

LaRe: Latent Refocusing for Multimodal Reasoning

Jizheng Ma^{1,2}, Xiaofei Zhou^{1,2†}, Geyuan Zhang^{1,2}, Yanlong Song^{1,2}, Han Yan^{1,2}

¹Institute of Information Engineering, Chinese Academy of Sciences

²School of Cyber Security, University of Chinese Academy of Sciences

zhouxiaofei@iie.ac.cn

Abstract

Chain of Thought (CoT) reasoning enhances logical performance by decomposing complex tasks, yet its multimodal extension faces a trade-off. The prevailing Thinking with Images paradigm achieves visual refocusing by explicitly cropping image regions, yet incurs rapidly growing computational overhead. The emerging line of latent space reasoning reduces token consumption, but lacks the capacity for dynamic refocusing. We argue that this trade-off stems from a tacitly accepted premise that effective visual refocusing must occur in the form of explicit tokens. Building on this, we propose Latent Refocusing (LaRe), a new multimodal reasoning paradigm in which visual refocusing takes place entirely within the latent space. We further design a semantic augmentation training strategy that ensures the semantic structure of the latent space through visual reconstruction objective. Experimental evaluations demonstrate that LaRe improves average accuracy by 7.6% compared to existing baselines while reducing the number of tokens required for inference by 59.7%. When scaled to a 8B-parameter Vision-Language Model backbone, LaRe achieves performance comparable to state-of-the-art methods, demonstrating the efficacy of our proposed latent refocusing paradigm for multimodal reasoning.

1 Introduction

The Chain of Thought (CoT) reasoning mechanism enhances the reasoning capabilities of large language models (LLMs) by decomposing complex problems through explicit intermediate steps (Wei et al., 2022; Zhou et al., 2024; Yao et al., 2023; OpenAI et al., 2024; Zhang et al., 2024; Wang et al., 2022; Brown et al., 2020; Minaee et al., 2024). With the rapid development of Multimodal Large

[†]Corresponding author.

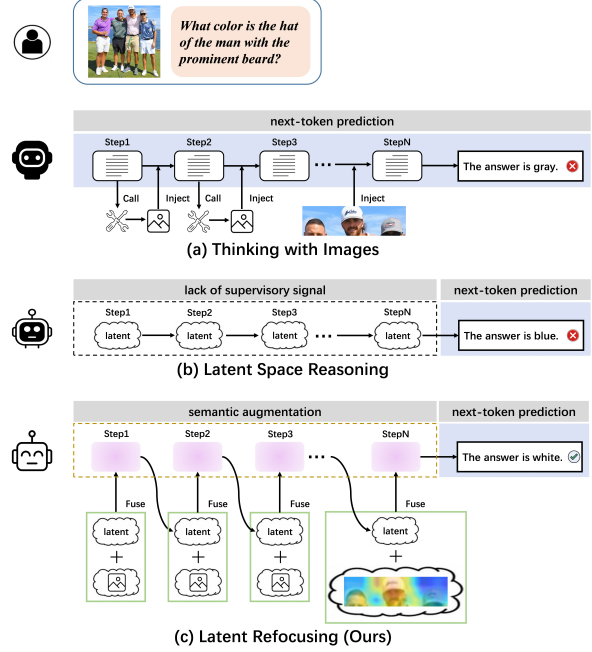


Figure 1: Illustrative comparison across Thinking with Images, Latent Space Reasoning, and our proposed Latent Refocusing (LaRe) framework. LaRe performs refocusing within latent space, achieving more flexible reasoning.

Language Models (MLLMs) in recent years, extending Chain of Thought to multimodal scenarios has become increasingly important (Xu et al., 2023; Li et al., 2025d). This trend has given rise to multimodal reasoning as an emerging research direction that is gradually becoming an important technique for achieving artificial general intelligence (AGI) (Wang et al., 2025b).

Currently, the predominant paradigm for multimodal reasoning is Thinking with Images (illustrated in Figure 1 (a)), which allows the model to explicitly crop and rescale image regions during inference so that the reasoning chain can refocus on visual evidence (Fu et al., 2025; Gao et al., 2025). Despite its effectiveness, this paradigm requires every visual refocus to be injected into the reasoning

chain in the form of explicit visual tokens, causing the number of reasoning tokens to grow rapidly with the number of refocusing (Wang et al., 2025b). In scenarios involving complex visual evidence, such overhead quickly becomes prohibitive.

To address this issue, a promising emerging direction is Latent Space Reasoning (illustrated in Figure 1 (b)) (Li et al., 2025c; Hao et al., 2024), which shifts the reasoning process from explicit tokens to latent representations. By performing reasoning within the latent space, such methods substantially reduce token consumption. However, current explorations of latent space reasoning remain at an early stage and lack the dynamic refocusing ability found in paradigms such as Thinking with Images. Consequently, the reasoning process cannot anchor itself to key visual information. Most existing methods either compress reasoning into a static sequence (Su et al., 2025; Xu et al., 2025) or treat latent visual features as static initialization inputs (Li et al., 2025a). Moreover, prevailing training objectives are largely confined to conventional autoregressive language modeling leaving the latent space without effective supervisory signals and rendering latent representations a black box.

These two lines of work reveal a fundamental tension in multimodal reasoning: under existing paradigms, the ability to refocus visual evidence and the token efficiency of the reasoning process cannot be achieved simultaneously. Thinking with Images trades token overhead for visual refocusing, whereas latent reasoning trades the capacity for visual refocusing for efficiency. We argue that this tension is not intrinsic, but stems from a tacitly accepted premise that effective visual refocusing must occur in the form of explicit tokens. Once this premise is abandoned, visual refocusing can take place within a continuous latent space, preserving the flexibility of multi-round refocusing while avoiding token inflation. Motivated by this observation, we propose Latent Refocusing (LaRe), a novel multimodal reasoning paradigm that realizes visual refocusing entirely within the latent space as depicted in Figure 1 (c).

Specifically, latent refocusing does not rely on explicit image tokens for refocusing. Instead it introduces a set of iteratively evolving latent representations. At each iteration a lightweight module refocuses on visual information conditioned on the current reasoning state and fuses it into the reasoning state to form latent tokens. These tokens then guide the next reasoning step producing an itera-

tive refocusing process. The latent tokens inhabit a latent space that encodes the evolving intermediate reasoning states. To shape this space into a semantically meaningful and interpretable structure, we further propose a semantic augmentation training strategy based on visual reconstruction. Inspired by the ability of denoising processes to encourage models to learn the underlying data manifold (Karas et al., 2022; Chen et al., 2023), this strategy employs a denoising objective for visual reconstruction, thereby guiding the latent space toward a structured semantic layout that is conducive to reasoning. The reconstruction objective ensures that the latent space retains the fine grained visual information required for reasoning. The contributions of this work are as follows:

- We propose Latent Refocusing, which enables the model to dynamically anchor critical visual evidence within a continuous latent space, thereby achieving more flexible and efficient inference. To the best of our knowledge, LaRe is among the first to investigate visual refocusing as an explicit mechanism within latent space for multimodal reasoning.
- We propose a training strategy tailored for latent space reasoning. By leveraging visual reconstruction with a denoising objective, we guide the model to produce semantic representations in the latent space that are beneficial for reasoning, thereby enhancing its ability to capture complex spatial layouts and implicit visual relationships.
- Across multiple benchmarks, LaRe improves average accuracy by 7.6% over strong baselines while reducing inference token consumption by 59.7%. Qualitative analysis further shows that the latent representations are semantically coherent and interpretable, offering new directions for future research on latent space reasoning.

2 Related Work

Multimodal Reasoning. Chain of Thought (CoT) reasoning was originally designed to augment the capabilities of Large Language Models (LLMs) through explicit reasoning steps (Wei et al., 2022) and has since been expanded into the multimodal domain (Wang et al., 2025b). Early methodologies primarily relied on textual rationales (Zhang et al., 2023), though such approaches frequently struggle to capture fine grained spatial

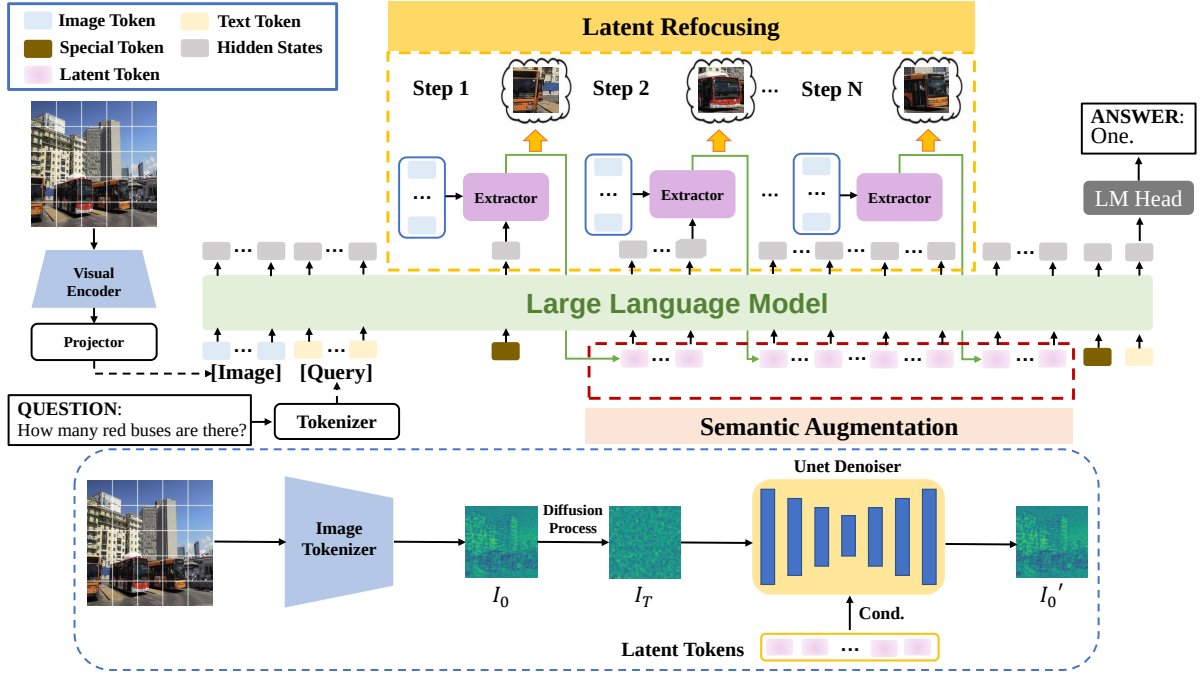


Figure 2: Overview of our proposed LaRe paradigm. The framework comprises Latent Refocusing (top), where the LLM and Extractor generate latent tokens by adaptively refocusing on visual evidence conditioned on reasoning states, and Semantic Augmentation (bottom), which leverages a diffusion-based visual reconstruction task to provide supervision, ensuring that latent tokens maintain both high-level semantics and fine-grained visual fidelity.

details. To bridge this gap, the Thinking with Images paradigm incorporates dynamic visual information into the reasoning process. For example, ReFocus (Fu et al., 2025) guides attention by highlighting structured regions, while Visual CoT (Rose et al., 2023) inserts image states into intermediate reasoning steps. Other methodologies such as Chain of Image (Meng et al., 2023) and VisualSketchpad (Hu et al., 2024) utilize auxiliary sketches for inference, whereas ICoT (Gao et al., 2025) constructs interleaved multimodal sequences. While these methods improve reasoning accuracy, they incur additional computational overhead.

Latent Space Reasoning. Recent research trends have increasingly shifted toward latent space reasoning, which processes intermediate reasoning steps within the internal states of a model without explicit verbalization (Zhu et al., 2025; Li et al., 2025c). Existing works explore latent optimization at the token level (Tack et al., 2025; Sun et al., 2025; Gong et al., 2025) or the trajectory level (Hao et al., 2024; Shen et al., 2025a,b; Zhang et al., 2025c; Kong et al., 2025) while incorporating signal guided control mechanisms (Herel and Mikolov, 2024; Li et al., 2025b) and recurrent execution structures across layers (Saunshi et al.,

2025; Mohtashami et al., 2025). Nevertheless, the majority of these approaches are restricted to the textual manifold.

Recent multimodal latent reasoning methods, including MCOU (Pham and Ngo, 2025), Mirage (Yang et al., 2025), LVR (Li et al., 2025a), and CrystaL (Zhang et al., 2026), often lack an effective refocusing mechanism. These methods attempt to emulate human mental imagery to reason about visual information. However, analysis reveals that the imagined content in the latent space is largely ineffective, as the latent tokens do not carry actual visual information (Li et al., 2026). This deficiency results in weak visual anchoring during reasoning, which limits their overall expressiveness and scalability.

3 Methodology

In this section, we present LaRe, a novel multimodal reasoning paradigm that performs refocusing operations within the latent space. LaRe comprises two core components: an iterative latent refocusing mechanism and a semantic augmentation training strategy. The mechanism introduces dynamically evolving latent representations to selectively integrate reasoning-critical visual information into the latent reasoning steps. The training

strategy is designed to encode fine-grained visual information in the latent space, which in turn facilitates the latent refocusing process.

3.1 Latent Refocusing

LaRe endows the model with flexible multimodal reasoning capabilities through an iterative refocusing mechanism as illustrated in Figure 2. We model the reasoning process as a set of latent representations that evolve within the latent space. These latent representations encapsulate reasoning information while simultaneously preserving the visual information under focus.

Specifically, at each iteration step, we adaptively focus on and fuse reasoning-relevant visual information based on latent textual information to obtain latent tokens. Formally, the process of the k -th iteration step is expressed as follows:

$$H_T^{(k)} = H_{\text{LLM}}^{(k)}([V_T; T; Z^{(1:k-1)}]), \quad (1)$$

$$Z^{(k)} = \text{Extractor}(H_T^{(k)}, V_T). \quad (2)$$

In these equations, $V_T \in \mathbb{R}^{m \times d}$ denotes the visual tokens obtained from the original image via a visual encoder and multimodal projector (Bai et al., 2023; Li et al., 2023a), $T \in \mathbb{R}^{n \times d}$ represents the text embeddings of the input query, $Z^{(k)} \in \mathbb{R}^{l \times d}$ represents the latent tokens generated at the current step, and $Z^{(1:k-1)}$ constitutes the historical sequence of latent tokens generated in previous iterations. In this context, m , n , and l denote the sequence lengths of visual tokens, textual tokens, and latent tokens, respectively, while d represents the hidden dimension of the LLM.

During this process, the Large Language Model (LLM) functions as the decision hub. The LLM processes the concatenated input sequence and extracts hidden states, denoted as $H_{\text{LLM}}^{(k)}$, to generate intermediate hidden representations $H_T^{(k)} \in \mathbb{R}^{l \times d}$ from every layer. These representations encapsulate the textual reasoning information of the model at the current step as described by (Hao et al., 2024). To exploit the representational characteristics inherent across all layers of the LLM (Skean et al., 2025), we set the length l to match the number of layers in the LLM. Subsequently, the Extractor, acting as a lightweight semantic aggregation module, focuses on reasoning-relevant visual information from V_T based on $H_T^{(k)}$ and fuses it into latent tokens. The resultant latent tokens $Z^{(k)}$ encapsulate

critical reasoning evidence and drive subsequent reasoning steps.

Extractor. The Extractor is a lightweight Transformer module designed to focus on task-relevant evidence from the visual features V_T through an N -layer process. Each layer performs internal logic modeling followed by cross-modal refocusing, both stabilized by residual connections to preserve reasoning context. Starting with $L^{(0)} = H_T^{(k)} \in \mathbb{R}^{l \times d}$, the update rules for the i -th layer and the final synthesis of latent tokens $Z^{(k)}$ are:

$$L_{\text{SA}}^{(i)} = \text{Self-Attn}(\text{Norm}(\text{Linear}(L^{(i-1)}))) + L^{(i-1)}, \quad (3)$$

$$L_{\text{CA}}^{(i)} = \text{Cross-Attn}(\text{Norm}(L_{\text{SA}}^{(i)}), V_T) + L_{\text{SA}}^{(i)}, \quad (4)$$

$$Z^{(k)} = \text{FFN}(\text{GLU}(L_{\text{CA}}^{(N)})). \quad (5)$$

By employing the latent state as a query to probe V_T , the Extractor extracts the precise visual signals required for the current reasoning step. The resulting $Z^{(k)}$ serves as a semantic carrier to drive the subsequent iteration of the large language model.

Inference Phase. After reaching the preset number of iterations K , we prepend and append the $< \text{bot} >$ and $< \text{eot} >$ special tokens to the latent sequence to delimit the boundaries of the continuous iteration, attach this sequence to the original image-text input, and then let the model generate text tokens autoregressively.

3.2 Semantic Augmentation

To prevent the loss of critical low-level details during latent compression and refocusing, we propose a semantic augmentation strategy centered on visual reconstruction. This objective requires the generated latent tokens to reconstruct fine-grained visual representations, which ensures that the latent representations maintain high-level semantics while preserving necessary evidential fidelity.

Inspired by diffusion models and denoising autoencoders, we adopt a denoising training scheme rather than simply reconstructing static features through latent tokens. This objective addresses significant spatial redundancy (He et al., 2022) and provides meaningful supervision signals centered on visual reconstruction. Denoising outperforms standard regression because the introduction of noise serves as implicit data augmentation and regularization, which compels the model to focus on the underlying data manifold (Karras et al., 2022; Chen et al., 2023) instead of memorizing specific instance values.

The reconstruction objective $\mathcal{L}_{\text{rec}}$ is formalized using a denoising score matching loss:

$$\mathcal{L}_{\text{rec}} = \mathbb{E} [\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}I_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t, Z)\|_2^2], \quad (6)$$

where $t \sim \mathcal{U}(0, 1)$ is the sampled timestep, I_0 denotes the visual representations encoded by the Image Tokenizer, $\epsilon \sim \mathcal{N}(0, I)$ represents the injected target noise, $\bar{\alpha}_t$ is the predefined noise schedule coefficient, and ϵ_θ is a lightweight denoising U-Net (Ronneberger et al., 2015) conditioned on all the latent tokens Z .

The complete training objective combines these losses with the standard next-token prediction loss:

$$\mathcal{L}_{\text{ntp}} = - \sum_{i=1}^{|Y|} \log P(y_i | V_T, T, Z, y_{<i}), \quad (7)$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ntp}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}}, \quad (8)$$

where λ_{rec} is hyperparameter.

4 Experiments

4.1 Datasets and Evaluation

To assess the performance of the proposed LaRe framework, we select seven benchmarks covering general multimodal understanding, fine-grained perception, hallucination detection, and complex reasoning. For general reasoning, we utilize MM-Bench (Liu et al., 2024) for comprehensive multimodal evaluation and ScienceQA (Lu et al., 2022) for interdisciplinary scientific reasoning, as these tasks require stable logical chains and effective vision-language alignment. To evaluate the refocusing capabilities of the model, we introduce MMVP (Tong et al., 2024) and V^* (Wu and Xie, 2024) for fine-grained perception and spatial reasoning, where V^* specifically targets visual grounding within complex environments. For complex logic and robustness, we include MMStar (Chen et al., 2024) to assess high-quality visual reasoning and MMMU-Pro (Yue et al., 2025) to examine multi-step logical inference in complex scenarios with significant modality gaps. Finally, we incorporate POPE (Li et al., 2023b) to measure model hallucination, ensuring a rigorous and thorough evaluation. Evaluation metrics include answer accuracy (Acc.) and the number of tokens generated when producing a correct answer (# Tokens).

4.2 Baselines and Implementation Details

We compare LaRe with seven representative methods, including standard Chain of Thought reasoning methods that generate final answers through explicit step by step reasoning, text only multimodal reasoning approaches such as MM-CoT (Zhang et al., 2023) and CCoT (Mitra et al., 2024), Thinking with Images paradigms including ICoT (Gao et al., 2025) and CoF (Zhang et al., 2025a), and latent space reasoning paradigms including MCOU (Pham and Ngo, 2025) and LVR (Li et al., 2025a).

Both the proposed method and the baselines are built upon the Qwen3-VL-2B/4B-Instruct (Bai et al., 2025) backbones. To ensure a fair comparison, we conduct single-stage supervised fine-tuning on these models using the Visual-CoT (Shao et al., 2024a) dataset, and only the post-training stages for CoF and LVR follow their original configurations. Throughout the training procedure, the visual encoder and the multimodal projector remain frozen, and only the parameters of the LLM are updated. More implementation details for LaRe are provided in the Appendix A.

4.3 Main Results

We evaluate LaRe on multiple multimodal benchmarks, comparing it against existing explicit reasoning methods and recent latent space reasoning approaches. Table 1 summarizes the results.

Improved Accuracy. LaRe achieves the highest accuracy on most benchmarks. At the 2B scale, it attains 82.5% on ScienceQA and 61.2% on MMStar, outperforming competitive baselines including CoF and LVR. While CoF maintains a slight advantage on MMBench within the 2B backbone setting, LaRe demonstrates superior robustness on perception-intensive tasks such as MMVP and V^* , surpassing the runner-up by approximately 3.2 and 3.9 percentage points respectively. Specifically, LaRe surpasses latent reasoning baselines (MCOU, LVR) by introducing dynamic refocusing for complex perception. Unlike image-injection methods like CoF, LaRe operates within the latent space to precisely capture fine-grained evidence, mitigating information loss from the modality gap.

Token Efficiency. LaRe demonstrates significant advantages in token efficiency. Unlike standard CoT or MM-CoT methods that typically require more than 150 tokens for inference, LaRe consistently maintains a token count below 100. For

Table 1: Comparison to various multimodal reasoning baselines. The best results are **bolded**. *Unlike LaRe and the other baseline methods, the post training of CoF and LVR follows their original settings.

Method	MMBench		MMStar		MMVP		ScienceQA		POPE		V*		MMMU-Pro	
	Acc.	# Tokens	Acc.	# Tokens	Acc.	# Tokens	Acc.	# Tokens	Acc.	# Tokens	Acc.	# Tokens	Acc.	# Tokens
<i>Base VLM: Qwen3-VL-2B-Instruct</i>														
Standard CoT	71.4	167.4	56.6	192.1	18.4	126.5	74.4	182.5	58.2	104.2	75.0	181.2	37.0	224.2
MM-CoT (Zhang et al., 2023)	72.7	158.3	57.8	120.2	24.0	107.7	80.1	195.7	51.2	100.9	66.5	192.3	39.8	198.8
CCoT (Mitra et al., 2024)	69.9	115.7	56.7	174.4	37.5	128.6	79.9	200.9	63.3	110.1	75.2	221.4	35.5	203.5
ICoT (Gao et al., 2025)	71.5	133.4	53.3	164.1	40.3	108.7	77.0	174.2	65.9	104.7	76.3	199.2	40.8	214.7
CoF* (Zhang et al., 2025a)	73.2	162.6	58.1	208.9	50.1	137.3	72.5	195.7	65.7	128.9	77.3	255.9	41.2	274.0
MCOU (Pham and Ngo, 2025)	62.6	65.3	58.2	95.4	37.3	67.9	76.8	104.3	64.0	37.2	76.5	99.2	39.0	120.1
LVR* (Li et al., 2025a)	67.5	75.0	56.6	108.2	43.9	72.4	79.2	93.8	64.5	29.5	77.4	112.7	38.9	134.7
LaRe (Ours)	71.9	34.6	61.2	70.5	53.3	27.2	82.5	55.0	68.0	18.9	81.3	87.2	44.2	102.4
<i>Base VLM: Qwen3-VL-4B-Instruct</i>														
Standard CoT	81.7	165.2	65.3	217.0	23.7	146.2	79.5	186.2	55.1	99.5	80.0	183.4	52.4	226.8
MM-CoT (Zhang et al., 2023)	75.9	170.6	67.8	210.9	27.5	113.0	82.7	203.1	55.7	102.7	74.5	187.2	49.2	204.2
CCoT (Mitra et al., 2024)	80.5	125.4	64.2	230.6	39.3	140.9	83.4	240.2	69.1	112.8	81.5	203.4	55.5	234.9
ICoT (Gao et al., 2025)	76.9	162.5	62.7	200.7	46.0	122.7	81.7	172.6	66.3	103.7	80.4	195.0	52.3	211.7
CoF* (Zhang et al., 2025a)	78.6	190.8	70.6	224.4	51.8	145.8	80.6	205.5	71.0	164.1	83.7	271.1	53.7	269.1
MCOU (Pham and Ngo, 2025)	73.8	72.3	61.4	145.5	45.7	69.2	78.8	97.5	64.5	56.0	82.5	104.8	52.0	124.5
LVR* (Li et al., 2025a)	79.0	80.7	68.5	125.3	56.2	72.9	80.3	101.4	69.4	42.4	83.7	110.9	53.9	144.4
LaRe (Ours)	82.1	49.0	72.8	82.7	57.6	32.3	87.5	65.2	77.4	29.0	87.0	93.5	58.4	103.5

Table 2: Comparison to state-of-the-art baselines. The best results are **bolded** and the second best results are underlined.

Method	MMBench	MMStar	MMVP	POPE	ScienceQA	V*	MMMU-Pro
Gemini2.5-Pro (Comanici et al., 2025)	84.2	74.7	45.5	–	–	81.3	59.6
o3 (OpenAI, 2025)	–	<u>75.3</u>	–	–	–	82.7	–
Janus-pro-7B (Chen et al., 2025a)	79.2	41.0	30.7	87.4	–	65.6	38.1
InternVL3.5-8B (Wang et al., 2025a)	79.5	69.3	50.6	<u>88.7</u>	79.3	66.2	50.7
Qwen3-VL-8B-Thinking (Bai et al., 2025)	87.5	75.2	–	–	85.7	77.5	60.4
<i>Base VLM: Qwen3-VL-8B-Instruct</i>							
Vision-R1 (Huang et al., 2025)	82.1	72.6	47.8	86.4	91.5	71.4	57.6
Thyme (Zhang et al., 2025b)	87.2	65.9	66.7	88.5	88.9	85.2	<u>64.8</u>
CaplImagine (Li et al., 2026)	86.8	74.0	72.4	86.3	90.2	<u>87.9</u>	59.4
LVR (Li et al., 2025a)	83.2	67.1	<u>72.7</u>	85.8	88.3	87.5	60.1
CrystaL (Zhang et al., 2026)	83.1	73.2	70.7	85.3	88.4	84.3	62.8
LaRe (Ours)	<u>87.4</u>	77.1	74.0	89.6	91.5	92.7	67.1

instance, on the POPE benchmark with a 2B backbone, LaRe utilizes only 18.9 tokens, which represents a nearly fivefold reduction compared to the 104.2 tokens consumed by standard CoT. This trend persists at the 4B scale, where LaRe achieves higher token efficiency than other latent space methods such as MCOU and LVR while preserving high accuracy. The reduction in token usage decreases operational costs while enhancing the efficiency of model inference. This advancement offers significant benefits for real-world applications that prioritize cost, speed, and scalability. In the Appendix B.2, we further analyze the inference latency of various methods. LaRe maintains excellent inference efficiency, which further demonstrates the efficiency of refocusing in the latent space.

4.4 Comparison With State-of-the-Arts

Illustrated in Table 2, we compare our LaRe based on Qwen3-VL-8B-Instruct (Bai et al., 2025) with both private models (Comanici et al., 2025; OpenAI, 2025), open-sourced alternatives (Chen et al., 2025a; Wang et al., 2025a; Bai et al., 2025), and some multimodal reasoning methods (Huang et al., 2025; Li et al., 2026; Zhang et al., 2025b; Li et al., 2025a; Zhang et al., 2026). Experimental results demonstrate that LaRe achieves superior performance across various challenging multimodal benchmarks and surpasses existing state-of-the-art baselines. On the MMMU-Pro task which evaluates complex visual reasoning, LaRe not only outperforms open-source reasoning models of a similar scale such as Qwen3-VL-8B-Thinking but also

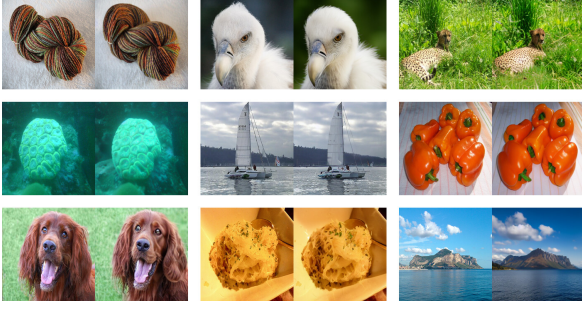


Figure 3: Visual reconstructions from latent sequences. In each panel, the left image shows the original input (from the ImageNet-1K validation set), and the right image shows its reconstruction.

exceeds the performance of larger closed-source models like Gemini 2.5 Pro and o3. Moreover, in comparison with other approaches based on the same VLM backbone such as Thyme and LVR, the performance lead of LaRe on the V^* benchmark confirms that the latent refocusing mechanism effectively extracts fine-grained visual evidence and mitigates the modality gap.

4.5 In-depth Analysis

Attention Analysis. We compute the mean attention scores from all latent tokens to all visual tokens using the MMStar dataset. We employ T-test (Student, 1908) to compare the means and the Mann-Whitney U test (Mann and Whitney, 1947) to compare the medians of the two distributions. As shown in Table 6, LaRe achieves higher attention scores than other latent reasoning methods. This indicates that LaRe effectively guides the model to focus more on the input visual signals.

Quantitative Analysis of Visual Refocusing. Unlike existing latent space reasoning methods such as LVR, our Extractor explicitly refocuses on visual evidence at every latent step through per-step cross-attention to V_T , rather than merely propagating a static visual context. To quantitatively verify this behavior, we introduce Visual Attention Entropy Drift (VAED) and two complementary diagnostics.

Let $\alpha^{(k)} \in \mathbb{R}^{|V_T|}$ denote the normalized cross-attention distribution over visual tokens at the k -th latent step, averaged over heads and latent positions. We define

$$\text{VAED} = \frac{1}{K-1} \sum_{k=2}^K \left| H\left(\alpha^{(k)}\right) - H\left(\alpha^{(k-1)}\right) \right|, \quad (9)$$

where $H(\cdot)$ denotes Shannon entropy. A low VAED indicates nearly stationary visual attention, whereas a high VAED reflects substantial redistribution of attention across reasoning steps. For methods without explicit per-step cross-attention, such as LVR, we derive an implicit attention distribution from the cosine similarity between latent states and visual token embeddings, followed by softmax normalization.

To distinguish purposeful refocusing from random attention reshuffling, we further measure the Top-5 coverage shift $\Delta_{\text{top-5}}$, defined as the mean Jaccard distance between Top-5 attended visual token sets across consecutive steps, and the average attention entropy $\bar{H}$, which characterizes per-step attention selectivity. We evaluate all methods on 200 randomly sampled instances from MMStar and V^* using the same Qwen3-VL-2B-Instruct backbone.

As shown in Table 3, LaRe achieves a VAED of 0.31, approximately $6\times$ that of MCOU (0.05) and $10\times$ that of LVR (0.03), demonstrating substantially greater attention migration across latent steps. This migration is also spatially meaningful: LaRe achieves a Top-5 coverage shift of 0.42, compared with only 0.08 for MCOU, indicating that it repeatedly switches the image regions being attended to. Importantly, this stronger drift is accompanied by *lower* average attention entropy (2.14 vs. 3.07 for MCOU), showing that attention remains selective rather than diffusely distributed. Thus, the combination of high inter-step drift and low per-step entropy provides evidence that LaRe performs purposeful, evidence-anchored visual refocusing rather than random attention reshuffling.

Taken together, these results support that LaRe’s performance gains stem from its structurally distinct mechanism of iterative and selective visual refocusing, rather than merely from introducing cross-attention.

4.6 Qualitative Results

Visualization of Each Iteration. To intuitively illustrate the refocusing process of LaRe, Figure 4 presents a side by side visualization of the attention heatmaps produced by the Extractor module and the corresponding reasoning text trajectories across four consecutive iterations. We observe a highly consistent coevolution between the model’s visual focus regions and the process of textual evidence accumulation. In the first step, attention is primarily concentrated on the most salient foreground

Method	Structural Refocusing?	VAED $\uparrow$	$\Delta_{\text{top-5}} \uparrow$	$\bar{H} \downarrow$	Acc. (MMStar)
Standard CoT	N/A	—	—	—	56.6
MCOUT	$\times$	0.05	0.08	3.07	58.2
LVR	$\times$	0.03	0.05	3.18	56.6
MCOUT + Cross-Attn [†]	Partial	0.12	0.19	2.61	59.6
LaRe (Ours)	$\checkmark$	0.31	0.42	2.14	61.2

Table 3: Quantitative analysis of visual refocusing. Higher VAED and Top-5 shift indicate stronger visual refocusing, while lower $\bar{H}$ indicates more selective attention. [†] MCOUT augmented with per-step cross-attention to V_T without the diffusion-style reconstruction objective.

Table 4: Ablation study of LaRe on Qwen3-VL-4B-Instruct backbone. The results demonstrate that each component contributes critically to the full model’s performance and efficiency.

Methods	Acc.	# Tokens
LaRe	74.7	65.0
w/o Extractor	69.3 (-5.4)	89.1 (+24.1)
w/o Semantic Augmentation	67.2 (-7.5)	68.5 (+3.5)
w/o Extractor & Semantic Augmentation	64.5 (-10.2)	95.9 (+30.9)

figures: the catcher and the umpire, which corresponds to the text marked in yellow. As reasoning progresses, the model refocuses on secondary regions: Step 2 localizes the batter and the players inside the dugout, corresponding to the blue text. In Steps 3–4, the model further refocuses on relatively concealed figures in the background, such as staff members dressed in black (corresponding to the brown text) and individuals seated near the entrance of the dugout (corresponding to the purple text). Through this sequential refocusing, the model arrives at the correct answer, indicated by the green text. This progressive transition from coarse grained salient targets to fine grained background details clearly indicates that LaRe does not perform a static single pass recognition, but instead implements a structured multi step evidence focusing process. Crucially, a high degree of synchronization is observed between the attention heatmaps and the color coded reasoning text. Each segment of textual reasoning, such as “two staff members dressed in black watching the game,” corresponds to the visual region emphasized at the current step.

From Latent to Images. Through image reconstruction visualization in Figure 3, we show that a U-net fine-tuned for five epochs on ImageNet-1K (Deng et al., 2009) can successfully recover clear scenes from the latent sequences. This result provides direct evidence that the Latent Refocusing process indeed focuses on visual evidence through

Table 5: Ablation results on Semantic Augmentation strategies. "Feature" refers to visual tokens from the image tokenizer, and "Pixel" refers to raw image pixels.

Variant	MMStar	MMU-Pro	MMVP	Avg.
LaRe	72.8	58.4	57.6	62.9
Regression-Pixel	69.2	55.1	52.4	58.9
Regression-Feature	70.8	56.4	54.1	60.4
Denoising-Pixel	71.5	57.2	55.9	61.5
w/o $\mathcal{L}_{rec}$	68.5	54.6	51.8	56.6

out the entire reasoning process.

4.7 Ablation Study

Role of Each Module. As shown in Table 4, the full LaRe configuration achieves the most competitive performance, striking a superior balance between reasoning accuracy (74.7%) and inference efficiency (65.0 tokens). The removal of Semantic Augmentation results in the sharpest decline in accuracy (-7.5%), highlighting its indispensable role in regularizing the latent space and preserving the fine-grained visual-semantic fidelity required for complex logic. Meanwhile, discarding the Extractor leads to a 5.4% drop in accuracy and a massive surge in token consumption (+24.1), demonstrating that the iterative refocusing mechanism is vital for pinpointing reasoning-relevant visual cues and preventing the model from generating verbose, redundant outputs. Finally, the simultaneous removal of both components leads to a compounded performance degradation, with accuracy falling to 64.5% and token usage spiking to 95.9, which validates the synergy between latent space structuring and dynamic refocusing in our proposed paradigm.

Impact of Reconstruction Objective. Table 5 presents the impact of different reconstruction objectives within the semantic augmentation strategy. Experimental results indicate that the removal of the reconstruction loss $\mathcal{L}_{rec}$ leads to a significant performance decline across all benchmarks, where

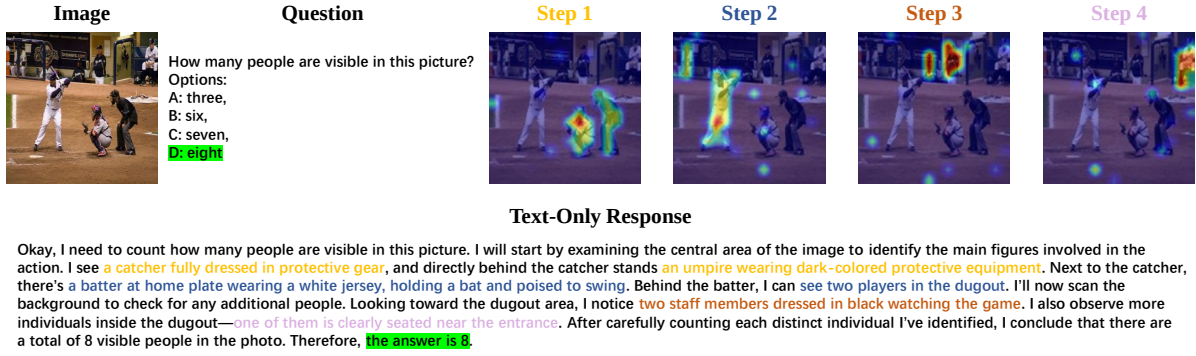


Figure 4: The evolution of attention in LaRe over four iterations and its alignment with the text-only reasoning trajectory.

the average accuracy drops from 62.9% to 56.6%. This observation underscores the role of the visual reconstruction task in guiding the model to establish structured latent space representations, which ensures that latent variables retain essential visual evidence throughout multiple reasoning iterations. Moreover, the feature-space denoising reconstruction mechanism utilized in LaRe demonstrates a clear advantage over ablation variants based on raw pixels or the standard regression objective. This suggests that the implicit regularization introduced by the denoising task allows the model to capture high-level spatial logic and fine-grained semantics more effectively, which substantially improves the reliability of multimodal reasoning.

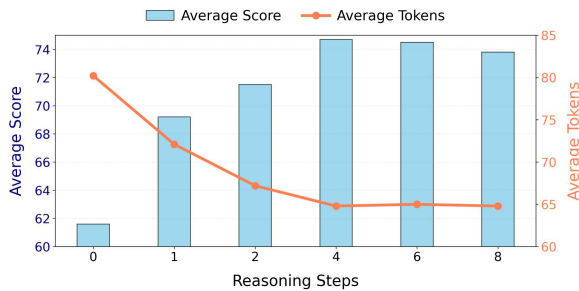


Figure 5: Impact of varying the number of reasoning steps on model performance and output efficiency.

Table 6: Quantitative comparison on attention values. The results indicated that LaRe achieves higher values than other latent reasoning methods.

Statistic ($\times 10^{-4}$)	MCOUT	LVR	LaRe	P-value
Mean	4.32	7.34	10.94	1.07×10^{-7}
25th Percentile	3.90	6.63	10.06	—
Median	4.29	7.33	10.87	3.79×10^{-9}
75th Percentile	4.71	8.01	11.79	—
95th Percentile	5.36	9.11	13.16	—

Impact of Reasoning Steps. The ablation study on the number of iterative steps (Figure 5) reveals a clear pattern of diminishing returns. Increasing the initial number of steps markedly improves performance and substantially reduces output length, validating the effectiveness of the iterative refocusing process. However, once the number of steps reaches a saturation threshold, both performance gains and output compression plateau. This finding indicates that the model can attain a stable reasoning state and offers guidance on the optimal number of steps to balance computational cost and reasoning effectiveness.

5 Conclusion

We introduced Latent Refocusing (LaRe), a multimodal reasoning paradigm in which visual refocusing takes place entirely within the latent space. LaRe realizes this idea through an iterative latent refocusing mechanism and a semantic augmentation training strategy. At each step, a lightweight Extractor refocuses on reasoning-relevant visual evidence conditioned on the current reasoning state, producing latent tokens that guide subsequent reasoning. A diffusion-based visual reconstruction objective then shapes the latent space into a semantically structured representation that retains both high-level semantics and fine-grained visual fidelity. Across seven multimodal benchmarks, LaRe improves average accuracy by 7.6% over strong baselines while reducing inference token consumption by 59.7%. These results point to a broader direction for multimodal reasoning: the latent space offers a productive venue for jointly achieving visual grounding and efficiency without committing to explicit visual tokens.

Limitations

A primary limitation of the current LaRe framework is its reliance on a predefined, static number of reasoning iterations K . While our empirical analysis identifies an optimal saturation threshold for performance, this "one-size-fits-all" approach does not account for the varying complexity across different multimodal tasks. For relatively straightforward visual queries, a fixed number of latent iterations may introduce redundant computational overhead and unnecessary latency. Conversely, for exceptionally intricate logical reasoning tasks, the model's performance may still be constrained by the preset iteration limit. Developing a dynamic halting mechanism, which allows the model to adaptively determine the required number of latent thoughts based on the evolving reasoning state, remains an important direction for further enhancing the efficiency and flexibility of the LaRe paradigm.

Ethical Considerations

The datasets utilized in this study are publicly available benchmarks widely recognized in the research community. These datasets consist of images and textual annotations sourced from public domains and are used strictly for academic evaluation. To the best of our knowledge, all visual and textual data have been curated to ensure they do not contain Personally Identifiable Information (PII) or sensitive personal content. Our experimental procedures strictly adhere to the licensing agreements of the respective datasets and align with the ethical guidelines for responsible artificial intelligence research. Therefore, we believe that this work fully complies with the ethical standards of the conference.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No.62176252).

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. [Qwen-vl: A frontier large vision-language model with versatile abilities](#). *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. [Qwen3-vl technical report](#). *arXiv preprint arXiv:2511.21631*.

Andrea Banino, Jan Balaguer, and Charles Blundell. 2021. [Pondernet: Learning to ponder](#). *arXiv preprint arXiv:2107.05407*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. [Language models are few-shot learners](#). *Advances in neural information processing systems*, 33:1877–1901.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024. [Are we on the right way for evaluating large vision-language models?](#) *Advances in Neural Information Processing Systems*, 37:27056–27087.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. 2023. [Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data](#). In *International Conference on Machine Learning*, pages 4672–4712. PMLR.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025a. [Janus-pro: Unified multimodal understanding and generation with data and model scaling](#). *arXiv preprint arXiv:2501.17811*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025b. [Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling](#). *Preprint*, arXiv:2412.05271.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *arXiv preprint arXiv:2507.06261*.

Corinna Cortes and Vladimir Vapnik. 1995. [Support-vector networks](#). *Machine learning*, 20(3):273–297.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. [Imagenet: A large-scale hierarchical image database](#). In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

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

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. 2025. [Interleaved-modal chain-of-thought](#). In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19520–19529.

- Zhuocheng Gong, Jian Guan, Wei Wu, Huishuai Zhang, and Dongyan Zhao. 2025. [Latent preference coding: Aligning large language models via discrete latent codes](#). *arXiv preprint arXiv:2505.04993*.
- Alex Graves. 2016. [Adaptive computation time for recurrent neural networks](#). *arXiv preprint arXiv:1603.08983*.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. [Training large language models to reason in a continuous latent space](#). *arXiv preprint arXiv:2412.06769*.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. [Masked autoencoders are scalable vision learners](#). In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.
- David Herel and Tomas Mikolov. 2024. [Thinking tokens for language modeling](#). *arXiv preprint arXiv:2405.08644*.
- Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. 2024. [Visual sketchpad: Sketching as a visual chain of thought for multimodal language models](#). *Advances in Neural Information Processing Systems*, 37:139348–139379.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. [Vision-r1: Incentivizing reasoning capability in multimodal large language models](#). *arXiv preprint arXiv:2503.06749*.
- Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. [Elucidating the design space of diffusion-based generative models](#). *Advances in neural information processing systems*, 35:26565–26577.
- Diederik P Kingma and Max Welling. 2013. [Auto-encoding variational bayes](#). *arXiv preprint arXiv:1312.6114*.
- Deqian Kong, Minglu Zhao, Dehong Xu, Bo Pang, Shu Wang, Edouardo Honig, Zhangzhang Si, Chuan Li, Jianwen Xie, Sirui Xie, and 1 others. 2025. [Scalable language models with posterior inference of latent thought vectors](#). *arXiv e-prints*, pages arXiv–2502.
- Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. 2025a. [Latent visual reasoning](#). *arXiv preprint arXiv:2509.24251*.
- Hengli Li, Chenxi Li, Tong Wu, Xuekai Zhu, Yuxuan Wang, Zhaoxin Yu, Eric Hanchen Jiang, Song-Chun Zhu, Zixia Jia, Ying Nian Wu, and 1 others. 2025b. [Seek in the dark: Reasoning via test-time instance-level policy gradient in latent space](#). *arXiv preprint arXiv:2505.13308*.
- Jindong Li, Yali Fu, Li Fan, Jiahong Liu, Yao Shu, Chengwei Qin, Menglin Yang, Irwin King, and Rex Ying. 2025c. [Implicit reasoning in large language models: A comprehensive survey](#).
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. [Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *International conference on machine learning*, pages 19730–19742. PMLR.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. [Evaluating object hallucination in large vision-language models](#). *arXiv preprint arXiv:2305.10355*.
- You Li, Chi Chen, Yanghao Li, Fanhu Zeng, Kaiyu Huang, Jinan Xu, and Maosong Sun. 2026. [Imagination helps visual reasoning, but not yet in latent space](#). *Preprint*, arXiv:2602.22766.
- Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025d. [A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges](#). *Preprint*, arXiv:2501.02189.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). *Preprint*, arXiv:2304.08485.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024. [Mmbench: Is your multi-modal model an all-around player?](#) In *European conference on computer vision*, pages 216–233. Springer.
- Ilya Loshchilov and Frank Hutter. 2017. [Decoupled weight decay regularization](#). *arXiv preprint arXiv:1711.05101*.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. [Learn to explain: Multimodal reasoning via thought chains for science question answering](#). *Advances in Neural Information Processing Systems*, 35:2507–2521.
- Henry B Mann and Donald R Whitney. 1947. [On a test of whether one of two random variables is stochastically larger than the other](#). *The annals of mathematical statistics*, pages 50–60.
- Fanxu Meng, Haotong Yang, Yiding Wang, and Muhan Zhang. 2023. [Chain of images for intuitively reasoning](#). *arXiv preprint arXiv:2311.09241*.
- Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. [Large language models: A survey](#). *arXiv preprint arXiv:2402.06196*.
- Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. [Compositional chain-of-thought prompting for large multimodal models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431.

- Amirkeivan Mohtashami, Matteo Pagliardini, and Martin Jaggi. 2025. [Cotformer: A chain-of-thought driven architecture with budget-adaptive computation cost at inference](#). In *The Thirteenth International Conference on Learning Representations*.
- OpenAI. 2025. Introducing openai o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>.
- OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, and 243 others. 2024. [Openai o1 system card](#). *Preprint*, arXiv:2412.16720.
- Tan-Hanh Pham and Chris Ngo. 2025. [Multimodal chain of continuous thought for latent-space reasoning in vision-language models](#). *arXiv preprint arXiv:2508.12587*.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. [High-resolution image synthesis with latent diffusion models](#). In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. [U-net: Convolutional networks for biomedical image segmentation](#). In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.
- Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. 2023. [Visual chain of thought: bridging logical gaps with multimodal infillings](#). *arXiv preprint arXiv:2305.02317*.
- David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. [Learning representations by back-propagating errors](#). *nature*, 323(6088):533–536.
- Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J Reddi. 2025. [Reasoning with latent thoughts: On the power of looped transformers](#). *arXiv preprint arXiv:2502.17416*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *arXiv preprint arXiv:1707.06347*.
- Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024a. [Visual cot: Advancing multimodal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning](#). *Advances in Neural Information Processing Systems*, 37:8612–8642.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024b. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *arXiv preprint arXiv:2402.03300*.
- Xuan Shen, Yizhou Wang, Xiangxi Shi, Yanzhi Wang, Pu Zhao, and Jiuxiang Gu. 2025a. [Efficient reasoning with hidden thinking](#). *arXiv preprint arXiv:2501.19201*.
- Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025b. [Codi: Compressing chain-of-thought into continuous space via self-distillation](#). *arXiv preprint arXiv:2502.21074*.
- Oscar SKEAN, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. [Layer by layer: Uncovering hidden representations in language models](#). *arXiv preprint arXiv:2502.02013*.
- Student. 1908. [The probable error of a mean](#). *Biometrika*, pages 1–25.
- DiJia Su, Hanlin Zhu, Yingchen Xu, Jiantao Jiao, Yuandong Tian, and Qingqing Zheng. 2025. [Token assorted: Mixing latent and text tokens for improved language model reasoning](#). *arXiv preprint arXiv:2502.03275*.
- Yuchang Sun, Yanxi Chen, Yaliang Li, and Bolin Ding. 2025. [Enhancing latent computation in transformers with latent tokens](#). *arXiv preprint arXiv:2505.12629*.
- Jihoon Tack, Jack Lanchantin, Jane Yu, Andrew Cohen, Ilia Kulikov, Janice Lan, Shibo Hao, Yuandong Tian, Jason Weston, and Xian Li. 2025. [Llm pre-training with continuous concepts](#). *arXiv preprint arXiv:2502.08524*.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. [Eyes wide shut? exploring the visual shortcomings of multimodal llms](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. [InternV3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency](#). *arXiv preprint arXiv:2508.18265*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. [Self-consistency improves chain of thought reasoning in language models](#). *Preprint*, arXiv:2203.11171.
- Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025b. [Multimodal chain-of-thought reasoning: A comprehensive survey](#). *arXiv preprint arXiv:2503.12605*.

- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). *Advances in neural information processing systems*, 35:24824–24837.
- Penghao Wu and Saining Xie. 2024. [V?: Guided visual search as a core mechanism in multimodal llms](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094.
- Peng Xu, Xiatian Zhu, and David A Clifton. 2023. [Multimodal learning with transformers: A survey](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132.
- Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025. [Softcot: Soft chain-of-thought for efficient reasoning with llms](#). *arXiv preprint arXiv:2502.12134*.
- Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. 2025. [Machine mental imagery: Empower multimodal reasoning with latent visual tokens](#). *arXiv preprint arXiv:2506.17218*.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of thoughts: Deliberate problem solving with large language models](#). *Advances in neural information processing systems*, 36:11809–11822.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2025. [Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186.
- Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and 1 others. 2025a. [Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl](#). *arXiv preprint arXiv:2505.15436*.
- Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. 2024. [Chain of preference optimization: Improving chain-of-thought reasoning in llms](#). *Advances in Neural Information Processing Systems*, 37:333–356.
- Yang Zhang, Danyang Li, Yuxuan Li, Xin Zhang, Tianyu Xie, Mingming Cheng, and Xiang Li. 2026. [Crystal: Spontaneous emergence of visual latents in mllms](#). *arXiv preprint arXiv:2602.20980*.
- Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, and 1 others. 2025b. [Thyme: Think beyond images](#). *arXiv preprint arXiv:2508.11630*.
- Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. 2025c. [Soft thinking: Unlocking the reasoning potential of llms in continuous concept space](#). *arXiv preprint arXiv:2505.15778*.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. [Multimodal chain-of-thought reasoning in language models](#). *arXiv preprint arXiv:2302.00923*.
- Xiongtao Zhou, Jie He, Lanyu Chen, Jingyu Li, Haojing Chen, Víctor Gutiérrez-Basulto, Jeff Z Pan, and Hanjie Chen. 2024. [Miceval: Unveiling multimodal chain of thought’s quality via image description and reasoning steps](#). *arXiv preprint arXiv:2410.14668*.
- Rui-Jie Zhu, Tianhao Peng, Tianhao Cheng, Xingwei Qu, Jinfa Huang, Dawei Zhu, Hao Wang, Kaiwen Xue, Xuanliang Zhang, Yong Shan, Tianle Cai, Taylor Kergan, Assel Kembay, Andrew Smith, Chenghua Lin, Binh Nguyen, Yuqi Pan, Yuhong Chou, Zefan Cai, and 14 others. 2025. [A survey on latent reasoning](#). *Preprint*, arXiv:2507.06203.

Appendix

Algorithm 1 Latent Refocusing (LaRe)

Require: Input image i , text prompt t , number of reasoning steps K , special tokens $\langle |bot| \rangle$ & $\langle |eot| \rangle$

Ensure: Final answer Y

- 1: $V_T \leftarrow \text{Projector}(\text{VisualEncoder}(i))$ $\triangleright$ Extract visual tokens
- 2: $T \leftarrow \text{Embed}(t)$ $\triangleright$ Project text to embedding space
- 3: $Z \leftarrow \emptyset$ $\triangleright$ Initialize latent token sequence
- 4: **for** $k = 1$ **to** K **do**
- 5: $H_T^{(k)} \leftarrow H_{\text{LLM}}([V_T; T; Z])$ $\triangleright$ Extract current reasoning state from LLM
- 6: $Z^{(k)} \leftarrow \text{Extractor}(H_T^{(k)}, V_T)$ $\triangleright$ Refocus on V_T and generate new latent token
- 7: $Z \leftarrow [Z; Z^{(k)}]$ $\triangleright$ Update the latent thought chain
- 8: **end for**
- 9: $\text{Input}_{\text{final}} \leftarrow [V_T; T; \langle |bot| \rangle; Z; \langle |eot| \rangle]$ $\triangleright$ Concatenate all modalities
- 10: $Y \leftarrow \text{LLM}_{\text{generate}}(\text{Input}_{\text{final}})$ $\triangleright$ Autoregressive answer generation
- 11: **return** Y

A Implementation Details

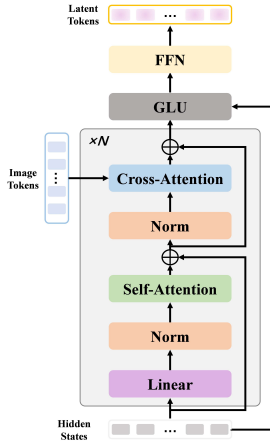


Figure 6: The detailed architecture of the Extractor. The module iteratively refines features through self-attention and cross-attention layers before passing them through a GLU and FFN.

Architecture of Extractor. The Extractor module serves as the core component for executing iterative refocusing and adopts a Transformer-like architectural design, as shown in figure 6. It consists of 2 iterative layers. In each layer, features

are first preprocessed through a linear transformation, followed by a multi-head self-attention mechanism with 12 attention heads to model the evolution of internal reasoning dynamics. On this basis, a cross-attention mechanism, also configured with 12 heads, is applied over visual features to perform precise refocusing, thereby extracting evidence that is highly relevant to the current reasoning step. To ensure sufficient representational capacity, the hidden dimension of the Extractor is uniformly set to 1536. After completing the iterative process, the refined features are sequentially passed through a gated linear unit (GLU) and a feed-forward network (FFN), ultimately producing a latent representation that drives subsequent reasoning.

Settings of Image Reconstruction. The diffusion process employs a linear noise schedule defined by $\beta_t \in [\beta_{start} = 1 \times 10^{-4}, \beta_{end} = 0.02]$ over $T = 1000$ iterations. At each training step, a random timestep $t \sim U(1, T)$ is sampled, and the input image x_0 is corrupted as $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon$, where $\alpha_t = \prod_{s=1}^t (1 - \beta_s)$. The sampling process follows an iterative denoising procedure using the predicted noise and the reparameterization of the posterior distribution. We adopt KL-16 (Rombach et al., 2022) as the image tokenizer, which is a continuous VAE (Kingma and Welling, 2013) with a Kullback–Leibler (KL) divergence regularization term. The architecture of our denoising U-Net is illustrated in Table 7.

Other Hyperparameters. We set the number of iterations K to 4 for LaRe. The training process employs the AdamW optimizer (Loshchilov and Hutter, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.95$ and weight decay of 0.05. The loss coefficients are set to $\lambda_{rec} = 1.0$ for the semantic augmentation, determined through grid search on the validation set. All models are trained using mixed precision (FP16) on eight NVIDIA A100 80GB GPUs with a global batch size of 128. To address GPU memory limitations, we implement gradient accumulation over four steps, which results in a micro-batch size of 4 per GPU. The memory workload is uniformly distributed across all devices through the integration of DeepSpeed ZeRO-3.

Image Resolution Configurations For image resolution configurations, all input images are resized to 224×224 pixels to match the standard input requirements of the CLIP-ViT-L/14 visual encoder. This resolution yields 256 visual tokens after

Table 7: U-Net configuration.

Parameter	Value
sample_size	img_size (input spatial size)
in_channels	self.channel (input feature channels)
out_channels	self.channel (output feature channels)
layers_per_block	3
block_out_channels	(96, 192, 384, 512)
down_block_types	(DownBlock2D, CrossAttnDownBlock2D, CrossAttnDownBlock2D, DownBlock2D)
up_block_types	(UpBlock2D, CrossAttnUpBlock2D, CrossAttnUpBlock2D, UpBlock2D)
cross_attention_dim	768
attention_head_dim	64
norm_num_groups	32

processing through the vision encoder (a 16×16 grid of patch tokens). The diffusion-based reconstructor operates on latent representations with a feature map shape of $[4, 28, 28]$, corresponding to the compressed spatial dimensions while preserving essential visual information. During reconstruction evaluation, the output images are upsampled back to the original 224×224 resolution for qualitative assessment.

Settings of Probing Analysis. To rigorously evaluate whether the latent tokens Z effectively encode joint linguistic reasoning and visual semantics, we conducted a comprehensive probing analysis following established methodologies in representation learning. The pooled latent vectors $z_{\text{agg}} = \phi_{\text{pool}}(Z)$, textual rationale embeddings $r = \phi_{\text{pool}}(R)$, and visual anchors $v = \phi_{\text{pool}}(V)$ were computed using the same hybrid pooling module (mean-max pooling) employed during training. All representations were subsequently projected to a 256-dimensional space via a lightweight, trainable linear layer and L2-normalized prior to classification to ensure scale-invariant comparison. For binary classification, positive pairs were constructed from matching (image, question) instances, while negative pairs were sampled in a balanced manner from three distinct strategies to prevent trivial solutions: (i) different-question/same-image, (ii) different-image/same-question, and (iii) random in-batch negatives. The final probing dataset comprised 1,000 labeled pairs (500 positive/500 negative) with a stratified 70/30 train/test split based on negative sampling type. We employed two distinct probe architectures: a Linear Support Vector Machine (SVM) with a linear kernel optimized via 5-fold cross-validation on the training set, and a single-layer Multi-Layer Perceptron (MLP) implemented in PyTorch consisting of one hidden layer, optimized with Adam ($\text{lr} = 1 \times 10^{-3}$) and early

stopping based on validation loss.

B Some Supplement Experiments Results

B.1 More Ablations

Table 8: Ablation study on Extractor design. "Dynamic" indicates using $H_T^{(k)}$ as the query source. All variants are evaluated on the 4B backbone.

Variant	Component	Acc.	# Tokens
LaRe	Full	74.7	65.0
w/o Self-Attn	Only Cross-Attn	71.1	72.7
Linear Proj.	Linear Bottleneck	66.5	82.3
GAP	Global Avg Pooling	67.8	80.1
Fixed Query	Learnable Latents	72.9	67.5

Ablation Study on Extractor Architecture. We investigate the architectural design of the Extractor through a series of ablation experiments on a 4B parameter backbone, as summarized in Table 8. The results demonstrate that the full LaRe configuration provides an optimal trade-off between inference accuracy (74.7%) and token consumption (65.0 tokens). Specifically, the omission of the self-attention mechanism within the Extractor leads to a significant 3.6% decrease in accuracy alongside an increase in the number of inference tokens, which confirms that logic modeling within the latent space is critical for maintaining reasoning coherence. Furthermore, the performance degradation observed when replacing the Transformer structure with simpler alternatives, such as a linear projection layer or global average pooling, indicates that basic bottleneck structures are insufficient for capturing the complex high-dimensional mappings between multimodal features, as accuracy drops to 66.5% and 67.8% respectively. Finally, a comparison with a fixed query baseline shows that employing learnable latent variables instead of dynamic queries

Table 9: Performance comparison of adaptive reasoning mechanisms using different convergence criteria.

Methods	Average Steps	Acc.	# Tokens
LaRe	4.0	74.7	65.0
+ Adaptive (Cosine)	5.9	74.5	67.8
+ Adaptive (L1 Distance)	2.8	73.1	71.3
+ Adaptive (L2 Distance)	2.9	74.2	70.5

driven by $H_T^{(k)}$ results in a 1.8% reduction in performance. This finding supports the central design principle that dynamic refocusing based on the current reasoning state is essential for the precise extraction of visual evidence.

Choices of the Image Tokenizer. To evaluate the influence of image tokenizer selection within the semantic augmentation strategy on reasoning performance, Table 10 presents a comparison of model performance across various pre-training paradigms. Experimental results indicate that KL-16, which utilizes a hybrid convolutional architecture, outperforms other variants across all metrics. It achieves an average accuracy of 62.9%, which markedly exceeds the generative task based MAE (59.3%), the discriminative DINOv2 (59.7%), and the contrastive learning model SigLIP (61.0%). In particular, KL-16 demonstrates a substantial lead on the MMMU-Pro (58.4%) and MMVP (57.6%) benchmarks, both of which require sophisticated visual logic. These results verify that the latent representations from KL-16 maintain a superior balance between high-level semantic consistency and low-level visual details during the denoising reconstruction process compared to pure semantic or pixel-level objectives. As a result, this provides high-quality supervision signals for the latent refocusing mechanism and enables the model to capture fine-grained visual evidence essential for multimodal reasoning.

Adaptive Reasoning Steps. To enhance efficiency, we investigate an adaptive stopping mechanism that determines the termination of reasoning based on the convergence of latent tokens. We evaluate this mechanism by tracking token evolution throughout the iterative process using cosine similarity, L_1 distance, and L_2 distance. As illustrated in Table 9, these metrics reveal a convergence trend where inter-step distances decrease monotonically as reasoning progresses. This observation confirms the intrinsic stability of latent reasoning and sup-

ports the conclusions drawn from the fixed-step ablation study in Figure 5. However, the results also demonstrate a significant trade-off between efficiency and performance. While adaptive stopping potentially mitigates redundancy and excessive computation, it results in a slight decline in overall performance metrics. We hypothesize that these heuristic adaptive stopping mechanisms are insufficient for accurately estimating the number of steps required for reasoning, and may instead lead to underthinking or overthinking. Consequently, future research focuses on developing more refined stopping criteria to achieve truly efficient adaptive inference without compromising model performance.

Schedule of β in Image Reconstruction. We investigate the sensitivity of Semantic Augmentation to the noise schedule β used in the diffusion-based reconstruction objective. As shown in Table 11, we compare linear, cosine, and squared cosine schedules across different maximum β values. The cosine schedule with $\beta_{\max} = 0.02$ achieves the best balance between reconstruction quality and reasoning performance, yielding the highest average score of 74.7%. While linear schedules demonstrate competitive performance, they exhibit slightly higher variance across tasks. The squared cosine schedule shows comparable results but requires more careful tuning of the maximum β value. These results indicate that LaRe is reasonably robust to different schedule choices, though smoother transitions in the noise schedule (as in cosine) appear to provide more stable training dynamics for the latent reasoning process.

On the choice of l . Setting l to the number of LLM layers is not a purely heuristic choice. Previous work (Skean et al., 2025) shows that different layers of an LLM encode semantically distinct representations. Inspired by this, our layer-aligned design assigns a dedicated latent token as a semantic anchor to the hidden state of each layer, enabling the extractor to capture reasoning cues at every level of abstraction. To further validate this design, we conduct an ablation study on l using Qwen3-VL-2B-Instruct, which has 28 LLM layers, where hidden states are selected sequentially starting from the last layer. As shown in Table 12, $l = L$ achieves the highest accuracy with only negligible additional latency compared to smaller l , indicating that aligning the number of latent tokens with the number of LLM layers yields the best

Table 10: Ablation study of different Image Tokenizers for Semantic Augmentation. All variants are trained using the same denoising reconstruction objective. The results demonstrate that KL-16 provides the most effective latent representations for multimodal reasoning.

Tokenizer	Pre-train Task	MMStar	MMMU-Pro	MMVP	Avg.
MAE	Generative	69.2	53.5	55.1	59.3
CLIP-L/336	Contrastive	72.5	55.8	51.2	59.2
DINOv2	Discriminative	71.1	54.2	55.8	59.7
SigLIP-L/384	Contrastive	72.8	56.9	54.6	61.0
KL-16 (Ours)	Hybrid/Conv	72.8	58.4	57.6	62.9

Table 11: Performance comparison of different β schedules in diffusion-based reconstruction.

Methods	Avg. Score	Avg. Tokens
Linear ($\beta_{\max} = 0.02$)	74.7	65.0
Linear ($\beta_{\max} = 0.05$)	73.9	65.7
Cosine ($\beta_{\max} = 0.02$)	74.7	65.2
Cosine ($\beta_{\max} = 0.05$)	74.1	64.8
Squared Cosine ($\beta_{\max} = 0.02$)	72.4	66.1
Squared Cosine ($\beta_{\max} = 0.05$)	73.6	65.3

accuracy-efficiency trade-off.

Table 12: Ablation study of the number of latent tokens l on Qwen3-VL-2B-Instruct.

l	Average Accuracy (%)	Token Count	Latency (s/step)
$L/4$ (7)	63.2	232	0.42
$L/2$ (14)	64.8	225	0.44
L (28)	66.2	218	0.47

B.2 More Analysis

Analysis of Computational Costs. To assess the computational feasibility of the LaRe paradigm, Table 13 provides a detailed comparison of training overhead across various model scales. The experimental results demonstrate that the computational burden introduced by the Extractor module and the reconstruction-based semantic augmentation remains marginal in comparison to the significant performance gains. In the case of the 2B backbone, the trainable parameters of the full LaRe model increase by approximately 2.9%, expanding from 1.72B to 1.77B, while the total training duration extends by only 1.2 hours with an iteration speed maintained at 2.88 seconds per step. Regarding GPU memory utilization, LaRe requires an additional 1.83 GB for the 2B scale and 3.05 GB for the 4B scale compared to Standard CoT, which suggests high engineering adaptability for mainstream computing devices. Crucially, as the semantic augmentation module operates exclusively during the training phase and the latent space iterations allevi-

ate the computational demands of explicit text generation, this marginal training investment achieves a competitive balance between efficiency and effectiveness by enhancing inference accuracy and reducing token overhead.

Analysis of Reasoning Efficiency. LaRe exhibits a favorable balance in terms of inference efficiency, as illustrated in Table 14. Although the iterative latent refocusing mechanism introduces additional forward pass overhead at each step, this strategy of substituting generation with latent computation effectively compresses the final textual path. Compared to conventional explicit Chain of Thought approaches, LaRe requires substantially fewer tokens, which directly diminishes the time spent on computationally intensive autoregressive decoding. While the latency advantage remains relatively modest in straightforward perceptual tasks such as POPE due to the inherent fixed iteration costs, the framework justifies its computational allocation in complex reasoning scenarios like MMMU-Pro and V^* . In these cases, the model prioritizes processing time for high-quality visual evidence extraction rather than generating redundant textual descriptions. This mechanism enables superior reasoning accuracy while maintaining competitive inference latency.

We further analyze the trade-off between reasoning accuracy and computational overhead by adjusting the number of extractor layers $N \in \{1, 2, 4, 8, 10, 12\}$. As illustrated in Figure 7, accuracy improves significantly as N increases from 1 to 2, after which performance gains begin to saturate. Specifically, increasing N beyond 4 yields marginal improvements of less than 0.5%, whereas inference latency continues to grow. To maintain optimal efficiency and a lightweight resource footprint that supports the claim of reducing token consumption, $N = 2$ is selected as the default configuration.

Table 13: Comparison of training computational overhead across different model scales.

Methods	Trainable Params	Speed (s/iter)	Time	GPU Memory
<i>Base VLM: Qwen3-VL-2B-Instruct</i>				
Standard CoT	1.72 B	2.62	12h 17min	24.25 GB
+ Extractor	1.73 B	2.79	13h 12min	25.57 GB
+ Semantic Aug.	1.76 B	2.71	12h 40min	25.13 GB
LaRe (Full)	1.77 B	2.88	13h 30min	26.08 GB
<i>Base VLM: Qwen3-VL-4B-Instruct</i>				
Standard CoT	4.02 B	4.59	21h 44min	38.15 GB
+ Extractor	4.03 B	4.84	22h 56min	40.21 GB
+ Semantic Aug.	4.07 B	4.73	22h 25min	39.70 GB
LaRe (Full)	4.08 B	4.96	23h 29min	41.20 GB

Table 14: Average inference latency (s/query) across different benchmarks. The total latency includes both latent reasoning and final text decoding. Lower is better.

Method	MMBench	MMStar	MMVP	ScienceQA	POPE	V*	MMMU-Pro	Avg.
<i>Base VLM: Qwen3-VL-2B-Instruct</i>								
Standard CoT	1.15	1.28	0.82	1.25	0.65	1.32	1.55	1.15
MM-CoT	1.12	1.15	0.78	1.32	0.62	1.28	1.45	1.10
CCoT	0.92	1.22	0.95	1.45	0.72	1.55	1.48	1.18
ICoT	1.05	1.18	0.85	1.22	0.68	1.42	1.62	1.15
CoF	1.18	1.35	0.92	1.38	0.85	1.82	1.95	1.35
MCOU	0.58	0.72	0.48	0.65	0.32	0.75	0.92	0.63
LVR	0.55	0.68	0.45	0.62	0.28	0.72	0.88	0.60
LaRe (Ours)	0.52	0.62	0.39	0.55	0.30	0.65	0.78	0.54
<i>Base VLM: Qwen3-VL-4B-Instruct</i>								
Standard CoT	2.15	2.45	1.65	2.12	1.15	2.25	2.85	2.09
MM-CoT	2.08	2.32	1.52	2.25	1.12	2.18	2.65	2.02
CCoT	1.75	2.38	1.72	2.45	1.28	2.65	2.72	2.14
ICoT	1.92	2.25	1.55	2.18	1.22	2.35	2.82	2.04
CoF	2.25	2.58	1.82	2.35	1.52	3.15	3.42	2.44
MCOU	1.05	1.25	0.68	0.85	0.58	0.95	1.38	0.96
LVR	1.02	1.18	0.62	0.78	0.52	0.88	1.32	0.90
LaRe (Ours)	0.98	1.14	0.57	0.73	0.52	0.85	1.25	0.82

Causal Intervention Analysis. To rigorously verify whether the attention weights generated by the extractor reflect the causal necessity of visual evidence or merely constitute superficial pseudo-alignment, we design a causal intervention experiment. During inference, we systematically perturb the visual tokens delivered to the model through four distinct masking protocols. These protocols involve suppressing the top 20% of visual patches with the highest attention scores, suppressing the bottom 20% with the lowest scores, applying random occlusion to 20% of patches across the full image, and completely masking all regions throughout every iteration step. By observing the performance degradation of the model on several challenging benchmarks, we directly quantify the semantic indispensability of the regions prioritized by LaRe, as shown in Table 15.

The experimental results exhibit a pronounced asymmetry across masking strategies, confirming

the high semantic fidelity of our latent space re-focusing mechanism. Masking the top 20% of patches with the highest attention scores under strategy A induces substantial performance degradation across all evaluation sets, with accuracy dropping by 11.4% on MMStar and 11.3% on MMVP. Conversely, discarding the bottom 20% of patches with the lowest attention scores under strategy B yields negligible decay, maintaining an average decrease below 1.0%. This behavior demonstrates that the extractor successfully filters task-irrelevant visual background noise. Moreover, the performance decline observed in strategy A substantially exceeds that of the random masking baseline in strategy C. This disparity verifies that LaRe isolates highly specific and non-trivial visual evidence critical for reasoning, rather than capturing generic saliency structures. Complete removal of the target regions across all iterations under strategy D triggers the most severe cumulative collapse in reasoning ac-

Table 15: Causal intervention results under various masking strategies using the Qwen3-VL-4B-Instruct backbone, where Δ denotes the performance drop relative to the full LaRe model without masking.

Masking strategy	MMStar	Δ	V^*	Δ	MMVP	Δ	POPE	Δ	MMMU-Pro	Δ
LaRe (Full)	72.8	-	87.0	-	57.6	-	77.4	-	58.4	-
Mask High-Attn	61.4	-11.4	77.8	-9.2	46.3	-11.3	71.1	-6.3	52.9	-5.5
Mask Low-Attn	71.7	-1.1	86.2	-0.8	56.3	-1.3	76.9	-0.5	57.8	-0.6
Random Mask	67.5	-5.3	82.9	-4.1	51.8	-5.8	74.3	-3.1	55.7	-2.7
Mask All Steps	58.9	-13.9	75.3	-11.7	43.6	-14.0	69.6	-7.8	51.4	-7.0

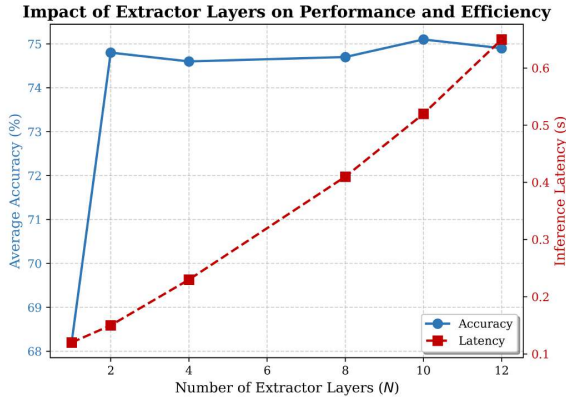


Figure 7: Impact of Extractor depth N on reasoning performance and efficiency. The blue solid line (left y-axis) represents the average accuracy (%), while the red dashed line (right y-axis) denotes the inference latency per step (seconds). Performance saturates beyond $N = 4$, justifying our choice of $N = 2$ for a lightweight and efficient implementation.

curacy, exemplified by a 14.0% drop on MMVP. Taken together, these empirical causal interventions demonstrate that the visual regions prioritized by our latent space refocusing mechanism are not merely correlated with textual generation. They constitute a causal necessity for accurate multi-modal reasoning.

Table 16: Probing analysis results for semantic alignment of latent thoughts.

Backbone	Module	Classifier	Accuracy	
			Vision	Text
Qwen3-VL-2B-Instruct	w/o Extractor	SVM	58.4	61.4
	with Extractor	SVM	94.9	95.2
	w/o Extractor	MLP	59.6	62.2
	with Extractor	MLP	96.2	96.4
Qwen3-VL-4B-Instruct	w/o Extractor	SVM	59.2	62.5
	with Extractor	SVM	95.7	95.3
	w/o Extractor	MLP	60.0	63.5
	with Extractor	MLP	94.8	95.0

Sequential Dependency Analysis. To more directly characterize the sequential dependency among latent tokens, we design a stepwise inter-

Table 17: Stepwise intervention results of latent tokens. Interventions on earlier steps cause larger accuracy drops, exhibiting a characteristic cascading effect of sequential dependency.

Ablated Step	Avg Acc (%)	Output Tokens
None (full)	74.7	65.0
Step 1	68.4	89.3
Step 2	70.9	78.6
Step 3	72.8	71.2
Step 4	73.6	67.4
All steps	61.2	102.7

vention experiment where at step i we remove the corresponding latent token and measure the downstream task accuracy as well as the output length of the generated responses. The results are summarized in Table 17.

The experimental results exhibit a clear cascading effect: interventions on earlier steps cause substantially larger accuracy drops than those on later steps (for example, step 1 leads to a 6.3% drop while step 4 only a 1.1% drop), accompanied by a corresponding increase in output length. This indicates that the model is forced to compensate for the missing latent information through text. This asymmetric, order-sensitive performance degradation is a hallmark of sequential dependency.

Sensitivity Analysis of Step Allocation. To further examine whether a fixed number of steps suffices for queries of varying complexity, we perform a sensitivity analysis of LaRe (4B) on seven benchmarks covering perception, mixed, and reasoning tasks, as shown in Table 18.

Across all seven benchmarks, the fixed configuration $K=4$ performs within 0.2% of the per-benchmark optimum K^* , irrespective of task type or complexity. This indicates that the potential improvement recoverable by an adaptive strategy is very limited, and further supports the conclusion

Table 18: Sensitivity analysis of LaRe (4B) with respect to the number of refocusing steps K across seven benchmarks. The fixed $K=4$ configuration performs within 0.2% of the per-benchmark optimum K^* , regardless of task type or complexity. The best result in each row is shown in **bold**.

Benchmark	Type	Number of Refocusing Steps K						K^*	Δ vs. $K=4$
		1	2	3	4	5	6		
MMBench	Mixed	79.1	80.7	81.5	82.1	82.2	82.1	5	+0.1
MMStar	Mixed	69.2	71.0	72.1	72.8	72.7	72.7	4	0.0
MMVP	Percep.	52.0	54.4	56.3	57.6	57.6	57.5	4	0.0
ScienceQA	Mixed	83.2	85.6	86.9	87.5	87.7	87.6	5	+0.2
POPE	Percep.	74.3	76.0	76.9	77.4	77.2	77.2	4	0.0
V*	Percep.	82.4	84.7	86.2	87.0	86.9	86.9	4	0.0
MMMU-Pro	Reason.	53.6	55.9	57.5	58.4	58.3	58.5	6	+0.1

that using a small fixed number of steps serves as a robust and effective choice for LaRe.

Probing Analysis. To examine whether the latent tokens jointly encode linguistic reasoning and visual semantics, we conduct a probing analysis with two weak classifiers including SVM (Cortes and Vapnik, 1995) and MLP (Rumelhart et al., 1986). As shown in Table 16, LaRe achieves latent-vision and latent-text matching accuracy close to 95% across different backbone models. In contrast, removing the Extractor leads to a drop in probing accuracy to approximately 60%. These results provide strong evidence for the critical role of the Extractor, demonstrating its effectiveness in structuring cross-modal inputs into unified semantic representations within the latent space.

B.3 Generalization across VLM Architectures

To further verify that the effectiveness of LaRe does not depend on a specific VLM architecture, we conduct supplementary experiments on two additional backbone models with markedly different designs: LLaVA-1.5-7B (Liu et al., 2023) and InternVL2.5-8B (Chen et al., 2025b). For each backbone, under identical training and inference settings, we compare LaRe against the standard CoT baseline and the latent space reasoning method LVR.

As shown in Table 19, on both additional architectures, LaRe consistently outperforms the two baseline methods across all seven benchmarks. Specifically, compared to standard CoT, LaRe improves the average accuracy by 6.2% on LLaVA-1.5-7B and by 5.9% on InternVL2.5-8B, while reducing the number of reasoning tokens by over 60%. This performance gain is highly consistent

with the results observed on the Qwen3-VL backbone, indicating that the proposed latent refocusing paradigm and semantic augmentation training strategy generalize well across VLMs that differ in visual encoders, projection modules, and pre-training schemes.

C Case Study

In this section, we present a qualitative attention comparison across three challenging perception-intensive tasks to empirically demonstrate the advantages of LaRe, as illustrated in Figure 8.

In the first case concerning long-range retrieval of a parachute color, the target occupies a minimal pixel footprint. The standard explicit CoT fails to localize the object during discrete token generation, causing attention to scatter across the foreground bicycle, the rider, and salient ground noise. In contrast, LaRe utilizes a lightweight semantic aggregation module driven by the latent decision states of the large language model. This Extractor enables dynamic visual focusing within the latent space, which concentrates high attention weights precisely on the small parachute pixels in the sky.

A similar pattern emerges in the second case, which evaluates item identification on a table within a cluttered indoor scene. The explicit CoT becomes distracted by large redundant background regions such as sofas, leading to substantial visual deviation. LaRe suppresses this background interference through a latent refocusing mechanism, directing attention exclusively to the round wooden table in the corner and its small items.

The third case assesses local attribute recognition for glove material. The standard explicit CoT

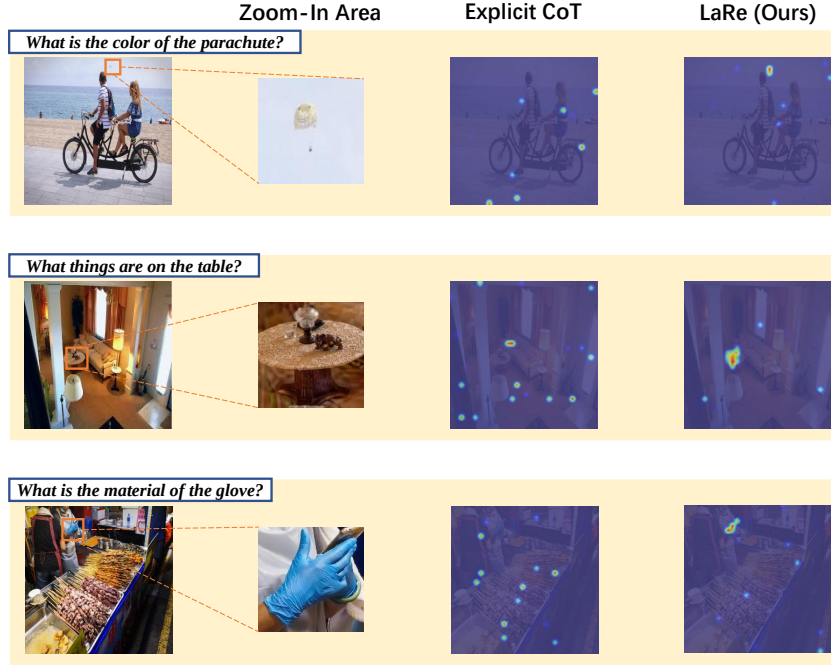


Figure 8: Comparison of attention heatmaps between explicit CoT and LaRe. Unlike conventional explicit CoT which is easily distracted by prominent or complex backgrounds, LaRe utilizes a latent refocusing mechanism to accurately isolate small or occluded reasoning evidence.

Table 19: Generalization of LaRe on Additional VLM Backbones. We report accuracy on seven multi-modal benchmarks and the average number of reasoning tokens (#Tok). The best result within each backbone group is shown in **bold**.

Base VLM	Method	MMBench	ScienceQA	MMVP	V*	MMStar	MMMUPro	POPE	Avg. #Tok
LLaVA-1.5-7B	CoT	66.4	69.8	22.0	47.3	36.2	16.8	85.4	49.1
	LVR	67.8	71.2	27.6	51.8	38.1	18.2	86.1	51.5
	LaRe	70.1	73.5	34.7	57.8	41.5	22.3	87.2	55.3
InternVL2.5-8B	CoT	79.6	95.4	42.7	63.5	60.1	48.2	87.6	68.2
	LVR	80.9	95.9	47.2	67.4	62.0	50.1	88.2	70.2
	LaRe	83.2	96.8	55.3	73.8	65.4	54.7	89.3	74.1

becomes anchored to the dense and highly salient barbecue skewers in the foreground, causing global attention to disperse across the food items without extracting the core answer. LaRe filters out low-level visual saliency interference and anchors attention precisely on the blue plastic gloves worn by the staff member at the top of the image.

These observations confirm that latent space reasoning effectively mitigates the limitations of explicit sequential generation in perception-intensive scenarios.

D Discussion

The proposed LaRe paradigm achieves dual advantages in token efficiency and reasoning per-

formance for multimodal tasks through an iterative latent refocusing mechanism and a diffusion model based semantic enhancement training strategy. However, as noted in the Limitations section, the current framework still leaves several issues worthy of deeper investigation in future work, and these issues also point to potential directions for further improvement.

The current LaRe framework relies on a preset, static number of latent iterations K . Although we have empirically determined an optimal threshold at which performance saturates, this one size fits all strategy struggles to accommodate the substantial variability in task complexity across different multimodal benchmarks. For simple visual ques-

tion answering tasks, a fixed number of iterations may introduce redundant computational overhead, whereas for complex tasks involving multi step spatial reasoning or fine grained perception, a preset upper bound may restrict the model’s reasoning potential. Future work can explore a dynamic halting mechanism, for instance by drawing on ideas from Adaptive Computation Time (ACT) (Graves, 2016) or PonderNet (Banino et al., 2021), allowing the model to adaptively decide when to terminate iterative reasoning based on the convergence of latent states, the confidence of hidden representations, or the trend of reconstruction loss.

Going a step further, we believe that a learnable stopping policy based on reinforcement learning is a particularly promising direction to explore. The iterative refocusing process of LaRe is inherently a sequential decision problem, where each iteration corresponds to a decision of whether to continue gathering more visual evidence. This aligns well with the Markov Decision Process (MDP) formulation of reinforcement learning. One can treat the hidden state at each step as a state, the continue or stop choice as a discrete action, and construct a composite reward function using the two natural supervision signals available in the LaRe framework, namely task accuracy and diffusion reconstruction loss. This would simultaneously constrain reasoning performance, computational efficiency, and latent space quality. Coupled with recent successful applications of policy optimization methods such as GRPO (Shao et al., 2024b) and PPO (Schulman et al., 2017) to control the reasoning length of large language models, this think on demand paradigm not only promises to further improve LaRe’s ability to balance efficiency and flexibility, but also may give rise to a multimodal reasoning framework with metacognitive capabilities. That is, the model would not only produce an answer, but also autonomously determine how much thinking is sufficient to arrive at a reliable answer.

E Usage of LLMs

In this paper, LLMs were used for coding assistance and writing support. Specifically, they were employed to generate code snippets, debug existing scripts, and optimize algorithms, significantly accelerating the development cycle. For writing support, they were primarily utilized for proofreading the text and formatting *LaTeX* code.