CoLVR: Enhancing Exploratory Latent Visual Reasoning via Contrastive Optimization

Ziyang Ding[1,2*], **Linjian Meng**[2], **Yiming Wu**[2], **Yuhan Li**[2,3], **Yuhao Liu**[4], **Zhen Zhao**[2†]
[1]Shanghai Jiao Tong University, [2]Shanghai Artificial Intelligence Laboratory
[3]Nanjing University, [4]Shandong University
zhen.zhao@outlook.com

Abstract

Due to the potential for exploratory reasoning of Latent Visual Reasoning, recent works tend to enable MLLMs (Multimodal Large Language Models) to perform visual reasoning by propagating continuous hidden states instead of decoding intermediate steps into discrete tokens. However, existing works typically rely on hard alignment objectives to force latent representations to match predefined visual features, thereby severely limiting the exploratory of latent reasoning process. To address this problem, we propose **CoLVR** (**C**ontrastive **O**ptimization for **L**atent **V**isual **R**easoning). To obtain a more exploratory visual reasoning, CoLVR introduces a latent contrastive training framework. Firstly, CoLVR learns diverse and exploratory representations with a latent contrastive objective guided by angle-based perturbation, which expands the semantic latent space and avoids over-constrained embedding. Then, CoLVR employs a latent trajectory contrastive reward for RL (Reinforcement Learning) post-training to enable fine-grained optimization of latent visual reasoning process and thus fostering diverse reasoning behaviors. Experiments demonstrate that CoLVR significantly enhances the exploratory capability of latent representations, achieving average improvements of **5.83%** on VSP and **8.00%** on Jigsaw, while also outperforming existing latent models on out of domain benchmarks, with a **3.40%** gain on MMStar. The data, codes, and models are released at https://github.com/Oscar-dzy/CoLVR.

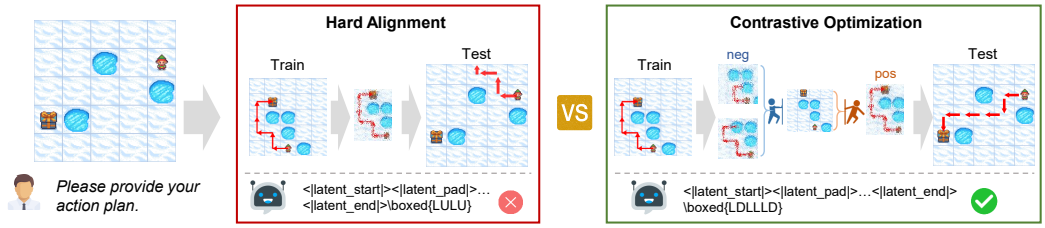
1 Introduction

Recent progress on VQA (Visual Question Answering) tasks in MLLMs (Multimodal Large Language Models) is significant [23, 17, 2, 5, 32]. These methods construct multimodal chains of thought with rich visual information through image editing and generation, which improves visual exploration and understanding [10, 30, 33]. However, they rely on autoregressive generation over discrete text tokens, which leads to long reasoning sequences and high latency. More importantly, discrete tokens are inherently limited in representing fine-grained and continuous visual features, which are essential for many visual reasoning tasks [31].

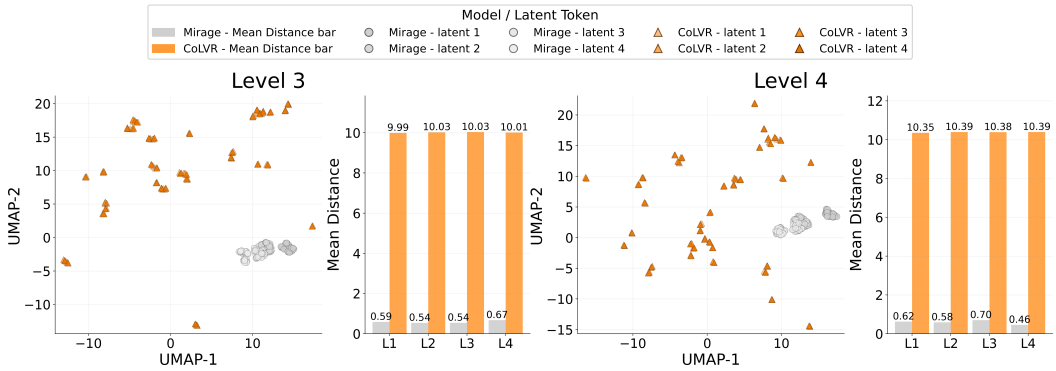
To address these limitations, recent studies introduce **Latent Visual Reasoning** [31, 15, 26], where intermediate reasoning steps are no longer decoded into text but propagated as continuous hidden states. This paradigm enables models to reason directly in the latent space, leading to more efficient and more exploratory processing of visual information. Despite its promise, existing approaches still lacks the flexibility since they often directly force latent states to match predefined visual representations, such as particular image patches or object-level regions, through hard alignment objectives, as shown in Figure 1(a).

*This work was done during his internship at Shanghai Artificial Intelligence Laboratory.
†Corresponding author.

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



(a) Latent Token Exploratory Comparison.



(b) UMAP Visualization of Latent Tokens.

Figure 1: (a) **Comparison of Latent Token Exploratory Capacity Under Hard Alignment and Contrastive Optimization**. Hard alignment enforces rigid feature matching and leads to fixed reasoning paths, while contrastive optimization encourages more flexible latent trajectories. (b) **UMAP-based 2D visualization of the first four latent tokens from Mirage and CoLVR**, where hidden states are projected into a two-dimensional space and cluster compactness is quantified via the mean distance to centroids. The results highlight that CoLVR learns more exploratory latent representations.

Such hard alignment further drives latent tokens into a homogeneous latent space, limiting their exploratory capability [28]. To empirically verify this effect, we analyze a classical Latent Visual Reasoning approach called Mirage [31], which explicitly aligns latent tokens with grouped image patches. By projecting latent tokens from different test samples into a 2D space using UMAP [16] as shown in Figure 1(b), we observe that Mirage latent tokens from the same reasoning step collapse into highly concentrated clusters, with an average distance to the cluster centroids of only $0.5 \sim 0.7$. This provides empirical evidence that hard alignments reduce the diversity of latent representations, implying a decline on the weaken exploratory reasoning. To mitigate this issue, recent work such as Laser [28] aligns latent tokens with future visual reasoning states. However, it still relies on predefined alignment targets, which implicitly constrain the latent semantic space, thereby only partially alleviating the limited flexibility and exploratory capacity of latent visual reasoning.

Furthermore, recent reinforcement learning methods for latent reasoning, such as VLPO [26], attempt to optimize latent tokens by modeling their probability distributions. However, these approaches still adopt predefined distribution assumptions and trajectory-level alignment objectives, forcing latent trajectories to match specific latent states observed during rollout. As a result, such optimization remains centered on matching predefined patterns instead of encouraging exploratory latent reasoning.

To address these challenges, we propose **CoLVR** (**C**ontrastive **O**ptimization for **L**atent **V**isual **R**easoning), a latent contrastive training framework toward more flexible and exploratory visual reasoning. In contrast to hard alignment methods that constrain latent visual tokens toward predefined patches, regions, or rollout states, CoLVR does not specify an explicit target pattern for each reasoning step. Instead, CoLVR optimizes relative contrastive relationships among latent visual states, thereby preserving the freedom of latent exploration while still providing task-relevant supervision. Specifically, after a warm-up stage, we construct positive and negative samples via angle-based perturbations and introduce a contrastive objective in the latent space. Positive samples consist of

image data paired only with result hints, while negative samples are generated by applying angular perturbations to latent representations. This design reduces the reliance on process-level visual annotations and encourages the model to learn more discriminative and diverse latent representations. Furthermore, we extend the contrastive principle to the reinforcement learning stage by introducing a latent-trajectory contrastive reward over hidden states, which provides relative preference signals among latent trajectories rather than enforcing predefined trajectory targets. This provides latent-level preference signals that encourage task-relevant latent rollouts during reinforcement learning.

Extensive experiments demonstrate the effectiveness of our method. In particular, CoLVR achieves an average improvement of **5.83%** on the VSP (Visual Spatial Planning) task and **8.00%** in Jigsaw task. It also significantly outperforming constraint-based latent models across multiple OOD benchmarks, including a **3.40%** improvement on MMStar. These results highlight the strong exploratory capability of the latent tokens learned by CoLVR.

Our contributions are summarized as follows:

1. We identify and analyze a fundamental limitation of existing latent visual reasoning methods: hard alignment objectives constrain latent representations into a homogeneous latent space, thereby limiting the exploratory capability of latent tokens.
2. We propose CoLVR, a latent visual reasoning training framework that introduces latent contrastive objectives in both the fine-tuning and reinforcement learning stages, in order to enhance the exploratory capabilities of latent tokens.
3. Extensive experiments demonstrate that our method effectively improves the exploratory behavior and flexibility of the model, thereby leading to better performance on visual understanding tasks.

2 Related Work

2.1 Latent Visual Reasoning

Latent reasoning was pioneered in the Coconut [8] framework, which eliminates the need for explicit decoding during intermediate steps. Rather than generating discrete tokens, the model propagates hidden states as a continuous reasoning medium, using the final state of one step as the input embedding for the next. Recent work has extended this paradigm to multimodal large language models to facilitate visual reasoning within the latent space. Mirage [31] aligns image patches with hidden states to integrate visual context, while LVR [14] employs region-of-interest indexing to link latent states with localized image features. Similarly, Chain-of-Visual-Thought [19] enhances fine-grained understanding by aligning hidden states with structured signals like segmentation and depth maps. However, these approaches rely on explicit alignment with preprocessed features, which constrains the latent space and often restricts latent tokens to a homogeneous feature set, ultimately limiting the exploration of diverse reasoning trajectories.

2.2 Contrastive Learning

Contrastive learning has been widely adopted in multimodal learning because it optimizes relative similarities by pulling positive samples closer while pushing negative samples apart in the feature space, thereby enhancing the discriminability and flexibility of learned representations [21]. This relative formulation provides a way to structure representation spaces without requiring each representation to match a fixed target feature. For example, HACL [12] constructs hallucinated captions as negative samples and applies contrastive learning to mitigate the modality gap between vision and language. MuCo [6] improves multimodal embeddings by modeling contrastive relationships across different query-image pairs. CogFlow [4] further incorporates contrastive principles into the reinforcement learning stage and introduces a knowledge initialization reward to guide optimization. However, the role of contrastive learning in optimizing latent reasoning trajectories and enhancing the exploratory capacity of latent tokens remains insufficiently explored. Our work investigates this direction by introducing latent contrastive objectives into both supervised fine-tuning and reinforcement learning, targeting latent representations and reasoning trajectories rather than multimodal embedding quality alone.

3 Methods

3.1 Problem Formulation

In this section, we present a unified formulation of latent visual reasoning. As shown in Figure 2(a), given an input query Q and an image I, the image is first encoded by a visual encoder f_{enc} and projected into a high-dimensional semantic space. The resulting representation is then combined with the textual input and fed into a multimodal large language model f_θ to generate an output sequence that contains a latent reasoning sequence. To formalize this process, we introduce latent reasoning tokens, where <|latent_start|> and <|latent_end|> denote the beginning and end of the reasoning segment, and <|latent_pad|> denotes intermediate latent tokens. The overall generation process is defined as:

$$\big\{\texttt{<|latent_start|>}, \underbrace{\texttt{<|latent_pad|>}}_{K}, \texttt{<|latent_end|>}, \text{Answer}\big\} = f_\theta\big(Q,\ \text{Proj}(f_{\text{enc}}(I))\big). \quad (1)$$

During latent reasoning, the model does not decode the last hidden state at intermediate steps. Instead, it directly feeds the hidden state as the input embedding for the next <|latent_pad|>, enabling iterative propagation over K steps in a continuous latent space. This mechanism allows the model to implicitly represent and reason over visual information without relying on explicit symbolic decoding.

3.2 Warm Up Alignment

In the initial training stage, we design a feature alignment based initialization strategy to enable the model to acquire preliminary reasoning capability in the visual latent space (as shown in Figure 2(b)). Specifically, the input image is processed by a visual encoder f_{enc} and a projector to obtain semantic features, followed by MeanPool$(\cdot)$ over the patch dimension to extract a global representation S:

$$S = \text{MeanPool}(\text{Proj}(f_{\text{enc}}(I))). \quad (2)$$

Based on this representation, we align the hidden states h_i of all latent tokens with S using cosine similarity, which provides an initial semantic constraint for latent reasoning. To preserve textual reasoning ability during alignment, we further introduce an auxiliary cross-entropy loss on the answer to stabilize language generation and reasoning performance:

$$\mathcal{L}_{\text{warmup}} = \frac{1}{K}\sum_i (1 - \text{cos_sim}(h_i, S)) + \lambda_1 \mathcal{L}_{\text{CE}}. \quad (3)$$

3.3 Latent Contrastive Supervised Fine-Tuning

In the fine-tuning stage, we introduce a latent contrastive objective via angle-based perturbations to replace the commonly used hard alignment strategy (Figure 2(c)). Specifically, we construct both positive and negative samples. The positive sample is defined as the augmented input image I_{hint}, which provides semantic guidance toward the correct answer and characterizes the feature distribution of the correct reasoning direction. The negative sample is designed to be structurally informative rather than random noisy perturbations. We aim to construct samples that are semantically close to the correct direction while remaining clearly deviated.

To this end, we follow the same feature extraction pipeline as in the warm up stage to obtain feature representations of the original input image and the image augmented with hints, denoted as S_I and $S_{I_{\text{hint}}}$. Their difference defines a direction vector $\Delta = S_{I_{\text{hint}}} - S_I$, which represents a semantic trajectory from the original perception S_I to the correct understanding $S_{I_{\text{hint}}}$. This trajectory represents the expected reasoning path in latent space. Accordingly, we view latent token learning as fitting this trajectory, where the model must identify and approach the direction that most effectively leads to the correct answer among multiple candidates.

Based on this trajectory, we construct structured perturbations to generate negative samples. We first sample a random perturbation vector ϵ from a standard normal distribution $\mathcal{N}(0, I)$ and compute its component orthogonal to the trajectory direction as

$$\eta = \epsilon - \frac{\epsilon \cdot \Delta}{\Delta^2} \cdot \Delta. \quad (4)$$

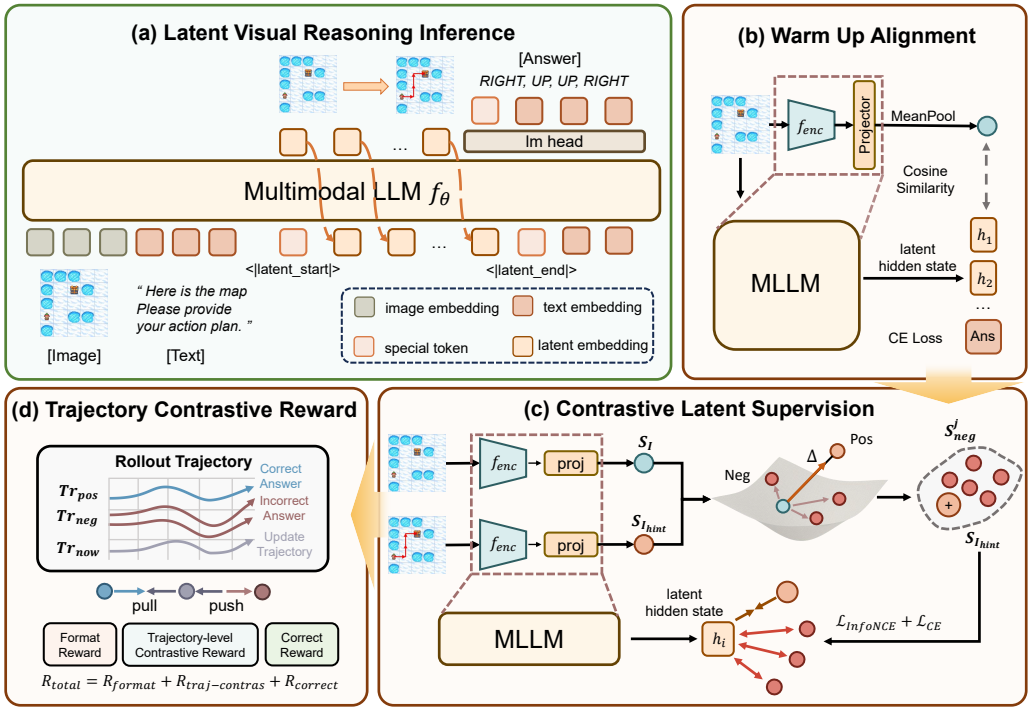


Figure 2: **The framework of CoLVR**. After a warm-up stage, CoLVR perform latent contrastive learning by constructing positive and negative samples via angular perturbations to encourage exploratory latent token representations. Additionally, we introduce a latent trajectory-based reward within the GRPO process to further optimize and sustain the exploration capability of latent tokens.

We then normalize both the trajectory vector Δ_{norm} and the perturbation component η_{norm}, and rotate the trajectory by a randomly sampled angle θ_{neg} within a predefined range to obtain a controlled deviation z. This operation introduces directional variation while keeping the perturbation magnitude bounded, resulting in a new trajectory that remains close to but deviates from the correct path. The perturbed trajectory is then rescaled to match the magnitude of the original trajectory and combined with it to form a misleading negative feature:

$$z = (\cos\theta_{\text{neg}} \cdot \Delta_{\text{norm}} + \sin\theta_{\text{neg}} \cdot \eta_{\text{norm}}) \cdot ||\Delta||, \tag{5}$$

$$S_{\text{neg}} = S_{I_{\text{hint}}} + z = S_I + \Delta + z. \tag{6}$$

This procedure produces a set of positive and negative samples distributed around the true reasoning trajectory. During training, we repeat the above process N_{neg} times and optimize the latent representations using the InfoNCE loss [9], which encourages closeness to positive samples and separation from the structured negative samples:

$$\mathcal{L}_{\text{contras}} = -\frac{1}{K}\sum_{i}^{K} \log\left(\frac{e^{\text{sim}(h_i, S_{I_{\text{hint}}})/\tau}}{e^{\text{sim}(h_i, S_{I_{\text{hint}}})/\tau} + \sum_{j}^{N_{\text{neg}}} e^{\text{sim}(h_i, S^j_{\text{neg}})/\tau}}\right), \tag{7}$$

where τ represents the temperature. In addition, we introduce a cross-entropy loss on the answer to preserve the model's text generation capability:

$$\mathcal{L}_{\text{contrasSFT}} = \lambda_2 \mathcal{L}_{\text{contras}} + \mathcal{L}_{\text{CE}}. \tag{8}$$

3.4 Latent-Trajectory Contrastive Reward

Finally, we further improve the reinforcement learning stage to alleviate the homogenization problem in the latent space. Unlike existing latent reasoning methods that mainly rely on format rewards and final answer rewards at the token distribution level, our method introduces latent-trajectory optimization to encourage more diverse and exploratory latent representations.

Specifically, the importance sampling term in the GRPO (Group Relative Policy Optimization) objective $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ indicates that the optimization aims to increase the generation probability of tokens corresponding to correct answers, thereby encouraging the model to produce higher-quality outputs in subsequent reasoning steps. Inspired by this observation, the optimization of hidden states should also move closer to the latent representations of correct samples. However, enforcing hard alignment with the hidden states corresponding to correct answers may suppress the exploration ability of the latent space. Therefore, it is necessary to introduce a contrastive learning mechanism that pulls positive samples closer while pushing negative samples apart, enabling the learning of more flexible latent representations.

Based on this motivation, this paper designs a latent-trajectory contrastive reward that provides relative preference signals among latent trajectories rather than enforcing predefined trajectory targets (as shown in Figure 2(d)). Specifically, the cosine similarity sim between latent states is first defined as

$$Tr_i = h_i^1, h_i^2, \ldots, h_i^K, \tag{9}$$

$$\overline{\text{sim}}(Tr_i, Tr_j) = \frac{1}{K}\sum_t \overline{\text{sim}}(h_t^i, h_t^j), \quad \text{where } \overline{\text{sim}} = \frac{1+\text{sim}}{2}, \tag{10}$$

where the similarity is obtained by averaging the similarities of latent tokens at corresponding positions, and is normalized to the interval $[0, 1]$ via the transformation $\overline{\text{sim}} = (1 + \text{sim})/2$ to ensure stable reward scaling. During the rollout phase, the model collects latent trajectories corresponding to both correct and incorrect samples generated under the current policy, and performs an additional forward pass to obtain latent representations Tr_{now} under consistent conditions. When correct answers are present, one latent trajectory is selected as the positive sample, while the remaining incorrect latent trajectories are treated as negative samples, and an InfoNCE-based reward is constructed at the hidden state level. When no correct answer exists, only a repulsion constraint from negative samples is applied. When no incorrect answer exists, only alignment with one latent trajectory is performed. The reward is defined as

$$R_{\text{latent}} = \begin{cases} \frac{e^{\overline{\text{sim}}(Tr_{\text{now}}, Tr_{\text{pos}})/\tau}}{e^{\overline{\text{sim}}(Tr_{\text{now}}, Tr_{\text{pos}})/\tau} + \sum_j^{n_{\text{neg}}} e^{\overline{\text{sim}}(Tr_{\text{now}}, Tr_j)/\tau}}, & \text{if pos \& neg mixed,} \\ \frac{1}{1+\sum_j^{n_{\text{neg}}} e^{\overline{\text{sim}}(Tr_{\text{now}}, Tr_j)/\tau}}, & \text{if only neg,} \\ \overline{\text{sim}}(Tr_{\text{now}}, Tr_{\text{pos}})/\tau, & \text{if only pos.} \end{cases} \tag{11}$$

Through this case-wise design, the proposed reward balances exploration and discrimination in the latent space during reinforcement learning. Finally, the proposed latent-trajectory contrastive reward is jointly incorporated into the GRPO training framework together with the original format R_{format} and textual rewards R_{correct}. The format reward combines a token count ratio term for `<|latent_pad|>` scaled by 0.5 with a `"\boxed{}"` format match bonus of 0.5. The textual reward is binary: 1 if the final answer matches the ground truth, 0 otherwise. This joint reward preserves generation quality while enabling effective optimization of the latent reasoning process.

$$R_{\text{total}} = R_{\text{format}} + R_{\text{correct}} + R_{\text{latent}}. \tag{12}$$

4 Experiments

4.1 Experiment Settings

Training Datasets and Benchmarks. This paper evaluates the model on diverse visual reasoning tasks. For VSP (Visual Spatial Planning), the goal is to measure path planning and spatial reasoning in maze environments. We use Mirage [31] VSP training data and benchmarks for base evaluation, and further introduce level 7 and level 8 datasets built on DiffThinker [10] to test generalization under higher complexity. For visual puzzle tasks, Jigsaw evaluates the ability to model global semantics and recover missing regions from incomplete input, while Tertis examines constrained spatial completion through a Tetris-like mechanism. For these tasks, we use two Zebra-CoT [13] datasets split into training and test sets for systematic analysis. To further assess cross-task generalization and CoLVR's exploratory advantage in latent visual reasoning, we also include multiple visual reasoning benchmarks, including VisPuzzle (from Thinkmorph [7]), MMVP [25], MMStar [3], V* [29] and CV-Bench [24], enabling comprehensive evaluation.

Table 1: **Main results across multiple visual task benchmarks**, comparing CoLVR with other VLMs and rigidly aligned latent baseline Mirage. Results are reported in accuracy (%), and best results are highlighted in **bold**.

Model	VSP									Jigsaw	Tertis
	Seen					Unseen			Total		
	Level 3	Level 4	Level 5	Level 6	Avg.	Level 7	Level 8	Avg.			
Proprietary Model											
GPT-4o [18]	67.00	46.00	24.00	17.00	38.50	11.00	6.00	8.50	28.50	36.00	30.67
Gemini-2.5-flash [22]	95.00	86.00	62.00	30.00	62.00	32.00	18.00	25.00	53.83	43.33	31.33
Open source model											
InternVL3.5-8B [11]	20.00	13.00	3.00	2.00	9.50	2.00	0.00	1.00	6.67	19.33	22.00
Qwen2.5-VL-7B [2]	11.00	7.00	7.00	1.00	6.50	1.00	0.00	0.50	4.50	24.00	24.00
w/ CoT	20.00	8.00	5.00	3.00	9.00	3.00	2.00	2.50	6.83	24.67	21.33
w/ vanilla SFT	91.00	85.00	71.00	50.00	74.30	36.00	29.00	29.00	59.17	67.33	42.67
w/ vanilla SFT + GRPO	89.00	80.00	66.00	49.00	71.00	34.00	14.00	24.00	55.33	68.00	41.33
Latent model											
Mirage [31]	90.00	84.00	67.00	55.00	74.00	39.00	25.00	32.00	60.00	70.00	43.33
CoLVR(Ours)	**94.00**	**92.00**	**80.00**	**57.00**	**80.70**	**43.00**	**29.00**	**36.00**	**65.83**	**78.00**	**44.67**
Δ vs Mirage	+4.00	+8.00	+13.00	+2.00	+6.70	+4.00	+4.00	+4.00	+5.83	+8.00	+1.34

Baselines. We use Qwen2.5-VL-7B-Instruct [2] (denoted as Qwen2.5-VL-7B) as the base model for both supervised fine-tuning and reinforcement learning. For evaluation, closed-source baselines include GPT-4o [18] and Gemini-2.5-flash [22], while the open-source comparison is InternVL3.5-8B-Instruct [11] (denoted as InternVL3.5-8B) . In addition, Qwen2.5-VL-7B is further trained on the same dataset with both supervised fine-tuning and GRPO [20], and the Mirage latent training paradigm is implemented under the same data setting for direct comparison, enabling a systematic assessment of the proposed method. To further examine generalization, LVR [14] is introduced as a complementary baseline, which achieves region-level alignment through ROI indexing.

Implementation Details. Training is conducted on four NVIDIA H200 GPUs with a global batch size of 4. To ensure reproducibility, the random seed is fixed at 42. For the VSP task, the learning rates across different training stages are set to 5e-5, 5e-5, and 5e-6, respectively, with a linear decay schedule. The number of latent tokens is consistently set to 8 and remains fixed during inference to maintain alignment between training and inference. During the SFT stage, θ is sampled from $[\frac{\pi}{2}, \pi]$, and the number of negative samples N_{neg} is set to 8. In addition, both supervised fine-tuning and reinforcement learning are implemented using the TRL framework.

4.2 Main Results

From the experimental results in Table 1, we observe that the proposed CoLVR framework consistently outperforms all competing baselines across multiple visual reasoning benchmarks, demonstrating stable and significant performance gains. Specifically, compared with both the vanilla SFT approach and the reinforcement learning based GRPO method, CoLVR achieves superior results on all evaluation metrics, indicating that latent space contrastive optimization effectively enhances the model's visual reasoning capability. Furthermore, when compared with Mirage, which relies on a hard alignment mechanism, CoLVR consistently surpasses it across all task settings, with overall improvements ranging from 1.34% to 8.00%. Notably, on the Jigsaw task, which requires stronger implicit visual imagination, the advantage of CoLVR becomes more pronounced, achieving an 8.00% improvement over Mirage. These results suggest that CoLVR introduces a more discriminative latent feature learning mechanism, which alleviates the representational constraints introduced by hard alignment in existing latent space models, thereby facilitating the learning of more exploratory latent tokens and enabling better modeling and reasoning of visual semantics.

4.3 Generalization on out of domain visual benchmarks

To further analyze the role of the exploratory latent contrastive objective in latent reasoning, we evaluate the CoLVR model on multiple out-of-domain benchmarks and compare it with two latent space models that adopt hard alignment constraints on visual features, namely Mirage and LVR. The results show that, CoLVR significantly improves the generalization ability of latent reasoning models and consistently achieves the best performance across all evaluated datasets. Specifically, as shown in Table 1, on the VSP benchmark, the CoLVR model trained on the VSP task demonstrates

Table 2: **Out-of-domain visual understanding benchmarks comparing CoLVR with other latent reasoning models**. LVR uses official checkpoints, while Mirage is trained on the same data as our method. Results are reported in accuracy (%), with the best performance highlighted in **bold**.

Model	VisPuzzle	V*	MMVP	MMStar	CV-Bench
CoLVR	**37.00**	**80.63**	**72.00**	**61.47**	**76.95**
Mirage	36.75	75.92	68.33	55.27	72.40
Δ	+0.25	+4.71	+3.67	+6.20	+4.55
LVR (8 steps)	**37.00**	80.10	69.33	58.07	75.51
Δ	+0.00	+0.53	+2.67	+3.40	+1.44

Table 3: **Ablation study of key design components**. *w/o Traj Reward GRPO* removes the final GRPO stage with latent trajectory rewards; + *w/o Contras SFT* retains only the warm-up training stage; *w/ Traj GRPO Only* directly applies GRPO with latent trajectory rewards after warm-up; *w/ Normal GRPO* applies GRPO without latent trajectory rewards.

Model	VSP									Jigsaw	Tertis
	Seen					Unseen			Total		
	Level 3	Level 4	Level 5	Level 6	Avg.	Level 7	Level 8	Avg.			
CoLVR(Ours)	94.00	**92.00**	**80.00**	**57.00**	**80.70**	**43.00**	**29.00**	**36.00**	**65.83**	**78.00**	**44.67**
w/o Traj Reward GRPO	**95.00**	91.00	75.00	54.00	78.70	39.00	27.00	33.00	63.50	78.00	42.67
+ w/o Contras SFT	92.00	87.00	72.00	51.00	75.50	35.00	24.00	29.50	60.17	74.70	40.00
w/ Traj GRPO Only	92.00	87.00	75.00	44.00	74.50	36.00	27.00	31.50	60.17	75.33	40.00
w/ Normal GRPO	92.00	87.00	69.00	48.00	74.00	39.00	27.00	33.00	60.33	76.00	38.67

strong robustness under the more challenging unseen settings, achieving an average score of 36.00%, which improves over Mirage by 4.00%. This indicates stronger transferability under distribution shift. As shown in Table 2, we evaluate the CoLVR model trained on the Jigsaw task. On the V* benchmark, CoLVR improves over Mirage by 4.71%, reaching an accuracy of 80.63%, and slightly outperforms LVR (8 steps), which achieves 80.10%. Moreover, on MMVP and MMStar, which require fine-grained visual perception and multimodal understanding, CoLVR achieves the highest scores of 61.47% and 72.00%, respectively.

These results suggest that latent tokens in CoLVR exhibit stronger flexibility in adapting to unseen visual scenarios, whereas Mirage and LVR, which rely on predefined feature spaces, show limited adaptability under distribution shift. This further indicates that the latent contrastive objective helps prevent the latent space from collapsing into a homogeneous semantic manifold. By encouraging more exploratory trajectories in the hidden state space, CoLVR effectively transfers the reasoning patterns learned from the training task to complex real-world visual understanding benchmarks.

4.4 Ablation Study

The latent contrastive framework improves visual reasoning across both seen and unseen settings. Table 3 evaluates the contributions of the key components in CoLVR. Compared with the warm-up-only variant, which removes both contrastive SFT and trajectory-reward GRPO, the full framework improves VSP Seen and Unseen accuracy by 5.20 and 6.50 percentage points, respectively, and raises the overall VSP score from 60.17% to 65.83%. It also improves Jigsaw and Tetris accuracy by 3.30 and 4.67 points, respectively. These consistent gains support the effectiveness of the overall training framework, with the larger improvement on unseen VSP tasks suggesting better generalization beyond the training settings.

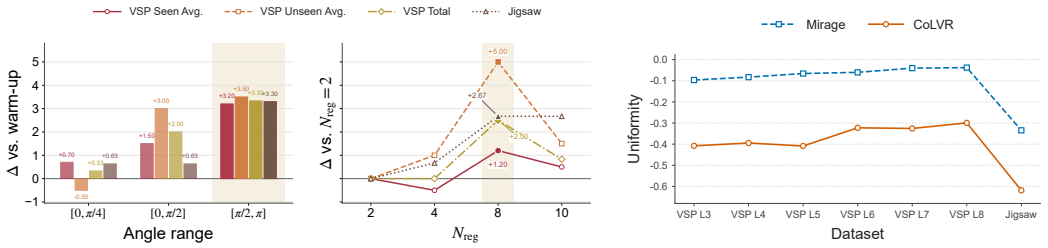
Contrastive SFT and trajectory-reward GRPO make complementary contributions. To examine the role of Stage 2 before GRPO optimization, we directly apply trajectory-reward GRPO after warm-up (*w/ Traj GRPO Only*). Compared with this variant in Table 3, the full framework improves VSP, Jigsaw, and Tertis scores by 5.66, 2.67, and 4.67 points, respectively, supporting the importance of learning task-relevant latent representations through contrastive SFT before subsequent trajectory-level optimization. Stage 3 provides a further benefit: compared with the variant without trajectory-reward GRPO (*w/o Traj Reward GRPO*), the full model achieves a 2.33 percentage-point improvement in the overall VSP score. By contrast, applying standard GRPO without the latent trajectory reward (*w/ Normal GRPO*) reduces this score to 60.33%. Together, these results indicate that the gains depend

Table 4: Comparison of different negative construction strategies for the latent contrastive objective.

Neg. construction	VSP Seen	VSP Unseen	VSP Total	Jigsaw
Random batch-pool	74.00	27.00	58.33	76.67
Incorrect-image-based	77.50	30.00	61.67	78.00
Noise-based	75.20	30.50	60.33	75.33
Angle-based	**78.70**	**33.00**	**63.50**	**78.00**

Table 5: Semantic association of constructed negatives.

Measure	VSP	Jigsaw	Tertis
Neg. Hint Cos.	0.97	0.89	0.91
RIA	0.57	0.23	0.17



(a) Negative construction sensitivity.

(b) Latent representation uniformity.

Figure 3: **Sensitivity of negative construction and analysis of latent-space homogenization.** (a) Effects of the angle range (left) and the number of negatives (right), measured as accuracy changes in percentage points relative to warm-up and $N_{\text{neg}} = 2$, respectively. Shaded regions indicate the selected settings. (b) Hyperspherical uniformity of normalized latent hidden states across VSP levels and Jigsaw; lower values indicate less concentrated representations. CoLVR consistently achieves lower scores than Mirage.

on both the preceding contrastive SFT stage and the trajectory-specific reward, supporting a training design that combines representation-level contrastive learning with trajectory-level optimization.

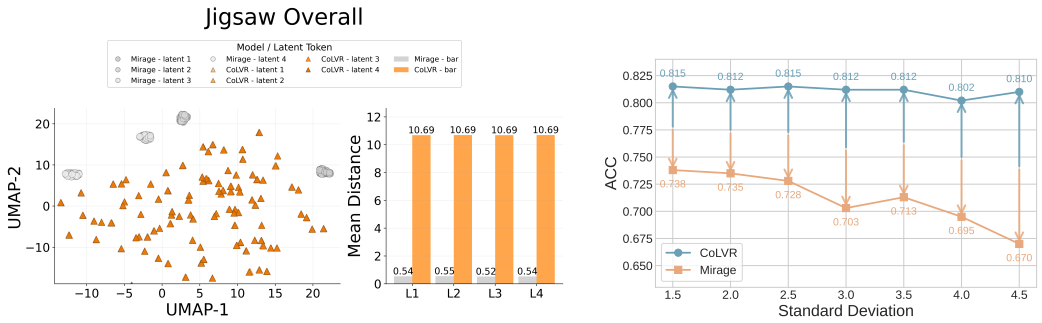
Angle-based perturbation provides effective negative supervision for contrastive SFT. Table 4 compares angle-based perturbations with batch-pool negatives from other samples, incorrect-hint image features, and Gaussian-noise perturbations. Our strategy achieves the highest VSP Total accuracy of 63.50%, exceeding these alternatives by 5.17, 1.83, and 3.17 percentage points, respectively, without requiring task-specific incorrect images during training. To assess the semantic relevance of the constructed negatives, we use incorrect solution-hint images generated with Gymnasium for VSP and provided by Zebra-CoT for the other tasks as references. Table 5 shows high cosine similarity between constructed negatives and incorrect-hint features. To account for their shared correct-hint component, we define Residual Incorrect Association (RIA) as the cosine similarity between their residuals after projecting out the correct-hint direction, following the principle of first-order partial correlation [1]. Positive RIA values across all three tasks indicate that the association persists beyond this shared component, supporting the semantic relevance of the constructed negatives.

Sensitivity to the angle range and number of negatives. Figure 3(a) (left) shows that $[\frac{\pi}{2}, \pi]$ yields the largest gains over warm-up across all evaluated metrics. With the noise component orthogonal to the correct trajectory direction Δ, this range ensures a non-positive projection onto Δ, producing orthogonal or backward deviations. For the number of negatives N_{neg} (right), increasing from two to eight improves all metrics, particularly unseen VSP, whereas increasing to ten reduces VSP performance without improving Jigsaw. These results support $\theta \in [\frac{\pi}{2}, \pi]$ and $N_{\text{neg}} = 8$ as our default configuration, which performs best overall among the tested settings.

4.5 Exploratory Capability Analysis

CoLVR learns a substantially more exploratory and discriminative latent space than rigid alignment-based methods. As shown in Figure 1(b) and 4(a), we compare the UMAP visualizations of the first four latent tokens across different VSP levels and Jigsaw for Mirage and CoLVR. The visualizations show more dispersed representations for CoLVR, with average distances to the corresponding cluster centroids of approximately 10.5–12.7, compared with 0.5–0.7 for Mirage.

To quantify concentration in the original latent space, we adopt hyperspherical uniformity [27]: $\mathcal{L}_{\text{uniform}} = \log \mathbb{E}_{i \neq j}[\exp(-t\|z_i - z_j\|_2^2)]$, where z_i is an L_2-normalized latent hidden state and $t = 2$.



(a) UMAP Visualization of Latent Tokens in Jigsaw task.

(b) Inference Noise Injection.

Figure 4: (a) **UMAP Visualization of Latent Tokens in Jigsaw task**, Mirage and CoLVR exhibit the same trend as observed on VSP. (b) **Inference Noise Injection**. We introduce random noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ into the latent tokens during the inference phase to evaluate their sensitivity and explorative capabilities.

Lower values indicate more uniformly distributed representations. As shown in Figure 3(b), CoLVR achieves consistently lower scores across all evaluated VSP levels and Jigsaw, supporting reduced latent-space homogenization and aligning with our objective of enabling more exploratory latent reasoning.

Contrastive latent optimization enables CoLVR to maintain robust reasoning under broader latent space exploration. To further evaluate the exploration capability learned by CoLVR in the latent space, we inject Gaussian noise $\mathcal{N}(0, \sigma^2)$ into the propagated last hidden state during latent reasoning, and assess the model performance on VSP under controlled perturbations (Figure 4(b)). Such noise forces the model to explore a broader region of the latent space. Under such perturbations, CoLVR maintains stable performance, whereas Mirage exhibits severe degradation. These results suggest that contrastive learning over positive and negative samples enables CoLVR to learn more exploratory and robust latent representations, sustaining effective reasoning under wider latent space exploration. In contrast, methods based on hard alignment objectives rely on highly compressed latent spaces, making them substantially more sensitive to perturbations.

5 Conclusion

In this work, we address the issue of latent space collapse in multimodal reasoning by examining how rigid hard-alignment objectives can limit the exploratory behavior of latent tokens. To mitigate this problem, we propose CoLVR, a framework that moves from static feature matching toward contrastive trajectory optimization. By introducing angle-based perturbations during supervised fine-tuning and incorporating latent trajectory contrastive reward, CoLVR helps improve the latent hidden state exploratory capability. Our experiments across visual reasoning and out-of-domain benchmarks, together with UMAP visualizations, hyperspherical uniformity measurements and noise-injection tests, consistently show that CoLVR leads to a more exploratory and flexible latent space. Overall, this approach provides a scalable way to improve the flexibility and generalization of latent reasoning, moving beyond hard alignment constraints toward more exploratory multimodal reasoning.

Limitations. Despite its effectiveness, CoLVR relies on result hints for positive construction, increasing supervision costs. Future work could bootstrap contrastive SFT with correctness-filtered pseudo-hints from model rollouts. Limited semantic grounding of latent tokens also hinders interpretability and diagnostic analysis, motivating lightweight concept-level probing.

Acknowledgments and Disclosure of Funding

This work is supported by Shanghai Artificial Intelligence Laboratory and Smart Grid-National Science and Technology Major Project (No. 2025ZD0803600, 2025ZD0803603). We also thank Zitang Gou from Fudan University for his support.

References

[1] Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, 2004.

[2] Shuai Bai, Ke qin Chen, Xue-Jing Liu, Jia-Lin Wang, Wenbin Ge, Si-Bo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Hu-Men Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jian-Qiang Wan, Peng-Fei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tian-Bao Xie, Zesen Cheng, Hang Zhang, Zhi-Bo Yang, Hai-Yang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arxiv:2502.13923*, 2025.

[3] 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 *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024.

[4] Shuhang Chen, Yunqiu Xu, Junjie Xie, Aojun Lu, Tao Feng, ZEYING HUANG, ZHANG NING, Yi Sun, Yi Yang, and Hangjie Yuan. Cogflow: Bridging perception and reasoning through knowledge internalization for visual mathematical problem solving. In *The Fourteenth International Conference on Learning Representations*, 2026.

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[6] Geonmo Gu, Byeongho Heo, Jaemyung Yu, Jaehui Hwang, Taekyung Kim, Sangmin Lee, HeeJae Jun, Yoohoon Kang, Sangdoo Yun, and Dongyoon Han. Muco: Multi-turn contrastive learning for multimodal embedding model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1749–1758, 2026.

[7] Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[8] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Second Conference on Language Modeling*, 2025.

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[10] Zefeng He, Xiaoye Qu, Yafu Li, Tong Zhu, Qipeng Guo, Muxin Fu, Siyuan Huang, and Yu Cheng. Diffthinker: Towards generative multimodal reasoning with diffusion models. In *Forty-third International Conference on Machine Learning*, 2026.

[11] Shanghai AI Laboratory InternVL Team. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[12] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024.

[13] Ang Li, Charles L. Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. Zebra-cot: A dataset for interleaved vision-language reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[14] Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[15] Jizheng Ma, Xiaofei Zhou, Geyuan Zhang, Yanlong Song, and Han Yan. Lare: Latent refocusing for multimodal reasoning. *arXiv preprint arXiv:2511.02360*, 2025.

[16] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2020.

[17] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.

[18] OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[19] Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. In *European Conference on Computer Vision*, pages 445–462. Springer, 2026.

[20] 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. *arXiv preprint arXiv:2402.03300*, 2024.

[21] Qingyong Su, Chong Feng, Ge Shi, Bo Wang, and Yan Zhuang. Enhancing discriminative ability in multimodal llms: A contrastive learning approach for ct report generation. *Information Fusion*, 123:103240, 2025.

[22] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[23] Kimi Team. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[24] Shengbang Tong, Ellis L Brown II, Penghao Wu, Sanghyun Woo, ADITHYA JAIRAM IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[25] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024.

[26] Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond image and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12030–12040, 2026.

[27] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.

[28] Yubo Wang, Juntian Zhang, Yichen Wu, Yankai Lin, Nils Lukas, and Yuhan Liu. Forest before trees: Latent superposition for efficient visual reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11272–11288, 2026.

[29] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094. IEEE, 2024.

[30] Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. In *The Fourteenth International Conference on Learning Representations*, 2026.

[31] Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 33510–33520, 2026.

[32] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Tsung-Yi Lin, Gordon Wetzstein, Ming-Yu Liu, and Donglai Xiang. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1702–1713, 2025.

[33] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and XingYu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

A Appendix

A.1 Detailed Experimental Settings

All training procedures for CoLVR are conducted using the Accelerate library to enable efficient multi-GPU parallel training. Across all stages, we uniformly set the warm-up steps to 10, gradient accumulation steps to 1, and weight decay to 0.01. The AdamW optimizer is adopted throughout, with training precision set to bf16 and gradient checkpointing enabled to reduce memory consumption. For a fair comparison, Mirage is trained under identical settings to CoLVR. The learning rate and number of training epochs for each stage in CoLVR are listed in the Table 6 below:

Table 6: **The detailed experiment settings of CoLVR training process.**

Stage	VSP			Jigsaw			Tertis		
	learning rate	epochs	batch size	learning rate	epochs	batch size	learning rate	epochs	batch size
Warm up	5e-5	10	4	2e-5	8	4	1e-5	8	4
Latent Contrastive SFT	5e-5	10	4	2e-5	8	4	1e-5	8	4
Latent Trajectory Contrastive GRPO	5e-6	5	4	2e-6	5	4	2e-6	5	4

During the warm-up stage, the weight of the cross-entropy (CE) loss is set to $\lambda_1 = 0.3$. In the latent contrastive SFT stage, the weight of the contrastive loss is set to $\lambda_2 = 2$. The remaining hyperparameters used in the GRPO stage are summarized in Table 7.

Table 7: **Hyperparameters for RL.**

Rollout size	Temperature	Max prompt length	Max completion length	β	Latent trajectory temperture τ
4	1.2	2048	1024	0.04	0.5

Meanwhile, during training, we froze the parameters of the visual encoder and trained all remaining parameters of the MLLM, ensuring that the pretrained visual representations remain intact while the model adapts to the target tasks. Regarding the data, the amounts of data used for training are shown in Table 8. For RL training, the VSP data are generated using DiffThinker code, while the Jigsaw and Tetris data are randomly sampled from Zebra-CoT. We ensure that the RL training data do not overlap with either the SFT training dataset or the test dataset.Notably, since the images provided in zebra-jigsaw have relatively large resolutions, we resized them proportionally to a maximum size of 1024 pixels to ensure training stability and prevent out-of-memory issues.

Table 8: **Dataset Statistics.**

Task	Warm-up/SFT	RL	Test
VSP	1000	500	400
VSP (unseen)	/	/	200
Jigsaw	1516	500	150
Tertis	1500	500	150

Mirage was trained on the same training data as CoLVR to ensure a fair comparison. During testing, we retained Mirage's original inference protocol: the upper limit of latent tokens was set to 8, and the model was allowed to freely determine how many latent tokens to output based on its internal halting mechanism. In contrast, CoLVR was fixed to output exactly 8 latent tokens, providing a controlled baseline for evaluating the effect of adaptive token allocation.

A.2 Dynamic Latent Inference Comparison

Given the slight difference in inference modes between CoLVR and Mirage, we adapted CoLVR's inference protocol to match that of Mirage and re-evaluated it on the VSP, Jigsaw, and Tetris tasks. The results are presented in Table 9 below. It can be observed that dynamically determining versus fixing the number of latent tokens has no measurable impact on CoLVR's performance, indicating that CoLVR learns latent representations that are robust to token allocation and do not rely on a precise token count to maintain reasoning quality.

To visualize how latent tokens explore an image, we compare the attention maps of the first four latent tokens in Mirage and CoLVR across the four levels of VSP. These maps, shown as heatmaps in Figure 5, reveal a stark contrast in spatial coverage and semantic richness. Mirage's latent tokens, constrained by rigid priors, attend almost exclusively to weak local cues near the image borders throughout the inference process, leading to a restricted and fragmented exploration pattern. In contrast, CoLVR imposes no such strict constraints; its latent

Table 9: **Main results across multiple visual task benchmarks on the same inference mode.**

Model	VSP									Jigsaw	Tertis
	Seen					Unseen			Total		
	Level 3	Level 4	Level 5	Level 6	Avg.	Level 7	Level 8	Avg.			
CoLVR	**94.00**	**92.00**	**80.00**	**57.00**	**80.70**	**43.00**	**29.00**	**36.00**	**65.83**	**78.00**	**44.67**
CoLVR (dynamic)	**94.00**	**92.00**	**80.00**	**57.00**	**80.70**	**43.00**	**29.00**	**36.00**	**65.83**	**78.00**	**44.67**
Mirage	93.00	83.00	76.00	51.00	76.00	39.00	25.00	32.00	61.17	70.00	43.33

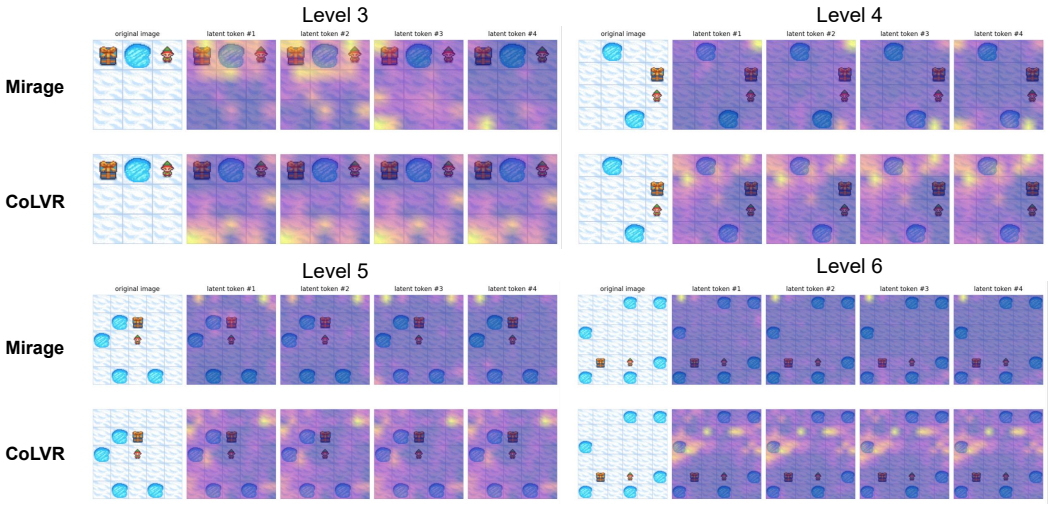


Figure 5: **Comparison of attention maps** across VSP levels for Mirage and CoLVR.

tokens learn more exploratory and discriminative features, yielding attention maps that distribute broadly across the global context. Consequently, CoLVR captures richer structural relationships and exhibits a markedly greater diversity in visual reasoning.

A.3 Visualization of UMAP for a Single Case with Latent Tokens Rolled Out K Times

To further investigate the exploratory capacity of latent tokens, we conduct an additional quantitative analysis at the level of individual cases. Specifically, we randomly sample 40 cases from the evaluation set. For each case, both Mirage and CoLVR are executed 20 times to generate multiple latent trajectories under stochastic conditions. We then employ UMAP to visualize the distribution of the resulting latent tokens and quantify their overall clustering behavior.

Concretely, for each model and each case, we compute the average distance from all latent tokens to their corresponding cluster centroid, which serves as a proxy for the degree of semantic dispersion. A larger average distance indicates stronger exploration in the latent space, while a smaller value suggests more concentrated representations. Finally, we average this metric across all sampled cases to obtain a global quantitative indicator of semantic exploration for each model. As shown in Fig. 6, CoLVR achieves an average clustering metric of 17.982, while Mirage reports a lower value of 16.544. This difference suggests that the latent tokens produced by CoLVR are more dispersed in the latent space, indicating a stronger ability to explore a wider range of semantic representations. Such behavior effectively expands the solution space and allows the model to capture alternative reasoning paths. By comparison, the lower clustering metric of Mirage indicates that its latent representations are more concentrated, which points to more limited exploration. This pattern is consistent with its use of stricter constraints, which restrict the evolution of latent tokens and reduce the diversity of semantic trajectories.

A.4 Attention Map

Figure 5 compares the attention maps of Mirage and CoLVR across different VSP levels. As the reasoning process progresses, Mirage exhibits highly concentrated and repetitive attention patterns, where the activated regions remain relatively fixed across latent reasoning steps. This behavior suggests limited exploration of the visual space and indicates that the latent tokens tend to collapse into similar attention distributions. In contrast, CoLVR consistently produces more diverse and spatially dynamic attention maps. The activated regions shift across different latent tokens and cover a broader range of informative areas in the image, especially around

task-relevant objects and surrounding contextual regions. This phenomenon is particularly evident in more complex levels such as Level 5 and Level 6, where CoLVR maintains exploratory attention behaviors while Mirage becomes increasingly static. These results demonstrate that CoLVR encourages more exploratory latent visual reasoning and enables the model to interact with visual content in a more flexible and comprehensive manner.

A.5 Task Details

Here we introduce specific examples corresponding to our training and evaluation tasks, including input images and image examples with hints (as shown in Figure 7, 8, 9).

Additionally, we provide illustrative examples from a range of out-of-domain benchmark tasks, including VisPuzzle, V*, MMVP, MMStar, and CV-Bench. These examples are intended to highlight the diversity and complexity of evaluation scenarios beyond the training distribution, offering a more comprehensive view of model generalization across varied visual reasoning and multimodal understanding tasks (As shown in Figure 10).

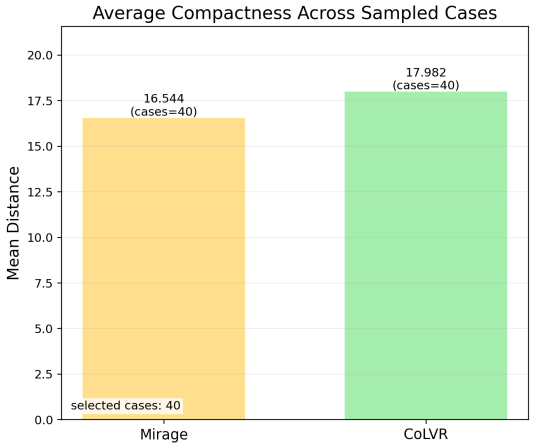


Figure 6: **Comparison of latent token dispersion across models**, measured by average distance to cluster centroids over 40 cases with 20 samples each.

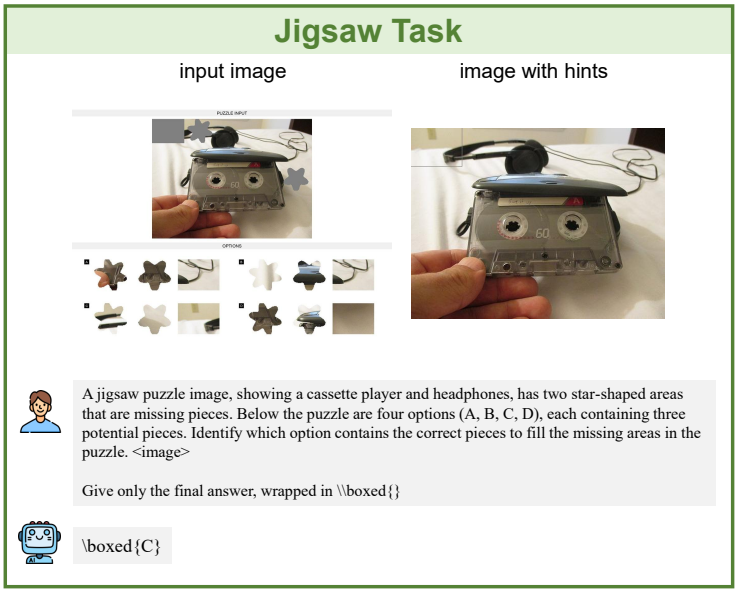


Figure 7: **Inference example: Jigsaw**.

Tertis Task

input image

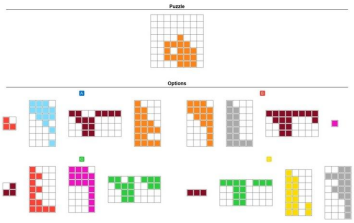
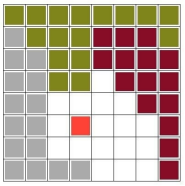


image with hints



Fill the exact yellow shape shown in the question grid. Choose the only option set whose pieces perfectly tile the shape without gaps or overlap.
<image>

Give only the final answer, wrapped in \\boxed{}

\boxed{B}

Figure 8: **Inference example: Tertis**.

VSP Task

input image



image with hints



As a professional maze solver, your task is to analyze a grid-based map and devise an action plan that enables a player to reach the goal from the starting point without falling into any holes, using the fewest possible moves.

Game Setup
- The game presents a fully observable grid-based map.
- The player starts at a specified grid square, with the goal located elsewhere on the map.
- Each grid square is either safe or contains a hole.\n- Your goal is to guide the player to the goal while avoiding holes.

Moving Rules
- The action plan involves a series of moves: 'L' (left), 'R' (right), 'U' (up), or 'D' (down).
- Each move transfers the player to the adjacent square in that direction, provided it is a safe square. The player cannot move more than one square at a time.
- Moving off the edge of the map has no effect. The player will remain at the same square.
- DO NOT MOVE INTO A HOLE! Falling into a hole results in defeat.
- Locating at the grid containing the goal results in victory.

Here is the map: <image>

Please provide your action plan.
The final answer MUST BE put in \\boxed{}.

\boxed{RIGHT , UP, UP, RIGHT}

Figure 9: **Inference example: VSP**.

VisPuzzle



The image below is divided vertically into two parts by a white strip, forming a 1×2 jigsaw puzzle. The left part is labeled \"1\" and the right part is labeled \"2\". These two parts are shuffled and may not be in their original order. Your task is to determine the correct arrangement of the two parts to reconstruct a natural-looking image.

Which of the following options (A, B, C, ...) is correct?
A. Part 1 should be to the left of Part 2
B. Part 1 should be to the right of Part 2
<image>
Give only the option letter, wrapped in \\boxed{}

\boxed{B}

V*



Is the dog on the left or right side of the scooter?
(A) right
(B) left
Answer with the option's letter from the given choices directly.

B

MMVP



Are the ears of the dog erect or drooping?

Options:
(A) Erect
(B) Drooping

<image>

Give only the final answer, wrapped in \\boxed{}

\boxed{A}

MMStar



Where is the largest body of water in the image located? Options: A: At the center of the image, B: On the bottom of the image, C: On the left side of the image, D: On the right side of the image

<image>

Give only the final answer, wrapped in \\boxed{}

\boxed{A}

CV-Bench



How many pillows are in the image? Select from the following choices.
(A) 2
(B) 1
(C) 4
(D) 0
(E) 3

(A)

Figure 10: **Inference examples: VisPuzzle, V*, MMVP, MMStar, and CV-Bench**.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Claims in abstract and introduction correctly reflect our contributions and scope

   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: Please refer to Sec. 5

   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: [N/A]

   Justification: This paper does not include theoretical results.

   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: Detailed instructions are provided in the experiment settings and appendix.

   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: We utilize publicly available models and datasets and have made our code open-source.

   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: We provide them in the experiment settings and appendix.

   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: Given the computational cost of post-training large models and the stable test results, we opt not to repeat identical experiments.

   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: We have claimed the used computer resources.

   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: We have checked it.

   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: [N/A]

    Justification: There is no societal impact of the work performed.

    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: The datasets and models are open-sourced and have been widely, which does not pose such a risk.

    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 models and datasets are properly used and cited.

    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: [N/A]

    Justification: This paper does not release new assets.

    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 nor research with human subjects.

    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 nor research with human subjects.

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: This paper do research related to multimodal LLMs.

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.