MACHINE MENTAL IMAGERY: EMPOWER MULTIMODAL REASONING WITH LATENT VISUAL TOKENS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) excel at multimodal understanding, yet their text-only decoding forces them to verbalize visual reasoning, limiting performance on tasks that demand visual imagination. Recent attempts train VLMs to render explicit images, but the heavy image-generation pre-training often hinders the reasoning ability. Inspired by the way humans reason with mental imagery—the internal construction and manipulation of visual cues—we investigate whether VLMs can reason through interleaved multimodal trajectories without producing explicit images. To this end, we present a Machine Mental Imagery framework, dubbed as **Mirage**, which augments VLM decoding with latent visual tokens alongside ordinary text. Concretely, whenever the model chooses to "think visually", it recasts its hidden states as next tokens, thereby continuing a multimodal trajectory without generating pixel-level images. Begin by supervising the latent tokens through distillation from ground-truth image embeddings, we then switch to text-only supervision to make the latent trajectory align tightly with the task objective. A subsequent reinforcement learning stage further enhances the multimodal reasoning capability. Experiments on diverse benchmarks demonstrate that Mirage unlocks stronger multimodal reasoning without explicit image generation.

1 INTRODUCTION

Vision–language models (VLMs) jointly encode images and text and attain impressive results on visual tasks through text-only decoding (Wang et al., 2024). Techniques such as chain-of-thought prompting and reinforcement-learning can lengthen these textual reasoning traces and yield extra gains. Nonetheless, VLMs still stumble on multimodal reasoning tasks such as spatial reasoning, which require active understanding and manipulation of visual elements beyond passive perception.

Consider the jigsaw puzzle in Fig. 1. Instead of textualizing every candidate piece, people picture how the two fragments might align and decide on the correct match. This reasoning unfolds in a native multimodal fashion, not through language alone. Recent studies (Team, 2024; Tong et al., 2024; Chern et al., 2024; Chen et al., 2025a) have pre-trained VLMs for large-scale image generation so a single model can produce both words and pictures. Yet the cognitive demands of logical reasoning differ sharply from the task of synthesizing pixels, and asking one model to master both goals often degrades its reasoning quality (Wang et al., 2025a). In addition, the image decoders cannot produce interleaved trajectories pertinent to input images. Consequently, fully exploiting the dormant multimodal reasoning capacity of VLMs remains an open challenge.

According to imagery theory, humans do not summon photorealistic pictures while thinking. We instead construct and manipulate mental images, simplified sketches that capture only task-relevant information, a process known as **mental imagery** (Shepard & Metzler, 1971; Kosslyn, 1996). In the jigsaw example, we examine fragment contours to decide whether two pieces fit. Likewise, when searching for misplaced keys, we recall the shelf rather than the full room. Inspired by this behavior, we ask whether VLMs can reason directly in their latent visual embedding space, weaving compact visual embeddings into the text stream and dispensing with the need for explicit image generation.

To this end, we present **Mirage**, a decoding mechanism that interleaves latent visual representations among text tokens. Prior studies have shown that LLMs can reason directly within the latent space. Building upon this insight, in our framework, when the model chooses to reason visually by producing a special token, it then reuses its current hidden state as a compact visual embedding and appends it to

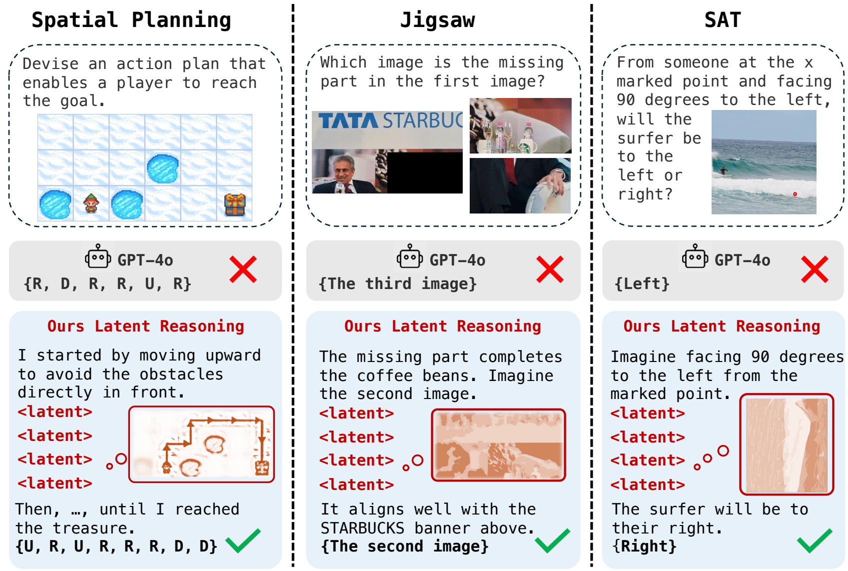

Figure 1: **Multimodal Reasoning Examples.** Mirage interleaves **latent visual tokens**, which represent compact imagery visual features, with explicit text tokens to solve diverse spatial reasoning multimodal tasks, boosting the reasoning performance without the full pixel-level image generation.

the context, skipping the language projection. These internal embeddings furnish focused visual cues for later reasoning steps. As shown in Fig. 1, Mirage yields a chain-of-thought trajectory without any external image decoder.

As illustrated in Fig. 2, we adopt a two-stage fine-tuning paradigm to equip the model with interleaved reasoning. In the first stage, with annotated interleaving trajectories, we supervise both modalities: the model predicts the next word while reconstructing a compact latent visual vector obtained from compressed image embeddings. This dual objective anchors the latent tokens in the visual subspace and teaches the model to weave visual cues into its output.

The second stage removes direct supervision on the latent vectors and optimizes only the text tokens, letting the model treat its autoregressively generated latent embeddings as priors that guide subsequent word generation. This relaxation yields a more flexible interleave reasoning trajectory without forcing the latent channel to match any predetermined embedding. After these two stages, we apply reinforcement learning to further boost the reasoning performance.

Extensive experiments and superior performance across multiple benchmarks demonstrate that our proposed Mirage significantly enhances the reasoning ability of VLMs compared with text-only decoding. More concretely, our contributions are threefold,

- We introduce Mirage, which enables VLMs to generate interleaved reasoning trajectories that mix latent visual tokens with ordinary text, without relying on external visual decoders.
- Our two-stage training paradigm empowers VLMs to produce stable yet flexible interleaved reasoning and shows that reinforcement learning can further boost performance.
- Mirage achieves consistent gains across diverse multimodal reasoning benchmarks. Further analysis reveals that the latent tokens embody meaningful visual cues, underscoring the potential to unlock deeper multimodal reasoning capabilities in VLMs.

2 Related Work

2.1 Multimodal Chain-of-Thought

Chain-of-Thought (CoT) prompting was first shown to elicit step-by-step reasoning in LLMs (Feng et al., 2023; Zhang et al., 2024a; Wei et al., 2023). Recent extensions of CoT to multimodal settings

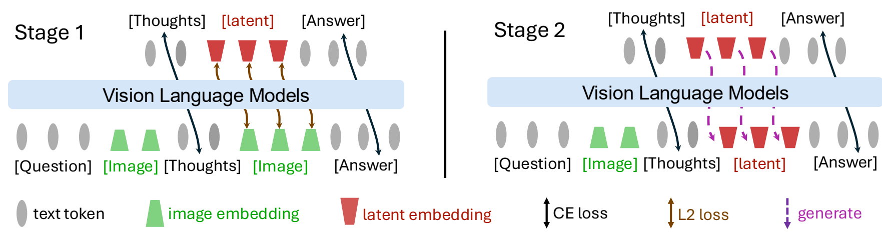

Figure 2: **Pipeline of Mirage Framework.** Stage 1 jointly supervises text and latent visual tokens, grounding the latter in the visual subspace; Stage 2 drops the latent supervision, anchoring the grounded latent tokens for subsequent text generation.

embed visual evidence directly into the reasoning trajectory. ICoT (Zhang et al., 2024b) interleaves attention-selected image crops with text tokens, yielding significant VQA gains, while Visual CoT (Shao et al., 2024a) supplies bounding-box rationales to train VLMs that emit explicit visual tokens. Recent works (Hu et al., 2024; Zhou et al., 2024; Yang et al., 2025c; Gao et al., 2024; Wu et al., 2025; Chern et al., 2025; Fang et al., 2025; Cheng et al., 2025a; Su et al., 2025b; Chen et al., 2025b) further leverage external tools to supply visual cues that enrich multimodal CoT reasoning..

Recent works (Chen et al., 2025a; Wang et al., 2025a) like Chameleon (Team, 2025; Chern et al., 2024) trains a unified token-based model that can emit arbitrary sequences of text and image tokens, but at the cost of large-scale pixel-level supervision and heavier decoding. In the VLM domain, multi-modal reasoning is a promising direction Zhang et al. (2025); Sarch et al. (2025); Xu et al. (2025); Wang et al. (2025b); Chen et al. (2025c); Lee et al. (2025); Zhan et al. (2025); Zheng et al. (2025); Pham & Ngo (2025). Multimodal-CoT Zhang et al. (2023) aligns text and image features to generate auxiliary images that enhance reasoning. MVoT Li et al. (2025a) urther trains a unified model to directly produce image and text interleaving trajectories, but absent of reasoning thoughts. Aurora Bigverdi et al. (2025) introduces an image de-tokenizer to explicitly generate perception tokens, thereby supporting multimodal perception. MMaDA Yang et al. (2025a) adopts a diffusion-based VLM to enable coherent reasoning and generation across modalities. In contrast, our Mirage framework differs by emitting compact latent visual tokens rather than real image patches or pixels, avoiding heavy image generation while still allowing fully interleaved visual–text reasoning.

2.2 Latent Reasoning in LLMs

Much recent work has highlighted the importance of intermediate hidden representations in LLMs (Biran et al., 2024; Yang et al., 2024a). To better guide the latent reasoning process, several approaches Wang et al. (2023); Goyal et al. (2023) introduce specialized tokens into the input sequence.

Another line of work Tan et al. (2025); Geiping et al. (2025); Shen et al. (2025b); Zhu et al. (2025b); Tang et al. (2025); Su et al. (2025a); Shi et al. (2025); Ruan et al. (2025) seeks to internalize reasoning behavior by distilling chain-of-thought rationales into latent representations. Deng et al. (2023) trains models to mimic CoT-style reasoning implicitly through hidden states, and (Deng et al., 2024) further improves inference efficiency by removing explicit intermediate steps altogether. Yu et al. (2024) proposes to distill latent reasoning capabilities into a model by supervising it with data generated for complex reasoning. More recently, Hao et al. (2024) go further by replacing CoT tokens with continuous latent embeddings, enabling unconstrained reasoning in the latent space to explore on complex tasks including math and logical reasoning. While prior work primarily focuses on enhancing efficiency or structural planning within the LLM's latent space, our approach takes a different perspective: we treat latent tokens as a *bridge for exploring visual information*.

3 Multimodal Reasoning with Latent Visual Tokens

Inspired by the cognitive process of mental imagery, we introduce Mirage, a framework that lets VLMs reason in interleaved multimodal trajectories. In contrast to prior unified models that integrate an external image decoder and pre-train on large-scale image generation, our method generates compact latent embeddings that serve as visual tokens. By sidestepping image generation, the model can devote its capacity to reasoning, producing only the essential visual cues and thereby echoing the concise, sketch-like representations humans employ during reasoning.

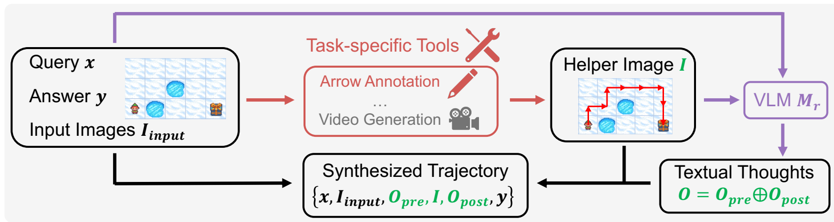

Figure 3: **Data-generation Pipeline.** For each question–answer pair, we first create a helper image with task-specific tools (here, annotate the map with arrows), then prompt a VLM to produce textual reasoning that embeds this image. The text and helper image together form the synthetic multimodal trajectory used for training.

In this section, we first explain how we synthesize informative multimodal reasoning data (Sec. 3.1). Next, we introduce our first joint supervision training stage in Sec. 3.2. Finally, we explain the second stage, which applies text-only supervision while relaxing the latent constraints (Sec. 3.3).

3.1 Data Generation

Consider the multimodal reasoning task where the VLMs need to generate responses y to the input that consists of one or more images and a textual query. For simplicity, we denote the input that contains both image and text as $\boldsymbol{x}$. Given VLMs naturally generating text tokens only, they require additional supervised fine-tuning to learn an interleaved reasoning pattern. We therefore begin by synthesizing a training corpus that pairs each input $\boldsymbol{x}$ with a task-specific helper image I (See Fig. 3). For example, in the navigation task, the helper image can be generated by taking the ground truth action list and manually drawing the corresponding path on the starting map with red arrows. Similarly, for the jigsaw task, we can concatenate the candidate fragments to form a composite image that captures the relationship among pieces. More details on the image generation procedures for different tasks can be found in the supplementary materials. In general, we obtain a help image that delivers precisely the visual cues needed to supervise latent reasoning.

With the helper image I prepared, we next synthesize a reasoning chain where the LLM incorporates the helper image to generate the final solution. Specifically, we first feed a large reasoning VLM M with the original input $\boldsymbol{x}$, the ground-truth answer $\boldsymbol{y}$, and the helper image I and prompt it to generate a step-by-step reasoning that incorporates the helper image. For example, the prompt can be `Generate a step-by-step reasoning that leads to the ground-truth answer while properly incorporating the helper image in reasoning.` Denote the model response as

$$\boldsymbol{o} = M\left(\boldsymbol{x}, \boldsymbol{y}, I\right).$$

Here $\boldsymbol{o}$ is a step-by-step reasoning with the helper image embedded in the reasoning process. Since the helper image is embedded in the reasoning chain, it splits the reasoning chain into two parts. Without loss of generality, we represent $\boldsymbol{o} = \boldsymbol{o}_{\text{pre}} \oplus I \oplus \boldsymbol{o}_{\text{post}}$, where $\oplus$ is the concatenation operation, $\boldsymbol{o}_{\text{pre}}$ is the reasoning chain before the helper image while $\boldsymbol{o}_{\text{post}}$ is the reasoning chain after the helper image. By prompt the large reasoning VLM with different inputs, we can thus collect a training dataset $\mathcal{D} = \{\boldsymbol{x}^{(i)}, I^{(i)}, \boldsymbol{o}^{(i)}, \boldsymbol{y}^{(i)}\}_{i=1}^{N}$, where each $\boldsymbol{o}^{(i)}$ is a synthesized reasoning chain with text and image interleaved.

3.2 Joint Supervision for Latent Grounding

To teach the model an interleaved style of reasoning, one naive solution is to directly train a VLM on the data collected above. However, the effectiveness can be negatively affected by the model's limited capability of synthesizing helper images. Therefore, we propose a novel training strategy: pass the helper images to the VLM first to convert the helper images in the synthetic training data into patch-level features; then fine-tune the VLM to output such features as latent reasoning tokens, thus eliminating the need to generate helper images by the VLM.

Specifically, for each training example $(\boldsymbol{x}, I, \boldsymbol{o}, \boldsymbol{y}) \in \mathcal{D}$, we pass the helper image I through the VLM $f_{\boldsymbol{\theta}}(\cdot)$ with parameter $\boldsymbol{\theta}$ to obtain its patch-level features $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\} = f_{\boldsymbol{\theta}}(I)$. Rather than asking the model to reproduce every patch, we mimic human mental imagery by compressing these features into k salient vectors, $\{\hat{\boldsymbol{e}}_1, \ldots, \hat{\boldsymbol{e}}_k\} = \text{Compress}\big(\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}\big)$, that retain only task-critical visual cues. In this work, we realize $\text{Compress}(\cdot)$ with simple average pooling over the original patch features, a lightweight yet surprisingly effective strategy that supplies a concise visual summary for supervision. We then train our model to (1) generate the response $\boldsymbol{o}_{\text{pre}}$ conditioned on the input $\boldsymbol{x}$, (2) generate the latent tokens $\{\hat{\boldsymbol{e}}_1, \ldots, \hat{\boldsymbol{e}}_k\}$ conditioned on $\boldsymbol{x}$ and $\boldsymbol{o}_{\text{pre}}$, where the last layer hidden states at corresponding positions will be regarded as the generated latent tokens, and (3) generate the response $\boldsymbol{o}_{\text{post}}$ conditioned on the proceeding content.

For the training objective for latent token generation, we adopt the cosine similarity between the last layer hidden states of the model and the target latent tokens:

$$\mathcal{L}_{\text{visual}} \;=\; \ell_{\cos}\Big(\hat{\boldsymbol{e}}_j,\, g_\theta\big(\boldsymbol{o}_{\text{pre}},\, \hat{\boldsymbol{e}}_{1:j-1}\big)\Big), \tag{1}$$

where $g_\theta\big(\boldsymbol{o}_{\text{pre}},\, \hat{\boldsymbol{e}}_{1:j-1}\big)$ denotes the model's prediction for the j-th latent token conditioned on the preceding context. This loss grounds the latent tokens firmly in the visual representation space.

Meanwhile, we train the surrounding textual tokens using the standard cross-entropy loss for next token prediction. For the left segment $\boldsymbol{o}_{\text{pre}}$ the model conditions only on earlier words, whereas for the right segment $\boldsymbol{o}_{\text{post}}$ it also attends to the k compressed visual embeddings.

$$\mathcal{L}_{\text{text}} = \sum_{i=1}^{|\boldsymbol{o}_{\text{pre}}|} \ell_{\text{CE}}\big(\boldsymbol{o}_{\text{pre},i},\, f_\theta(\boldsymbol{x}, \boldsymbol{o}_{\text{pre},<i})\big) + \sum_{i=1}^{|\boldsymbol{o}_{\text{post}}|} \ell_{\text{CE}}\big(\boldsymbol{o}_{\text{post},i},\, f_\theta(\boldsymbol{x}, \boldsymbol{o}_{\text{pre}}, \{\hat{\boldsymbol{e}}_j\}_1^k, \boldsymbol{o}_{\text{post},<i}\big). \tag{2}$$

Here $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ denotes the next token prediction probability conditioned on the input and $\{\hat{\boldsymbol{e}}_j\}_1^k$ is the set of the ground-truth latent tokens. The overall training objective in this stage combines this term with the visual-alignment loss $\mathcal{L}_1 = \mathcal{L}_{\text{visual}} + \gamma\mathcal{L}_{\text{text}}$, where the γ is the loss coefficient, thereby anchoring the latent tokens in visual space while teaching the model to weave them naturally into its textual thoughts.

3.3 Text-Only Supervision with Latent Relaxation

The first stage grounds the latent tokens by forcing the model to reconstruct the compressed image embeddings. Although effective for visual alignment, this can over-constrain the model, diverting capacity from its primary goal of answering the question correctly, degrading the reasoning performance. Therefore, in the second stage, we remove the cosine loss altogether and keep only the cross-entropy loss over text tokens.

Although the latent tokens no longer carry an explicit loss, we still anchor them so that they meaningfully guide the following thoughts. For each training instance, the model autoregressively produces its own latent tokens $\{\boldsymbol{e}_i\}_{i=1}^k$, with

$$\boldsymbol{e}_j = f_{\boldsymbol{\theta}}\big(\boldsymbol{x}, \boldsymbol{o}_{\text{pre}},\, \boldsymbol{e}_{<j}\big). \tag{3}$$

These self-generated embeddings replace the compressed image vectors used in Stage 1 and serve as priors for the tokens that follow the image placeholder. Therefore, the training objective becomes

$$\mathcal{L}_{\text{text}} = \sum_{i=1}^{|\boldsymbol{o}_{\text{pre}}|} \ell_{\text{CE}}\big(\boldsymbol{o}_{\text{pre},i},\, f_\theta(\boldsymbol{x}, \boldsymbol{o}_{\text{pre},<i})\big) + \sum_{i=1}^{|\boldsymbol{o}_{\text{post}}|} \ell_{\text{CE}}\big(\boldsymbol{o}_{\text{post},i},\, f_\theta(\boldsymbol{x}, \boldsymbol{o}_{\text{pre}}, \{\boldsymbol{e}_j\}_1^k, \boldsymbol{o}_{\text{post},<i}\big). \tag{4}$$

Due to the continuous property of $\{\boldsymbol{e}_i\}_{i=1}^k$, these self-generated latent tokens are fully differentiable. Since the next token prediction of $\boldsymbol{o}_{\text{post}}$ is a function of the latent tokens, the gradient can be propagated to these latent tokens when minimizing the above loss on the textual tokens. This allows us to optimize the generation of the latent tokens within the learned visual subspace, acting as flexible priors that guide subsequent text generation and yield a more adaptive, task-focused reasoning.

The overall framework of our two-stage pipeline is shown in Fig. 2. Two stages jointly endow VLMs with the ability to generate interleaved multimodal reasoning with latent visual tokens. Empirical results in Sec. 4.2 further validate the effectiveness of our latent reasoning over text-only decoding.

3.4 Reinforcement Learning

After the two supervised fine-tuning stages, the model has already learned to reason using both interleaved text and latent tokens. Here, we further explore whether the model's performance can be improved using reinforcement learning (RL), inspired by recent long-CoT language models (Xie et al., 2025; Shen et al., 2025a). Specifically, we adopt group relative policy optimization (GRPO) (Shao et al., 2024b) for RL training. For each input query in the training set, we sample multiple responses from the model. During RL, we explicitly optimize the probabilities of textual tokens while allowing gradients to flow through the latent tokens. Following LMM-R1 (Peng et al., 2025), we adopt two types of rewards: accuracy and format. We consider both accuracy and format rewards. More implementation details are provided in Appx D.1.

Table 1: **Results on Visual-Spatial Planning (VSP) tasks.** Mirage outperforms text-only baselines and achieves superior performance compared to interleave reasoning models.

VSP	Spatial Reasoning					Spatial Planning				
	Level 3	Level 4	Level 5	Level 6	Avg.	Level 3	Level 4	Level 5	Level 6	Avg.
Zero-Shot	0.32	0.23	0.40	0.32	0.32	0.10	0.08	0.05	0.01	0.06
Direct SFT	0.83	0.81	0.85	0.86	0.83	0.88	0.81	0.73	0.47	0.72
CoT SFT	0.88	0.86	0.80	0.83	0.84	0.68	0.53	0.35	0.31	0.47
GRPO	0.54	0.49	0.76	0.67	0.62	0.42	0.35	0.26	0.08	0.28
CoT SFT + GRPO	0.89	0.85	0.84	0.80	0.85	0.65	0.58	0.43	0.38	0.51
Anole	0.46	0.51	0.49	0.63	0.52	0.02	0.01	0.00	0.00	0.01
MVoT	0.53	0.64	0.67	0.59	0.61	0.21	0.11	0.08	0.03	0.11
Aurora	0.64	0.78	0.71	0.71	0.71	0.26	0.11	0.11	0.04	0.13
Ours (Direct)	0.86	0.84	**0.88**	0.87	0.86	**0.93**	**0.83**	**0.76**	**0.51**	**0.76**
Ours (CoT)	0.87	**0.92**	0.86	0.84	0.87	0.75	0.63	0.53	0.39	0.58
+ w/ GRPO	**0.92**	0.90	0.86	**0.88**	**0.89**	0.78	0.65	0.52	0.43	0.60

4 Experiments

4.1 Experimental Settings

Benchmarks. We evaluate our approach on four different benchmarks. **VSP** (Wu et al., 2024) measures spatial planning in a simulated maze-navigation environment. In addition to its main task, we adopt its spatial reasoning subtask, which asks the model to predict the outcome of a prescribed action sequence. **BLINK-Jigsaw** (Fu et al., 2024) systematically evaluates the capacity of multimodal large language models to extrapolate global structural and semantic information from incomplete visual inputs, thereby assessing their reasoning about spatial organization and maintaining perceptual coherence. **SAT** (Ray et al., 2024) evaluates both static and dynamic spatial relations. Additionally, we include the Mathematical Geometry subset of the recent **COMT-Geometry** (Cheng et al., 2025b) to assess formal spatial reasoning in mathematical contexts. Details are provided in Appx C.

Data Synthesis. For each task, we sample 1k training instances for fine-tuning and 2k instances for reinforcement learning. COMT provides interleaved multimodal reasoning trajectories, which we directly use as both helper images and reasoning supervision. For the other benchmarks, we synthesize data following the procedure outlined in Sec. 3.1. For VSP, the helper image is either the start map annotated with the red-arrow path or the agent's current state snapshot. In Jigsaw, we concatenate one candidate patch beside the reference image. For SAT, following MindJourney Yang et al. (2025b), we prompt fine-tuned `CogVideoX-5B` (Yang et al., 2024b) to render a scene that matches the textual description. Full synthesis details are provided in Appx C.

Baseline Models. First, we fine-tune the model directly with answer labels and evaluate zero-shot reinforcement learning without any supervised warm-up. Next, using synthetic data, we perform CoT fine-tuning (CoT SFT) and then add reinforcement learning. In addition, we benchmark against a unified model **Anole** (Chern et al., 2024), training with the same multimodal supervision, and **MVoT** (Li et al., 2025a), which generates action and state images. We also compare against reasoning models that interleave either generated or processed images into reasoning trajectories, including Aurora Bigverdi et al. (2025), ViGoRL Sarch et al. (2025), MindJourney Yang et al. (2025b), and MINT-CoT Chen et al. (2025b). For more baseline details, please refer to Appx D.2.

Implementation Details. Unless stated otherwise, we use `Qwen2.5-VL 7B` as the base model, and we use a latent token size of $k = 4$ and a loss coefficient of $\gamma = 0.1$. The random seed is fixed at 42 to ensure reproducibility. For more implementation details, please refer to Appx D.1.

4.2 Experimental Results

We first evaluate the effectiveness of our method on the VSP benchmark. The results are shown in Tab. 1. First, adding latent visual tokens to the reasoning process significantly improves the reasoning capability of VLMs compared to text-only baselines. Compared to directly fine-tuning the VLM with the synthesized data, our method achieves 3% higher accuracy on the spatial reasoning task and 11% on the spatial planning task. Also, Mirage improves the CoT SFT + GRPO, by 2% and 7%, respectively, demonstrating the effectiveness of our two-stage training method. Moreover,

Table 2: **Experimental Results with Qwen2.5-VL 3B on Jigsaw and SAT tasks.** For each baseline, we report the best performance across zero-shot and fine-tuning settings, with and without CoT. Baseline implementations are described in Sec. D.2, and detailed results are provided in Sec. E.1.

Method	Jigsaw	SAT Synthetic			SAT Real
		GoalAim	ObjM	Avg.	
Zero-Shot	0.45	0.50	0.38	0.44	0.51
Direct SFT	0.80	0.82	0.83	0.83	0.55
CoT SFT	0.59	0.73	0.88	0.71	0.54
GRPO	0.54	0.78	0.80	0.79	0.54
SFT + GRPO	0.72	0.82	0.85	0.84	0.52
ViGoRL	0.56	0.75	0.58	0.67	0.59
MindJourney	-	0.84	0.62	0.73	0.57
Ours	**0.85**	**0.85**	**0.93**	**0.89**	**0.64**

reinforcement learning can further improve the performance of our method. By weaving latent visual tokens within the text trajectories, instead of placing them at the start, our model can naturally explore diverse sequences. After optimizing with GRPO, Mirage achieves extra gains (+2% accuracy) on VSP tasks. These results further confirm that interleaved latent cues provide informative guidance with flexible reasoning, highlighting the potential of our framework.

Additionally, baselines such as Aurora and Anole, despite explicitly generating image tokens, perform poorly on interleaved text–image reasoning. After fine-tuning on the same data, they reach only 71% accuracy on the spatial reasoning task and 13% on the spatial planning task. We attribute this to the overhead of explicit image generation. Aurora produces fixed image tokens per image, hindering reasoning performance. Notably, Anole struggles to even generate valid answers for the spatial planning task post fine-tuning. Following Li et al. (2025a), we construct interleaved trajectories by combining textual thoughts with simulated state images. Our reproduced results are lower than those reported in their paper, likely due to training data differences. To ensure a fair comparison, we reproduce MVoT results using same data as used in our framework. Nonetheless, our model still outperforms their reported results, highlighting the advantage of our latent design.

We also evaluate our model on COMT tasks in Tab. 3. Mirage achieves about 5% accuracy than best baselines. Furthermore, we report results on Jigsaw and SAT tasks in Tab. 2 using the `Qwen2.5-VL 3B`, showing that Mirage transfers effectively to smaller models, with performance improvements consistent with 7B models. Across both benchmarks, our model outperforms all relevant text and image interleaved reasoning baselines, including MINT-CoT, which is explicitly trained on large-scale math data, and ViGoRL, which is trained on large-scale spatial reasoning data. Consistent improvements underscore that interleaving compact visual cues consistently strengthens reasoning ability. More detailed results are provided in Appx E.1.

Table 3: **Results on COMT.**

Method	COMT
Zero-Shot	0.63
Direct SFT	0.71
CoT SFT	0.74
GRPO	0.69
SFT + GRPO	0.72
R1-VL	0.69
MINT-CoT	0.72
Ours	**0.77**

We notice that on VSP spatial planning task, fine-tuning with synthesized reasoning thoughts performs significantly worse than training directly on answer labels, both with and without our latent design. Two factors likely contribute to this outcome. First, as noted in prior work (Li et al., 2025b), certain visual tasks that rely heavily on perception may not benefit from explicit reasoning during fine-tuning. Second, the synthesized thoughts are generated by `Qwen2.5-VL-32B`; although generally sound, they are not flawless, and any imperfections propagate into the base model. Likely, in SAT, the helper images are produced by a video generation model without ground-truth annotations, which can introduce further noise to the latent prior. Despite these challenges, our latent reasoning pipeline still closes much of the performance gap, highlighting its practical robustness.

4.3 Ablation Study

In this section, we first conduct an ablation study to evaluate the influence of the two stages of our framework. Tab. 4 reports the effect of removing each phase. Training with only the first phase,

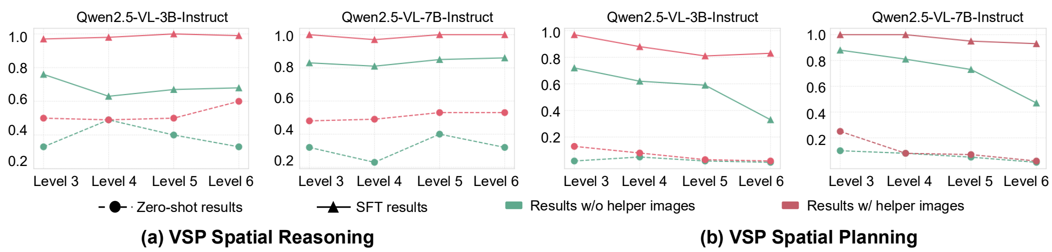

Figure 4: **Performance with Helper Images as Input Priors.** We evaluate model accuracy using synthesized helper images under both zero-shot and fine-tuned settings. The results highlight the informativeness of the generated images and confirm their high data quality.

which jointly supervises text and latent visual tokens, anchors the latent embeddings but leaves them constrained and lowers performance, similar to the plight of unified models.

Table 4: **Ablation study of training stages on VSP Spatial Planning.** Both stages jointly improve reasoning performance.

	VSP Spatial Planning				
Method	3	4	5	6	Avg.
Ours	0.75	0.63	0.53	0.39	0.58
– w/o Stage 1	0.69	0.58	0.46	0.36	0.52
– w/o Stage 2	0.38	0.19	0.16	0.09	0.21

Training with only the second stage, which relies on text loss alone while letting latent tokens evolve freely, performs slightly better the text-only baseline. Without the grounding supplied by the first stage, the latent vectors drift into regions of the multimodal embedding space that do not aid reasoning. This outcome contrasts with findings on LLMs in Coconut (Hao et al., 2024), where unsupervised latent vectors can benefit subsequent reasoning. The difference indicates that visual and textual subspaces in VLMs remain heterogeneous enough that a grounding phase is effective. This confirms that the first stage aligns latent tokens with visual features, the second stage allows them to adapt, and both steps are necessary.

Table 5: **Ablation study of latent size k and loss coefficient γ on VSP Spatial Reasoning.** The pipeline remains robust across hyperparameters.

k	γ	VSP Spatial Reasoning				
		3	4	5	6	Avg.
2	0.1	0.85	0.86	0.89	**0.93**	0.86
4	0.1	**0.87**	**0.92**	0.86	0.84	0.87
6	0.1	0.85	0.90	**0.91**	0.87	**0.88**
8	0.1	0.77	0.77	0.74	0.70	0.75
4	0.5	0.84	0.91	0.84	0.78	0.84
4	1.0	0.77	0.85	0.85	0.87	0.83

To delve deeper into the robustness of our framework, we investigate the influence of hyperparameters: latent token size k and the multimodal loss coefficient γ. As Tab. 5 shows, adjusting the loss coefficient γ has a moderate effect. A larger γ weights the latent-alignment loss less in the first stage. When γ approaches infinity, the first stage becomes equivalent to skipping visual supervision entirely, in other words, the second stage. This gives a poor initialization for subsequent training. Even so, after the second stage, each γ tested still obtains over 80% accuracy, which attests to the overall robustness of the framework. Additionally, we observe that varying the latent size k from 2 to 6 yields consistently strong performance, with $k = 6$ showing a slight improvement—highlighting the resilience of our latent design. However, increasing k to 8 results in a significant performance drop around 13%, likely due to error accumulation in longer latent sequences under autoregressive non-decoding generation. These observations are consistent with prior findings that optimal latent reasoning performance in LLMs typically occurs with fewer than 6 latent tokens (Hao et al., 2024).

5 Analysis

Synthesized Data Quality. Data quality plays a critical role in model performance. In this section, we investigate whether the helper images generated by various tools are genuinely informative for VLM reasoning. For the two VSP tasks, we supply the helper image as prior input and evaluate model performance in both zero-shot and fine-tuned settings. As shown in Fig. 4, providing the helper image leads both models to achieve nearly 100% accuracy on both tasks. Even in the zero-shot setting, we observe substantial performance gains on the spatial reasoning task. However, improvements on the spatial planning task are limited to simpler map layouts in the zero-shot setting. We attribute this to the inherent difficulty of extracting and leveraging spatial information from the helper image without

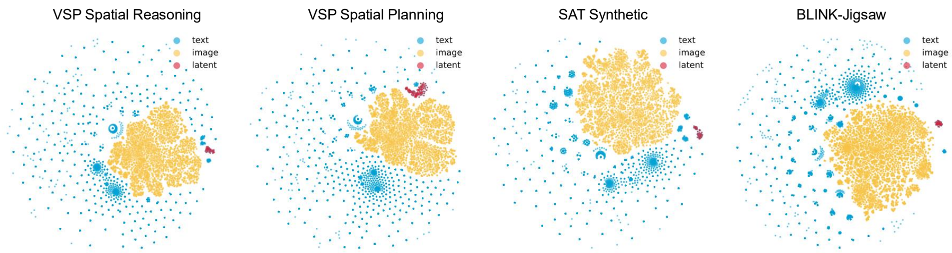

Figure 5: **Visualization of Latent Embeddings.** We visualize our latent tokens along with text and image embeddings with t-SNE. Latent tokens cluster near, yet just outside, the visual representation subspace, consistent with the two-stage training design.

task-specific fine-tuning. These results suggest that the synthesized helper images do indeed enhance VLM reasoning. Moreover, if the model's latent thoughts can fully internalize the information encoded in these images, it would represent a strong performance upper bound for our Mirage.

Latent Behavior Analysis. The model learns to reproduce compressed image embeddings in the first stage, anchoring its latent tokens in the visual subspace. However, after the second stage, these latent tokens receive no direct supervision. Therefore, it is unclear whether they still encode visual representations. By sampling 100 examples, we obtain the corresponding latent token vectors alongside the text and visual embeddings. We use t-SNE to embed all vectors into two dimensions for better visualization with a perplexity of 30, and initialize the embeddings via PCA. As shown in Fig. 5, the text embeddings (blue dots) fill the entire plot in a radial scattering pattern, while the visual token embeddings (yellow dots) cluster tightly inside a distinct visual subspace, consistent with previous findings. Our latent embeddings (red dots) form a compact cloud that sits just outside that visual cluster, shifted by the second training stage, which tailors the latent embeddings to answer generation. However, we notice that our latent tokens remain clearly separated from the text distribution and closer to the visual embedding across diverse tasks. This pattern shows that even without an explicit decoder, the latent tokens stay close to the visual manifold while retaining the flexibility introduced in the second stage, echoing the way mental imagery abstracts rather than reproduces visual input.

Inference Time Efficiency. During inference, our framework forwards latent embeddings directly, bypassing decoding and thereby reducing computation. The only additional overhead arises from latent mode checking and, in some cases, slightly longer reasoning sequences. We compared inference speed against text-only 7B baselines on both VSP tasks. Our framework matches GPU time usage and is slightly faster on the VSP spatial planning task. These results confirm that our approach introduces no extra inference cost while maintaining efficiency. For detailed training and inference time comparisons, see Appx. E.3.

Multiple Latent Reasoning Steps. Beyond using a single helper image, we investigate the case where multiple steps of latent tokens appear in the reasoning trajectory. Interestingly, models trained with only one helper image naturally learn to produce multiple latent visual reasoning steps occasionally during inference on structured, multi-step tasks such as VSP, reflecting the flexibility of our framework. We further extend experiments by explicitly adding two helper images to the trajectory. According to the results in Appx. E.2, 3B models achieve an average 7% improvement on the VSP spatial reasoning task, suggesting that additional visual support can further benefit.

6 Conclusion

In this work, mimicking human mental imagery, we propose Mirage, a lightweight framework that interleaves compact latent visual tokens with text so a vision–language model can reason multimodally without ever generating pixel-level images. Specifically, our framework is trained in two stages: a joint supervision stage that anchors latent tokens to visual embeddings while learning the surrounding text, followed by a text-only supervision stage that lets those tokens adapt freely to support answer generation. A brief reinforcement-learning refinement further aligns the entire trajectory with task goals. Across four spatial-reasoning benchmarks, Mirage consistently outperforms text-only baselines, underscoring the effectiveness and potential of latent visual reasoning for multimodal models.

Ethics Statement This work does not involve human subjects, biased, private or sensitive data. All datasets used are publicly available and have established research usage licenses. This work complies with the general ethical principles of ICLR, and the authors confirm that the research has been conducted responsibly.

Reproducibility Statement We present all training and inference details clearly in the manuscript. Full model architectures and training procedures are provided in Section 4.1. Hyperparameters and evaluation protocols are documented in Appx. D.1. Dataset curation and detailed prompt usage are described in Appx. C. All reported results were obtained under the same environment and with a fixed random seed to ensure fairness and reproducibility. We plan to make the full code and datasets public upon acceptance.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3836–3845, 2025.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. *arXiv preprint arXiv:2406.12775*, 2024.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. *arXiv preprint arXiv:2506.05331*, 2025b.

Zhangquan Chen, Xufang Luo, and Dongsheng Li. Visrl: Intention-driven visual perception via reinforced reasoning. *arXiv preprint arXiv:2503.07523*, 2025c.

Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi Wang, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, et al. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *arXiv preprint arXiv:2505.15510*, 2025a.

Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23678–23686, 2025b.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024.

Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with generated images. *arXiv preprint arXiv:2505.22525*, 2025.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023.

Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*, 2024.

Irving Fang, Juexiao Zhang, Shengbang Tong, and Chen Feng. From intention to execution: Probing the generalization boundaries of vision-language-action models. *arXiv preprint arXiv:2506.09930*, 2025.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36:70757–70798, 2023.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, et al. Cantor: Inspiring multimodal chain-of-thought of mllm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9096–9105, 2024.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*, 2025.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. *arXiv preprint arXiv:2310.02226*, 2023.

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

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*, 2024.

Stephen M Kosslyn. *Image and brain: The resolution of the imagery debate*. MIT press, 1996.

Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025a.

Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, and Kaipeng Zhang. Think or not think: A study of explicit thinking in rule-based visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.16188*, 2025b.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

Tan-Hanh Pham and Chris Ngo. Multimodal chain of continuous thought for latent-space reasoning in vision-language models. *arXiv preprint arXiv:2508.12587*, 2025.

Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, Kuo-Hao Zeng, et al. Sat: Spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Yangjun Ruan, Neil Band, Chris J Maddison, and Tatsunori Hashimoto. Reasoning to learn from latent thoughts. *arXiv preprint arXiv:2503.18866*, 2025.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. *arXiv preprint arXiv:2505.23678*, 2025.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning, 2024a. URL `https://arxiv.org/abs/2403.16999`.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024b.

Maohao Shen, Guangtao Zeng, Zhenting Qi, Zhang-Wei Hong, Zhenfang Chen, Wei Lu, Gregory Wornell, Subhro Das, David Cox, and Chuang Gan. Satori: Reinforcement learning with chain-of-action-thought enhances llm reasoning via autoregressive search. *arXiv preprint arXiv:2502.02508*, 2025a.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*, 2025b.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171 (3972):701–703, 1971.

Teng Shi, Weicong Qin, Weijie Yu, Xiao Zhang, Ming He, Jianping Fan, and Jun Xu. Bridging search and recommendation through latent cross reasoning. *arXiv preprint arXiv:2508.04152*, 2025.

DiJia Su, Hanlin Zhu, Yingchen Xu, Jiantao Jiao, Yuandong Tian, and Qinqing Zheng. Token assorted: Mixing latent and text tokens for improved language model reasoning. *arXiv preprint arXiv:2502.03275*, 2025a.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025b.

Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Jian Luan, and Ruihua Song. Think silently, think fast: Dynamic latent compression of llm reasoning chains. *arXiv preprint arXiv:2505.16552*, 2025.

Jiakai Tang, Sunhao Dai, Teng Shi, Jun Xu, Xu Chen, Wen Chen, Jian Wu, and Yuning Jiang. Think before recommend: Unleashing the latent reasoning power for sequential recommendation. *arXiv preprint arXiv:2503.22675*, 2025.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2025. URL `https://arxiv.org/abs/2405.09818`.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024.

Dianyi Wang, Wei Song, Yikun Wang, Siyuan Wang, Kaicheng Yu, Zhongyu Wei, and Jiaqi Wang. Autoregressive semantic visual reconstruction helps vlms understand better. *arXiv preprint arXiv:2506.09040*, 2025a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Xinyi Wang, Lucas Caccia, Oleksiy Ostapenko, Xingdi Yuan, William Yang Wang, and Alessandro Sordoni. Guiding language model reasoning with planning tokens. *arXiv preprint arXiv:2310.05707*, 2023.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025.

Qiucheng Wu, Handong Zhao, Michael Saxon, Trung Bui, William Yang Wang, Yang Zhang, and Shiyu Chang. Vsp: Assessing the dual challenges of perception and reasoning in spatial planning tasks for vlms. *arXiv preprint arXiv:2407.01863*, 2024.

Zhihui Xie, Liyu Chen, Weichao Mao, Jingjing Xu, Lingpeng Kong, et al. Teaching language models to critique via reinforcement learning. *arXiv preprint arXiv:2502.03492*, 2025.

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. *arXiv preprint arXiv:2505.11409*, 2025.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025a.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? *arXiv preprint arXiv:2402.16837*, 2024a.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. Mindjourney: Test-time scaling with world models for spatial reasoning. *arXiv preprint arXiv:2507.12508*, 2025b.

Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3d-mem: 3d scene memory for embodied exploration and reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17294–17303, 2025c.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024b.

Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling system 2 into system 1. *arXiv preprint arXiv:2407.06023*, 2024.

Yufei Zhan, Ziheng Wu, Yousong Zhu, Rongkun Xue, Ruipu Luo, Zhenghao Chen, Can Zhang, Yifan Li, Zhentao He, Zheming Yang, et al. Gthinker: Towards general multimodal reasoning via cue-guided rethinking. *arXiv preprint arXiv:2506.01078*, 2025.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Advances in Neural Information Processing Systems*, 37:333–356, 2024a.

Yizhen Zhang, Yang Ding, Shuoshuo Zhang, Xinchen Zhang, Haoling Li, Zhong-zhi Li, Peijie Wang, Jie Wu, Lei Ji, Yelong Shen, et al. Perl: Permutation-enhanced reinforcement learning for interleaved vision-language reasoning. *arXiv preprint arXiv:2506.14907*, 2025.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024b. URL https://arxiv.org/abs/2302.00923.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. Image-of-thought prompting for visual reasoning refinement in multimodal large language models. *arXiv preprint arXiv:2405.13872*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a.

Rui-Jie Zhu, Tianhao Peng, Tianhao Cheng, Xingwei Qu, Jinfa Huang, Dawei Zhu, Hao Wang, Kaiwen Xue, Xuanliang Zhang, Yong Shan, et al. A survey on latent reasoning. *arXiv preprint arXiv:2507.06203*, 2025b.