy optimization. In *European Conference on Computer Vision (ECCV)*, 2026. doi: 10.1007/978-3-032-37592-6_35.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. *CoRR*, abs/2404.12390, 2024. doi: 10.48550/arXiv.2404.12390.

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. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385, 2024.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *CoRR*, abs/2412.06769, 2024. doi: 10.48550/arXiv.2412.06769.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2024)*, 2024. URL `http://papers.nips.cc/paper_files/paper/2024/hash/fb82011040977c7712409fbdb5456647-Abstract-Conference.html`.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-R1: Incentivizing reasoning capability in multimodal large language models. *CoRR*, abs/2503.06749, 2025. doi: 10.48550/arXiv.2503.06749.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. *arXiv preprint arXiv:2509.07969*, 2025. URL `https://arxiv.org/abs/2509.07969`.

Ang Li et al. Zebra-cot: A dataset for interleaved vision language reasoning. *CoRR*, abs/2507.16746, 2025a. doi: 10.48550/arXiv.2507.16746.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. *CoRR*, abs/2509.24251, 2025b. doi: 10.48550/arXiv.2509.24251.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, pages 19730–19742. PMLR, 2023.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual reinforcement fine-tuning. *CoRR*, abs/2503.01785, 2025. doi: 10.48550/arXiv.2503.01785.

Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. John Wiley & Sons, 2000.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. MM-Eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *CoRR*, abs/2503.07365, 2025. doi: 10.48550/arXiv.2503.07365.

OpenAI. GPT-4o system card. *CoRR*, abs/2410.21276, 2024. doi: 10.48550/arXiv.2410.21276.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. LMM-R1: Empowering 3B LMMs with strong reasoning abilities through two-stage rule-based RL. *CoRR*, abs/2503.07536, 2025. doi: 10.48550/arXiv.2503.07536.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/arXiv.2402.03300.

Jintao Tong, Jiaqi Gu, Yujing Lou, Lubin Fan, Yixiong Zou, Yue Wu, Jieping Ye, and Ruixuan Li. Sketch-in-latents: Eliciting unified reasoning in MLLms. *CoRR*, abs/2512.16584, 2025. doi: 10.48550/arXiv.2512.16584.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, 2024. doi: 10.48550/arXiv.2401.06209.

Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *CoRR*, abs/2502.14786, 2025. doi: 10.48550/arXiv.2502.14786.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond images and language. *CoRR*, abs/2511.21395, 2025a. doi: 10.48550/arXiv.2511.21395.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2025)*, 2025b. doi: 10.48550/arXiv.2408.15556.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. URL `http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html`.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, pages 13084–13094. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01243.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. LLaVA-CoT: Let vision language models reason step-by-step. *CoRR*, abs/2411.10440, 2024. doi: 10.48550/arXiv.2411.10440.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *CoRR*, abs/2506.17218, 2025. doi: 10.48550/arXiv.2506.17218.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *CoRR*, abs/1910.07467, 2019. doi: 10.48550/arXiv.1910.07467.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. MME-RealWorld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations (ICLR)*, 2025a.

Yi-Fan Zhang et al. Thyme: Think beyond images. *CoRR*, abs/2508.11630, 2025b. doi: 10.48550/arXiv.2508.11630.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024, 2024. URL `https://openreview.net/forum?id=y1pPWFVfvR`.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. *CoRR*, abs/2505.14362, 2025. doi: 10.48550/arXiv.2505.14362.

Jinguo Zhu et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479, 2025. doi: 10.48550/arXiv.2504.10479.

A Derivation of the vMF Policy Optimization

For completeness, we present the normalized hybrid policy formulation and mode-referenced density ratio following DePO [Cheng et al., 2026], together with the standard shared-concentration vMF KL identity. Our additional SFT-anchored trajectory Gram regularizer is described in §3.3. §A.1 sets up the unified policy log-probability over the hybrid discrete–continuous action space, which is the theoretical basis on which both branches of GEPO become a single policy-optimization algorithm. §A.2 derives the mode-referenced density ratio Eq. (3). §A.3 derives the closed-form vMF KL used as the gaze-branch trust region. §A.4 then fills in the remaining GEPO details—advantage estimator, reward, caching, text-branch KL, algorithm, and hyperparameters—that the main-paper section omits for space.

A.1 Unified policy log-probability over the hybrid action space

A common source of confusion in latent-reasoning RL is whether text tokens and continuous latents are optimized under one policy or two. Under this formulation, both are modeled by a single policy π_θ that differs only in which action space its head operates on at each step. At time t, the action is either a discrete vocabulary token $a_t \in \mathcal{V}$ or a continuous direction $a_t \in \mathcal{S}^{D-1}$, and the log-probability is

$$\log \pi_\theta(a_t \mid s_t) \;=\; \begin{cases} \log \operatorname{softmax}\big(\mathbf{W}_m \, \mathbf{h}_t^\theta\big)_{a_t}, & a_t \in \mathcal{V}, \\ \log C_D(\kappa) \;+\; \kappa \, \big(\boldsymbol{\mu}_t^\theta\big)^\top \tilde{\mathbf{z}}_t, & a_t \in \mathcal{S}^{D-1}, \end{cases} \tag{7}$$

where $\mathbf{W}_m$ is the LM head, $\mathbf{h}_t^\theta$ is the decoder hidden state at position t, $\boldsymbol{\mu}_t^\theta = L_2\text{-normalize}(\mathrm{RMSNorm}(\mathbf{h}_t^\theta))$ is the current-policy direction on the unit hypersphere, $\tilde{\mathbf{z}}_t$ is the latent direction stored from the old-policy rollout (see §A.2 for its role), κ is a fixed shared concentration, and $C_D(\kappa)$ is the vMF normalization constant in dimension D.

Two points deserve emphasis. First, both cases in Eq. (7) are parameterized by the *same* θ through the backbone that produces $\mathbf{h}_t^\theta$; the only difference is whether the head is a categorical softmax over $\mathcal{V}$ or a vMF density on $\mathcal{S}^{D-1}$. In this sense text and gaze are the two action-space branches of a single policy, not two independent policies. Second, because κ is fixed and shared between the current and old policies, the additive term $\log C_D(\kappa)$ cancels whenever the same action a_t is scored under two parameter settings (current θ vs. old θ_{old}). This cancellation is what reduces the continuous-action log-ratio to the mode-referenced angular form derived in §A.2, and what lets a single PPO-style clipped surrogate (Eq. (5) in the main body) act symmetrically on both branches.

A.2 Mode-referenced density ratio

Starting from the unified log-probability in Eq. (7), the gaze branch corresponds to the vMF density with current-policy mean direction $\boldsymbol{\mu}_i^\theta$ and shared concentration κ:

$$\pi_\theta(\mathbf{a}_i \mid s_i) \;=\; C_D(\kappa)\,\exp\!\big(\kappa\,(\boldsymbol{\mu}_i^\theta)^\top \mathbf{a}_i\big), \qquad \mathbf{a}_i \in \mathcal{S}^{D-1}, \tag{8}$$

where $C_D(\kappa)$ is the vMF normalization constant in dimension D. In our formulation the vMF density is *not* used to sample stochastic continuous actions during rollout; it plays the role of a directional density model that scores how well the current-policy direction aligns with the direction produced by the old policy π_{old}. Concretely, during rollout we cache the old-policy direction $\tilde{\mathbf{z}}_i := \boldsymbol{\mu}_i^{\theta_{\mathrm{old}}}$ and, in the subsequent optimization step, evaluate both the old and the current policies at this same reference point—that is, we take $\mathbf{a}_i \equiv \tilde{\mathbf{z}}_i$. This is the mode-referenced evaluation point introduced in the main body around Eq. (3). Under the old policy, the density at this point is

$$\pi_{\theta_{\mathrm{old}}}(\mathbf{a}_i \mid s_i) \;=\; \pi_{\theta_{\mathrm{old}}}(\tilde{\mathbf{z}}_i \mid s_i) \;=\; C_D(\kappa)\,\exp\!\big(\kappa\,\tilde{\mathbf{z}}_i^\top \tilde{\mathbf{z}}_i\big) \;=\; C_D(\kappa)\,\exp(\kappa), \tag{9}$$

where we used $\|\tilde{\mathbf{z}}_i\|_2 = 1$. Under the current policy, evaluated at the same reference point, it is

$$\pi_\theta(\mathbf{a}_i \mid s_i) \;=\; \pi_\theta(\tilde{\mathbf{z}}_i \mid s_i) \;=\; C_D(\kappa)\,\exp\!\big(\kappa\,(\boldsymbol{\mu}_i^\theta)^\top \tilde{\mathbf{z}}_i\big). \tag{10}$$

Because current and old policies share κ, the normalization $C_D(\kappa)$ cancels in the ratio, giving

$$r_i \;=\; \frac{\pi_\theta(\mathbf{a}_i \mid s_i)}{\pi_{\theta_{\mathrm{old}}}(\mathbf{a}_i \mid s_i)} \;=\; \exp\!\Big(\kappa\big((\boldsymbol{\mu}_i^\theta)^\top \tilde{\mathbf{z}}_i - 1\big)\Big). \tag{11}$$

Since $\boldsymbol{\mu}_i^\theta$ and $\tilde{\mathbf{z}}_i$ are both ℓ_2-normalized, their inner product coincides with the cosine similarity $\cos(\boldsymbol{\mu}_i^\theta, \tilde{\mathbf{z}}_i)$. The log-ratio therefore depends purely on the *angular deviation* of the current policy direction from the old-policy mode,

$$\log r_i \;=\; \kappa\,\big(\cos(\boldsymbol{\mu}_i^\theta, \tilde{\mathbf{z}}_i) - 1\big) \;\le\; 0, \tag{12}$$

with equality iff $\boldsymbol{\mu}_i^\theta = \tilde{\mathbf{z}}_i$, which matches Eq. (3) in the main paper. Eq. (12) should be read as a deterministic, mode-referenced policy update in latent space rather than as a standard importance-sampling ratio over sampled stochastic actions; replacing it with a Gaussian-style Euclidean ratio would force the decoder off the hyperspherical manifold its hidden states actually live on after RMSNorm.

A.3 Closed-form vMF KL divergence

We next derive the gaze-branch trust-region regularizer used in the RL objective Eq. (6). Consider two vMF distributions on $\mathcal{S}^{D-1}$ sharing the concentration κ:

$$p(\mathbf{x}) \;=\; \mathrm{vMF}(\boldsymbol{\mu}_i^\theta, \kappa), \qquad q(\mathbf{x}) \;=\; \mathrm{vMF}(\boldsymbol{\mu}_i^{\mathrm{ref}}, \kappa). \tag{13}$$

Their Kullback–Leibler divergence is

$$\begin{aligned} D_{\mathrm{KL}}(p \,\|\, q) \;&=\; \mathbb{E}_{\mathbf{x}\sim p}\left[\log \frac{C_D(\kappa)\,\exp\!\big(\kappa\,(\boldsymbol{\mu}_i^\theta)^\top \mathbf{x}\big)}{C_D(\kappa)\,\exp\!\big(\kappa\,(\boldsymbol{\mu}_i^{\mathrm{ref}})^\top \mathbf{x}\big)}\right] \\ &=\; \kappa\,(\boldsymbol{\mu}_i^\theta - \boldsymbol{\mu}_i^{\mathrm{ref}})^\top\, \mathbb{E}_{\mathbf{x}\sim p}[\mathbf{x}]. \end{aligned} \tag{14}$$

The normalization constants cancel as in §A.2. For a vMF distribution, the mean resultant vector has a standard closed form [Mardia and Jupp, 2000],

$$\mathbb{E}_{\mathbf{x}\sim p}[\mathbf{x}] \;=\; A_D(\kappa)\,\boldsymbol{\mu}_i^{\theta}, \qquad A_D(\kappa) \;:=\; \frac{I_{D/2}(\kappa)}{I_{D/2-1}(\kappa)}, \tag{15}$$

where $I_\nu(\cdot)$ is the modified Bessel function of the first kind and $A_D(\kappa) \in (0,1)$ is the mean resultant length. Substituting Eq. (15) into Eq. (14) and using $\|\boldsymbol{\mu}_i^{\theta}\|_2 = 1$,

$$\begin{aligned} D_{\mathrm{KL}}(p \,\|\, q) \;&=\; \kappa A_D(\kappa)\,\big((\boldsymbol{\mu}_i^{\theta})^\top \boldsymbol{\mu}_i^{\theta} - (\boldsymbol{\mu}_i^{\mathrm{ref}})^\top \boldsymbol{\mu}_i^{\theta}\big) \\ &=\; \kappa A_D(\kappa)\,\big(1 - (\boldsymbol{\mu}_i^{\mathrm{ref}})^\top \boldsymbol{\mu}_i^{\theta}\big) \\ &=\; W_\kappa\,\big(1 - \cos(\boldsymbol{\mu}_i^{\theta}, \boldsymbol{\mu}_i^{\mathrm{ref}})\big), \end{aligned} \tag{16}$$

with constant $W_\kappa := \kappa A_D(\kappa)$. Under hyperspherical normalization and shared concentration, the exact vMF KL therefore reduces to a *scaled cosine distance* between the current policy direction and the reference direction. This is the closed form used as the gaze-branch KL penalty $\mathrm{KL}_{\mathrm{vMF}}$ in Eq. (6); it avoids the high-variance sample-based KL estimate required on the text branch and aligns the trust region with the same angular geometry as the ratio Eq. (12). Both the log-ratio and the KL penalty are thus natively consistent with the spherical structure of the decoder's latent space.

A.4 GEPO: remaining details

With the vMF ratio (§A.2) and KL (§A.3) in place, the rest of GEPO is standard PPO/GRPO machinery applied to the hybrid discrete–continuous action space of §3.3.

Group-relative advantage estimator. Following [Shao et al., 2024], for each prompt we draw $G{=}8$ rollouts from $\pi_{\theta_{\mathrm{old}}}$ and compute group-normalized advantages

$$A^{(g)} \;=\; \frac{r^{(g)} - \mu_G}{\sigma_G + 10^{-8}}, \qquad \mu_G \;=\; \tfrac{1}{G}\sum_{g} r^{(g)}, \quad \sigma_G \;=\; \mathrm{std}\big(r^{(1)}, \ldots, r^{(G)}\big), \tag{17}$$

and broadcast $A^{(g)}$ uniformly to every action position (text token and gaze slot) of rollout g.

Scalar reward. Each rollout receives $r \;=\; r_{\mathrm{acc}} + \lambda_{\mathrm{fmt}}\, r_{\mathrm{fmt}}$. The accuracy component uses a *rule-then-API* judge: when the answer can be assessed by deterministic rules, we use exact matching or, for numeric items with gold $y^\star$, the tolerance $|\hat{y} - y^\star| \le \max(10^{-3}|y^\star|, 10^{-6})$. Cases that cannot be resolved by these rules are passed to an external LLM judge through an OpenAI-compatible API to assess correctness against the gold answer. We set $r_{\mathrm{acc}} = 1$ for an answer judged correct and 0 otherwise. The format reward is $r_{\mathrm{fmt}} = 1$ if the response contains a `\boxed{...}` block and exactly one pair of `<|gaze_start|>`/`<|gaze_end|>` sentinels, and 0 otherwise. We set $\lambda_{\mathrm{fmt}} = 0.1$ so format never dominates accuracy. No process-level reward is attached to gaze tokens.

Caching policy for old and reference quantities. For each rollout, we store the text tokens, the gaze actions $\mathbf{z}_i^{\mathrm{roll}} \in \mathbb{R}^D$ with their original magnitudes, and their positions. We then replay this stored trajectory under π_{old} without gradients to cache the old text log-probabilities and gaze scores $\ell_i^{\mathrm{old}} = \kappa(\mathbf{v}_i^{\mathrm{old}})^\top \mathbf{z}_i^{\mathrm{roll}}$. Here $\mathbf{v}_i$ is the decoder gaze output before any additional L_2 normalization; the old replay output need not equal the recorded action. A separate frozen π_{ref} forward on the same stored trajectory provides the reference text log-probabilities and normalized directions used by the SFT-reference KL in Eq. (16). These rollout-conditioned quantities are distinct from the offline, prompt-level Gram anchor $\mathbf{Z}_{\mathrm{SFT}}$. All recorded actions, old scores, reference outputs, and SFT anchors are held fixed during the M inner epochs; the current actor is re-evaluated with gradients at every update.

Text-branch KL estimator. For the discrete text branch we use Schulman's "k3" estimator—a non-negative, low-variance unbiased estimator of $\mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})$, standard in GRPO implementations:

$$\widehat{\mathrm{KL}}^{\mathrm{k3}}_{\mathrm{tok}}(s_t) \;=\; \rho_t - \log \rho_t - 1, \qquad \rho_t \;=\; \frac{\pi_{\mathrm{ref}}(y_t \mid s_t)}{\pi_\theta(y_t \mid s_t)}, \tag{18}$$

which is pointwise non-negative and has lower variance than the naive $\log \pi_\theta - \log \pi_{\mathrm{ref}}$ term. We average $\widehat{\mathrm{KL}}^{\mathrm{k3}}_{\mathrm{tok}}$ over the text positions $\mathcal{T}$. Strictly, the k3 estimator is unbiased when samples are

drawn from π_θ; we evaluate it on tokens from stored $\pi_{\theta_{\text{old}}}$ rollouts and therefore use it as an off-policy approximation, in line with standard GRPO practice [Shao et al., 2024]. With $M{=}2$ inner PPO epochs the discrepancy between π_θ and $\pi_{\theta_{\text{old}}}$ remains small under the clipping trust region, so this approximation introduces negligible bias.

GEPO algorithm. Algorithm 1 assembles the practical score ratios and the differentiable Gram objective. It is stated for one gaze block of $K > 1$ ordered slots per rollout and a cached SFT anchor with the same number of slots, compared by slot index. The anchor is shared across the G rollouts of a prompt. Within each rollout, we suppress (b, g) on position-level quantities for readability. The raw gaze output $\mathbf{v}_i^\theta$ is used for scoring, while $\boldsymbol{\mu}_i^\theta = \mathbf{v}_i^\theta / \|\mathbf{v}_i^\theta\|_2$ is used for cosine geometry. Geometry drift enters the actor loss directly, not the scalar reward.

For branch $j \in \{\text{tok}, \text{gaze}\}$, let $q_j(r, A) = \min\big(rA, \text{clip}(r, 1 - \epsilon_j^-, 1 + \epsilon_j^+)A\big)$. The dual-clip surrogate to be maximized is

$$\Phi_j(r, A) = \begin{cases} \max\big(q_j(r, A), c_j A\big), & A < 0, \\ q_j(r, A), & A \geq 0, \end{cases} \qquad c_j > 1.$$

Before either branch exponentiates its log-ratio, we clip that difference to $[-10, 10]$ for numerical stability; this is separate from the PPO clipping above.

All per-run numerical settings (group size G, clip ranges, KL weights, vMF κ, batch sizes, data, hardware) are collected in the implementation-details section below; see Table 6.

B Implementation Details

The configurations and compute measurements in this section refer to our main experiments with Qwen2.5-VL-7B. Additional results on InternVL3-8B are reported in Appendix C.

B.1 SFT training data and sample structure

Zebra-CoT subset selection. Our SFT training corpus is constructed from the Zebra-CoT dataset [Li et al., 2025a], which comprises 182K samples of logically coherent interleaved text–image reasoning traces organized into several top-level task categories. Following the protocol of SkiLa [Tong et al., 2025], we *include* samples whose reasoning traces contain natural, spatially meaningful intermediate visuals that a vision–language decoder can be expected to reproduce, and *exclude* categories whose intermediate visuals are either too abstract, too symbolic, or too far from natural images to be faithfully reconstructed in the decoder's continuous latent space. Table 4 lists the included and excluded subsets; after filtering, the SFT training set contains 101K samples.

Table 4: **Zebra-CoT subset selection for GELVR SFT training.** We retain categories whose intermediate visuals are natural and spatially grounded, and drop categories whose intermediate visuals are 3D, highly symbolic, or weakly related to natural visual reasoning.

Included subsets	**Excluded subsets**
2D Visual Reasoning — Visual Search	3D Visual Reasoning
2D Visual Reasoning — Visual Jigsaw	Chemistry
Visual Logic & Strategic Games — Chess	Competitive Problems
Visual Logic & Strategic Games — Maze	Graph Theory
Visual Logic & Strategic Games — RPM	Physics
Visual Logic & Strategic Games — Checkers	ARC-AGI
Scientific Reasoning — Geometry	Ciphers
	Tetris
	Connect-Four
Resulting SFT training set: $\sim$101K samples from the above included subsets.	

Sample format. Each filtered Zebra-CoT trace is reformatted into an interleaved sequence of textual thoughts and gaze-anchored visual thoughts, with the latter bracketed by our gaze sentinels and supervised against target embeddings produced by the Gaze Perceiver (§3.2). The resulting training sample has the structure shown in Figure 4.

Algorithm 1 One outer iteration of GEPO (Geometry-preserving Policy Optimization).

Require: current parameters θ; frozen SFT policy π_{ref}; cached anchors $\{\mathbf{Z}_{\mathrm{SFT}}(x,q)\}$; score scale κ; branch clips $\{\epsilon_j^-, \epsilon_j^+, c_j\}_{j\in\{\mathrm{tok},\mathrm{gaze}\}}$; weights $\alpha, \eta_{\mathrm{tok}}, \eta_{\mathrm{gaze}}, \lambda_{\mathrm{geo}}$; group size G; inner epochs M; batch size B; step size η

Sample a batch of prompts $\{(x^{(b)}, q^{(b)})\}_{b=1}^{B}$
Snapshot $\theta_{\mathrm{old}} \leftarrow \theta$; let $\pi_{\mathrm{old}} = \pi_{\theta_{\mathrm{old}}}$
Rollout and caching phase (no gradient)
for each prompt $b = 1, \ldots, B$ **do**
  for $g = 1, \ldots, G$ **do**
    Generate $o^{(b,g)}$ under π_{old}; store its text, raw gaze actions $\{\mathbf{z}_i^{\mathrm{roll}}\}$, and positions
    Replay $o^{(b,g)}$ under π_{old} on the recorded text and gaze inputs
    Cache $\ell_i^{\mathrm{old}} = \kappa(\mathbf{v}_i^{\mathrm{old}})^\top \mathbf{z}_i^{\mathrm{roll}}$ and $\log \pi_{\mathrm{old}}(y_t \mid s_t)$
    Replay $o^{(b,g)}$ under frozen π_{ref}; cache $\boldsymbol{\mu}_i^{\mathrm{ref}}$ and $\log \pi_{\mathrm{ref}}(y_t \mid s_t)$
    Parse the final answer; compute $r^{(b,g)} = r_{\mathrm{acc}} + \lambda_{\mathrm{fmt}} r_{\mathrm{fmt}}$
  end for
  $\mu_G^{(b)} \leftarrow \frac{1}{G}\sum_g r^{(b,g)}$; $\sigma_G^{(b)} \leftarrow \mathrm{std}(r^{(b,\cdot)})$; $A^{(b,g)} \leftarrow (r^{(b,g)} - \mu_G^{(b)})/(\sigma_G^{(b)} + 10^{-8})$
end for
Inner update phase (gradient through current actor only)
for inner epoch $m = 1, \ldots, M$ **do**
  for each rollout $o^{(b,g)}$ **do**
    Replay $o^{(b,g)}$ under π_θ with gradients enabled; obtain current log-probabilities and $\{\mathbf{v}_i^\theta\}$
    Hold recorded inputs and all cached quantities fixed; set $A \leftarrow A^{(b,g)}$
    for each text position $t \in \mathcal{T}^{(b,g)}$ **do**
      $\delta_t^{\mathrm{tok}} \leftarrow \log \pi_\theta(y_t \mid s_t) - \log \pi_{\mathrm{old}}(y_t \mid s_t)$
      $r_t^{\mathrm{tok}} \leftarrow \exp\big(\mathrm{clip}(\delta_t^{\mathrm{tok}}, -10, 10)\big)$
    end for
    for each gaze position $i \in \mathcal{S}^{(b,g)}$ **do**
      $\ell_i^\theta \leftarrow \kappa(\mathbf{v}_i^\theta)^\top \mathbf{z}_i^{\mathrm{roll}}$; $r_i^{\mathrm{gaze}} \leftarrow \exp\big(\mathrm{clip}(\ell_i^\theta - \ell_i^{\mathrm{old}}, -10, 10)\big)$
      $\boldsymbol{\mu}_i^\theta \leftarrow \mathbf{v}_i^\theta / \|\mathbf{v}_i^\theta\|_2$ ▷ Direction for geometry; retain gradients
    end for
    $\mathcal{L}_{\mathrm{tok}}^{(b,g)} \leftarrow \frac{1}{|\mathcal{T}^{(b,g)}|}\sum_{t\in\mathcal{T}^{(b,g)}} \Phi_{\mathrm{tok}}(r_t^{\mathrm{tok}}, A)$; $\mathcal{L}_{\mathrm{gaze}}^{(b,g)} \leftarrow \frac{1}{K}\sum_{i\in\mathcal{S}^{(b,g)}} \Phi_{\mathrm{gaze}}(r_i^{\mathrm{gaze}}, A)$
    $\rho_t \leftarrow \pi_{\mathrm{ref}}(y_t \mid s_t)/\pi_\theta(y_t \mid s_t)$ for $t \in \mathcal{T}^{(b,g)}$
    $\mathrm{KL}_{\mathrm{tok}}^{(b,g)} \leftarrow \frac{1}{|\mathcal{T}^{(b,g)}|}\sum_t (\rho_t - \log \rho_t - 1)$ ▷ Eq. (18)
    $\mathrm{KL}_{\mathrm{vMF}}^{(b,g)} \leftarrow \frac{1}{K}\sum_i W_\kappa\big(1 - (\boldsymbol{\mu}_i^\theta)^\top \boldsymbol{\mu}_i^{\mathrm{ref}}\big)$ ▷ SFT-reference term, Eq. (16)
    Look up $\mathbf{Z}_{\mathrm{SFT}}(x^{(b)}, q^{(b)}) = [\boldsymbol{\mu}_1^{\mathrm{SFT}}, \ldots, \boldsymbol{\mu}_K^{\mathrm{SFT}}]$ (shared across g)
    $\mathcal{L}_{\mathrm{geo}}^{\mathrm{RL},(b,g)} \leftarrow \frac{1}{K(K-1)}\sum_{i\neq j}\big((\boldsymbol{\mu}_i^\theta)^\top \boldsymbol{\mu}_j^\theta - (\boldsymbol{\mu}_i^{\mathrm{SFT}})^\top \boldsymbol{\mu}_j^{\mathrm{SFT}}\big)^2$
    $\mathcal{L}_{\mathrm{RL}}^{(b,g)} \leftarrow -\big[\mathcal{L}_{\mathrm{tok}}^{(b,g)} + \alpha \mathcal{L}_{\mathrm{gaze}}^{(b,g)}\big] + \eta_{\mathrm{tok}} \mathrm{KL}_{\mathrm{tok}}^{(b,g)} + \eta_{\mathrm{gaze}} \mathrm{KL}_{\mathrm{vMF}}^{(b,g)} + \lambda_{\mathrm{geo}} \mathcal{L}_{\mathrm{geo}}^{\mathrm{RL},(b,g)}$
  end for
  $\mathcal{L}_{\mathrm{RL}} \leftarrow \frac{1}{BG}\sum_{b,g} \mathcal{L}_{\mathrm{RL}}^{(b,g)}$; $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\mathrm{RL}}$
end for
return updated θ

B.2 SFT hyperparameters

The SFT stage is trained for one epoch on the 101K Zebra-CoT subset (§B.1). The native ViT and the vision-to-decoder merger are frozen; the LLM is updated in full, and the gaze module (Perceiver + projector) is trained from scratch with a larger learning rate. Full settings are listed in Table 5.

B.3 RL hyperparameters

The RL stage implements GEPO (Algorithm 1) on top of a decoupled hybrid PPO. The actor and reference are initialized from the SFT checkpoint, and the reference policy remains frozen throughout RL. For each newly collected rollout, we replay its recorded text and gaze inputs through the frozen reference policy to obtain the reference text log-probabilities and latent directions $\boldsymbol{\mu}_i^{\mathrm{ref}}$ used by the KL terms. These outputs are cached and held fixed during the inner policy updates for that rollout. Separately, before RL begins, we cache one deterministic SFT trajectory $\mathbf{Z}_{\mathrm{SFT}}$ and its Gram matrix

GELVR Training Sample

Interleaved Textual and Visual Thoughts

<think> *[Textual Thought 1]* </think>
<|gaze_start|> *[Gaze Target 1]* <|gaze_end|>
<think> *[Textual Thought 2]* </think>
<|gaze_start|> *[Gaze Target 2]* <|gaze_end|>
... ...
<think> *[Textual Thought N]* </think>
<|gaze_start|> *[Gaze Target N]* <|gaze_end|>
<think> *[Final Textual Thought]* </think>
<answer> *[Final Answer]* </answer>

Figure 4: **Format of a GELVR training sample.** Discrete textual thoughts (wrapped by <think>) and continuous gaze segments (bracketed by <|gaze_start|> and <|gaze_end|>) interleave within a single autoregressive sequence; each gaze segment's hidden states are supervised against target visual embeddings produced by the Gaze Perceiver at that step.

Table 5: **SFT hyperparameters.** One epoch on the Zebra-CoT subset.

Category	Parameter	Value
Backbone	Model	Qwen2.5-VL-7B-Instruct
	Gaze encoder	`google/siglip2-so400m-patch14-384` (frozen)
	Visual tokens / image	8192 (max 6,422,528 px; min 3,136 px)
Gaze module	Gaze slot count K	8/16/24/32
	Gaze Perceiver	$2\times$ CrossAttn, 8 heads
	Gaze projector	Linear $1152 \to 3584$
	Insertion layer	L28 (Qwen2.5-VL)
Loss	CE (language)	1.0
	Gaze MSE α	0.5
	Geometry MSE β	0.5 (detached target, ℓ_2-normalized)
Optimizer	LR (LLM)	10^{-5}
	LR (gaze module)	10^{-4}
	Scheduler	cosine, warmup ratio 0.1
	Weight decay	0.1
	Gradient clip	1.0
	Precision	bf16
Batching	GPUs	$8\times$ H20
	Per-device batch	1
	Gradient accumulation	16
	Global batch	**128**
	Epochs	1
Data	Dataset	Zebra-CoT subset ($\sim$101K samples)
	Freeze ViT / merger / LLM	True / True / False
Runtime	DeepSpeed	ZeRO-2
	Gradient checkpointing	True
	Seed	42

for each RL prompt. This offline anchor is shared across rollouts of the same prompt and remains fixed throughout RL for Eq. (4). Full settings are listed in Table 6.

B.4 Hyperparameter sensitivity

Table 7 reports sensitivity to the gaze score scale κ, the RL geometry weight λ_{geo}, and the gaze clipping range ϵ_{gaze} on V* and HR-Bench-4K. Our default settings are $\kappa = 0.01$, $\lambda_{\text{geo}} = 0.05$, and $\epsilon_{\text{gaze}} = 0.1$, consistent with Table 6. These values attain the highest observed scores in their respective sweeps; they are the best among the tested settings, rather than evidence of a global

Table 6: **RL hyperparameters.** Decoupled hybrid PPO with the GEPO objective of §3.3.

Category	Parameter	Value
Framework	Base	`verl` 0.4.1, decoupled hybrid PPO
Reward	Task reward	rule-then-API judge (OpenAI-compatible)
	Repetition penalty	false
Geometry loss	Actor-loss term	$+\lambda_{\text{geo}}\mathcal{L}_{\text{geo}}^{\text{RL}}$ (Eq. (6))
	Normalization	$1/[K(K-1)]$ over off-diagonal pairs
	λ_{geo}	0.05
Rollout	n (samples / prompt)	8
	Temperature	0.9
	TP size	1
	Chunked prefill	true
PPO / KL	KL coefficient (text)	0.05
	Latent vMF KL	enabled
	Latent KL coefficient	10^{-2}
	Gaze score scale / vMF κ	0.01 (shared)
	Latent clip (low / high / dual)	$0.1/0.1/3.0$
	Latent loss α	0.5
Batching	Rollout batch	64
	Train samples	$49{,}664$
	Val samples	800
	GPUs	$8\times$ H20
Online filter	Enabled	true
	Filter key	accuracy
	Low / high	$0.1/0.9$
	Answer-tag filter	false
Latent	Latent size	32
	Gaze token id	151665
References	KL outputs	Frozen SFT replay on stored rollout
	Gram anchor	Cached once per prompt before RL

optimum. For the clipping sweep, ϵ_{gaze} denotes the symmetric lower and upper PPO clip range applied to the practical gaze score ratio in §A.4; it is distinct from the dual-clip coefficient.

Score scale. Among the three tested values, $\kappa = 0.01$ performs best on both benchmarks. Either decreasing it to 0.001 or increasing it to 0.1 lowers the reported scores, indicating sensitivity to the scale of the gaze-branch updates.

Geometry weight. The default $\lambda_{\text{geo}} = 0.05$ improves V* and HR-Bench-4K by 1.04 and 0.88 points over $\lambda_{\text{geo}} = 0$, respectively. Weights between 0.01 and 0.1 yield similar scores, whereas increasing the weight to 1 reduces performance on both benchmarks. This is consistent with a trade-off between preserving the SFT relational structure and allowing task-driven adaptation.

Gaze clipping. The tested ranges 0.05 and 0.1 both attain 85.86 on V*; 0.1 yields 77.13 on HR-Bench-4K, compared with 77.00 at 0.05. We therefore use the default lower/upper clip ranges $0.1/0.1$, with the dual-clip coefficient 3.0 as listed in Table 6. The zero-width setting and the larger tested ranges (0.2 and 0.3) yield lower scores. These observations concern the practical inner-product score ratios, not clipping of the normalized theoretical ratio in Eq. (11).

B.5 Compute and wall-clock time

The main Qwen2.5-VL-7B experiments are run on a single server with the specifications in Table 8. Under this configuration, one SFT epoch on the 101K Zebra-CoT subset takes approximately 30 hours, and one full GEPO RL run on the DeepEyes+Thyme mixture takes approximately 24 hours. End-to-end, a complete GELVR training run (SFT followed by RL) therefore costs about 54 hours on $8\times$H20, or roughly 432 GPU-hours.

Table 7: **Hyperparameter sensitivity.** Scores on V^* and HR-Bench-4K. Bold scores mark the highest value within each sweep, including ties; bold parameter values indicate the default configuration.

Parameter value	V^*	HR-Bench-4K
Gaze score scale κ		
0.001	83.77	74.13
0.01	**85.86**	**77.13**
0.1	84.82	75.63
RL geometry weight λ_{geo}		
0	84.82	76.25
0.01	85.34	76.63
0.05	**85.86**	**77.13**
0.1	85.34	77.00
1	82.72	74.13
Gaze clipping range ϵ_{gaze}		
0	79.06	74.63
0.05	**85.86**	77.00
0.1	**85.86**	**77.13**
0.2	79.58	76.25
0.3	78.01	73.38

Table 8: **Compute environment for the Qwen2.5-VL-7B experiments.**

Component	Specification
OS	CentOS
CPU	AMD EPYC 9K84, 96 cores
System memory	2.2 TiB
GPUs	8× NVIDIA H20, 96 GB HBM each
GPU driver	570.158.01
SFT wall-clock	∼30 hours (1 epoch, 101K samples)
RL wall-clock	∼24 hours (GEPO, DeepEyes + Thyme)
Total per full run	∼54 hours (≈ 432 GPU-hours)

C Generality across MLLM backbones

To assess transfer beyond Qwen2.5-VL-7B, we apply the complete GELVR pipeline, including geometry-aligned SFT and GEPO, to InternVL3-8B [Zhu et al., 2025]. Table 9 reports the results on the four primary benchmarks.

Table 9: **Generality across MLLM backbones.** Overall scores of the base InternVL3-8B model and its GELVR variants. Bold scores indicate the best value within this comparison. MME-Lite denotes MME-RealWorld-Lite.

Model	V^*	HR-4K	HR-8K	MME-Lite
InternVL3-8B	70.20	70.00	69.30	48.60
GELVR-SFT	74.35	72.25	70.88	52.94
GELVR-SFT+GEPO	**76.96**	**75.38**	**73.38**	**55.86**
Δ *vs. InternVL3-8B*	+6.76	+5.38	+4.08	+7.26

Both SFT and GEPO improve all four scores over the preceding stage. Relative to the base InternVL3-8B model, the full pipeline gains 6.76, 5.38, 4.08, and 7.26 percentage points on V^*, HR-Bench-4K, HR-Bench-8K, and MME-RealWorld-Lite, respectively. These results provide additional evidence that the two-stage training recipe transfers to a second MLLM family.

D Limitations

We briefly discuss the scope of our study and the failure modes most likely to appear outside it.

Fixed gaze-policy scale and clipping. Our GEPO formulation uses a *single, globally fixed* $\kappa = 0.01$ across gaze slots and training updates (Table 6). In the normalized theoretical vMF model, a shared concentration permits cancellation of $C_D(\kappa)$ in Eq. (11) and the closed-form KL in Eq. (16). In practical optimization, κ scales raw inner-product scores, which depend on both latent directions and magnitudes; the gaze clipping range $\epsilon_{\text{gaze}} = 0.1$ acts on the resulting score ratios. The sensitivity study in §B.4 supports these defaults among the tested settings, but does not establish that they transfer unchanged to other backbones or latent scales. A fixed score scale and clipping range cannot adapt to changes across latent positions or training stages. Learning position- or time-dependent scales and studying adaptive clipping are directions for future work.

Backbone and scale. Our main experiments use Qwen2.5-VL-7B [Bai et al., 2025], with an additional evaluation of the complete pipeline on InternVL3-8B [Zhu et al., 2025] in Appendix C. These results provide evidence on two model families at the 7B–8B scale; transfer to a broader range of architectures and substantially larger or smaller decoders remains to be established. The spherical vMF policy assumes the decoder's hidden states are effectively directional after RMSNorm [Zhang and Sennrich, 2019]; backbones that apply LayerNorm-with-bias or omit the final normalization may violate this assumption and require a different policy parameterization.

Second-order relational geometry. The object we install and protect—the pairwise cosine structure of the latent trajectory, $\mathrm{Gram}(\mathbf{Z}_\theta)$—is a *second-order* relational quantity. Failure modes that depend on higher-order relations (e.g., strict triple-product or ordering constraints among gaze slots, or on explicit temporal dynamics within the gaze block) lie outside what $\mathcal{L}_{\text{geo}}^{\text{SFT}}$ and $\mathcal{L}_{\text{geo}}^{\text{RL}}$ can enforce, and would require richer structural objectives. Within this second-order regime, we also inherit the biases of the visual target manifold produced by SigLIP2 [Tschannen et al., 2025]: our framework preserves whatever structure SFT installs, so systematic failures of the visual encoder (e.g., mis-localization of extremely small targets) propagate rather than self-correct.

Training-data scope and evaluation. The SFT corpus is restricted to the ~ 101K 2D-reasoning subset of Zebra-CoT [Li et al., 2025a] (§B.1), and the RL corpus uses only DeepEyes [Zheng et al., 2025] and Thyme [Zhang et al., 2025b] prompts. Categories we explicitly exclude—3D visual reasoning, highly symbolic puzzles (Tetris, Connect-Four, Ciphers), and non-visual scientific problems (Chemistry, Physics, Graph Theory)—have not been tested, so we do not expect the pairwise-cosine preservation to be informative in those settings. Evaluation is likewise concentrated on high-resolution perception and real-world visual search; tasks with heavy temporal dynamics (e.g., long video understanding) or extensive external-knowledge retrieval are outside the scope of this paper.

Reward model and API dependency. During RL, we use a rule-then-API hybrid judge (§B.3): exact-match and numeric tolerance when ground truth admits an unambiguous parse, falling back to an external OpenAI-compatible LLM judge otherwise. The fallback component introduces a dependency on a proprietary scoring model whose behavior may change over time; this affects reproducibility and should be kept in mind when porting our RL recipe to new benchmarks.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three concrete claims that exactly match the body: (i) a geometry-consistent paradigm unifying geometry alignment in SFT and geometry preservation in RL (§3.2, §3.3); (ii) the GEPO algorithm with a spherical vMF policy, decoupled trust regions, and an SFT-anchored Gram regularizer, derived in §A.2–§A.3; and (iii) state-of-the-art latent-reasoning performance on high-resolution perception and real-world reasoning benchmarks, supported by Tables 1–3 and Figure 3. The main results and ablations use Qwen2.5-VL-7B, with an additional evaluation of the complete pipeline on InternVL3-8B in Appendix C. Our evidence is limited to these two backbones and does not establish generalization across model scales or unbounded latent spans.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.

- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We devote a dedicated section (Appendix D) to limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper contains two analytical results, both restated with full assumptions and complete derivations in the appendix. The mode-referenced vMF density ratio (Eq. (3) in the main body) is derived in Appendix A.2, under the explicit assumptions that the actor and old policies share a fixed concentration κ, that the latent direction $\boldsymbol{\mu}_i^\theta$ is ℓ_2-normalized, and that the ratio is evaluated at the old-policy mode $\tilde{\mathbf{z}}_i$. The closed-form vMF KL divergence used as the gaze-branch trust region (Eq. (6)) is derived in Appendix A.3 under the same shared-κ assumption, using the standard mean-resultant formula for vMF distributions [Mardia and Jupp, 2000]. Every equation is numbered and cross-referenced between the main body and the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the complete source code in the supplementary material to ensure full reproducibility. Furthermore, all necessary implementation details—including are fully detailed in the Appendix. All utilized datasets and base models are open-source.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All training and evaluation datasets used in this paper are third-party open-source releases: Zebra-CoT [Li et al., 2025a] for SFT (with the exact included/excluded subsets listed in Table 4), DeepEyes [Zheng et al., 2025] and Thyme [Zhang et al., 2025b] prompts for RL, and V*Bench [Wu and Xie, 2024], HR-Bench-4K/8K [Wang et al., 2025b], and MME-RealWorld-Lite [Zhang et al., 2025a] for evaluation; no new dataset is introduced. Our implementation is provided as an anonymized code archive in the supplemental material,

containing the training and evaluation scripts, the exact hyperparameter configuration files corresponding to Tables 5–6, and a README with the commands needed to reproduce the main-table runs.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix B describes the training data and hyperparameter settings for the main Qwen2.5-VL-7B experiments. The additional InternVL3-8B evaluation is reported separately in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer:[No]

   Justification: Due to the substantial compute cost of SFT and RL on a 7B-parameter MLLM, each configuration in Table 1 and the ablation table is reported from a single seed, so we do not provide formal error bars or confidence intervals. We note however that the observed gains are large (e.g., $+9$ to $+14$ absolute points over the Qwen2.5-VL-7B baseline on V^*, HR-Bench-4K/8K, and MME-RealWorld-Lite) and stable across four benchmarks and multiple sub-metrics, which reduces the likelihood that they are driven by seed variance. We explicitly acknowledge this single-seed limitation in Appendix D and recommend caution when interpreting small numerical differences (e.g., ≤ 0.5 points on MME-RealWorld-Lite). The full hyperparameter tables (Tables 5–6) make multi-seed replication straightforward for readers with sufficient compute.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B.5 reports the server specifications and approximate SFT and RL wall-clock times for the main Qwen2.5-VL-7B experiments. These measurements are not presented as runtime estimates for the additional InternVL3-8B evaluation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research was conducted in accordance with the NeurIPS Code of Ethics. The paper does not involve human-subject experiments, crowdsourcing, collection of personally identifiable information, or deployment to real users. All experiments are performed on existing datasets and standard benchmarks, and the work is intended to improve model reasoning and grounding rather than enable harmful applications.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: This work focuses on fundamental algorithmic advancements in visual latent reasoning. It does not introduce specific downstream applications or unique negative societal impacts beyond those already inherent to existing open-source multimodal large language models.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Our work provides algorithmic improvements (fine-tuning and RL) on top of an already publicly released and safety-aligned open-source model (Qwen2.5-VL-7B). We evaluate and train on standard, public academic benchmarks without scraping new sensitive data. Thus, our released weights and code do not introduce novel high-risk capabilities that necessitate additional, specialized safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All pre-trained models (e.g., Qwen2.5-VL-7B, SigLIP2) and public benchmarks (e.g., V*, HR-Bench, MME-RealWorld) utilized in this work are open-source. We strictly adhere to their respective licenses and terms of use. Proper citations and credits to the original creators are explicitly provided throughout the main text and references.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer:[Yes]

Justification: Justification: The primary new assets introduced in this work are the GELVR codebase, the implementation of the GEPO algorithm, and the customized data subset configurations. The complete source code is provided in the supplementary material, accompanied by a README file for setup and execution. Furthermore, the mathematical formulation, architecture modifications, and data selection strategies are comprehensively documented in the main text and Appendix.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All experiments are conducted using existing datasets and automated evaluation protocols, so participant instructions, screenshots, and compensation details are not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:The paper does not involve crowdsourcing, user studies, or research with human subjects. All experiments are conducted on existing datasets and automated evaluations, so IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer:[Yes]

    Justification: Our main experiments build on the Qwen2.5-VL-7B Multimodal Large Language Model (MLLM). Its role as the reasoning engine and its integration with the gaze module and latent reasoning process are described in Section 3.2. Appendix C additionally evaluates the complete pipeline on InternVL3-8B.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.