REASONING WITHIN THE MIND: DYNAMIC MULTIMODAL INTERLEAVING IN LATENT SPACE

Anonymous authors
Paper under double-blind review

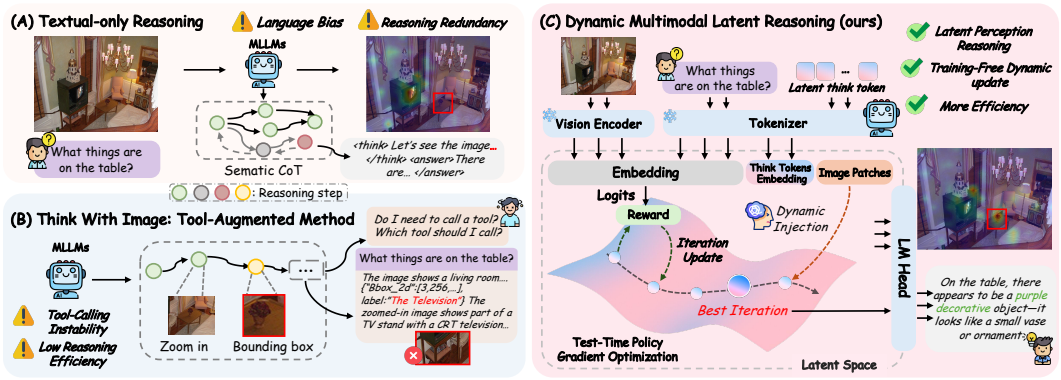

Figure 1: Comparison between DMLR and two reasoning paradigms. (A) Text-only reasoning: relies solely on explicit CoT, often causing visual grounding errors and redundant steps. (B) Think-with-Image reasoning: depends on external perception tools, leading to unstable tool calls and extra overhead. (C) DMLR (ours): refines latent think tokens in the latent space through confidence-guided optimization and dynamically injects visual information, achieving self-improving reasoning without additional training while maintaining high efficiency.

ABSTRACT

Recent advancements in Multimodal Large Language Models (MLLMs) have significantly enhanced cross-modal understanding and reasoning by incorporating Chain-of-Thought (CoT) reasoning in the semantic space. Building upon this, recent studies extend the CoT mechanism to the visual modality, enabling models to integrate visual information during reasoning through external tools or explicit image generation. However, these methods remain dependent on explicit step-by-step reasoning, unstable perception–reasoning interaction and notable computational overhead. Inspired by human cognition, we posit that thinking unfolds not linearly but through the dynamic interleaving of reasoning and perception within the mind. Motivated by this perspective, we propose **DMLR**, a test-time **Dynamic Multimodal Latent Reasoning** framework that employs confidence-guided latent policy gradient optimization to refine latent think tokens for in-depth reasoning. Furthermore, a Dynamic Visual Injection Strategy is introduced, which retrieves the most relevant visual features at each latent think token and updates the set of best visual patches. The updated patches are then injected into latent think token to achieve dynamic visual–textual interleaving. Experiments across seven multimodal reasoning benchmarks and various model architectures demonstrate that DMLR significantly improves reasoning and perception performance while maintaining high inference efficiency.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) (Bai et al., 2025a; Wang et al., 2025a; Team et al., 2025; Li et al., 2024) have achieved remarkable breakthroughs in integrating visual and linguistic information. This progress has facilitated the incorporation of Chain-of-Thought (CoT) reasoning into multimodal tasks, enabling models to construct structured reasoning paths across visual and

textual modalities. Current multimodal reasoning approaches can be broadly categorized into three types: *(1) Textual-only Reasoning* (Mondal et al., 2024; Su et al., 2025a; Huang et al., 2025a), which generates intermediate reasoning steps in the sematic space. Such methods explicitly express reasoning logic through language generation but often suffer from language bias and insufficient visual grounding, as shown in Figure. 1(a). *(2) Think with Images* attempts to directly manipulate or augment images during reasoning, such as local zooming (Su et al., 2025b; Zheng et al., 2025), region highlighting (Fu et al., 2025; Fan et al., 2025), or generating intermediate reasoning steps via diffusion models (Li et al., 2025a; Zhang et al., 2025a) to enhance visual alignment. Despite their effectiveness in improving reasoning to a certain extent, they still face challenges such as unstable tool invocation and high inference overhead, as reflected in Figure 1(b). Recently, latent-space reasoning has emerged as a promising paradigm for enhancing reasoning capabilities in large language models, as exemplified by approaches such as CoCoNut (Hao et al., 2025) and LatentSeek (Li et al., 2025b). Its core idea is to perform implicit reasoning in the latent space, replacing explicit textual steps with latent vectors to reduce redundant generation and capture more compact information. However, recent studies (Li et al., 2025c; Yang et al., 2025a; Pham and Ngo, 2025; Zhang et al., 2025b) still rely on extra training to enforce latent reasoning triggered at fixed positions (via special tokens). This rigidity prevents the model from adaptively allocating reasoning effort.

Inspired by human cognition, we argue that reasoning is not fixed. Instead, humans dynamically revisit visual information, specifically when they encounter uncertainty. Drawing on this intuition, we empirically analyze the interplay between the model's visual reliance and its internal confidence. Our analysis reveals two key phenomena: ***(i)*** *Visual information is used only at a few specific stages of the reasoning process rather than at fixed positions*, and ***(ii)*** *Internal confidence serves as a natural indicator for the need of visual grounding as it strongly correlates with reasoning correctness*. These findings suggest that effective multimodal reasoning relies on dynamic visual usage guided by internal confidence.

In light of these observations, we propose **DMLR**, a Test-time **D**ynamic **M**ultimodal **L**atent **R**easoning Framework, as shown in Figure 1(c). Specifically, it introduces optimizable latent think tokens to serve as a mental draft, which are iteratively refined through confidence-guided policy gradient updates. Crucially, we design a confidence-driven dynamic visual injection strategy. At each step, the model autonomously determines whether to revisit visual information and which contents to select (ranging from none to a few specific patches). This mechanism allows the model to naturally skip visual injection when internal confidence is sufficient, or actively integrate targeted visual clues when necessary, all driven by the objective of maximizing reasoning confidence, effectively mimicking the human cognitive process of checking visual clues to build confidence. After several iterations, the optimized latent tokens are decoded with the input without extra inference cost. Extensive experiments demonstrate that DMLR consistently outperforms existing methods across diverse architectures and tasks while maintaining high efficiency. The main contributions can be summarized as follows:

- ❶ We reveal two key phenomena: Visual information contributes only at specific reasoning steps; and confidence reflects both reasoning quality and visual grounding.
- ❷ We propose DMLR, a test-time framework for multimodal latent reasoning that integrates confidence-guided latent optimization with dynamic visual injection.
- ❸ Extensive evaluations show that DMLR consistently outperforms other methods across diverse architectures and multimodal tasks, while maintaining high efficiency.

2 Related Work

Explicit Reasoning. Many prior works have explored visual reasoning. Early approaches mainly relied on semantic CoT, where the model performs all inference in the text space after a one-time visual encoding (Mondal et al., 2024; Li et al., 2024; Chen et al., 2025a; Wang et al., 2025b). However, this separation between perception and reasoning often leads to misalignment and hallucination (Su et al., 2025a; Liu et al., 2025a;b; Tong et al., 2024; Liu et al., 2025a; Tang et al., 2025). To address these limitations, recent studies adopt a Thinking-with-Images paradigm, where the model can draw auxiliary elements (Zhang et al., 2025a; Su et al., 2025c; Hu et al., 2024), zoom or crop regions (Fan et al., 2025; Su et al., 2025b; Hong et al., 2025; Zhang et al., 2025c), or generate intermediate visual

cues (Zou et al., 2025; Fu et al., 2025; Zheng et al., 2025), enabling it to reason directly over visual structures.

Latent Reasoning. Recently, an increasing number of studies have begun to shift reasoning from the explicit token space to the model's latent representation space. Some methods introduce dedicated training frameworks that optimize latent representations to support more effective internal reasoning (Liu et al., 2025c; Hao et al., 2025; Huang et al., 2025b; Mi et al., 2025; Deng et al., 2025; Huang et al., 2025c), while others propose training-free approaches that manipulate latent activations during inference to refine the reasoning process (Li et al., 2025b; Zhang et al., 2025d; Butt et al., 2025; Ye et al., 2025; Li et al., 2025d). In addition, several recent works explore injecting visual information into the latent space (Li et al., 2025c; Yang et al., 2025a; Sun et al., 2025; Gao et al., 2025; Pham and Ngo, 2025), enabling models to iteratively operate over both latent semantic features and latent visual cues, thereby supporting a more flexible form of interleaved multimodal reasoning.

3 Preliminary and Motivation

As shown in Figure 12, existing reasoning paradigms commonly suffer from insufficient visual grounding, unstable tool invocation, and high computational overhead. These limitations motivate a fundamental question: why can't MLLMs reason like humans do, dynamically deciding how to reason and which visual information to pay attention on during the thinking process? To this end, we organize the section around two research questions: ***(RQ1)*** *Whether multimodal models require visual perception at every step of reasoning?* ***(RQ2)*** *If not, can their internal representations indicate when visual perception and reasoning is required?*

3.1 Dynamic Perception-Reasoning is Necessary

Definition 3.1 ***(Visual Dependency Score).*** *Let the visual input be denoted as I, and its perturbed version as $\tilde{I}$. Given a query q, the model's dependence on visual information can be quantified by measuring the output discrepancy between the original and perturbed visual inputs. Specifically, for the i-th generated sequence $\mathcal{X}_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,t}\}$, the visual dependency score at position t is defined as:*

$$S_{i,t} = \log \pi_\theta(x_{i,t} \mid x_{i,<t}, I, q) - \log \pi_\theta(x_{i,t} \mid x_{i,<t}, \tilde{I}, q) \tag{1}$$

where $\pi_\theta(\cdot)$ denotes the token-level conditional probability distribution of the model. A larger $S_{i,t}$ indicates a stronger dependency of the generated token on visual information. Building upon the above metric, we analyze visual dependency on the MathVision benchmark using the Qwen2.5-VL-7B (Bai et al., 2025b) at two levels. First, for individual reasoning chains, we compute token-level visual dependency scores, capturing how much each generated token relies on visual information, as illustrated in Figure 2(a). Second, as shown in Figure 2(b), we aggregate these scores across full reasoning trajectories to obtain chain-level visual dependency, which reveals how different reasoning paths vary in their reliance on visual perception. These results reveal that:

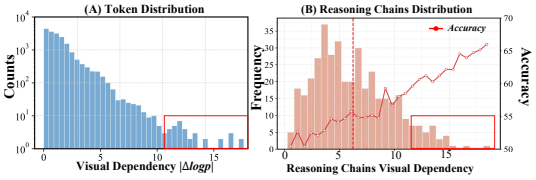

Figure 2: Analysis of visual dependency in reasoning. (A) Token-level distribution shows visual sensitivity is concentrated in a few tokens. (B) Chain-level distribution reveals large variation in visual reliance across reasoning trajectories.

✦ ***Takeaway 1.*** *The dependency on visual input across the reasoning process is highly uneven: only a small subset of tokens show strong sensitivity to visual features, while the majority operate independently of the image.*

✦ ***Takeaway 2.*** *Across reasoning chains sampled from the same model, visual dependency varies substantially. Chains exhibiting stronger visual reliance consistently yield higher accuracy.*

3.2 INTERNAL CONFIDENCE AFFECTS MULTIMODAL REASONING

Definition 3.2 ***(Confidence Gain).*** *Let I denote the visual input, q the query, and $\mathcal{T}_t$ denotes the reasoning at step t. The Confidence Gain at step t is defined as the change in the probability of the ground-truth answer Y_{gt} after adding step x_t. A positive $\mathcal{G}_t$ suggests that step x_t strengthens the confidence, whereas a negative value indicates the opposite.*

$$\mathcal{G}_t = \log \pi_\theta(Y_{gt} \mid I, q, \mathcal{T}_{\leq t}) - \log \pi_\theta(Y_{gt} \mid I, q, \mathcal{T}_{<t}) \tag{2}$$

❖ ***Observation 1: Higher Confidence Tends to Indicate Higher Reasoning Accuracy.*** We analyze reasoning chains generated by various reasoning models across four benchmarks, where all chains are partitioned into a correct set $\mathcal{T}^+$ and an incorrect set $\mathcal{T}^-$ based on their answer correctness. We then compute the proportion of reasoning steps for each chain that obtain a positive confidence reward. As shown in Figure 3(a), reasoning chains in $\mathcal{T}^+$ exhibit a substantially higher proportion of positive confidence increments compared to those in $\mathcal{T}^-$, *indicating that the reasoning leading to correct answers tends to exhibit more stable and higher confidence.*

❖ ***Observation 2: Confidence Reflects Reasoning Chains Quality.*** We investigate whether confidence dynamics reflect reasoning quality by evaluating reasoning chains within the correct set $\mathcal{T}^+$ using the evaluator GPT-4o (OpenAI et al., 2024). Each chain is assessed for logical validity and factual consistency, and categorized into Faithful and Spurious groups. As shown in Figure 3(b), faithful reasoning chains exhibit a higher proportion of positive confidence increments, *suggesting that confidence improvement not only correlates with answer accuracy but also reveals the intrinsic quality of the reasoning process.*

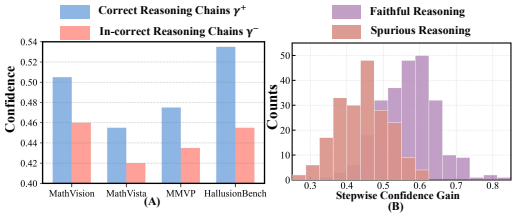

Figure 3: Analysis of the relationship between confidence and reasoning quality. (A) Correct reasoning chains exhibit substantially higher frequencies of positive confidence gains than incorrect ones. (B) Faithful reasoning shows consistently stronger confidence improvement than spurious reasoning.

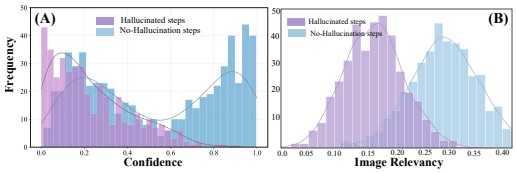

Figure 4: Analysis of the relationship between confidence and visual grounding. (A) Hallucinated steps show lower confidence than non-hallucinated ones. (B) Hallucinated steps exhibit weaker image relevancy than their counterparts.

❖ ***Observation 3: High Confidence Aligns with Stronger Visual Grounding.*** We further evaluate various reasoning models on the perception benchmark to analyze the relationship between confidence and visual grounding. Each step in the reasoning chain is categorized as *hallucinated* or *non-hallucination* based on whether it refers to an object actually present in the image. As shown in Figure 4, hallucinated steps exhibit lower confidence and weaker visual grounding, while non-hallucinatory steps maintain higher and more stable confidence with stronger visual alignment. *The results indicate that confidence acts as an intrinsic signal of visual faithfulness, with higher confidence consistently associated with more reliable reasoning.*

4 METHODOLOGY

4.1 PROBLEM FORMULATION

Given a text input sequence $\mathcal{Q} = (q_1, \ldots, q_k)$ and a set of visual embeddings $\mathcal{Z} = (z_1, \ldots, z_I)$ extracted by a visual encoder, the MLLM π_θ encodes the text sequence into embeddings and incorporates visual features to generate the reasoning sequence $\mathcal{X} = (x_1, x_2, \ldots, x_N)$.

$$\pi_\theta(\mathcal{X} \mid q, z) = \prod_{n=1}^{N} \pi_\theta(x_n \mid \mathcal{X}_{<n}, q, z) \tag{3}$$

where $x_{<n}$ denotes the sequence of tokens preceding position n. Different from approaches that use the last hidden state of the previous reasoning step as latent think tokens (Chen et al., 2025b; Pham and Ngo, 2025), we introduce L learnable *latent think tokens* into the input sequence, whose embeddings after projection are denoted as $\mathcal{T} = [\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_L]$. These tokens are concatenated with the original inputs and fed into the model. During test-time inference, our core idea is to keep model parameters fixed and improve reasoning solely by optimizing the embeddings of the latent think tokens. Motivated by the observations in Section 3, we define a reward function $\mathcal{R}$ to quantify the confidence of the current latent reasoning state. This leads to the following test-time optimization objective:

$$\mathcal{T}^* = \arg\max_{\mathcal{T}} \mathcal{R}(\mathcal{T}, \mathcal{Q}, \mathcal{Z}), \tag{4}$$

In practice, the model iteratively update the latent think tokens for T steps, allowing them to progressively evolve toward directions that maximize the reward.

4.2 Dynamic Multimodal Latent Reasoning

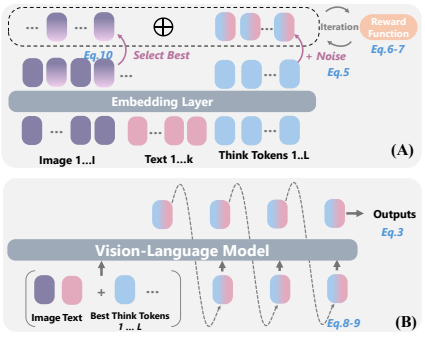

Figure 5: Overview of DMLR framework.

In light of the observations in Section 3, DMLR comprises two key processes: dynamic visual injection strategy for ***RQ1***, and confidence-guided optimization of latent think tokens for ***RQ2***, as shown in Figure 5 and Algorithm 1.

Latent Think Tokens Initialization. We initialize the latent think tokens before each iteration to facilitate exploration in the latent space. To this end, we adopt a stochastic perturbation strategy that adds controlled randomness while preserving representation stability. Specifically, multiplicative noise sampled from a Gaussian distribution is applied as a local perturbation to the current latent state:

$$\mathcal{T}'^{(t)} = \mathcal{T}^{(t)} + \xi^{(t)}, \quad \xi^{(t)} \sim \mathcal{N}(0, \sigma^2 I) \tag{5}$$

where σ^2 is a variance hyperparameter that controls the magnitude of exploration and $\xi^{(t)}$ is the multiplicative Gaussian noise sampled at iteration t. More analyses and results are shown in Section 5.3.

Reward Formulation. We propose a confidence-guided reward that dynamically optimizes latent think tokens during reasoning. In contrast to prior approaches (Zhi et al., 2025; Zou et al., 2025) that use confidence ony for post-hoc evaluation, we treats it as an intrinsic feedback signal that continuously guides latent reasoning optimization. Given the latent think state $\mathcal{T}^{(t)}$, the query q, and visual features z, the model π_θ generates token-level probability distributions $\mathcal{P}_i^{(t)}$ over the vocabulary w. We further quantify the model's confidence for each latent think token by computing the truncated entropy over its top-k most probable tokens, defined as:

$$\mathcal{H}_k\big(\mathcal{P}_i^{(t)}\big) = -\sum_{w \in \mathrm{Top}_k\left(\mathcal{P}_i^{(t)}\right)} \mathcal{P}_i^{(t)}(w)\, \log\big(\mathcal{P}_i^{(t)}(w)\big) \tag{6}$$

where $\mathrm{Top}_k(\cdot)$ denotes the set of the k tokens with the highest probabilities. A lower value of the entropy $\mathcal{H}_k(\cdot)$ corresponds to higher confidence in the model's prediction at that position. The reward for the entire latent reasoning sequence is defined as the complement of the mean truncated entropy computed over all L latent think tokens:

$$\mathcal{R}(\mathcal{T}^{(t)}) = 1 - \frac{1}{L}\sum_{i=1}^{L} \mathcal{H}_k(\mathcal{P}_i^{(t)}) \tag{7}$$

Test-Time Latent Optimization. Recent works (Li et al., 2025b; Zhang et al., 2025e; Ye et al., 2025) have explored test-time gradient optimization to enable adaptation in language tasks, whereas we focus on optimization processes for multimodal latent reasoning. Specifically, during the test-time inference, guided by the objective defined in Equation 7, we adopt a *REINFORCE-based* (Williams, 1992) direct policy gradient method to adaptively optimize the latent think tokens

Algorithm 1: ***Dynamic Multimodal Latent Reasoning***

Require: Image embeddings $\mathcal{Z}$, text embeddings $\mathcal{Q}$, latent tokens $\mathcal{T}_l$, learning rate η, iterations T, best visual patch $\mathcal{V}_{best}$, top-k probability $\mathrm{Top}_k(\mathcal{P}_i)$, the number of candidate patches m

$\mathrm{Top}_k(\mathcal{P}_i) = \pi_\theta([\mathcal{Q}, \mathcal{Z}, \mathcal{T}])$; $r \leftarrow \mathcal{R}(\mathcal{P}_i)$ ▷ reward

 # Latent Policy Gradient Optimization

for $T = 1 \dots t$ **do**

- $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ ▷ latent perturbation
- $\mathcal{T}^{(t)\prime} \leftarrow \mathcal{T}^{(t)} + \epsilon$
- $\mathcal{T}^{(t)} \leftarrow \mathcal{T}^{(t)} + \eta \nabla_{\mathcal{T}^{(t)}} \mathcal{J}(\mathcal{T}^{(t)})$ ▷ latent update
- *# Dynamic Visual Injection*
- $\mathcal{V}_{best} \leftarrow \mathrm{Initialize}(\mathcal{T}^{(0)}, m)$ ▷ initialize best patch
- **for** $L = 1 \dots l$ **do**
  - $\mathcal{Z}_{cand} \leftarrow \mathrm{AttentionSelect}(\mathcal{T}_l^{(t)}, m)$ ▷ select m candidate visual patches
  - $\tilde{\mathcal{T}}_l^{(t)} \leftarrow [\mathcal{T}_l^{(t)}, \mathcal{Z}_{cand}, \mathcal{V}_{best}]$
  - $r \leftarrow \mathcal{R}(\mathcal{Q}, \mathcal{Z}, \tilde{\mathcal{T}}_l^{(t)})$
  - **if** $r > r_{best}$ **then**
    - $\mathcal{V}_{best} \leftarrow \mathcal{V}_{best} \cup \mathcal{Z}_{cand}$;
    - $\mathcal{T}_l^{(t)} \leftarrow \tilde{\mathcal{T}}_l^{(t)}$ ▷ update best
  - **else**
    - $\mathcal{T}_l^{(t)} \leftarrow [\mathcal{T}_l^{(t)}, \mathcal{V}_{best}]$ ▷ revert to previous best

$\mathcal{X} \leftarrow \mathrm{Decode}(\mathcal{T}^{(t)}, \mathcal{Z}, \mathcal{Q})$

return $\mathcal{X}$

$\mathcal{T}^{(t)}$. Assuming that each latent think token is independent, the update rule is formulated as:

$$\mathcal{T}^{(t)} \leftarrow \mathcal{T}^{(t)} + \eta \nabla_{\mathcal{T}^{(t)}} \mathcal{J}(\mathcal{T}^{(t)}) \tag{8}$$

where η denotes the learning rate. According to the Policy Gradient Theorem and Equation 5, the gradient can be formulated and further expressed as:

$$\nabla_{\mathcal{T}} \mathcal{J}(\mathcal{T}) = \mathbb{E}_{\mathcal{T}' \sim \pi(\cdot|\mathcal{T})}[\mathcal{R}(\mathcal{T}')\, \nabla_{\mathcal{T}} \log \pi(\mathcal{T}' \mid \mathcal{T})] = \mathbb{E}\left[\mathcal{R}(\mathcal{T}')\, \frac{\xi}{\sigma^2}\right]. \tag{9}$$

Visual Injection Strategy. Different from methods that directly inject high-attention regions (Gao et al., 2025), our strategy updates the most informative visual patches based on the reward at each iteration and injects them as latent visual tokens. As illustrated in Algorithm 1, we first use the initial attention of the latent think token to collect m highly relevant image patches (see Section 5.1), which serve as the initial best patch $\mathcal{V}_{best}$. At each iteration, the model resamples m candidate patches $\mathcal{Z}_{cand} = \{\mathcal{Z}_1, \dots, \mathcal{Z}_m\}$ based on the updated attention and injects them together with the previous best patch into the latent sequence for reward, as formulated in Equation 10. If the reward $r > r_{best}$, indicating that the candidate patches provide enhanced visual evidence, the best patch $\mathcal{V}_{\text{best}}$ is updated; otherwise, the previous best is retained.

$$r = \mathcal{R}\Big(\mathcal{Z}, \mathcal{Q}, [\mathcal{T}^{(t)}, \mathcal{V}_{best}, \mathcal{Z}_{cand}]\Big) \tag{10}$$

As the iterations progress, the best visual patch converges to the regions most relevant to the latent think state, guiding the latent reasoning toward more effective optimization.

4.3 Theoretical Analysis

To further understand why DMLR achieves high efficiency and robust performance, we provide theoretical explanations through the following two theorems.

Theorem 4.1 (Confidence Reflects Reasoning Quality)**.** *Let h denote the latent reasoning state in DMLR, where $C(h)$ represents the model's confidence level and $Q(h)$ denotes the corresponding reasoning quality. If and only if the gradients of $C(h)$ and $Q(h)$ are positively aligned, the DMLR update along the confidence ascent direction will consequently improve the reasoning quality:*

$$\nabla C(h) \cdot \nabla Q(h) > 0 \tag{11}$$

Theorem 4.2. (Visual Injection Enhances Confidence)**.** *Let $\mathcal{T}$ be the latent reasoning states, $\hat{\mathcal{T}}$ denote the updated states after visual injection, and z_v be the visual features. Visual injection in DMLR increases the mutual information between latent states and visual features, thereby enhancing the expected confidence $J_{conf}(\mathcal{T})$, satisfying:*

$$I(\hat{\mathcal{T}}; z_v) \geq I(\mathcal{T}; z_v) \;\Rightarrow\; J_{\text{conf}}(\hat{\mathcal{T}}) \geq J_{\text{conf}}(\mathcal{T}) \tag{12}$$

5 Experiments

5.1 Experiment Setup

Baselines. We evaluate the proposed DMLR using two types of baselines: model-based and method-based. For the model baselines, we consider six representative MLLMs, including two reasoning models, R1-OneVision (Yang et al., 2025b) and VLAA-Thinking (Chen et al., 2025c), as well as four non-reasoning models, Qwen2.5-VL-3B/7B (Bai et al., 2025b) and Qwen3-VL-4B/8B (Team, 2025). For method baselines, we consider two reasoning paradigms: *Text-only Reasoning* (CCoT (Mitra et al., 2024)) and *Vision-Text Involved Reasoning* (ICoT (Gao et al., 2025), Multimodal-CoT (Zhang et al., 2024)). We further include a Vanilla baseline where non-reasoning models answer directly and reasoning models use their default prompts.

Evaluation Benchmarks. We evaluate our method on three tasks across six benchmarks: (1) *Mathematics Reasoning* ($\text{MathVista}_{\text{mini}}$ (Lu et al., 2024), $\text{MathVision}_{\text{mini}}$ (Wang et al., 2024), MM Math (Sun et al., 2024)); (2) *Visual Reasoning* (HallusionBench (Guan et al., 2024), MMVP (Tong et al., 2024)); (3) *Multimodal Composition* (MMStar (Chen et al., 2024), ScienceQA (Lu et al., 2022)). Details are provided in Appendix A.1.

Implementation Details. All frameworks adopt the eager attention mode to enable access to internal attention maps. A total of 4 latent think tokens $\mathcal{T}$ are used, with $m = 2$ visual candidate patches injected at each iteration. The defaultnumber of optimization iterations is set to 15, with a learning rate of 10^{-3}. To ensure stable exploration in the latent space, the perturbation magnitude σ is set to 10%. All experiments are conducted on four NVIDIA H100 GPUs, with further detailed parameter analyses in Appendix A.3.

5.2 Main Results

Overall Results. As shown in Table 1, models integrated with DMLR achieve the best performance on over 95% of tasks. On mathematical and visual reasoning benchmarks, Qwen2.5-VL-7B achieves average improvements of +1.5% in mathematics and +0.9% in visual reasoning, while the reasoning counterpart R1-OneVision attains average gains of +4.5% and +3.45% on the two domains, respectively. These results indicate that DMLR generalizes robustly across diverse model paradigms and scales. Unlike other baseline methods that often involve trade-offs between reasoning and perception, DMLR consistently improves performance in both domains. For instance, while ICoT yields noticeable gains on mathematical tasks but provides only limited improvements on visual reasoning (e.g., MMVP), DMLR delivers more stable cross domain enhancements, with DMLR-integrated VLAA-Thinking averaging +2.43% higher across all benchmarks.

5.3 Ablation Study

Impact of Visual Injection Strategies. We evaluate various visual injection strategies to assess their effects on reasoning performance. As shown in Table 2, removing visual injection maintains stable reasoning results but leads to a clear drop in perceptual accuracy, underscoring the necessity of visual cues during latent optimization. Injecting all visual patches enhances perception but introduces instability due to redundant visual information. In contrast, DMLR exhibits consistently more stable performance, indicating that its continuously selects more relevant and stable visual information throughout the iterative optimization.

Impact of Iteration Number. As shown in Figure 6, increasing the number of iterations leads to a steady improvement on both reasoning and perception tasks, indicating that iterative optimization effectively enhances latent reasoning. Morever, the reasoning model maintains consistently higher

Table 1: Comparison of different reasoning methods and DMLR across various benchmarks. All metrics are reported in Accuracy (%). Results are evaluated over a diverse suite of mathematics reasoning, visual reasoning, and multimodal composition tasks under multiple backbone models.

		Mathematics Reasoning ↑			Visual Reasoning ↑		Multimodal Composition ↑	
Method	**Model**	$\text{MathVista}_{\text{mini}}$	$\text{MathVision}_{\text{mini}}$	MM-Math	HallusionBench	MMVP	MMStar	ScienceQA
Vanilla	Qwen2.5 VL 7B	58.7	21.6	37.5	65.4	68.7	59.3	49.7
Multimodal COT		56.4	21.8	35.6	63.6	68.1	57.9	49.5
CCOT		57.8	22.5	36.3	64.9	69.0	58.7	50.2
ICoT		58.9	23.3	37.0	65.5	69.3	60.4	50.4
+DMLR (Ours)		**59.1** ↑0.40%	**24.4** ↑2.8%	38.8 ↑1.3%	65.8 ↑0.4%	**70.1** ↑1.4%	**60.1** ↑0.8%	**51.3** ↑1.6%
Vanilla	Qwen2.5 VL 3B	48.2	15.7	29.0	64.2	55.6	50.2	44.1
Multimodal COT		47.3	14.3	28.5	63.8	54.4	48.5	42.9
CCOT		48.0	15.6	30.2	64.0	55.5	49.3	44.5
ICoT		49.8	16.0	30.6	64.7	55.9	49.6	45.0
+DMLR (Ours)		**51.0** ↑2.80%	**17.7** ↑2.74%	**33.3** ↑4.3%	**64.7** ↑0.5%	**56.8** ↑1.26%	**51.2** ↑1.00%	**46.9** ↑2.8%
Vanilla	VLAA Thinking 7B	61.1	23.5	41.5	62.0	68.3	58.9	50.6
Multimodal COT		59.6	23.1	40.6	62.8	67.2	57.1	48.2
CCOT		60.5	24.8	41.8	64.6	68.0	59.0	49.4
ICoT		61.4	25.0	42.3	65.9	68.3	58.2	50.6
+DMLR (Ours)		**62.9** ↑1.80%	**27.6** ↑4.10%	**43.9** ↑2.41%	**67.9** ↑5.94%	**69.4** ↑1.1%	**59.2** ↑0.3%	**51.98** ↑1.38%
Vanilla	R1 OneVision 7B	51.2	18.7	40.7	62.1	67.0	52.1	50.9
Multimodal COT		52.5	18.9	39.6	62.5	68.0	51.6	51.7
CCOT		53.4	20.3	40.8	63.0	68.9	53.5	52.8
ICoT		55.6	21.5	41.7	63.8	69.6	54.0	54.4
+DMLR (Ours)		**58.0** ↑6.81%	**23.3** ↑4.56%	**42.9** ↑2.21%	**64.1** ↑2.09%	**71.9** ↑4.93%	**56.2** ↑4.14%	**55.4** ↑4.52%
Vanilla	Qwen3 VL 8B	66.0	32.9	66.2	73.2	71.9	68.1	54.1
Multimodal COT		64.8	32.8	65.1	73.0	69.6	66.9	53.2
CCOT		66.5	33.3	65.5	73.5	70.3	68.8	54.4
ICoT		66.2	34.9	66.8	74.5	71.8	69.3	55.8
+DMLR (Ours)		**66.9** ↑0.9%	**36.2** ↑3.34%	**67.7** ↑1.51%	**74.6** ↑1.48%	**72.8** ↑0.97%	**70.0** ↑1.91%	**55.6** ↑1.48%
Vanilla	Qwen3 VL 4B	64.7	24.2	65.4	71.6	71.3	57.4	52.4
Multimodal COT		62.3	24.8	63.9	70.0	69.6	57.7	53.0
CCOT		64.5	26.6	64.8	71.5	71.2	58.8	52.9
ICoT		64.5	27.5	65.0	72.2	72.5	59.3	53.7
+DMLR (Ours)		**65.6** ↑0.93%	**29.4** ↑5.20%	**65.9** ↑0.5%	**72.7** ↑1.12%	**72.3** ↑0.98%	**60.3** ↑2.88%	**54.9** ↑2.48%

Table 2: Ablation on Latent Visual Injection. We compare different injection strategies across multiple benchmarks. *All* injects all visual patches at every iteration, while *Ours* injects the best visual patches. Refer to Section 5.1 for detailed settings.

Method	MathVista	MathVision	MMStar	ScienceQA
w/o Injection	0.627	0.321	0.687	0.536
+ Injection (All)	0.621	0.327	0.676	0.527
+ DVI (Ours)	**0.634**	**0.340**	**0.694**	**0.549**

accuracy throughout the process and continues to yield gains even after multiple iterations, demonstrating a stronger ability to benefit from iterative refinement.

Impact of Noise Scale. We further analyze the influence of the perturbation magnitude σ on latent optimization. As shown in Figure 7(b), increasing the initial noise scale promotes effective exploration, allowing the model to cover a wider range of latent trajectories and identify higher-confidence reasoning paths. However, when σ becomes excessively large, the injected perturbation makes the updates unstable, leading to a subsequent drop in performance. This indicates that latent reasoning benefits from only a modest level of perturbation.

Impact of Visual Patch Number. As shown in Figure 7(a), performance improves when a moderate number of candidate visual patches are injected, whereas injecting an excessive number of patches leads to a clear decline. This trend indicates that a limited number of candidates is sufficient for effective updates, while excessive patches introduce redundant visual information that negatively affect optimization. Furthermore, Figure 8 shows that as the iterations progress, the reward steadily increases and the selected best patch becomes increasingly stable, exhibiting a clear convergence trend. This trend indicates that the dynamic injection strategy does not continually introduce additional visual patches into the latent space, but instead converges toward a small set of highly relevant patches during optimization.

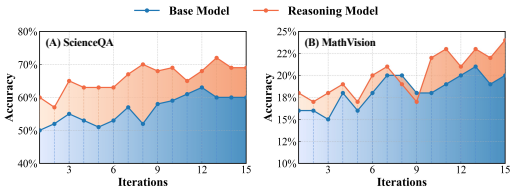

Figure 6: Effect of iterations on performance. For both the base model and the reasoning model, accuracy on both datasets increases as the number of iterations grows.

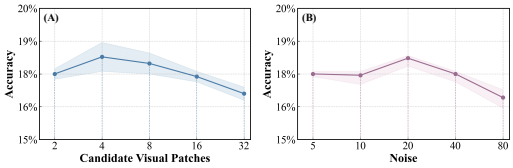

Figure 7: (A) Effect of the number of injected candidate visual patches on performance. (B) Impact of noise magnitude (%) on performance. All results are evaluated on the MathVision dataset.

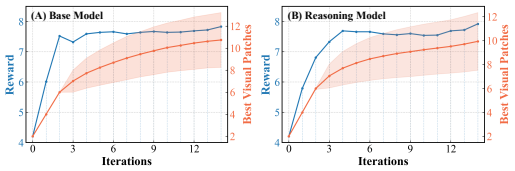

Figure 8: Confidence reward and best visual patch injection across iterations. Both the base model and the reasoning model exhibit a clear positive correlation.

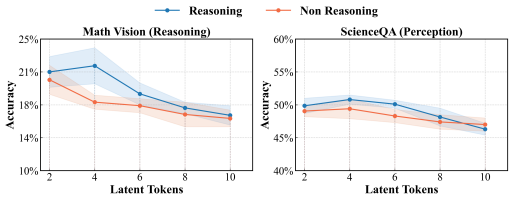

Figure 9: Effect of the number of latent tokens. Increasing the number of latent tokens initially improves performance, but excessive tokens lead to noticeable degradation.

Number of Latent Think Tokens. We further evaluate the impact of the number of latent think tokens on overall performance. As shown in Figure 9, setting the number of latent tokens to a small range (2–4) yields stable improvements on both reasoning and perception tasks. However, as the number of tokens continues to increase, performance on both tasks begins to decline, with the reasoning model exhibiting more pronounced fluctuations. This overall trend indicates that increasing the number of latent tokens beyond a moderate level does not provide additional benefits and instead makes the optimization process less stable.

5.4 Quantitative Analysis

Visual Grounding Analysis. We visualize the attention heatmaps of VLAA-Thinking during the reasoning process. As shown in Figure 10(a), the explicit CoT baseline often shifts its attention toward task-irrelevant regions, whereas DMLR maintains a stable focus on task-relevant areas. This demonstrates that latent multimodal reasoning produces more consistent and reliable visual grounding throughout the reasoning process. Figure 10(b) further shows the evolution of attention across iterations. The attention distribution gradually converges toward task-relevant regions in models integrated with DMLR, reflecting a more stable and consistent focus throughout the optimization.

Latent Behavior Analysis. We visualize the final distributions of latent think tokens, text tokens, and image tokens using t-SNE (Maaten and Hinton, 2008) to analyze the effect of the iterative optimization on the latent reasoning. As shown in Figure 10(c), the latent think tokens form a tight cluster that is well separated from both text and visual embeddings, and are located in a stable intermediate region between the two modalities. This distribution suggests that the optimized latent tokens become modality-independent, forming a unified cross-modal semantic representation. The compactness of the cluster further indicates that the optimization process yields more stable and consistent latent reasoning states.

Inference Efficiency Analysis. As shown in Figure 11, different reasoning paradigms exhibit distinct trade-offs between accuracy and efficiency. The explicit methods such as Multimodal CoT rely on long-chain text generation, incurring substantial computational overhead. Although ICoT enhances reasoning to some extent, it injects a large volume of visual information during decoding, which significantly slows inference. In contrast, DMLR performs optimization entirely within the latent space, introducing no additional sequence generation cost. Moreover, its dynamic visual injection strategy selects only the relevant visual patches to the current latent state at each iteration,

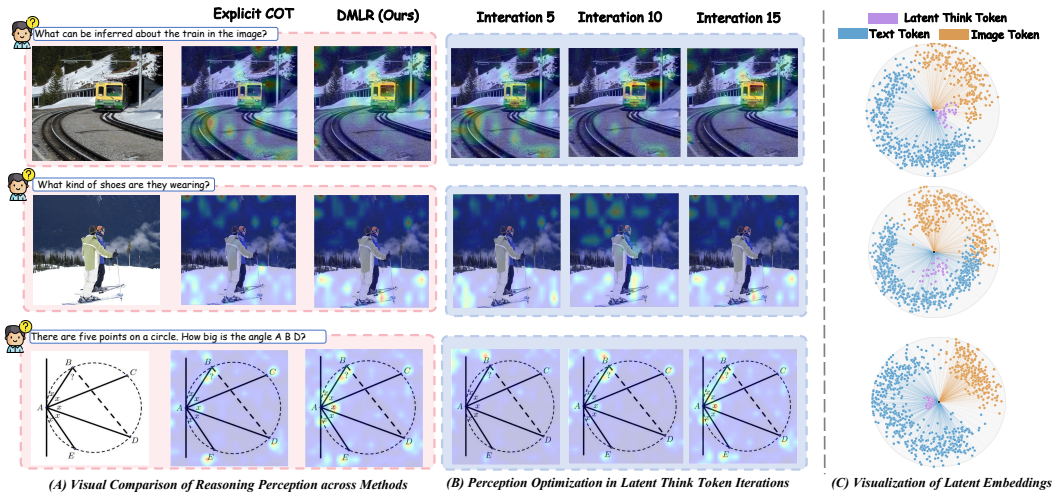

Figure 10: Qualitative analysis of our DMLR framework. (A) Visual comparison of visual grounding behaviors between Explicit CoT and DMLR across diverse queries. DMLR produces more focused and stable visual grounding than explicit CoT. (B) Perception optimization across latent think token iterations, where visual attention becomes progressively sharper and better aligned with relevant regions. (C) Visualization of latent embeddings showing the geometric separation of latent think tokens, text tokens, and image tokens, illustrating the structured organization of the latent reasoning space.

eliminating redundant visual computation. By preserving accuracy gains while reducing inference overhead, DMLR achieves a more favorable balance between efficiency and performance.

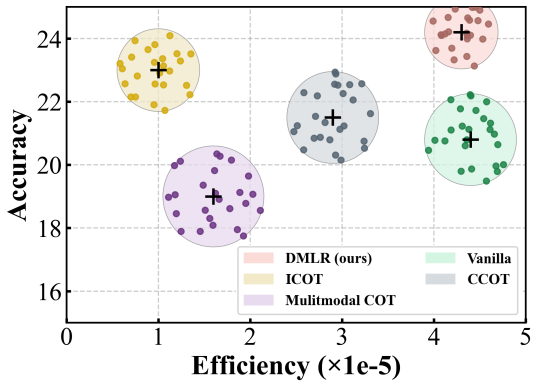

Figure 11: Comparison of efficiency and accuracy across various reasoning methods on the MathVision Benchmark. DMLR achieves the best overall trade-off, delivering higher accuracy while maintaining strong inference efficiency. The x-axis reports the efficiency metric $(\text{Acc}/\text{AvgBatchTime})^2$.

6 Conclusion

In this work, we analyze how MLLMs utilize visual information and confidence during reasoning. Based on these observations, we introduce DMLR, a test-time multimodal latent reasoning framework that integrates confidence-guided latent optimization with dynamic visual injection.This method enables models to refine their reasoning, retrieve visual evidence only when need without training. Extensive experiments across various tasks show that DMLR consistently boosts both reasoning and perception tasks, offering a stable and training-free alternative to other methods.

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*, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18798–18806, 2024.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025a.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025a.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025b.

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.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*, 2025.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images, 2025. URL https://arxiv.org/abs/2505.15879.

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.

Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, Tieniu Tan, and Furu Wei. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in mllms, 2025a. URL https://arxiv.org/abs/2510.24514.

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, 2025. URL https://arxiv.org/abs/2412.06769.

Hengli Li, Chenxi Li, Tong Wu, Xuekai Zhu, Yuxuan Wang, Zhaoxin Yu, Eric Hanchen Jiang, Song-Chun Zhu, Zixia Jia, Ying Nian Wu, and Zilong Zheng. Seek in the dark: Reasoning via test-time instance-level policy gradient in latent space, 2025b. URL https://arxiv.org/abs/2505.13308.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning, 2025c. URL https://arxiv.org/abs/2509.24251.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens, 2025a. URL https://arxiv.org/abs/2506.172182.

Tan-Hanh Pham and Chris Ngo. Multimodal chain of continuous thought for latent-space reasoning in vision-language models, 2025. URL https://arxiv.org/abs/2508.12587.

Guibin Zhang, Muxin Fu, and Shuicheng Yan. Memgen: Weaving generative latent memory for self-evolving agents, 2025b. URL https://arxiv.org/abs/2509.24704.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025a.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning, 2025b. URL https://arxiv.org/abs/2504.08837.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models, 2025a. URL https://arxiv.org/abs/2505.21523.

Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. URL https://arxiv.org/abs/2401.06209.

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, Wenxue Li, Yulong Li, Wenxuan Song, Shiyan Su, Wei Feng, Jionglong Su, Mingquan Lin, Yifan Peng, Xuelian Cheng, Imran Razzak, and Zongyuan Ge. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26147–26159, June 2025.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, and Yu Cheng. Openthinkimg: Learning to think with images via visual tool reinforcement learning, 2025c. URL https://arxiv.org/abs/2505.08617.

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, 2024. URL https://arxiv.org/abs/2406.09403.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model, 2025. URL https://arxiv.org/abs/2511.05271.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv preprint arXiv:2505.15436*, 2025c.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models, 2025. URL https://arxiv.org/abs/2410.03577.

Sheng Liu, Tianlang Chen, Pan Lu, Haotian Ye, Yizheng Chen, Lei Xing, and James Zou. Fractional reasoning via latent steering vectors improves inference time compute. *arXiv preprint arXiv:2506.15882*, 2025c.

Chi-Pin Huang, Yueh-Hua Wu, Min-Hung Chen, Yu-Chiang Frank Wang, and Fu-En Yang. Thinkact: Vision-language-action reasoning via reinforced visual latent planning, 2025b. URL https://arxiv.org/abs/2507.16815.

Yapeng Mi, Hengli Li, Yanpeng Zhao, Chenxi Li, Huimin Wu, Xiaojian Ma, Song-Chun Zhu, Ying Nian Wu, and Qing Li. Milr: Improving multimodal image generation via test-time latent reasoning, 2025. URL https://arxiv.org/abs/2509.22761.

Jingcheng Deng, Liang Pang, Zihao Wei, Shichen Xu, Zenghao Duan, Kun Xu, Yang Song, Huawei Shen, and Xueqi Cheng. Latent reasoning in llms as a vocabulary-space superposition, 2025. URL https://arxiv.org/abs/2510.15522.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. Spotlight on token perception for multimodal reinforcement learning, 2025c. URL https://arxiv.org/abs/2510.09285.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space, 2025d. URL https://arxiv.org/abs/2505.15778.

Natasha Butt, Ariel Kwiatkowski, Ismail Labiad, Julia Kempe, and Yann Ollivier. Soft tokens, hard truths, 2025. URL https://arxiv.org/abs/2509.19170.

Wengao Ye, Yan Liang, and Lianlei Shan. Thinking on the fly: Test-time reasoning enhancement via latent thought policy optimization, 2025. URL https://arxiv.org/abs/2510.04182.

Zihao Li, Xu Wang, Yuzhe Yang, Ziyu Yao, Haoyi Xiong, and Mengnan Du. Feature extraction and steering for enhanced chain-of-thought reasoning in language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10904–10924, Suzhou, China, November 2025d. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.552. URL https://aclanthology.org/2025.emnlp-main.552/.

Guohao Sun, Hang Hua, Jian Wang, Jiebo Luo, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Latent chain-of-thought for visual reasoning, 2025. URL https://arxiv.org/abs/2510.23925.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought, 2025. URL https://arxiv.org/abs/2411.19488.

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

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia,

Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Chao Chen, Zhixin Ma, Yongqi Li, Yupeng Hu, Yinwei Wei, Wenjie Li, and Liqiang Nie. Reasoning in the dark: Interleaved vision-text reasoning in latent space, 2025b. URL https://arxiv.org/abs/2510.12603.

Zhuo Zhi, Chen Feng, Adam Daneshmend, Mine Orlu, Andreas Demosthenous, Lu Yin, Da Li, Ziquan Liu, and Miguel R. D. Rodrigues. Seeing and reasoning with confidence: Supercharging multimodal llms with an uncertainty-aware agentic framework, 2025. URL https://arxiv.org/abs/2503.08308.

Guibin Zhang, Fanci Meng, Guancheng Wan, Zherui Li, Kun Wang, Zhenfei Yin, Lei Bai, and Shuicheng Yan. Latentevolve: Self-evolving test-time scaling in latent space, 2025e. URL https://arxiv.org/abs/2509.24771.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization, 2025b. URL https://arxiv.org/abs/2503.10615.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models, 2025c. URL https://arxiv.org/abs/2504.11468.

Qwen Team. Qwen3-vl: Sharper vision, deeper thought, broader action. Blog post, https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef, Sept 2025.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14420–14431, June 2024.

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

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. URL https://arxiv.org/abs/2310.02255.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset, 2024. URL https://arxiv.org/abs/2402.14804.

Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. Mm-math: Advancing multimodal math evaluation with process evaluation and fine-grained classification, 2024. URL https://arxiv.org/abs/2404.05091.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 2024. URL https://arxiv.org/abs/2310.14566.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024. URL https://arxiv.org/abs/2403.20330.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. URL https://arxiv.org/abs/2209.09513.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

APPENDIX

A MORE DETAILED ABOUT EVALUATION

A.1 DATASETS

- **MathVista$_{\text{mini}}$** is a benchmark for mathematical reasoning in visual contexts, aggregating diverse multimodal math tasks that require fine-grained visual understanding and compositional numerical reasoning.
- **MathVision$_{\text{mini}}$** is a curated benchmark of competition-level visual math problems spanning multiple disciplines and difficulty levels to assess multimodal models' mathematical reasoning under challenging and diverse settings.
- **MM Math** is a benchmark of open-ended math problems with visual contexts that supports both outcome and process evaluation, enabling detailed analysis of multimodal reasoning behaviors and typical error patterns.
- **HallusionBench** is a benchmark for image-context reasoning that uses carefully structured question pairs to diagnose hallucination, visual illusion, and logical inconsistency in large vision-language models.
- **MMVP** is a benchmark built from multimodal visual patterns designed to expose "CLIP-blind" image–text pairs, revealing systematic visual perception failures and hallucinated explanations in multimodal LLMs.
- **MMStar** is a vision-indispensable multimodal benchmark composed of carefully human-filtered samples that ensure true visual dependency while evaluating core multimodal capabilities along multiple fine-grained axes.
- **ScienceQA** is a multimodal multiple-choice science benchmark with rich textual and visual contexts, lectures, and explanations that spans diverse subjects and skills, supporting evaluation of both answer accuracy and explanation quality.

For all datasets, we limit the maximum sample size to 1000 instances.

A.2 EVALUATION SETTING

We adopt a unified prompting setup for all models. Unless otherwise stated, we use greedy decoding (`do_sample=False`) for all generation tasks.

System Prompt.

> A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within `<think> </think>` and `<answer> </answer>` tags, respectively, i.e., `<think>` reasoning process here `</think> <answer>` answer here `</answer>`.

Task Prompt.

> Please analyze the image carefully and solve this problem step by step. Show your reasoning process clearly, then put your final answer within `\boxed{}`.
>
> **Question:** *[Problem Text]*

For all benchmarks considered in our experiments, the ground-truth answers are verifiable; we use regular expressions to extract the content within `\boxed{}` from the model outputs and then match it against the correct answers.

A.3 PARAMETERS SETUP

- **Latent Think Tokens $\mathcal{T}$:** We set the number of latent think tokens to 4. During generation, after each latent token the model dynamically injects a visual patch into the latent stream to refresh its internal perception state.

- **Image Patches m:** We dynamically insert visual patches into the latent stream. At initialization, we inject 2 patches; at each subsequent iteration, we select $m = 2$ patches with the highest attention scores and append them after each latent think token, with at most 16 patches inserted per iteration. Additionally, we set the image processor's max pixel size to 256 for all inputs.
- **Optimization Parameters:** We perform 15 latent optimization steps with a learning rate of 1×10^{-3}. To ensure stable exploration in the latent space, the perturbation magnitude is set to $\sigma = 0.1$ (10%) with a decay factor of 0.95, and we fix the random seed to 42 for all experiments.
- **Environment Setup:** All experiments are conducted on NVIDIA H100 GPUs. For a fair comparison, we run all baselines and our method in `float32` precision and use the `eager` attention backend for inference.

B Case Study

In this section, we provide qualitative examples to illustrate the behavior of the DMLR-enhanced model across different task settings. As shown in the following example, the model demonstrates more reliable and fine-grained visual perception, supported by stable and semantically coherent internal representations. In reasoning scenarios, the model produces accurate and well-structured logical steps. Importantly, the dynamic latent reasoning mechanism suppresses unnecessary expansions, resulting in concise and task-focused reasoning chains rather than overly long or speculative explanations. Taken together, these examples highlight that DMLR simultaneously strengthens perceptual quality and reasoning reliability, while ensuring that the generated reasoning remains efficient and free from redundant elaboration.

Perception Example

Figure 12: Question: Is the door of the truck cab open?
Options: (a) Yes (b) No
Ground Truth: (a)

Qwen2.5-VL-7B-Instruct + DMLR:
The image shows a garbage truck with its door on the side of the cab open. This is evident from the visible interior and the fact that the door is not aligned with the rest of the vehicle's structure, indicating it has been opened. The answer is $\boxed{a}$.

Qwen2.5-VL-7B-Instruct:
To determine if the door of the truck cab is open, let's analyze the image step by step:

1. Observe the side of the truck where the door would be located.
2. Look for any visible gap or opening that would indicate the door is not closed.
3. Check the overall appearance of the door; if it is open, there should be a noticeable gap or difference in the alignment compared to when it is closed.

Upon examining the image:

- The door appears to be aligned with the rest of the truck's body.
- There is no visible gap or difference in the alignment that would suggest the door is open.

Based on this analysis, the door of the truck cab does not appear to be open.
<think> The door of the truck cab is not open.</think> <answer> No</answer>
$\boxed{\text{No}}$

Reasoning Example

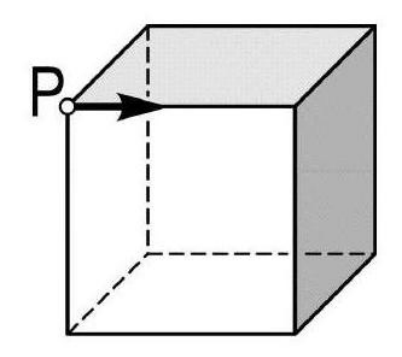

Figure 13: Question: A beetle walks along the edges of a cube. Starting from point P it first moves in the direction shown. At the end of each edge it changes the direction in which it turns, turning first right then left, then right etc. Along how many edges will it walk before it returns to point P?
Ground Truth: 6

Qwen2.5-VL-7B-Instruct + DMLR:
The beetle starts at point P and moves along the edges of a cube, changing direction according to the pattern: right, left, right. To determine how many edges it will walk before returning to point P, we need to trace its path.

1. The beetle starts at P.
2. It moves right to an adjacent vertex.
3. It turns left to another adjacent vertex.
4. It turns right again to reach a third adjacent vertex.
5. It continues this pattern until it returns to P.

By examining the cube's structure, we can see that after 6 edges, the beetle will return to point P because each face of the cube has 4 edges, and the beetle covers all edges of one face before repeating the pattern on the next face.
Thus, the beetle will walk along 6 edges before returning to point P.

Qwen2.5-VL-7B-Instruct:
To solve this problem, we need to track the beetle's path as it moves along the edges of the cube, changing its direction according to the specified pattern (right, left, right, etc.).

1. **Initial Position and Direction**: The beetle starts at point P and moves in the direction indicated by the arrow, which is along one edge of the cube.
2. **First Edge**: The beetle moves along the first edge and changes direction to the right.
3. **Second Edge**: Moving right, the beetle now faces a new direction and continues along the next edge.
4. **Third Edge**: After moving along the second edge, the beetle changes direction again to the left.
5. **Fourth Edge**: Moving left, the beetle now faces another new direction and continues along the next edge.
6. **Fifth Edge**: After moving along the third edge, the beetle changes direction to the right again.
7. **Sixth Edge**: Moving right, the beetle now faces yet another new direction and continues along the next edge.
8. **Seventh Edge**: After moving along the fourth edge, the beetle changes direction to the left once more.
9. **Eighth Edge**: Moving left, the beetle now faces a new direction and continues along the next edge.

At this point, the beetle has completed a full cycle around the cube, returning to point P.
The beetle will have walked along 8 edges before it returns to point P. The answer is 8.