Contrastive-Guided Self-Supervised Latent Visual Reasoning for Hallucination Mitigation

Aoxue Dai and Ningning Wang[†]

Dalian University of Technology, Dalian, China
{dax3387,nnwang}@mail.dlut.edu.cn

Abstract. Recent research on Large Vision-Language Models (LVLMs) has explored visual chain-of-thought (CoT) mechanisms to enhance visual-centric reasoning, effectively mitigating the hallucination issues that plague current models. However, existing paradigms typically rely on external visual tools and multi-turn interactions. While recent latent visual reasoning approaches achieve tool-free, single-turn inference by supervising intermediate representations with auxiliary images, they incur significant annotation costs. To address this, we propose Contrastive-Guided Self-Supervised Latent Visual Reasoning (CoLVR), a novel framework for self-supervised training of visual latent tokens in an annotation-free manner. Specifically, we leverage visual contrastive principles to identify key visual regions and employ Reinforcement Learning (RL) to align the generated latent tokens with these salient regions. Extensive experiments on hallucination benchmarks demonstrate that our method significantly outperforms mainstream de-hallucination approaches and even surpasses fully supervised visual reasoning baselines.

Keywords: Self-supervised learning · Latent visual reasoning · De-hallucination

1 Introduction

Large Vision-Language Models (LVLMs) have demonstrated remarkable performance across various visual tasks [1, 2, 7, 23, 37]. However, object hallucination remains a significant obstacle [3, 12, 53], severely compromising the reliability of current models [9, 16]. Mainstream de-hallucination approaches predominantly focus on post-processing output logits [8, 13, 14, 17, 25, 39] or steering latent features [4, 24, 45]. Recently, Visual Chain-of-Thought (Visual-CoT) methods have proven more effective than the aforementioned strategies by grounding reasoning in visual details [6, 15, 27, 29, 31, 32, 42, 44, 50–52]. Nevertheless, current Visual-CoT paradigms ("thinking with images") rely heavily on multi-turn interactions and external visual tools, such as croppers, depth estimators, or plotting agents [6, 32, 42]. This reliance on external modules limits their applicability and lacks an intrinsic mechanism for perceptual exploration within the model itself.

Drawing inspiration from latent reasoning [11], recent research has shifted towards internalizing visual reasoning directly within the model's continuous

[†] Corresponding author.

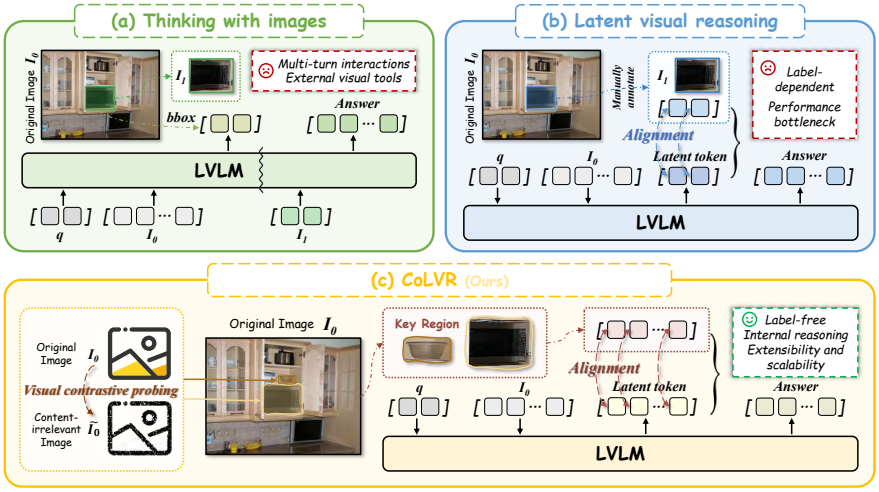

Fig. 1: Comparisons of different visual reasoning paradigms. (a) Thinking with images relies on external tools (*e.g.*, croppers) and multi-turn interactions, lacking intrinsic autonomy. (b) Existing latent visual reasoning internalizes thoughts but depends on costly manual annotations of auxiliary images for alignment. (c) Our `CoLVR` enables self-supervised latent reasoning. By leveraging visual contrastive probing to identify key regions, it aligns latent tokens with salient visual evidence in a label-free and tool-free manner.

space. By generating latent embeddings as intermediate visual thoughts, LVLMs can effectively circumvent the reliance on external tools and explicit image outputs [10, 18, 19, 22, 34, 38, 46, 48]. Despite this progress, the majority of existing approaches still heavily rely on strong supervision. They typically construct explicit visual chains-of-thought by aligning latent features with the embeddings of manually annotated auxiliary images via Supervised Fine-Tuning (SFT). This reliance on supervision introduces two critical limitations: (1) Scalability issues, as the dependence on manual annotation of auxiliary images incurs prohibitive labor costs; and (2) Performance bottlenecks, where the reasoning capability of the model is inherently bounded by the quality and granularity of the human-curated visual chains.

Recent studies have further traced the origins of hallucination through the lens of attention mechanisms [5, 13, 25, 40]. It has been observed that during hallucinated generation, the model's self-attention exhibits an over-reliance on textual tokens while neglecting actual visual evidence. To address this, methods such as OPERA [13] and PAI [25] propose mitigating hallucinations by reinforcing attention towards critical visual information. Inspired by the philosophy of VCD [17] and ICD [39], recent research has elevated this contrastive paradigm from the output logit space to the internal attention mechanism (such as ASCD [40]). By contrasting attention responses between the original image and content-irrelevant images, these methods effectively filter out language pri-

ors and positional biases, thereby precisely localizing the critical visual patches required for de-hallucination.

To address limitations (1) and (2) and leverage the insights of visual contrastive probing [4,14,25], we propose Contrastive-Guided Self-Supervised Latent Visual Reasoning (CoLVR). This novel training framework enables visual latent reasoning without requiring any annotated data and ground truth labels. Specifically, our training pipeline consists of two stages: SFT and Reinforcement Learning (RL). The SFT stage aims to equip the model with the fundamental capability to generate and utilize latent tokens. In the RL stage, we design a Visual Contrastive-Guided RL mechanism based on Group Relative Policy Optimization (GRPO) [30], which encourages the generated latent tokens to activate the model's attention towards critical visual regions, thereby effectively mitigating hallucinations. Crucially, our approach eliminates the dependency on human-annotated visual chains-of-thought and ground truth labels. As illustrated in Fig. 1, our CoLVR distinguishes itself from "thinking with images" and existing label-dependent supervised latent visual reasoning methods. We provide further details in Sec. 3.2 and Sec. 3.3.

Experimental results on standard hallucination benchmarks [9, 20, 33, 35] demonstrate that CoLVR effectively mitigates object hallucination. Implemented on the Qwen2.5-VL-7B backbone, our method outperforms the base model and existing de-hallucination approaches [13,17,25] and visual reasoning methods [18, 31, 38, 52]. Furthermore, our method exhibits strong extensibility. The latent reasoning patterns learned through self-supervision—specifically the alignment with critical visual regions—provide a solid foundation for subsequent supervised RL. Consequently, CoLVR achieves superior performance on General Visual Understanding and Multimodal Reasoning benchmarks [26, 36, 43, 49]. We provide further details in Sec. 4.2.

Our contributions are summarized as follows:

- We propose CoLVR, an annotation-free framework for self-supervised latent training, comprising SFT and RL stages. Guided by visual contrastive probing, our method effectively steers latent tokens to autonomously focus on critical visual regions.
- Extensive experiments on hallucination benchmarks demonstrate that our CoLVR significantly outperforms mainstream de-hallucination approaches and existing supervised latent visual reasoning baselines.
- Our framework exhibits strong extensibility for post-training. In the same post-training settings, our CoLVR achieves superior performance on General Visual Understanding and General Multimodal Reasoning benchmarks compared to baselines.

2 Related Work

2.1 Latent Visual Reasoning

Thinking with images. Recent advancements in Large Vision-Language Models (LVLMs) [1, 2, 7, 23, 37] have explored explicit visual reasoning paradigms,

commonly referred to as Visual Chain-of-Thought (Visual-CoT) ("thinking with images") [15, 27, 29, 44, 50–52]. These methods decompose complex tasks into intermediate visual steps grounded in specific image regions, such as Deepeyes [52], Chain of Focus [50] and Pixel-Reasoner [31]. Despite their efficacy, these approaches typically hinge on multi-turn interactions or external visual tools (*e.g.*, detectors, croppers), lacking an intrinsic perceptual mechanism for autonomous exploration within the model itself.

Latent visual reasoning. To circumvent the dependency on external tools, research has shifted toward internalizing reasoning within the continuous hidden space of LVLMs [11]. By generating latent embeddings as intermediate "visual thoughts," models can achieve tool-free, single-turn inference [10,18,19,22,34,38, 46, 48]. However, existing paradigms predominantly rely on strong supervision, where latent representations are aligned with manually annotated auxiliary images or region-level tokens via SFT. This dependence not only incurs prohibitive annotation costs but also imposes a performance bottleneck dictated by the quality of human-curated data. In contrast, our `CoLVR` introduces a self-supervised alignment mechanism that eliminates the need for any manual annotations.

2.2 Mitigation of Object Hallucination

Object hallucination remains a primary obstacle for LVLMs, where generated content lacks grounding in the input image. Existing mitigation strategies include: (i) Logit-level post-processing by contrastive decoding (*e.g.*, VCD [17], ICD [39] and M3ID [8]), which adjusts output distributions by contrasting original and perturbed inputs to suppress language priors; (ii) Visual feature steering (*e.g.*, VTI [24] and VaLSe [4]), which alleviates over-reliance on textual tokens by reinforcing attention toward visual evidence; and (iii) Preference optimization, where DPO-based methods (*e.g.*, TPO [21] and RLAIF-V [47]) construct preference signals to favor visually grounded responses over hallucinated ones.

Our method is inspired by the ideas of visual contrastive and latent steering. We use contrastive visual signals to guide the training of latent representations by contrastive visual probing, which localizes query-relevant regions by comparing the attention matrix between original and content-irrelevant images. Fundamentally, our approach shares the same underlying philosophy as methods like PAI [25] and VaLSe [4]: mitigating hallucinations by explicitly reinforcing the model's attention toward visual evidence. While DPO-based methods are orthogonal to our latent visual reasoning framework, they are potentially complementary to our extension.

3 Methods

We begin by outlining the problem setup and the proposed latent inference mechanism in Sec. 3.1. In Fig. 2, the training framework of `CoLVR` consists of two stages: (i) the SFT stage (Sec. 3.2), which trains the model to generate latent tokens that encode localized visual regions; and (ii) the RL stage (Sec. 3.3),

which aligns these latent embeddings with query-relevant visual regions via visual contrastive probing to mitigate hallucinations.

3.1 Preliminaries

Problem setup. Given an input image $\mathbf{I}$ and a user query $\mathbf{q}$, an LVLM generates an answer sequence $\mathbf{y} = (y_1, \ldots, y_{T_y})$. Our goal is to endow the LVLM with an intrinsic visual reasoning mechanism that reduces hallucinations by explicitly grounding intermediate reasoning in visual evidence, without invoking external visual tools.

Auto-regressive model decoding. We consider a decoder-only LVLM where visual and text tokens are processed as a unified sequence [1, 2, 23, 37]. Let $\mathbf{v}$ denote the visual tokens encoded from the input image $\mathbf{I}$. We define the unified sequence as $\mathbf{s} = [\mathbf{v}, \mathbf{q}, \mathbf{y}] = (s_1, \ldots, s_T)$, consists of the visual tokens $\mathbf{v}$, the query tokens $\mathbf{q}$, and the response tokens $\mathbf{y}$. Let $\mathbf{H}^{(\ell)} = (\mathbf{h}_1^{(\ell)}, \ldots, \mathbf{h}_T^{(\ell)})$ be the hidden states at layer $\ell \in \{1, \ldots, L\}$. The model produces next-token logits $\mathbf{o}_t \in \mathbb{R}^{|\mathcal{V}|}$ and predicts $p_\theta(s_{t+1} \mid s_{\leq t}) = \text{Softmax}(\mathbf{o}_t)$.

Latent mode inference. We extend the vocabulary with three special tokens: ⟨`latent`⟩, ⟨/`latent`⟩, and ⟨`latent_pad`⟩. When the model predicts ⟨`latent`⟩ during text inference, the inference transitions into latent mode. The model begins generating latent token embeddings, denoted as $\mathbf{z}$. These embeddings are derived from the LVLM's last-layer hidden states (details in Sec. 3.2). Specifically, when predicting the k-th latent token, the embedding $\mathbf{z}_{k-1}$ generated at the previous step serves as the input embedding. Each latent token position corresponds to a ⟨`latent_pad`⟩ placeholder in the autoregressive sequence, which serves only as a slot for continuous latent embeddings rather than a semantic text token to be generated. When the sequence length reaches $k = K$, we append ⟨/`latent`⟩ to terminate the latent mode and switch back to text inference. We adopt a fixed-length design (K is a hyper-parameter) for simplicity and stability, following the effective practice in prior work [18, 38].

3.2 Latent Alignment Supervision via SFT

The SFT stage equips the LVLM with the capability to (i) decide when to enter/exit latent mode by generating ⟨`latent`⟩ and ⟨/`latent`⟩, and (ii) make latent embedding encode localized visual evidence. We optimize a standard language modeling loss together with a latent-to-region alignment loss without any annotated labels.

Region pseudo-labels without manual annotation. To avoid reliance on human-annotated visual chains-of-thought, we construct coarse region supervision via a fixed partition of the image into a small set of canonical regions (*e.g.*, `top-left`, `top-right`, `bottom-left`, `bottom-right`, `center`), each associated with a bounding box $\mathbf{b} = (x_1, y_1, x_2, y_2)$ [1]that fluctuates within a certain range

[1] y is temporarily reused for coordinates; it otherwise denotes answer sequence.

in pixel coordinates. These localized regions serve as supervision signals to train the latent tokens in encoding visual information.

Latent embeddings from hidden states. Latent token embeddings correspond to continuous embeddings extracted from the LVLM forward pass. Specifically, for each $t \in \mathcal{P}_{\text{lat}}$, where $\mathcal{P}_{\text{lat}} = \{t \mid s_t = \langle \texttt{latent\_pad} \rangle\}$ denotes the set of latent token positions, we treat the last-layer hidden state $\mathbf{h}_t^{(L)} \in \mathbb{R}^d$ as the continuous embedding at that latent token. This choice makes latent supervision and optimization compatible with standard autoregressive decoding, while keeping all reasoning internal to the LVLM.

We extract latent embeddings from last-layer hidden states:

$$\mathbf{Z} = \{\mathbf{z}_k\}_{k=1}^{K}, \qquad \mathbf{z}_k = \mathbf{h}_{t_k}^{(L)},\ t_k \in \mathcal{P}_{\text{lat}}, \tag{1}$$

and use their mean as a compact latent summary $\bar{\mathbf{z}} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{z}_k$.

Region embedding from image-token hidden states. Let $\mathcal{P}_{\text{img}}$ denote the positions of image tokens in $\mathbf{s}$, and let $\mathbf{H}_{\text{img}}^{(L)} = \{\mathbf{h}_t^{(L)}\}_{t \in \mathcal{P}_{\text{img}}}$ be their last-layer states. For region-based supervision, we represent a rectangular region by a bounding box $\mathbf{b}$. We define a mapping $\phi(\mathbf{b})$ that converts $\mathbf{b}$ to a subset of image-token positions $\mathcal{P}_{\mathbf{b}} \subseteq \mathcal{P}_{\text{img}}$.

Given a bounding box $\mathbf{b}$, we map it to image-token positions $\mathcal{P}_{\mathbf{b}} = \phi(\mathbf{b})$ and pool their hidden states:

$$\mathbf{r}(\mathbf{b}) = \frac{1}{|\mathcal{P}_{\mathbf{b}}|} \sum_{t \in \mathcal{P}_{\mathbf{b}}} \mathbf{h}_t^{(L)}. \tag{2}$$

Importantly, we compute $\mathbf{r}(\mathbf{b})$ in the LVLM hidden space (rather than directly using vision-encoder features), so that latent and region representations lie in the same semantic space.

Latent-to-region alignment loss. We encourage latent tokens to match the corresponding region representation by maximizing cosine similarity:

$$\mathcal{L}_{\text{align}} = \frac{1}{N} \sum_{i=1}^{N} \Big(1 - \cos\big(\bar{\mathbf{z}}^{(i)}, \mathbf{r}^{(i)}(\mathbf{b}^{(i)})\big)\Big), \tag{3}$$

where N is the batch size and $\cos(\mathbf{a}, \mathbf{c}) = \frac{\mathbf{a}^\top \mathbf{c}}{\|\mathbf{a}\|\|\mathbf{c}\|}$.

Latent-only backpropagation. A naive optimization of Eq. (3) can admit shortcut solutions where the model reduces $\mathcal{L}_{\text{align}}$ by altering non-latent pathways (*e.g.*, image-token representations) instead of making the latent token carry region information. To prevent such degeneration, we restrict gradients from $\mathcal{L}_{\text{align}}$ to flow only through the latent token embeddings $\{\mathbf{z}_k\}$, following the latent-only update principle in previous work [38]. Concretely, let $g_k = \nabla_{\mathbf{z}_k} \mathcal{L}_{\text{align}}$ and $\text{sg}(\cdot)$ denote stop-gradient. We optimize a proxy objective:

$$\widetilde{\mathcal{L}}_{\text{align}} = \sum_{k=1}^{K} \mathbf{z}_k^\top \text{sg}(g_k), \tag{4}$$

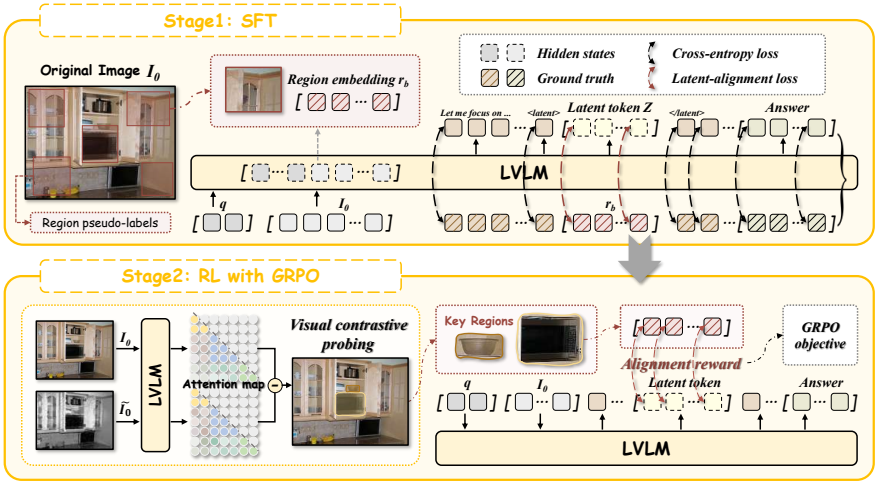

Fig. 2: Overview of CoLVR. Top (SFT): Latent tokens ($\mathbf{Z}$) are aligned with coarse region embeddings ($\mathbf{r}_b$); Bottom (RL with GRPO): Visual contrastive probing ($\mathbf{I}_0$ vs. $\tilde{\mathbf{I}}_0$) identifies query-relevant key regions. GRPO then maximizes an alignment reward, explicitly grounding the latent tokens to these key regions to achieve annotation-free training for de-hallucination.

which preserves the desired latent-only gradient contribution while blocking gradients through other states. We report $\mathcal{L}_{\text{align}}$ for monitoring (interpretable cosine-based metric) and use $\widetilde{\mathcal{L}}_{\text{align}}$ for optimization.

Next-token prediction loss. We also optimize the standard next-token cross-entropy loss $\mathcal{L}_{\text{CE}}$ on text tokens.

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{P}_{\text{txt}}|}\sum_{t\in\mathcal{P}_{\text{txt}}}\log p_\theta\left(s_t^{(i)} \mid s_{<t}^{(i)}\right), \tag{5}$$

where $\mathcal{P}_{\text{txt}}$ denotes the set of text tokens positions in response tokens.

Specially, we do not mask ⟨`latent`⟩ and ⟨/`latent`⟩, so the model learns to generate these control tokens. In contrast, we mask ⟨`latent_pad`⟩ in $\mathcal{L}_{\text{CE}}$ because it is a placeholder for continuous latent embeddings rather than a linguistic token to be predicted.

Total SFT loss. The final SFT objective is

$$\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{CE}} + \lambda\,\widetilde{\mathcal{L}}_{\text{align}}, \tag{6}$$

where λ balances linguistic accuracy and latent grounding, and we fix $\lambda = 1$ in all experiments. This stage provides a functional latent interface for the subsequent contrastive-guided RL alignment to query-relevant salient regions (Sec. 3.3).

3.3 Contrastive-Guided Reinforcement Learning with GRPO

The SFT stage (Sec. 3.2) provides a usable latent interface, but it does not ensure that the latent embeddings $\mathbf{Z}$ are grounded on the query-relevant visual evidence. We therefore introduce a contrastive-guided RL stage that (i) discovers query-relevant salient visual tokens via visual contrastive probing, and (ii) aligns the generated latent embeddings to these salient regions by optimizing a GRPO objective, without requiring ground-truth answers.

Visual contrastive probing for query-relevant saliency. Given an input pair $(\mathbf{I}, \mathbf{q})$, we construct a content-irrelevant contrastive image $\tilde{\mathbf{I}}$ (*e.g.*, blank or corrupted) and feed them into the model separately. For each image-token position $t \in \mathcal{P}_{\text{img}}$ (defined in Sec. 3.2), we compute a contrastive value

$$a_t \;=\; g_\theta(\mathbf{I}, \mathbf{q})_t \;-\; g_\theta(\tilde{\mathbf{I}}, \mathbf{q})_t, \tag{7}$$

where $g_\theta(\cdot)_t$ is a patch-level scalar function. We instantiate g_θ based on attention matrix. Let $\mathbf{A}^{(\ell,h)} \in \mathbb{R}^{T\times T}$ be the self-attention matrix at layer ℓ and head h, where $\mathbf{A}^{(\ell,h)}_{i,t}$ denotes attention from token i to token t. Let $\mathcal{P}_{\text{q}}$ denote the positions of query tokens $\mathbf{q}$ in $\mathbf{s} = [\mathbf{v}, \mathbf{q}, \mathbf{y}]$. We define

$$g^{\text{att}}_\theta(\mathbf{I}, \mathbf{q})_t \;=\; \frac{1}{LH}\sum_{\ell=1}^{L}\sum_{h=1}^{H}\left(\frac{1}{|\mathcal{P}_{\text{q}}|}\sum_{i\in\mathcal{P}_{\text{q}}}\mathbf{A}^{(\ell,h)}_{i,t}\right). \tag{8}$$

We then obtain a non-negative saliency distribution over image tokens by rectification and normalization:

$$w_t \;=\; \frac{\max(a_t, 0)}{\sum_{j\in\mathcal{P}_{\text{img}}}\max(a_j, 0)}, \qquad \sum_{t\in\mathcal{P}_{\text{img}}} w_t = 1. \tag{9}$$

Reward design. For each sampled completion $\mathbf{y}$, we extract the mean latent embedding $\bar{\mathbf{z}}$ from latent token positions $\mathcal{P}_{\text{lat}}$ as in Sec. 3.2. Given contrastive saliency weights $\{w_t\}_{t\in\mathcal{P}_{\text{img}}}$ from Eq. (9), we first define a per-patch similarity

$$s_t \;=\; \frac{1+\cos\left(\bar{\mathbf{z}},\, \mathbf{h}^{(L)}_t\right)}{2} \in [0,1], \qquad t \in \mathcal{P}_{\text{img}}. \tag{10}$$

We consider two variants of the latent-to-saliency alignment reward: (i) *soft* alignment,

$$r_1^{\text{soft}}(\mathbf{I}, \mathbf{q}, \mathbf{y}) \;=\; \sum_{t\in\mathcal{P}_{\text{img}}} w_t\, s_t, \tag{11}$$

and (ii) *hard* alignment, where $\mathcal{S} \subseteq \mathcal{P}_{\text{img}}$ is the top-M salient set under w_t,

$$r_1^{\text{hard}}(\mathbf{I}, \mathbf{q}, \mathbf{y}) \;=\; \frac{1}{|\mathcal{S}|}\sum_{t\in\mathcal{S}} s_t. \tag{12}$$

To ensure the model actually enters latent mode, we use a simple format reward that only checks for the existence of a valid latent block:

$$r_2(\mathbf{I}, \mathbf{q}, \mathbf{y}) = \begin{cases} 0, & \text{if } \mathbf{y} \text{ contains a latent block,} \\ -1, & \text{otherwise.} \end{cases} \tag{13}$$

The total reward is

$$R(\mathbf{I}, \mathbf{q}, \mathbf{y}) \;=\; r_1(\mathbf{I}, \mathbf{q}, \mathbf{y}) + r_2(\mathbf{I}, \mathbf{q}, \mathbf{y}), \tag{14}$$

where r_1 is instantiated as either r_1^{soft} or r_1^{hard} and we use r_1^{hard} in main experiments. Equivalently, one may write $R = r_1 + \beta r_2$ with a single coefficient β, and we fix $\beta = 1$ in all experiments.

GRPO objective. For each prompt $(\mathbf{I}^{(i)}, \mathbf{q}^{(i)})$, we sample G completions $\{\mathbf{y}^{(i,g)}\}_{g=1}^{G}$ from π_θ and compute rewards $R^{(i,g)}$. GRPO constructs group-relative advantages

$$A^{(i,g)} = \frac{R^{(i,g)} - \frac{1}{G}\sum_{g'=1}^{G} R^{(i,g')}}{\sigma^{(i)} + \epsilon}, \tag{15}$$

where $\sigma^{(i)}$ is the standard deviation of $\{R^{(i,g)}\}_{g=1}^{G}$ and ϵ is a small constant. The GRPO loss is

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{i,g}\left[A^{(i,g)} \sum_{t \in \mathcal{P}_{\text{txt}}} \log p_\theta\Big(y_t^{(i,g)} \mid \mathbf{I}^{(i)}, \mathbf{q}^{(i)}, \mathbf{y}_{<t}^{(i,g)} \Big) \right] + \beta\, \mathbb{E}_{i,g}[\mathrm{KL}(\pi_\theta \,\|\, \pi_{\text{ref}})]\,, \tag{16}$$

where β optionally regularizes the policy to a reference π_{ref} (set to 0 when no reference is used). Optimizing Eq. (16) increases the probability of completions whose latent embeddings better align with contrastively identified salient visual evidence, thereby improving visual grounding and mitigating hallucinations.

4 Experiments

4.1 Setup

Evaluation benchmarks and metrics. We conduct experiments on CHAIR [33], PoPE [20], Hallbench [9] and MMHal [35] to evaluate the effectiveness of our `CoLVR`. For CHAIR, we report three metrics: CHAIR_s (C_s), CHAIR_i (C_i) and F1 scores (F1); For PoPE, we report accuracy (Acc.) and F1; For Hallbench, we report Acc.; For MMHal, we report score and hallucination rate (Hal.) evaluated by GPT-4 [28].

To verify the extensibility of our model on general visual tasks, we also evaluate on MMVP [36], MathVista [26], MathVerse [49] and V^* [43] benchmarks. For all benchmarks, we report their accuracy as metric.

Baselines. We adopt Qwen2.5-VL-7B and Qwen2.5-VL-3B as base model and compare `CoLVR` against the following baselines: (1) Base model [2]. (2) Train-free

methods, such as VCD [17], OPERA [13], PAI [25] and HALC [5]. (3) "Thinking with images" methods, including Deepeyes [52] and Pixel-Reasoner [31]. (4) Existing latent visual reasoning methods, including LVR [18] and Monet [38].

Training details. We train our models with Qwen2.5-VL-7B and Qwen2.5-VL-3B. The training data for both the SFT and RL stages are derived from Visual-CoT data [29], which is a large-scale VQA dataset. In the SFT stage, we randomly sample 2K raw instances. The ground truth for training is constructed by combining pseudo-region labels with the responses generated by the original model, resulting in a total of 6K SFT samples. In the RL stage, we sample a larger-scale dataset of approximately 30K instances.

We employ the Low-Rank Adaptation (LoRA) fine-tuning strategy, freezing the vision encoder while updating the remaining parameters. During the SFT stage, the batch size is 32, the learning rate is 1×10^{-5}, and the model is trained for 1 epoch. In the RL stage, we use the Hugging Face TRL framework [41] for GRPO training. The sampling group size G is set to 8, the KL coefficient β is 0.01, the batch size is 16, and the learning rate is 1×10^{-6} for 3 epochs. The full CoLVR training takes approximately 18 hours on 4 A-800 GPUs. Further implementation details are provided in Sec. A.

4.2 Main Results

As presented in Tab. 1, we can draw the following key observations:

- **Superior overall performance.** Our proposed `CoLVR` significantly mitigates object hallucination, achieving the highest performance of 80.1% (F1, +1.8%) on CHAIR and 89.6% (Acc., +4.7%) on PoPE on the 7B backbone. While train-free approaches (*e.g.*, VCD, OPERA) yield marginal improvements by post-processing logits or steering attention during inference, they fail to inherently correct the model's visual grounding. In contrast, `CoLVR` internalizes visual perception through self-supervised latent training, leading to an absolute improvement over those baselines.
- **Surpassing supervised baselines without annotations.** Remarkably, `CoLVR` outperforms existing supervised latent visual reasoning paradigms (such as Monet and LVR) without relying on any human-annotated labels. Specifically, `CoLVR` reduces the C_s error metric to 20.2% (-10.6%) and achieves a leading score of 5.12 (+0.26) on the MMHal benchmark. This compellingly verifies that our contrastive-guided RL effectively circumvents the performance bottlenecks inherent to human-curated Visual-CoT, enabling the model to autonomously align with salient visual evidence.
- **Consistent scalability across model sizes.** The de-hallucination efficacy of `CoLVR` generalizes seamlessly across different model sizes. Implemented on the lighter Qwen2.5-VL-3B backbone, our method still consistently outperforms train-free baselines and effectively suppresses the MMHal hallucination rate (Hal.) to 30.2% (-4.2%). This demonstrates the robustness and broad applicability of our annotation-free framework.

Table 1: Main results on evaluation hallucination benchmarks. Our method, `CoLVR`, is compared against all baseline models and methods described in Sec. 4.1. 'Tr.' and 'Lb.' denote whether the method requires training and human-annotated labels, respectively. ↑ indicates higher is better, and ↓ indicates lower is better. Our proposed methods are highlighted in gray and best results are in **bold**.

Method	Tr.	Lb.	CHAIR			POPE		Hallbench	MMHal	
			C_s↓	C_i↓	F1↑	Acc.↑	F1↑	Acc.↑	Score↑	Hal.↓
Qwen2.5-VL-7B										
Base	×	×	30.8	8.1	78.3	84.9	84.0	52.5	4.86	32.3
Train-free										
VCD	×	×	34.6	8.7	78.6	86.0	84.4	52.9	4.85	33.3
OPERA	×	×	31.6	8.5	78.7	85.4	83.4	52.9	5.02	31.3
PAI	×	×	34.4	8.8	77.8	88.2	87.3	53.2	5.00	37.5
HALC	×	×	27.0	9.9	65.9	87.7	**89.1**	52.5	4.27	38.5
Thinking with images										
Deepeyes-7B	✓	✓	27.6	7.3	78.6	86.1	85.1	53.9	5.10	**29.2**
Pixel-Reasoner-7B	✓	✓	23.2	6.7	77.1	84.8	82.9	42.6	4.77	33.3
Latent visual reasoning										
LVR-7B	✓	✓	33.2	9.8	79.4	84.8	84.4	50.4	4.88	37.5
Monet-7B	✓	✓	34.1	9.9	67.8	86.2	87.3	52.1	5.02	31.3
Ours										
`CoLVR`-7B (ours)	✓	×	**20.2**	**6.1**	**80.1**	**89.6**	86.3	**54.3**	**5.12**	30.2
Qwen2.5-VL-3B										
Base	×	×	38.4	**8.3**	79.0	85.9	84.6	40.6	4.69	34.4
Train-free										
VCD	×	×	48.6	9.6	79.4	85.2	84.0	39.3	4.46	38.5
OPERA	×	×	43.2	9.6	**79.9**	85.9	84.6	40.8	**4.91**	31.3
PAI	×	×	48.8	9.7	**79.9**	86.0	84.8	40.4	4.67	46.9
HALC	×	×	32.2	11.5	64.8	87.8	**87.6**	41.1	4.28	34.4
Ours										
`CoLVR`-3B (ours)	✓	×	**30.8**	8.5	78.6	**88.2**	85.9	**42.6**	4.78	**30.2**

4.3 Extensibility of `CoLVR`

We believe that models trained with our method possess a unified latent visual reasoning capability that extends beyond de-hallucination, which can be further scaled with annotated data to achieve excellent performance across downstream and diverse visual tasks. Compared to the "thinking with images" paradigm and supervised latent visual reasoning, our self-supervised method achieves superior performance. The former is limited by external tools and multi-turn interactions, while the latter is constrained by the quality of annotated Visual-CoT data. Therefore, our method possesses stronger extensibility and scalability.

Table 2: Extensibility evaluation on general visual understanding and multimodal reasoning benchmarks. All methods are based on the Qwen2.5-VL-7B backbone. 'Tr.' and 'Lb.' denote whether the method requires training and human-annotated labels, respectively. `CoLVR` * denotes the variant subjected to further extension training initialized from our base `CoLVR`. All metrics are accuracy (↑ indicates higher is better). Our proposed methods are highlighted in gray and best results are in **bold**.

Method	Tr.	Lb.	MMVP↑	MathVista↑	MathVerse↑	V^*↑
Base	×	×	68.3	68.6	49.2	78.5
SFT	✓	✓	65.7	68.9	50.1	79.1
Thinking with images						
Deepeyes-7B	✓	✓	72.0	70.1	47.3	82.7
Pixel-Reasoner-7B	✓	✓	67.0	69.2	46.2	80.1
Latent visual reasoning						
LVR-7B	✓	✓	71.7	69.2	50.1	80.6
Monet-7B	✓	✓	**72.1**	71.2	51.2	83.3
Ours						
`CoLVR` (ours)	✓	×	68.0	68.2	49.3	78.5
`CoLVR` *(ours)	✓	✓	71.2	**73.4**	**51.4**	**83.8**

Table 3: Ablation study on the core training stages of `CoLVR`. All models are evaluated on the Qwen2.5-VL-3B backbone. 'SFT' and 'RL' denote whether the SFT and RL stages are applied, respectively. Best results are in **bold**.

Method	SFT	RL	CHAIR			POPE		Hallbench	MMHal	
			C_s↓	C_i↓	F1↑	Acc.↑	F1↑	Acc.↑	Score↑	Hal.↓
Base	×	×	38.4	8.3	**79.0**	85.9	84.6	40.6	4.69	34.4
SFT-only	✓	×	38.6	**8.5**	78.3	86.0	84.4	40.8	4.46	34.4
`CoLVR` (SFT+RL)	✓	✓	**30.8**	**8.5**	78.6	**88.2**	**85.9**	**42.6**	**4.78**	**30.2**

To verify this hypothesis, we continue to train our model on LVR-SFT data [18] using SFT. We evaluate it on general visual benchmarks (see Sec. 4.1) and compare it with baseline methods from both "thinking with images" and latent visual reasoning paradigms.

As reported in Tab. 2, our `CoLVR` * consistently outperforms both "thinking with images" and supervised latent visual reasoning baselines across all benchmarks. Notably, it achieves competitive accuracies of 73.4% on MathVista and 51.4% on MathVerse, yielding an absolute improvement of 2.2% and 0.2% over the strongest baseline, respectively. These results compellingly validate that our self-supervised latent alignment paradigm provides a superior foundation for generalized multimodal reasoning.

4.4 Ablation Study and Visualization

Effect of main components. We ablate the core training stages of `CoLVR` to validate their individual contributions: (1) SFT-only, which equips the model with basic latent tokens guided by coarse region pseudo-labels; and (2) SFT+RL (Full `CoLVR`), which further aligns these latent embeddings with salient visual evidence using contrastive-guided GRPO. As shown in Tab. 3, relying solely on SFT yields a moderate performance of 78.3% (F1.) on CHAIR and 84.4% (Acc.) on PoPE. By introducing the RL stage, our full `CoLVR` framework effectively bridges the visual grounding gap, significantly boosting the score (+1.3% F1 on PoPE, +2.0% accuracy on Hallbench and +0.09 Score on MMHal) and demonstrating that contrastive reward optimization is effective for de-hallucination.

Effect of latent length (K). We further analyze the sensitivity of our framework to the latent reasoning length K by varying it across $\{1, 2, 4, 8, 16, 32\}$. As illustrated in Fig. 3, the model achieves its optimal dehallucination performance at $K = 4$. Notably, `CoLVR` maintains consistently effective performance across the broader range of $K \in [2, 32]$. This performance plateau demonstrates the strong robustness of our self-supervised latent alignment to this hyperparameter. Intuitively, a singular latent token ($K = 1$) provides insufficient representational capacity to encapsulate complex regional semantics. Conversely, extending the latent sequence beyond necessary bounds yields diminishing returns and marginally complicates the GRPO optimization without providing further reasoning benefits. Therefore, $K = 4$ serves as an ideal balance between reasoning cost and optimization efficiency.

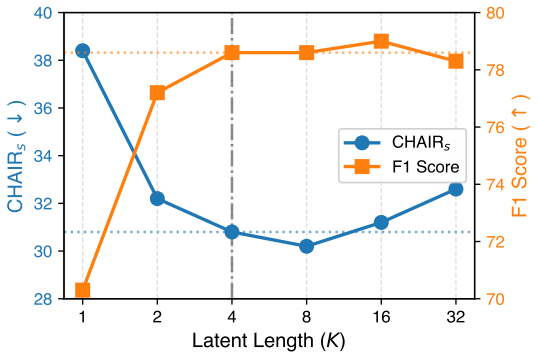

Fig. 3: Ablation study on the latent length K on CHAIR benchmark with Qwen2.5-VL-3B backbone. `CoLVR` achieves ideal balance at $K = 4$.

Effect of latent-only backpropagation. We investigate the critical role of the stop-gradient operator (sg) introduced in Eq. (4) on Qwen2.5-VL-3B backbone. We conduct a variant experiment where gradients from the alignment loss are allowed to backpropagate freely into the standard image tokens. Removing this constraint results in a severe performance degradation, reducing the PoPE Accuracy to 84.6% (-3.6%) and the MMHal score to 4.28 (-0.50). Without latent-only gradient routing, the network exploits optimization shortcuts: it alters the foundational visual representations to match the initialized latent tokens, rather than forcing the latent embeddings to actively encode salient visual regions.

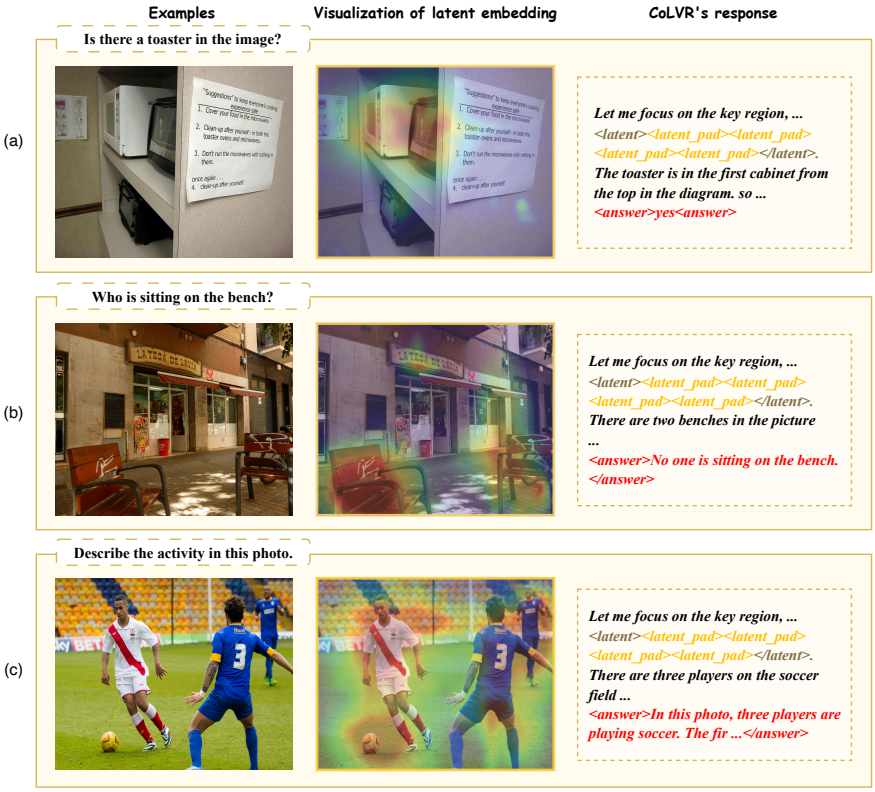

Fig. 4: Inference examples of CoLVR. The middle column displays a visualization generated by CoLVR, mapping latent tokens to the image by calculating cosine similarity.

This causes severe representation degeneration, thereby confirming that sg is indispensable for stable latent visual reasoning.

Table 4: Ablation on contrastive visual probing on CHAIR using Qwen2.5-VL-3B. CoLVR-3B results are **bold**.

Variant	$C_s \downarrow$	$C_i \downarrow$	F1↑
Gaussian contrast	**30.8**	**8.5**	**78.6**
Blank contrast	36.9	8.9	76.5
Original-image attn.	33.3	8.3	75.9
Random contrast	38.1	9.6	74.1
Reverse contrast	46.9	12.8	67.6

Effect of contrastive visual probing. We further ablate the saliency source used in contrastive visual probing on CHAIR. As shown in Tab. 4, replacing our Gaussian-noise contrast with original-image attention, random contrast, or reverse contrast consistently hurts performance, with reverse contrast causing the most severe degradation. This confirms that both the contrastive formulation and its direction are important for identifying reliable visual evidence. Moreover, the blank contrast image performs worse than Gaus-

sian noise, suggesting that Gaussian perturbation provides a more effective content-irrelevant reference. These results support our claim that contrastive attention suppresses language priors and provides more reliable query-relevant saliency for latent alignment.

Visualization. In this section, we provide some selected examples of the inference of `CoLVR`. As shown in Fig. 4, we visualize the image regions that are most relevant to the latent tokens generated by `CoLVR`. It can be observed that the latent tokens produced by our method not only accurately focus on query-relevant key objects but also attend to multiple objects simultaneously. Furthermore, these focused regions are at the patch level, which is superior to the latent visual reasoning methods trained with traditional paradigms.

5 Conclusion

In this paper, we proposed `Co`ntrastive-Guided Self-Supervised `L`atent `V`isual `R`easoning (`CoLVR`), an annotation-free framework for self-supervised latent training, comprising SFT and RL stages. Guided by visual contrastive probing, our method effectively steers latent tokens to autonomously focus on critical visual regions. Extensive experiments demonstrate that `CoLVR` outperforms mainstream de-hallucination approaches, "thinking with images" paradigms, and fully supervised latent visual reasoning methods. Finally, our work overcomes the limitations of manual annotation and performance bottlenecks in latent visual reasoning, inspiring new training paradigms for future development.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J.L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Bińkowski, M.a., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: NeurIPS (2022)
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. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey (2025)
4. Chen, B., Zheng, Z., Yang, L., Geng, Z., Zhao, Z., Lin, C., Shen, C.: Seeing it or not? interpretable vision-aware latent steering to mitigate object hallucinations (2025)
5. Chen, Z., Zhao, Z., Luo, H., Yao, H., Li, B., Zhou, J.: HALC: Object hallucination reduction via adaptive focal-contrast decoding. In: ICML (2024)
6. Chern, E., Hu, Z., Chern, S., Kou, S., Su, J., Ma, Y., Deng, Z., Liu, P.: Thinking with generated images. arXiv preprint arXiv:2505.22525 (2025)

7. Dai, W., Li, J., LI, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. In: NeurIPS (2023)
8. Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., Soatto, S.: Multi-modal hallucination control by visual information grounding. In: CVPR (2024)
9. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., Zhou, T.: Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: CVPR (2024)
10. Han, F., Jiao, Y., Chen, S., Xu, J., Chen, J., Jiang, Y.G.: Controlthinker: Unveiling latent semantics for controllable image generation through visual reasoning. arXiv preprint arXiv:2506.03596 (2025)
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. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**, 1 – 55 (2023)
13. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: CVPR. pp. 13418–13427 (2024)
14. Huo, F., Xu, W., Zhang, Z., Wang, H., Chen, Z., Zhao, P.: Self-introspective decoding: Alleviating hallucinations for large vision-language models. In: ICLR (2025)
15. Jiang, C., Heng, Y., Ye, W., Yang, H., Xu, H., Yan, M., Zhang, J., Huang, F., Zhang, S.: Vlm-r: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. arXiv preprint arXiv:2505.16192 (2025)
16. Kaul, P., Li, Z., Yang, H., Dukler, Y., Swaminathan, A., Taylor, C.J., Soatto, S.: Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In: CVPR (2024)
17. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: CVPR (2024)
18. Li, B., Sun, X., Liu, J., Wang, Z., Wu, J., Yu, X., Chen, H., Barsoum, E., Chen, M., Liu, Z.: Latent visual reasoning (2025)
19. Li, K., Shang, C., Karlinsky, L., Feris, R., Darrell, T., Herzig, R.: Latent implicit visual reasoning (2025)
20. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: EMNLP (Dec 2023)
21. Liao, W., Chu, X., Wang, Y.: Tpo: Aligning large language models with multi-branch & multi-step preference trees (2025)
22. Liu, C., Yang, Y., Fan, Y., Wei, Q., Liu, S., Wang, X.E.: Reasoning within the mind: Dynamic multimodal interleaving in latent space (2025)
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
24. Liu, S., Ye, H., Zou, J.Y.: Reducing hallucinations in large vision-language models via latent space steering. In: ICLR (2025)
25. Liu, S., Zheng, K., Chen, W.: Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In: ECCV (2024)

26. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In: ICLR (2024)
27. Mitra, C., Huang, B., Darrell, T., Herzig, R.: Compositional chain of thought prompting for large multimodal models. In: CVPR (2024)
28. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., et al.: Gpt-4 technical report (2024)
29. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In: NeurIPS (2024)
30. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y.K., Wu, Y., Guo, D.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models (2024)
31. Su, A., Wang, H., Ren, W., Lin, F., Chen, W.: Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. arXiv preprint arXiv:2505.15966 (2025)
32. Su, Z., Li, L., Song, M., Hao, Y., Yang, Z., Zhang, J., Chen, G., Gu, J., Li, J., Qu, X., et al.: Openthinkimg: Learning to think with images via visual tool reinforcement learning. arXiv preprint arXiv:2505.08617 (2025)
33. Sun, A.: Chair-classifier of hallucination as improver. In: IJCNN (2025)
34. Sun, G., Hua, H., Wang, J., Luo, J., Dianat, S., Rabbani, M., Rao, R., Tao, Z.: Latent chain-of-thought for visual reasoning (2025)
35. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L., Wang, Y.X., Yang, Y., Keutzer, K., Darrell, T.: Aligning large multimodal models with factually augmented RLHF. In: ACL (2024)
36. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: CVPR (2024)
37. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution (2024)
38. 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 (2025)
39. Wang, X., Pan, J., Ding, L., Biemann, C.: Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In: ACL (2024)
40. Wang, Y., Bi, J., Ma, Y., Pirk, S.: Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm (2025)
41. von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., Huang, S., Rasul, K., Gallouédec, Q.: TRL: Transformers reinforcement learning (2020)
42. Wu, J., Guan, J., Feng, K., Liu, Q., Wu, S., Wang, L., Wu, W., Tan, T.: Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In: NeurIPS (2025)
43. Wu, P., Xie, S.: V*: Guided visual search as a core mechanism in multimodal llms. In: CVPR (2024)
44. Xu, G., Jin, P., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. In: ICCV (2025)
45. Yang, L., Zheng, Z., Chen, B., Zhao, Z., Lin, C., Shen, C.: Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In: CVPR (2025)

46. Yang, Z., Yu, X., Chen, D., Shen, M., Gan, C.: Machine mental imagery: Empower multimodal reasoning with latent visual tokens (2025)
47. Yu, T., Zhang, H., Li, Q., Xu, Q., Yao, Y., Chen, D., Lu, X., Cui, G., Dang, Y., He, T., Feng, X., Song, J., Zheng, B., Liu, Z., Chua, T.S., Sun, M.: Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness (2025)
48. Zhang, H., Wu, W., Li, C., Shang, N., Xia, Y., Huang, Y., Zhang, Y., Dong, L., Zhang, Z., Wang, L., Tan, T., Wei, F.: Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in mllms. arXiv preprint arXiv:2510.24514 (2025)
49. Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.W., Gao, P., et al.: Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In: ECCV (2024)
50. Zhang, X., Gao, Z., Zhang, B., Li, P., Zhang, X., Liu, Y., Yuan, T., Wu, Y., Jia, Y., Zhu, S.C., Li, Q.: Adaptive chain-of-focus reasoning via dynamic visual search and zooming for efficient vlms (2025)
51. Zhao, K., Zhu, B., Sun, Q., Zhang, H.: Unsupervised visual chain-of-thought reasoning via preference optimization. In: ICCV (2025)
52. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing "thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025)
53. Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., Yao, H.: Analyzing and mitigating object hallucination in large vision-language models. In: ICLR (2024)