HyLaR: Hybrid Latent Reasoning with Decoupled Policy Optimization

Tao Cheng[1], Shi-Zhe Chen[1,*,†], Hao Zhang[1], Yixin Qin[1]
Jinwen Luo[2], and Zheng Wei[1,*]

[1] Tencent PCG, China [2] Tencent CSIG, China
{elvistcheng,shizhechen,jeffhzhang,wadeqin,jamsluo,hemingwei}@tencent.com

Abstract. Chain-of-Thought (CoT) reasoning significantly elevates the complex problem-solving capabilities of multimodal large language models (MLLMs). However, adapting CoT to vision typically discretizes signals to fit LLM inputs, causing early semantic collapse and discarding fine-grained details. While external tools can mitigate this, they introduce a rigid bottleneck, confining reasoning to predefined operations. Although recent latent reasoning paradigms internalize visual states to overcome these limitations, optimizing the resulting hybrid discrete-continuous action space remains challenging. In this work, we propose HyLaR (Hybrid Latent Reasoning), a framework that seamlessly interleaves discrete text generation with continuous visual latent representations. Specifically, following an initial cold-start supervised fine-tuning (SFT), we introduce DePO (Decoupled Policy Optimization) to enable effective reinforcement learning within this hybrid space. DePO decomposes the policy gradient objective, applying independent trust-region constraints to the textual and latent components, alongside an exact closed-form von Mises-Fisher (vMF) KL regularizer. Extensive experiments demonstrate that HyLaR outperforms standard MLLMs and state-of-the-art latent reasoning approaches across fine-grained perception and general multimodal understanding benchmarks. Code is available at https://github.com/EthenCheng/HyLaR.

Keywords: Multimodal Large Language Models · Visual Latent Reasoning · Policy Optimization

1 Introduction

The integration of explicit Chain-of-Thought (CoT) reasoning has fundamentally transformed how multimodal large language models (MLLMs) approach intricate vision-language tasks [1, 5, 13, 14, 46, 48]. However, most prevailing MLLMs suffer from a critical architectural bottleneck: *early semantic collapse*. As illustrated in Fig. 1(A), standard text-only CoT forces high-bandwidth, continuous visual signals to be prematurely compressed into discrete text tokens. This early

* Corresponding author.
† Project lead.

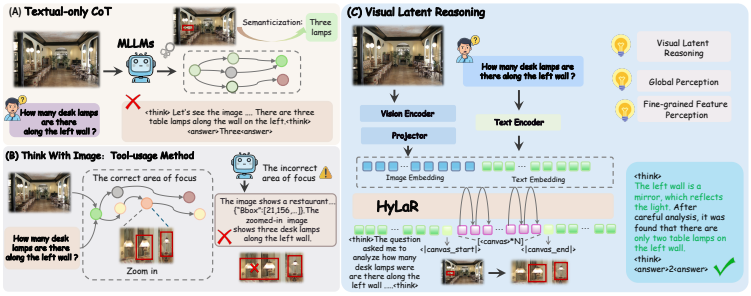

Fig. 1: HyLaR vs. two reasoning paradigms. **(A) Text-only CoT:** visual grounding errors and redundant steps. **(B) Think-with-Images:** external tools cause unstable invocations and latency. **(C) HyLaR (ours):** refines latent think tokens in latent space, preserving fine-grained visual evidence.

discretization inevitably discards fine-grained visual evidence that is difficult to verbalize, leaving the model to rely on linguistic priors rather than grounded visual facts.

To mitigate this, recent studies have explored two primary alternatives. The first is the "Think-with-Images" paradigm (Fig. 1(B)), which relies on external tools to re-perceive the image, yet introduces rigid bottlenecks and inference latency [12, 44, 47]. The second alternative, including recent pioneering works, such as LVR [17], SkiLa [28] and Monet [32], shifts reasoning into a continuous latent space to preserve visual fidelity. While promising, optimizing the resulting hybrid discrete-continuous action space remains profoundly challenging. Current approaches predominantly relying on supervised fine-tuning (SFT) or vanilla reinforcement learning (RL) often yield sub-optimal optimization, as conventional Gaussian assumptions and uniform clipping fundamentally mismatch the native hyperspherical geometry of MLLMs and the distinct variance of hybrid actions.

To overcome these challenges, we propose **HyLaR** (**Hy**brid **La**tent **R**easoning), an elegant and highly effective framework that seamlessly interleaves discrete text generation with continuous visual latent representations (Fig. 1(C)). Following a straightforward cold-start SFT phase that teaches the model to dynamically alternate between logical text tokens and continuous visual working memory, we focus on unlocking the full potential of this hybrid space through reinforcement learning. The core innovation of this work is **DePO** (**De**coupled **P**olicy **O**ptimization), an RL algorithm tailored specifically for hybrid action spaces. Standard policy optimization applies uniform constraints and assumes a flat continuous space, which fails in our setting due to two critical mismatches. First, the importance sampling ratios of discrete tokens and continuous vectors exhibit vastly different variance characteristics. DePO resolves this via *decoupled trust-region clipping*, applying position-specific clipping ranges to stabilize continuous updates without hindering text convergence. Second, and more importantly, rather than relying on standard Gaussian distributions that imply a Euclidean geometry, we explicitly model the continuous latent policy using the vMF

distribution [3, 6, 33]. Because layer-normalized LLM representations inherently reside on a high-dimensional hypersphere, the vMF formulation perfectly aligns with this native geometry. This elegant spherical modeling gracefully translates intractable probability estimations and KL divergences into exact, closed-form cosine distances. Consequently, DePO renders the latent RL pipeline remarkably simpler, more geometrically rigorous, and easier to implement than prior arts.

Our main contributions are summarized as follows:

- We present HyLaR, a hybrid reasoning framework that overcomes early semantic collapse by interleaving discrete textual logic with continuous visual latent representations.
- We introduce DePO, a novel reinforcement learning algorithm for hybrid action spaces. By identifying the geometric and variance mismatches in standard RL, DePO leverages hyperspherical vMF modeling and decoupled clipping to achieve highly effective and exact latent policy optimization.
- Extensive experiments validate HyLaR's superiority in high-resolution perception and complex visual reasoning. Furthermore, rigorous ablations on latent scaling, geometric modeling, and decoupled clipping firmly establish the stability and necessity of our framework.

2 Related Work

2.1 Explicit Reasoning in MLLMs

A substantial body of work has explored visual reasoning in vision-language models. Early approaches [10, 18, 31, 46] primarily relied on textual Chain-of-Thought (CoT) prompting, where the model performed a single inference after encoding both the image and the question. However, subsequent studies revealed a fundamental disconnect: the reasoning process often fails to properly leverage perceptual outputs, leading to scenarios where models correctly perceive visual information yet still produce erroneous answers. To address these limitations, recent work has shifted toward "Think-with-Images" paradigms [12, 44, 47], which employ explicit tool invocations to zoom in on or crop specific regions of interest. Other works incorporate generative visual modules [20, 21, 39] to produce pixel-level intermediate outputs. While many of these approaches have successfully employed reinforcement learning (RL) to optimize discrete textual trajectories or tool invocations [47], they still fundamentally rely on externalized visual processing, introducing tool invocation errors, inference latency, and rigid operational bottlenecks.

2.2 Latent Space Reasoning

To avoid early discretization and external overhead, recent studies have increasingly shifted toward latent space reasoning by replacing discrete tokens with continuous embedding representations during autoregressive generation [11, 26,

37, 42, 45]. This paradigm exploits the richer information encoded in latent vectors to enhance reasoning flexibility and compress lengthy reasoning chains.

Beyond language-only models, latent reasoning has been actively extended to MLLMs through various strategies. For instance, Ray et al. [22] introduce modality-agnostic latent tokens as an internal scratchpad to facilitate free-form reasoning. Other approaches, such as LVR [17] and Mirage [40], propose aligning generated latent embeddings with those of auxiliary images. While effective, these designs involve inherent trade-offs: Mirage [40] further compresses image embeddings via mean pooling before alignment, whereas LVR [17] focuses primarily on cropped image regions, limiting its capacity to encode visual operations over the entire image. To preserve global context, Laser [36] enforces a "forest-before-trees" cognitive hierarchy via dynamic windowed alignment, maintaining a probabilistic superposition of global features to prevent premature semantic collapse. From a training perspective, SkiLa [28] offers a straightforward SFT approach similar to LVR, yet it lacks RL optimization. Conversely, Monet [32] presents a highly comprehensive framework combining a complex three-stage SFT pipeline with subsequent RL, though its intricate design can be challenging to implement and susceptible to accumulated training bias.

While RL can further unlock hybrid latent reasoning, standard methods are not fully aligned with the hyperspherical geometry and hybrid action variance of MLLM representations, motivating our Decoupled Policy Optimization (DePO).

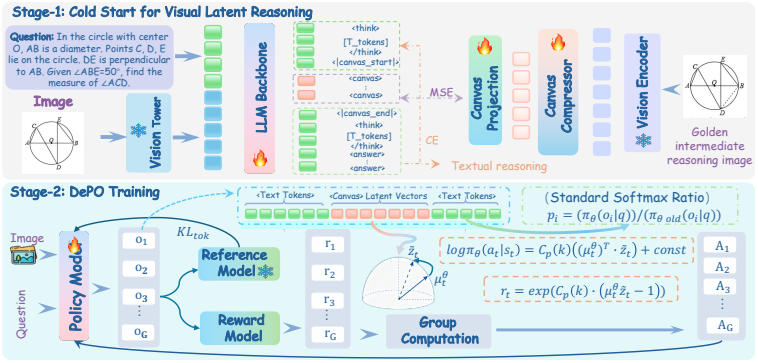

Fig. 2: Overview of the **HyLaR** framework. **Stage-I (SFT):** jointly optimizes text via cross-entropy ($\mathcal{L}_{CE}$) and aligns hidden states with compressed ground-truth canvases via MSE ($\mathcal{L}_{Canvas}$). **Stage-II (DePO):** refines the hybrid trajectory with RL—text via probability ratios, latent vectors via vMF cosine similarity on a hypersphere.

3 Method

In this section, we present **HyLaR** (**Hy**brid **La**tent **R**easoning), a novel framework that empowers MLLMs to seamlessly interleave discrete textual reasoning

with continuous visual latent representations. We first provide an overview of the HyLaR architecture, including its hybrid generation mechanism and the distinct paradigms for training and inference (§3.1). Subsequently, we detail the two-stage training methodology: the cold-start SFT with canvas compression and alignment strategy (§3.2), followed by the **DePO** (**De**coupled **P**olicy **O**ptimization) to further unlock and stabilize latent reasoning capabilities via RL (§3.3).

3.1 Overview of the HyLaR Framework

HyLaR is built upon a standard MLLM architecture, but fundamentally extends the traditional discrete decoding space into a hybrid discrete-continuous action space. To achieve this, we introduce specialized control tokens, `<|canvas_start|>` and `<|canvas_end|>`, to explicitly bound the visual reasoning process. During inference, when the model requires fine-grained spatial or visual deduction, it transitions into the canvas mode. In this mode, textual generation is suspended, and the hidden state from the previous layer is recurrently fed back as the input embedding for the next step, bypassing the discrete token vocabulary. This continuous latent recursion acts as an internal visual working memory and continues autonomously until the `<|canvas_end|>` token is emitted or reaches the maximum canvas length budget, effectively enabling deep visual thinking while avoiding the substantial latency of pixel-level image generation.

HyLaR is trained in two stages. We first design a cold-start SFT phase that teaches the model to alternate between text tokens and latent canvas steps, supervising the latter with compressed ground-truth canvases instead of high-resolution images for efficiency (detailed in §3.2); crucially, this auxiliary Canvas Compressor is entirely discarded after SFT, ensuring zero architectural overhead at inference. Building on this alignment, we further refine the hybrid generation via Decoupled Policy Optimization (DePO), which applies independent trust-region constraints to optimize reasoning trajectories against task-specific rewards without any explicit intermediate visual supervision (§3.3).

3.2 Stage I: Supervised Fine-Tuning

The supervised fine-tuning (SFT) stage equips the MLLM to seamlessly alternate between textual reasoning and latent visual thinking by approximating the semantic representations of ground-truth intermediate canvases.

Canvas Compression and Injection. During training, ground-truth canvas images are first processed by a frozen SigLIP2 encoder [30] to extract $P = 729$ patch tokens. To avoid the exorbitant sequence length caused by these raw high-resolution patches, a learnable cross-attention compressor is adopted. Configured with $L = 2$ layers and $N = 16$ query tokens, the compressor attentively aggregates these patches into N compact canvas embeddings $\mathbf{e}$. Within the training reasoning trajectory, original intermediate images are replaced by `<canvas>` placeholders bounded by `<|canvas_start|>` and `<|canvas_end|>`, into which the compressed target embeddings $\mathbf{e}$ are directly injected via masked scattering.

Joint End-to-End Optimization. The model is jointly trained to predict both the next discrete text tokens and the next continuous latent canvas embeddings. We apply the standard autoregressive cross-entropy loss ($\mathcal{L}_{CE}$) for text generation. Simultaneously, we optimize an MSE-based canvas prediction loss:

$$\mathcal{L}_{\text{Canvas}} = \frac{1}{|\mathcal{S}|} \sum_{t \in \mathcal{S}} \|\mathbf{h}_t - \mathbf{e}_{t+1}\|_2^2 , \tag{1}$$

where $\mathbf{h}_t$ is the LLM's hidden state at the canvas position $t \in \mathcal{S}$. Crucially, we do not detach the target embeddings $\mathbf{e}_{t+1}$ during backpropagation. This formulation allows gradients to flow end-to-end from $\mathcal{L}_{Canvas}$ into both the LLM backbone and the external Canvas Compressor, actively pulling the extracted visual features into the LLM's native semantic space. The overall SFT objective is defined as $\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{Canvas}}$.

3.3 Stage II: Reinforcement Learning with DePO

Limitations of Standard RL Objectives. Standard reinforcement learning objectives, such as those used in PPO [23], GRPO [24], and DAPO [41], typically apply a shared surrogate loss with a uniform clipping rule and sample-based KL regularization across all action steps. While this design is effective for purely discrete text generation, it is not well calibrated for our hybrid reasoning trajectory, where discrete tokens and continuous latent states exhibit fundamentally different statistical and geometric properties. In particular, directly treating these two action types in the same way leads to two key limitations.

(1) Optimization mismatch between discrete and continuous actions. The importance ratio $r_t = \pi_\theta(a_t \mid s_t)/\pi_{\theta_{\text{old}}}(a_t \mid s_t)$ behaves differently in discrete token spaces and in high-dimensional continuous latent spaces. For discrete text generation, action probabilities are defined over a normalized categorical distribution, and token-level updates are typically relatively well behaved under standard clipping schemes. By contrast, for continuous latent actions, the log-density often depends on distances or directional deviations in a high-dimensional space, so even small policy changes can induce disproportionately large changes in likelihood ratios (see Fig. 3). This issue is further amplified by autoregressive latent recursion: each latent action is fed into subsequent reasoning steps, so small perturbations can alter future states and accumulate along the rollout. As a result, a single clipping range or regularization strength that is adequate for token

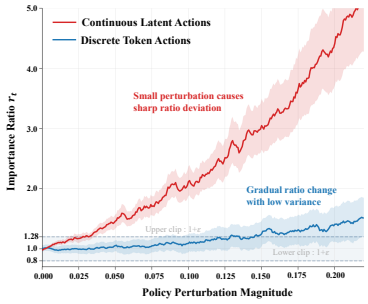

Fig. 3: Importance ratio r_t vs. policy perturbation magnitude $\|\theta - \theta_{\text{old}}\|_2 / \|\theta_{\text{old}}\|_2$ for discrete token vs. continuous latent actions. Latent ratios deviate far more sharply under small policy updates, motivating tighter clipping for latent positions.

updates may be too loose or poorly calibrated for latent updates, leading to unstable optimization.

(2) Geometric mismatch in continuous latent policy modeling. Standard continuous RL typically assumes a Gaussian action space, which implies a flat Euclidean geometry. However, modern MLLMs heavily employ normalization layers (e.g., RMSNorm [43]) that inherently constrain hidden states to a high-dimensional hyperspherical manifold [33]. As highlighted by prior representation learning studies [6], applying sample-based Euclidean Gaussian estimation to these hyperspheres suffers from severe geometric distortion. In high-dimensional spaces, this mismatch results in prohibitively high sampling variance for the KL penalty, failing to effectively regularize the policy.

To address these issues, we propose *Decoupled Policy Optimization* (DePO), which explicitly models continuous latent reasoning steps with a hyperspherical policy, while decoupling stable rollout execution from stochastic policy optimization.

Hyperspherical vMF Modeling. We model continuous latent actions using the von Mises–Fisher (vMF) distribution, which is the natural probability distribution on the unit hypersphere and has recently shown promise in continuous reinforcement learning. In our setting, the model performs latent recursion between the `<|canvas_start|>` and `<|canvas_end|>` delimiters:

$$\mathbf{h}_t^\theta = f_\theta(\mathbf{h}_{t-1}^\theta), \tag{2}$$

where $\mathbf{h}_t^\theta \in \mathbb{R}^D$ denotes the hidden state at step t. This induces a hybrid action space in which, at step t, the action a_t is either a discrete token $a_t \in \mathcal{V}$ or a continuous latent vector $a_t \in \mathbb{S}^{D-1}$. For discrete token positions, the policy follows a standard categorical distribution over the vocabulary $\mathcal{V}$:

$$\pi_\theta(a_t \mid s_t) = \mathrm{softmax}(\mathbf{W}_m \mathbf{h}_t^\theta)_{a_t}, \tag{3}$$

where $\mathbf{W}_m$ denotes the output projection matrix. For continuous latent positions, a categorical policy is no longer applicable because the action space is continuous. Instead, we define the latent policy density of the vMF policy on the hypersphere as

$$\pi_\theta(a_t \mid s_t) = C_D(\kappa) \exp\left(\kappa (\boldsymbol{\mu}_t^\theta)^\top \tilde{\mathbf{z}}_t\right), \qquad \tilde{\mathbf{z}}_t \in \mathbb{S}^{D-1}, \tag{4}$$

where $\boldsymbol{\mu}_t^\theta \in \mathbb{S}^{D-1}$ is the ℓ_2-normalized hidden state $\mathbf{h}_t^\theta$, $\kappa > 0$ is a fixed concentration parameter and $C_D(\kappa)$ is the vMF normalization constant in dimension D.

Combining the discrete and continuous cases, we write the unified policy log-probability as

$$\log \pi_\theta(a_t \mid s_t) = \begin{cases} \log \pi_\theta(a_t \mid s_t), & a_t \in \mathcal{V}, \\ \log C_D(\kappa) + \kappa (\boldsymbol{\mu}_t^\theta)^\top \tilde{\mathbf{z}}_t, & a_t \in \mathbb{S}^{D-1}. \end{cases} \tag{5}$$

Since κ is fixed, the normalization term $\log C_D(\kappa)$ cancels when computing policy ratios between the old and new policies.

Decoupled Surrogate Objectives. Because the unified log-probability in Eq. (5) places discrete token actions and continuous latent actions on the same sequence, the importance-sampling ratio r_t can exhibit very different magnitudes at text positions versus latent positions. In particular, small directional shifts of the hidden-state μ_t^θ can cause disproportionately large swings in the vMF ratio at latent positions, destabilizing training.

To address this, we partition each response into the set of *text positions* $\mathcal{Z}$ and *latent positions* $\mathcal{S}$, and apply the dual-clip PPO surrogate independently to each set with separate clipping ranges:

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathcal{L}_{\text{tok}}(\theta \mid \mathcal{Z}; \epsilon_l^{\text{tok}}, \epsilon_h^{\text{tok}}) + \alpha \mathcal{L}_{\text{lat}}(\theta \mid \mathcal{S}; \epsilon_l^{\text{lat}}, \epsilon_h^{\text{lat}}), \tag{6}$$

where θ denotes the model parameters, and $\mathcal{L}_{\text{tok}}(\theta \mid \mathcal{Z}; \epsilon_l^{\text{tok}}, \epsilon_h^{\text{tok}})$ represents the dual-clip PPO objective evaluated over the token positions $\mathcal{Z}$. Here, ϵ_l^{tok} and ϵ_h^{tok} are the asymmetric lower and upper clip thresholds, respectively. The latent loss $\mathcal{L}_{\text{lat}}$ is defined analogously over latent positions $\mathcal{S}$, with α balancing the two terms. Crucially, we use a substantially tighter clipping range for latent positions ($\epsilon^{\text{lat}} \ll \epsilon^{\text{tok}}$) to counteract the higher ratio volatility inherent to continuous vMF actions. We set $\epsilon_l^{\text{tok}} = 0.2$, $\epsilon_h^{\text{tok}} = 0.28$, and $\epsilon_l^{\text{lat}} = \epsilon_h^{\text{lat}} = 0.05$ in our experiments.

Closed-form vMF KL Regularization. To prevent policy degradation, we apply position-specific KL penalties. For text token positions, we utilize the standard sample-based KL penalty against the reference policy π_{ref}. For latent positions, the vMF assumption allows us to completely bypass the high-variance sample-based estimation. As proven in supplemental materials, the exact KL divergence between two vMF distributions sharing the same concentration κ elegantly reduces to a scaled cosine distance. Thus, the latent KL loss is computed in closed-form:

$$\mathcal{L}_{\text{KL}}^{\text{lat}} = \frac{1}{|\mathcal{S}|} \sum_{t \in \mathcal{S}} \mathbf{W}_\kappa \cdot \left(1 - \cos\left(\boldsymbol{\mu}_t^\theta, \tilde{\mathbf{z}}_t\right)\right), \tag{7}$$

where μ_t^θ is the current policy's normalized hidden state (retaining gradients), $\tilde{\mathbf{z}}_t$ is the previously sampled latent vector from the old policy, and $\mathbf{W}_\kappa$ is a constant weight derived from κ. This objective provides an exact, zero-variance KL estimate that natively respects the spherical geometry of LLM representations.

Total Objective. The overall training objective combines the decoupled surrogate losses with the position-specific KL penalties:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{PPO}} + \beta_{\text{tok}} \mathcal{L}_{\text{KL}}^{\text{tok}} + \beta_{\text{lat}} \mathcal{L}_{\text{KL}}^{\text{lat}}, \tag{8}$$

where β_{tok} and β_{lat} control the regularization strength for text tokens and latent tokens, respectively. In practice, we set $\beta_{\text{tok}} = 0.01$ and $\beta_{\text{lat}} = 0.005$.

4 Experiments

We conduct extensive experiments across multiple benchmarks to evaluate our method. Specifically, our experiments aim to: (1) demonstrate the effectiveness

Table 1: Results on High-Resolution Image Perception and Visual Search Benchmarks. The best-performing latent reasoning model is highlighted in **bold**. (*Reproduced via our evaluation pipeline for fair comparison; original reported scores are in gray.)

Methods	V*			HRBench-4K			HRBench-8K		
	Overall	Attribute	Spatial	Overall	FSP	FCP	Overall	FSP	FCP
Proprietary Model									
GPT-4o [15]	67.5	72.2	60.5	59.0	70.0	48.0	55.5	62.0	49.0
Gemini-3-Flash [8]	86.4	-	-	87.9	-	-	85.0	-	-
Open-Source Model									
LLaVA-OneVision-7B [18]	71.25	73.48	67.89	62.50	74.25	50.75	58.25	67.50	49.00
Qwen2.5-VL-7B [2]	76.44	77.39	75.00	68.00	80.25	55.75	63.75	73.75	53.75
InternVL3-8B [48]	70.2	67.8	73.7	70.0	78.8	61.3	69.3	78.8	59.8
Thinking-with-Images Agent Model									
ZoomEye [25]	79.85	80.52	78.82	68.75	81.25	56.25	64.50	75.00	54.00
Thyme [44]	82.2	83.5	80.3	77.0	91.0	63.0	72.0	86.5	57.5
DeepEyes [47]	85.6	-	-	75.1	-	-	72.6	-	-
DeepEyesV2 [12]	81.8	-	-	77.9	-	-	73.8	-	-
Visual-Latent Model									
LVR [17]	80.6	81.7	79.0	-	-	-	-	-	-
Laser [36]	-	-	-	72.50	-	-	-	-	-
SkiLa* [28]	78.53	-	-	72.12	-	-	66.5	-	-
SkiLa [28]	84.3	-	-	72.0	-	-	66.5	-	-
Monet* [32]	80.1	81.73	77.63	67.37	78.25	**56.50**	64.37	74.25	**54.49**
Monet [32]	83.25	83.48	82.89	71.00	85.25	56.75	68.00	79.75	56.25
HyLaR-SFT (Ours)	80.63	81.73	78.95	71.50	93.00	50.00	67.19	87.01	47.11
HyLaR-7B (Ours)	**83.77**	**82.61**	**85.53**	**75.00**	**93.75**	56.25	**70.50**	**88.25**	52.75
Δ (vs Qwen2.5-VL 7B)	↑ 7.33	↑ 5.22	↑ 10.53	↑ 7.00	↑ 13.50	↑ 0.50	↑ 6.75	↑ 14.50	↓ 1.00

of the proposed approach against competitive baselines, (2) analyze the contribution of individual components through ablation studies, and (3) provide insights into visual latent reasoning.

4.1 Experimental Settings

Training Data. In the SFT stage, following prior work, we use the Zebra-CoT [16] dataset for cold-start training. Zebra-CoT covers four major task categories: scientific problems, 2D visual reasoning, 3D visual reasoning, and visual logic and strategy games, thereby providing diverse and logically coherent multimodal reasoning trajectories across a broad range of domains. In the RL stage, we curate and combine samples from the DeepEyes [47], Thyme [44] and CodeDance [27] datasets to construct the training set.

Benchmarks. We report results on three high-resolution benchmarks targeting ultra-high-definition perception and visual search: HRBench (4K & 8K) [35] and V* [38]. Additionally, we evaluate on a robust suite of general and diagnostic

VQA benchmarks to cover various specialized capabilities: MMStar [4] (visual dependency), MMVP [29] (fine-grained visual patterns), SeedBench2Plus [19] (multimodal reasoning), BLINK [7] (core visual perception), and HallusionBench [9] (illusion and hallucination diagnostics).

Baselines. To rigorously evaluate our approach, we benchmark it against four categories of representative baselines: (1) **Proprietary Frontier MLLMs**, including GPT-4o [15] and Gemini-3-Flash [8]; (2) **Open-source General MLLMs**, including LLaVA-OneVision-7B [18] and Qwen2.5-VL-7B [2]; (3) **Thinking-with-Images Agent Models**, including ZoomEye [25], DeepEyes [47], DeepEyesV2 [12], and Thyme [44]; (4) **Visual Latent Reasoning Models**, including LVR [17], Laser [36], Monet [32], and SkiLa [28].

Implementation Details. In the SFT stage, we limit training to a single epoch to mitigate overfitting. The learning rate is set to 10^{-5}, the per-GPU batch size to 1, and the gradient accumulation steps to 16. Following this cold-start phase, we continue with DePO-based reinforcement learning, reducing the learning rate to 10^{-6}. More implementation details can be found in supplemental materials.

Table 2: Results on Multimodal VQA, Reasoning and Hallucination Benchmarks. The best-performing latent reasoning model for each dataset is highlighted in **bold**.

Methods	MMVP	MMStar	SeedBench2Plus	BLINK	HallusionBench
Open-Source Model					
Qwen2.5-VL-7B [2]	65.67	59.70	65.31	53.60	56.57
LLaVA-OneVision [18]	73.00	59.13	61.22	49.34	51.10
InternVL3.5-8B [34]	57.67	53.33	69.78	54.81	56.15
Thinking-with-Images Agent Model					
ZoomEye [25]	72.67	63.20	70.27	55.55	71.08
Thyme [44]	-	65.90	-	56.10	55.60
DeepEyes [47]	70.00	58.73	69.08	51.08	62.57
Visual-Latent Model					
LVR [17]	64.00	57.93	47.39	53.60	65.19
Laser [36]	72.00	60.27	70.05	56.92	**67.72**
Monet [32]	68.00	60.33	65.88	50.71	56.36
HyLaR-SFT (Ours)	71.00	60.47	69.39	54.50	60.23
HyLaR-7B (Ours)	**73.67**	**62.00**	**70.32**	**57.14**	63.68
Δ (vs Qwen2.5-VL 7B)	↑ 8.00	↑ 2.30	↑ 5.01	↑ 3.54	↑ 7.11

4.2 Main Results

High-Resolution Image Perception and Visual Search Benchmarks. In the benchmarks reported in Table 1, the target regions occupy an extremely small fraction of ultra high resolution images, typically only 100 to 200 pixels.

Accurately localizing such tiny targets in large-scale, visually cluttered imagery poses not only a substantial perceptual challenge for large models but also readily triggers severe visual hallucinations. As shown in the table, our model achieves significant performance gains across all three high-resolution benchmarks, and on some tasks it matches or surpasses Agent-based models that rely on external tools. Concretely, compared with the baseline Qwen2.5-VL-7B [2], our model improves by 7.33% and 7.00% on V* [38] and HRBench-4K [35], respectively. Notably, our model achieves state-of-the-art performance in visual latent space reasoning. On V* [38], it outperforms SkiLa (78.53%) [28] and Monet (80.10%) [32] by margins of 5.24% and 3.67%, respectively. This superiority extends to other high-resolution evaluations [35]: on HRBench-4K, our model exceeds SkiLa (72.12%) and Monet (67.37%) by 2.88% and 7.63%; similarly, on HRBench-8K, it surpasses their respective scores of 66.50% and 64.37% by 4.00% and 6.13%.

General VQA and Hallucination Benchmarks. Beyond fundamental capabilities, we further validate the effectiveness of our model on general VQA and hallucination evaluation benchmarks. Experimental results in Table 2 demonstrate that HyLaR substantially improves VQA accuracy while significantly mitigating hallucination phenomena in model generation. This can be attributed to HyLaR's capability of leveraging latent space representation guidance to enable deep exploration and fine-grained decomposition of specific local regions during inference, thereby achieving precise verification of visual content. These findings indicate that HyLaR successfully integrates high-resolution perception with an efficient verification mechanism, providing robust guarantees for the reliability of multimodal models.

Efficiency Analysis. Benchmarked on an H20 GPU, HyLaR adds just 16 latent tokens and discards the SFT canvas extractor at inference, incurring zero architectural overhead. Unlike tool-based methods requiring detokenization, re-encoding, and external API calls, it reasons within a single forward pass. As shown in Table 3, HyLaR runs at 0.96× latency and 0.98× memory of the Qwen2.5-VL-7B long-CoT baseline, whereas DeepEyes costs 2.28× and 3.64×.

Table 3: Inference efficiency (normalized to Qwen2.5-VL-7B).

Method	Latency	Memory
Qwen2.5-VL-7B	1.00×	1.00×
DeepEyes (multi-turn)	2.28×	3.64×
HyLaR (N=16)	**0.96**×	0.98×

Scalability to Smaller Model Sizes. To validate generalizability beyond the 7B scale, we evaluate a 3B variant against its backbone and the only available latent reasoning baseline at this scale. As shown in Table 4, HyLaR-3B consistently outperforms both Qwen2.5-VL-3B and LVR-3B across all benchmarks, confirming that our framework scales effectively without loss of advantage.

Table 4: Results of 3B-parameter models. **Bold** denotes the best latent reasoning model at the 3B scale.

Method	V*	HRBench-4K	HRBench-8K	MMVP	MMStar	SeedBench2Plus
Qwen2.5-VL-3B	**70.68**	61.75	60.25	62.00	52.40	58.82
LVR-3B	67.00	-	-	58.00	-	-
HyLaR-3B (Ours)	68.58	**62.88**	**61.25**	**63.00**	**54.20**	**59.21**

4.3 Ablation Studies

In this section, we conduct comprehensive ablation studies to evaluate the necessity and effectiveness of each component within our framework.

Ablation on the Number of Latent Steps. To comprehensively understand the scaling behavior and generalization of latent reasoning, we investigate the interplay between training (K_{train}) and inference (K_{test}) reasoning steps across both SFT and RL stages. For the SFT stage, we train models with $K_{\text{train}} \in \{8, 16, 24\}$. To further explore if RL can extrapolate reasoning beyond SFT priors, we initialize RL models from the SFT-8-steps checkpoint and optimize them with varying horizons: $K_{\text{train}} \in \{8, 16, 32\}$. We systematically evaluate all models by varying K_{test} from 0 to 64 across two benchmarks: V* and HRBench-8K (as illustrated in Fig. 4). This unified setup reveals several key findings: **(1) Latent reasoning consistently yields significant gains.** Across all benchmarks, introducing latent tokens ($K_{\text{test}} > 0$) strictly outperforms the zero-latent baseline ($K_{\text{test}} = 0$). For instance, on V*, SFT-latent-16 improves from 74.87% ($K_{\text{test}} = 0$) to 80.63% ($K_{\text{test}} = 16$), a substantial gain of $+$**5.76%**. This demonstrates that latent tokens provide crucial "thinking time" for deeper implicit computation. **(2) SFT models suffer from "over-thinking" degradation.** In the SFT stage, performance consistently degrades when K_{test} significantly exceeds K_{train}. This indicates that excessive latent steps introduce noise or drift in hidden representations. Among SFT configurations, moderate training depth ($K_{\text{train}} = 16$) provides the most balanced trade-off between peak accuracy and generalization stability compared to $K_{\text{train}} = 8$ or 24. **(3) RL unlocks reasoning extrapolation and mitigates over-thinking.** Unlike SFT models that overfit to their training length, RL optimization significantly enhances length generalization. Models trained with RL, particularly with longer horizons ($K_{\text{train}} \in \{16, 32\}$), maintain robust performance even when $K_{\text{test}} \gg K_{\text{train}}$, effectively mitigating the representation drift seen in SFT. Furthermore, RL pushes the peak accuracy beyond the upper bound of SFT, demonstrating that reinforcement learning can intrinsically align latent representations to better utilize extended inference budgets, an extrapolation capability that pure supervised learning fails to achieve.

Ablation on Different RL Algorithms. To validate the effectiveness of our latent reasoning training approach, we conduct an ablation study by masking the intermediate thinking images in the SFT training data and performing standard

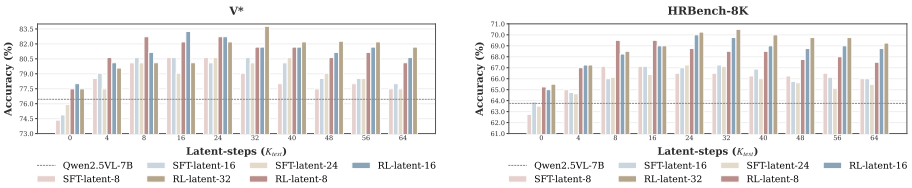

Fig. 4: Ablation on inference latent steps (K_{test}). SFT and RL models trained with varying horizons (K_{train}), evaluated on V* and HRBench-8K. The dashed line marks the Qwen2.5-VL-7B baseline.

SFT. As shown in Table 5, the model trained with conventional SFT consistently underperforms our proposed method across various benchmarks, demonstrating the superiority of the latent approach. Furthermore, we compare DePO against three alternative optimization methods: GRPO [24], DAPO [41], and VLPO [32], with evaluations conducted on V*, HRBench (4K & 8K), and MMVP. The results demonstrate that DePO, with its hybrid optimization approach, achieves superior performance compared to these baseline RL algorithms. Notably, the models trained using GRPO and DAPO even showed a significant decline on the V* benchmark (from 80.63% to 78.53% and 79.06%), while VLPO only achieves marginal improvements over HyLaR-SFT. These compelling results highlight the remarkable effectiveness of DePO in dramatically advancing the performance of visual latent reasoning models on fine-grained feature understanding tasks.

Table 5: Ablation on different RL algorithms. All RL methods start from the same HyLaR-SFT checkpoint trained with 8 latent steps.

Methods	V*	HRBench-4K	HRBench-8K	MMVP	Avg.
Supervised Fine-tuning					
Text-only SFT	69.11	65.88	61.00	70.20	66.55
HyLaR-SFT	80.63	71.50	67.19	71.00	72.58
Reinforcement Learning (from HyLaR-SFT)					
+ GRPO [24]	78.53	71.50	67.50	71.40	72.23
+ DAPO [41]	79.06	71.75	68.00	70.20	72.25
+ VLPO [32]	80.10	72.84	68.00	69.67	72.65
+ DePO (Ours)	**83.77**	**75.00**	**70.50**	**73.67**	**75.74**

Ablation on DePO Hyper-parameters. We ablate two key hyper-parameters in DePO's decoupled policy loss $\mathcal{L}_{\text{PPO}} = \mathcal{L}_{\text{tok}} + \alpha \cdot \mathcal{L}_{\text{lat}}$: the latent loss weight α and the latent clipping ratio $[\epsilon_l^{\text{lat}}, \epsilon_h^{\text{lat}}]$, evaluating on V*, HRBench-8K, and MMStar.

Table 6: Ablation on latent loss weight α. We vary the latent loss weight α in $\mathcal{L}_{\text{PPO}} = \mathcal{L}_{\text{tok}} + \alpha \cdot \mathcal{L}_{\text{lat}}$ while keeping all other hyper-parameters fixed. During evaluation, we set the latent step to 32.

α	V*	HRBench-8K	MMStar
0.0	80.63	68.63	60.17
0.2	81.15	69.38	60.04
0.3	81.68	70.00	60.71
0.5	**83.77**	**70.50**	**62.00**
1.0	83.25	70.50	61.51
2.0	80.63	69.50	61.04

Table 7: Effect of latent clipping ratio ϵ^{lat}. We set $\epsilon_l^{\text{lat}} = \epsilon_h^{\text{lat}} = \epsilon^{\text{lat}}$ (symmetric clipping). Token clipping is fixed at $(\epsilon_l^{\text{tok}}, \epsilon_h^{\text{tok}}) = (0.20, 0.28)$. During evaluation, we set the latent step to 32.

ϵ^{lat}	V*	HRBench-8K	MMStar
0.0	79.58	69.37	61.00
0.01	82.72	70.00	61.40
0.025	82.72	**70.75**	61.67
0.05	**83.77**	70.50	**62.00**
0.1	83.77	70.00	61.47
0.2	80.10	69.00	60.87

Latent loss weight α. As shown in Table 6, $\alpha = 0$ (no RL signal to latents) consistently underperforms across all three benchmarks, confirming that latent tokens must be actively optimized. Increasing α to 2.0 also degrades results; although our vMF formulation naturally mitigates sampling variance, an excessively large weight causes the continuous latent updates to disproportionately overwhelm the discrete text gradients, disrupting the delicate balance of the hybrid policy. Setting $\alpha = 0.5$ strikes the best balance between sufficient latent optimization and training stability.

Latent clipping ratio $[\epsilon_l^{\text{lat}}, \epsilon_h^{\text{lat}}]$. Table 7 shows that applying a clipping ratio equivalent to the lower bound of the text tokens ($\epsilon_l^{\text{lat}} = \epsilon_h^{\text{lat}} = \epsilon_l^{\text{tok}} = 0.2$) leads to clear performance drops, since the continuous density ratio inherently exhibits different variance characteristics and requires a tighter trust region. Conversely, overly aggressive clipping ($\epsilon_l^{\text{lat}} = \epsilon_h^{\text{lat}} = 0.01$) suppresses useful policy updates. Setting $\epsilon_l^{\text{lat}} = \epsilon_h^{\text{lat}} = 0.05$ achieves the best overall performance across V*, HRBench-8K, and MMStar, validating the necessity of decoupled clipping.

Ablation on Latent Distribution Assumptions. Our RL framework models continuous latent embeddings with the **von Mises-Fisher (vMF) distribution**, optimizing the policy via cosine similarity (directional alignment) between generated and rollout embeddings. To justify this choice, we replace vMF with a standard **Gaussian distribution**, under which embeddings are sampled around the policy output $\mathbf{h}_{i,t}^{\theta}$ and the probability ratio $r_{i,t}(\theta)$ is driven by Euclidean (ℓ_2) distance rather than angular similarity:

$$r_{i,t}^{\text{Gaussian}}(\theta) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{h}_{i,t}^{\text{old}} - \mathbf{h}_{i,t}^{\theta}\|^2\right) \tag{9}$$

where σ controls the variance, tuned so the initial gradient scales match our vMF setting for fair comparison. Results are summarized in Table 8.

Based on the results in Table 8, we draw the following conclusions: **(1) Directional alignment is more effective than spatial distance for latent**

Table 8: Ablation on the distribution assumptions for latent embeddings during RL. The vMF distribution (our main method) consistently outperforms the Gaussian variant across all benchmarks.

Methods / Distribution Assumption	V*	HRBench-4K	HRBench-8K	SeedBench2Plus	MMStar	MMVP
Baseline (SFT only, w/o RL)	80.63	71.50	67.19	65.31	60.47	71.00
RL with **Gaussian** ($\sigma = 10.0$)	83.25	74.00	69.50	67.28	61.27	72.33
RL with **vMF** (Ours, $\kappa = 0.01$)	**83.77**	**75.00**	**70.50**	**70.32**	**62.00**	**73.67**

space. While the Gaussian assumption improves upon the SFT baseline, it consistently underperforms our vMF-based approach across all benchmarks. In high-dimensional spaces (such as the hidden states of LLMs), semantic information is primarily encoded in the direction of the vectors rather than their magnitude. By explicitly optimizing cosine similarity, the vMF distribution aligns perfectly with the nature of dense representations, leading to more meaningful policy updates. **(2) vMF provides natural regularization against norm explosion.** Optimizing the Euclidean distance under the Gaussian assumption can lead to training instability. Specifically, to maximize the advantage $\hat{A}_{i,t}$, the policy might trivially increase the magnitude (norm) of the latent embeddings to push them further apart, rather than learning better semantic representations. The vMF distribution, which inherently operates on the unit hypersphere, acts as a natural regularizer. It forces the model to focus solely on angular updates, thereby stabilizing the RL training process.

5 Conclusion

In this work, we propose HyLaR (Hybrid Latent Reasoning), a novel framework designed to overcome early semantic collapse in MLLMs by seamlessly interleaving discrete textual logic with continuous visual working memory. To effectively optimize this hybrid discrete-continuous action space, we introduce Decoupled Policy Optimization (DePO). By identifying the geometric and variance mismatches inherent in standard RL, DePO leverages von Mises-Fisher (vMF) spherical modeling and decoupled trust-region clipping to achieve exact, stable, and geometrically rigorous latent policy optimization. Extensive experiments demonstrate that HyLaR significantly outperforms existing text-only and tool-augmented paradigms on fine-grained perception and reasoning benchmarks, exhibiting remarkable robustness to visual hallucinations. In future work, we aim to extend this hybrid RL paradigm to open-ended agentic environments and explore scaling laws for longer-horizon latent planning.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)

2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025)
3. Bendada, W., Salha-Galvan, G., Hennequin, R., Bontempelli, T., Bouabça, T., Cazenave, T.: Exploring large action sets with hyperspherical embeddings using von mises-fisher sampling. In: Forty-second International Conference on Machine Learning (2025)
4. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? Advances in Neural Information Processing Systems **37**, 27056–27087 (2024)
5. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
6. Davidson, T.R., Falorsi, L., De Cao, N., Kipf, T., Tomczak, J.M.: Hyperspherical variational auto-encoders. In: Conference on Uncertainty in Artificial Intelligence (UAI) (2018)
7. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)
8. Google DeepMind: Gemini 3 flash model card. Tech. rep., Google (2025), `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf`
9. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: 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. pp. 14375–14385 (2024)
10. Gupta, T., Kembhavi, A.: Visual programming: Compositional visual reasoning without training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14953–14962 (2023)
11. Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., Tian, Y.: Training large language models to reason in a continuous latent space. arXiv preprint arXiv:2412.06769 (2024)
12. Hong, J., Zhao, C., Lu, W., Zhu, C., Xu, G., XingYu: Deepeyesv2: Toward agentic multimodal model. In: The Fourteenth International Conference on Learning Representations (2026)
13. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025)
14. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025)
15. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)

16. Li, A., Wang, C.L., Fu, D., Yue, K., Cai, Z., Zhu, W.B., Liu, O., Guo, P., Neiswanger, W., Huang, F., Goldstein, T., Goldblum, M.: Zebra-cot: A dataset for interleaved vision-language reasoning. In: The Fourteenth International Conference on Learning Representations (2026)
17. Li, B., Sun, X., Liu, J., Wang, Z., Wu, J., Yu, X., Chen, H., Barsoum, E., Chen, M., Liu, Z.: Latent visual reasoning. arXiv preprint arXiv:2509.24251 (2025)
18. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
19. Li, B., Ge, Y., Chen, Y., Ge, Y., Zhang, R., Shan, Y.: Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. arXiv preprint arXiv:2404.16790 (2024)
20. Luo, R., Li, Y., Chen, L., He, W., Lin, T.E., Liu, Z., Zhang, L., Song, Z., Alinejad-Rokny, H., Xia, X., Liu, T., Hui, B., Yang, M.: DEEM: Diffusion models serve as the eyes of large language models for image perception. In: The Thirteenth International Conference on Learning Representations (2025)
21. Mi, Z., Wang, K.C., Qian, G., Ye, H., Liu, R., Tulyakov, S., Aberman, K., Xu, D.: I think, therefore i diffuse: Enabling multimodal in-context reasoning in diffusion models. In: Forty-second International Conference on Machine Learning (2025)
22. Ray, A., Abdelkader, A., Mao, C., Plummer, B.A., Saenko, K., Krishna, R., Guibas, L., Chu, W.S.: Mull-tokens: Modality-agnostic latent thinking. arXiv preprint arXiv:2512.10941 (2025)
23. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
24. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
25. Shen, H., Zhao, K., Zhao, T., Xu, R., Zhang, Z., Zhu, M., Yin, J.: Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 6613–6629 (2025)
26. Shen, Z., Yan, H., Zhang, L., Hu, Z., Du, Y., He, Y.: CODI: Compressing chain-of-thought into continuous space via self-distillation. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 677–693 (2025)
27. Song, Q., Li, H., Yu, Y., Zhou, H., Yang, L., Bai, S., She, Q., Huang, Z., Zhao, Y.: Codedance: A dynamic tool-integrated mllm for executable visual reasoning. arXiv preprint arXiv:2512.17312 (2025)
28. Tong, J., Gu, J., Lou, Y., Fan, L., Zou, Y., Wu, Y., Ye, J., Li, R.: Sketch-in-latents: Eliciting unified reasoning in mllms. arXiv preprint arXiv:2512.16584 (2025)
29. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9568–9578 (2024)
30. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
31. Wang, H., Qu, C., Huang, Z., Chu, W., Lin, F., Chen, W.: Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. arXiv preprint arXiv:2504.08837 (2025)

32. Wang, Q., Shi, Y., Wang, Y., Zhang, Y., Wan, P., Gai, K., Ying, X., Wang, Y.: Monet: Reasoning in latent visual space beyond images and language. arXiv preprint arXiv:2511.21395 (2025)
33. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International conference on machine learning. pp. 9929–9939. PMLR (2020)
34. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
35. Wang, W., Ding, L., Zeng, M., Zhou, X., Shen, L., Luo, Y., Yu, W., Tao, D.: 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. pp. 7907–7915 (2025)
36. Wang, Y., Zhang, J., Wu, Y., Lin, Y., Lukas, N., Liu, Y.: Forest before trees: Latent superposition for efficient visual reasoning. arXiv preprint arXiv:2601.06803 (2026)
37. Wei, X., Liu, X., Zang, Y., Dong, X., Cao, Y., Wang, J., Qiu, X., Lin, D.: SIM-cot: Supervised implicit chain-of-thought. In: The Fourteenth International Conference on Learning Representations (2026)
38. Wu, P., Xie, S.: V*: Guided visual search as a core mechanism in multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13084–13094 (2024)
39. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. In: The Thirteenth International Conference on Learning Representations (2025)
40. Yang, Z., Yu, X., Chen, D., Shen, M., Gan, C.: Machine mental imagery: Empower multimodal reasoning with latent visual tokens. arXiv preprint arXiv:2506.17218 (2025)
41. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., YuYue, Dai, W., Fan, T., Liu, G., Liu, J., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, M., Zhang, R., Zhang, W., Zhu, H., Zhu, J., Chen, J., Chen, J., Wang, C., Yu, H., Song, Y., Wei, X., Zhou, H., Liu, J., Ma, W.Y., Zhang, Y.Q., Yan, L., Wu, Y., Wang, M.: DAPO: An open-source LLM reinforcement learning system at scale. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
42. Yu, X., Chen, Z., He, Y., Fu, T., Yang, C., Xu, C., Ma, Y., Hu, X., Cao, Z., Xu, J., et al.: The latent space: Foundation, evolution, mechanism, ability, and outlook. arXiv preprint arXiv:2604.02029 (2026)
43. Zhang, B., Sennrich, R.: Root mean square layer normalization. Advances in neural information processing systems **32** (2019)
44. Zhang, Y.F., Lu, X., Yin, S., Fu, C., Chen, W., Hu, X., Wen, B., Jiang, K., Liu, C., Zhang, T., et al.: Thyme: Think beyond images. arXiv preprint arXiv:2508.11630 (2025)
45. Zhang, Z., He, X., Yan, W., Shen, A., Zhao, C., Wang, X.E.: Soft thinking: Unlocking the reasoning potential of LLMs in continuous concept space. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
46. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2023)
47. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., XingYu: Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In: The Fourteenth International Conference on Learning Representations (2026)

48. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)