Reasoning in the Dark: Interleaved Vision-Text Reasoning in Latent Space

Chao Chen[1], Zhixin Ma[2], Yongqi Li[1†], Yupeng Hu[3], Yinwei Wei[3], Wenjie Li[1], Liqiang Nie[4]

[1]The Hong Kong Polytechnic University
[2]Singapore Management University
[3]Shandong University
[4]Harbin Institute of Technology (Shenzhen)
ochenchaoo@outlook.com, liyongqi0@gmail.com

Abstract

Multimodal reasoning aims to enhance the capabilities of MLLMs by incorporating intermediate reasoning steps before reaching the final answer. It has evolved from text-only reasoning to the integration of visual information, enabling the thought process to be conveyed through both images and text. Despite its effectiveness, current multimodal reasoning methods depend on explicit reasoning steps that require labor-intensive vision-text annotations and inherently introduce significant inference latency. To address these issues, we introduce multimodal latent reasoning with the advantages of multimodal representation, reduced annotation, and inference efficiency. To facilicate it, we propose Interleaved Vision-Text Latent Reasoning (IVT-LR), which injects both visual and textual information in the reasoning process within the latent space. Specifically, IVT-LR represents each reasoning step by combining two implicit parts: latent text (the hidden states from the previous step) and latent vision (a set of selected image embeddings). We further introduce a progressive multi-stage training strategy to enable MLLMs to perform the above multimodal latent reasoning steps. Experiments on M^{3}CoT and ScienceQA demonstrate that our IVT-LR method achieves an average performance increase of 5.45% in accuracy, while simultaneously achieving a speed increase of over 5 times compared to existing approaches.

1 Introduction

Over the past few years, the capabilities of large language models (LLMs) have been further unlocked through advancements in reasoning. Researchers have sought to enhance the reasoning abilities of LLMs, initially through prompting techniques such as Chain-of-Thought prompting (Wei et al., 2022), and more recently by developing large reasoning models using reinforcement learning, like GPT-4o (OpenAI, 2024) and Deepseek-R1 (DeepSeek-AI, 2025). Building on the success of reasoning in LLMs, there has been growing interest in the research community to extend these reasoning capabilities to multimodal LLMs (MLLMs). This has brought the promising topic of multimodal reasoning, aiming to improve the performance of models on multimodal tasks, such as VQA, through reasoning.

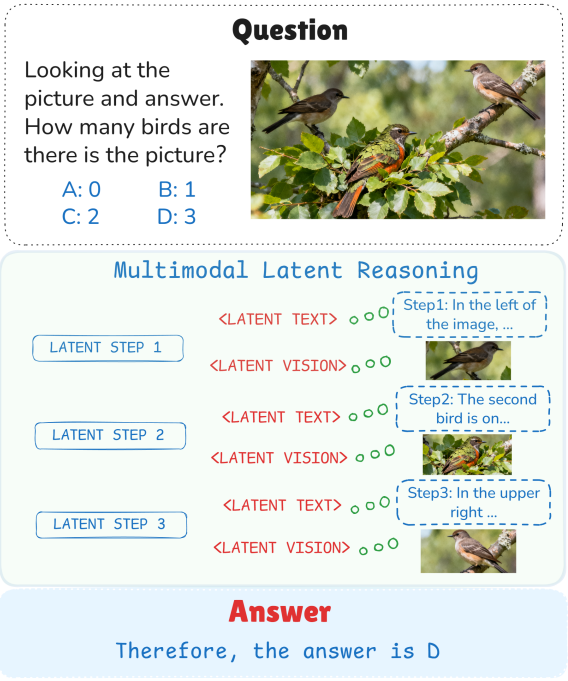


Figure 1: An example of interleaved vision-text latent reasoning, where the intermediate reasoning steps are carried out entirely within the multimodal latent space.

Current multimodal reasoning approaches can be broadly categorized into the following progressive steps: 1) Text-only reasoning. Early multimodal reasoning methods primarily focused on pure text-based reasoning, where MLLMs perform textual reasoning before generating the final answer. These approaches (Hu et al., 2022; Mondal et al., 2024)

†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 39117–39129

could seamlessly apply LLMs methodologies to MLLMs. 2) Vision-text involved reasoning. Some studies (Hu et al., 2024; Liu et al., 2025; Chern et al., 2025) highlight that the intermediate reasoning steps also require the involvement of visual information. For instance, Gao et al. (2025) enhances reasoning by generating sequential steps that interleave visual information with textual rationales via selected image patches. In a related direction, Zheng et al. (2025) trains models through end-to-end reinforcement learning to autonomously zoom in on image regions for fine-grained visual inspection during the reasoning process. Alternatively, Li et al. (2025b) enables MLLMs to actively "think visually" by generating explicit image visualizations of their reasoning traces, thereby significantly enhancing performance on complex spatial reasoning tasks.

Recently, latent reasoning has emerged as a new paradigm in LLMs, which eliminates the need for explicit and lengthy textual reasoning by leveraging implicit latent vectors (Hao et al., 2024). Inspired by this, we believe that latent reasoning holds even greater potential for facilitating vision-text interleaved intermediate reasoning steps due to the following reasons: 1) Multimodal representation potential. Latent reasoning enables the reasoning process to occur entirely within a hidden space, offering a greater capacity to represent rich, multimodal information during reasoning. 2) Reduced Annotation. Introducing latent reasoning will lessen reliance on heavily annotated vision-text interleaved reasoning data, as reasoning steps no longer need to be fully observable or linguistically aligned. 3) Inference efficiency. By avoiding long chains of explicit multimodal representation in the reasoning step, it will significantly improve efficiency.

In this work, we propose the Interleaved Vision-Text Latent Reasoning (IVT-LR) method, which enables both textual and visual modalities to perform reasoning entirely in latent space. As shown in Figure 1, in our framework, each latent reasoning step consists of two parts: *latent text* and *latent vision*. At each reasoning step, we use the hidden state from the previous step to replace explicit text as the *latent text* component. Afterwards, for *latent vision* part, a certain number of image embeddings are selected based on their attention scores then concatenated with the hidden state to serve as input for the subsequent reasoning step. To effectively blend the *latent text* and *latent vision* components for joint reasoning in the latent space, we introduce a progressive, multi-stage training strategy that gradually substitutes explicit CoT steps with latent reasoning steps, where supervision is focused on the remaining future steps and the final answer to ensure accurate inference.

The key contributions are summarized:

- We introduce IVT-LR, the first framework to achieve fully unified multimodal latent reasoning. Unlike prior methods, our approach enables both textual and visual information to be reasoned with in the latent space, eliminating the need for intermediate explicit text or image generation.
- Our method presents a novel training paradigm that is both data-efficient and computationally efficient, without requiring explicit annotations for intermediate visual reasoning steps. By reasoning in latent space, it also drastically reduces the number of autoregressive steps required for inference.
- We validate the effectiveness of IVT-LR through extensive experiments on challenging visual question answering benchmarks, including M^3COT and ScienceQA, where our model establishes new state-of-the-art performance in accuracy and significantly improves inference efficiency, as measured by fewer autoregressive steps and lower inference latency.

2 Related Works

2.1 Multimodal Reasoning

Multimodal reasoning focuses on enabling models to reason over information from different modalities to solve complex tasks. Existing approaches can be roughly divided into text-only reasoning and interleaved reasoning.

Text-only reasoning. Early works attempt to convert visual information into text before reasoning, using tools or visual experts to generate textual representations to guide LLMs. Hu et al. (2022) first introduced the concept of captions, extracting visual content as textual captions and concatenating them to the input to enhance reasoning. Inspired by this, subsequent works pursued finer-grained understanding of images to improve textual expressiveness. Zheng et al. (2023) generates a rationale that incorporates image information from visual-text inputs, which is then used for reasoning. Other

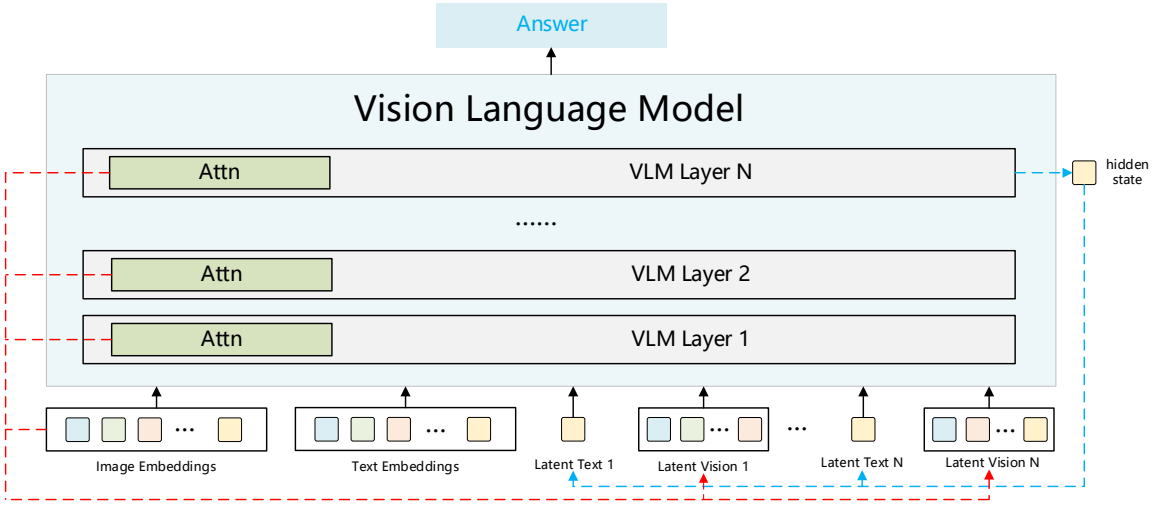


Figure 2: Overview of our Interleaved Vision-Text Latent Reasoning (IVT-LR) framework. At each step, reasoning is performed entirely in the latent space by fusing *latent text* (the hidden state from the previous step) and *latent vision* (dynamically selected image embeddings based on attention scores).

works (Mitra et al., 2024; Mondal et al., 2024) leverage graph structures to identify entities in images and construct relationships among them, enhancing reasoning based on these inter-entity connections.

Vision-text involved reasoning. This line of work emphasizes using images together with text during the rationale generation and reasoning process. Building on the reasoning paradigm of large language models, Zhang et al. (2024) first proposed decoupling rationale generation from answer generation in the Vision-Text Reasoning field. Subsequently, Shao et al. (2024) annotates key regions of the original image in intermediate steps, training models to focus on image regions relevant to the answer. While some works (Gao et al., 2025; Zhang et al., 2025) further extract key image regions progressively during reasoning, combining visual information with textual reasoning to generate the final answer. Moreover, new methods (Hu et al., 2024; Liu et al., 2025) emulate human thought by sketching images during reasoning, focusing on core concepts, structures, and relationships while ignoring redundant details. Other works (Li et al., 2025b; Chern et al., 2025) generate new images during reasoning, combining them with text to improve reasoning in complex scenarios. To completely decouple reasoning from language and amplify the role of images, Xu et al. (2025) proposes reasoning solely with newly generated images, achieving substantial improvements in visual navigation tasks.

2.2 Latent Reasoning

Latent reasoning refers to internal, non-linguistic thinking performed in a hidden latent space before generating the final answer. Early methods used special tokens to guide latent reasoning. Goyal et al. (2024) introduces learnable <pause> tokens, giving the model opportunities to internally update information before generating an answer, while Wang et al. (2024b) uses <plan> tokens to guide reasoning.

Later, some works exploit the model's continuous hidden states to replace explicit reasoning steps. Hao et al. (2024) pioneers continuous latent space reasoning by feeding the last hidden states as input embeddings for the next step without generating intermediate tokens, significantly reducing reasoning tokens and improving efficiency. Inspired by this, subsequent methods improve the quality of intermediate representations. Cheng and Van Durme (2024) uses variable-length contemplation tokens for latent reasoning, addressing quality degradation caused by fixed-length embeddings. Shen et al. (2025) employs self-distillation to align student and teacher hidden activations under CoT supervision, constraining latent reasoning paths.

In the multimodal domain, latent reasoning has also been introduced. Unlike traditional LLMs, VLMs emphasize how image features interact with the latent space. Some efforts (Yang et al., 2025; Li et al., 2025a; Pham and Ngo, 2025) have been made to integrate visual "thoughts" into the latent space

for reasoning. However, these existing works focus solely on single-modal latent reasoning. Combining text and vision for multimodal latent reasoning in the latent space remains unexplored.

3 Method

In this section, we present IVT-LR, the first VLM framework that unifies textual and visual representations in the latent space and implements multimodal latent reasoning. Given a text sequence $\mathcal{X} = (x_1, \ldots, x_I)$ and a set of visual embeddings $\mathcal{Z} = (z_1, \ldots, z_J)$ from a visual encoder, a standard VLM encodes the text sequence into embeddings, incorporates visual features, and predicts a conditional distribution over the next token:

$$e_{1:t}^{\text{text}} = g(x_{1:t}) \in \mathbb{R}^{t \times d},$$

$$e_t^{\text{fused}} = f(e_{1:t}^{\text{text}}, \mathcal{Z}) \in \mathbb{R}^d,$$

$$\mathcal{M}(x_{t+1} \mid x_{1:t}, \mathcal{Z}) = \text{softmax}(W \cdot e_t^{\text{fused}}),$$

where $g(\cdot)$ denotes the text embedding function, $f(\cdot)$ is a function that generates the hidden state for the next token based on the textual embeddings and visual features, and $W \in \mathbb{R}^{|V| \times d}$ is trained to project the fused representation to a distribution over the vocabulary. This formulation illustrates how a VLM predicts the next token conditioned on both textual context and visual information.

3.1 Multimodal Latent Reasoning

Figure 2 provides an overview of our approach. In IVT-LR, the latent reasoning is conducted over both *latent text* and *latent vision*. Following (Hao et al., 2024), the textual modality bypasses explicit token prediction: instead of using the embedding of the previous explicit text token, we represent the latent text with the hidden state h_{t-1}^{hidden}, which effectively encodes the necessary reasoning logic and preserves richer intermediate information in a continuous latent space. Meanwhile, the latent vision is designed to model the dynamic focus on the visual features at each step. Specifically, we extend latent reasoning to visual modality by selecting the k most relevant visual features from the image embedding set. Thus, an attention-based selection mechanism is designed to choose a fixed number of image embeddings from the full set $[z_1, z_2, \ldots, z_J]$. We utilize the sum of attention weights across all layers to identify the k image embedding positions with the highest cumulative scores. The selected features are

Algorithm 1 IVT-LR

Input: Text input embeddings $\mathcal{E} = [e_1, \ldots, e_I]$, Image input embeddings $\mathcal{Z} = [z_1, \ldots, z_J]$, Whole input embeddings $\mathcal{Q} = \mathcal{E} + \mathcal{Z} = [q_1, \ldots, q_m]$, Latent step positions $\mathcal{L} = [l_1, \ldots, l_N]$, Number of selected embeddings k
for $i = 1$ to N **do**
  $h_i \leftarrow \text{LastHiddenState}(q_{1:l_i-1})$
  $\mathcal{Z}_{sel} \leftarrow \text{AttentionSelect}(\mathcal{Z}, k)$
  $latent[i] \leftarrow [h_i, \mathcal{Z}_{sel}]$
  $\mathcal{Q} \leftarrow \text{Concat}(\mathcal{Q}, latent[i])$
  $l_n \leftarrow l_n + (k+1) \quad \text{for } n > i$
end for
$q_{:end} \leftarrow \text{PredictToEnd}(q_{:l_N})$
Answer $\leftarrow \text{Decode}(q_{l_N+1:})$
return Answer

appended to the hidden states h_{t-1}^{hidden}, resulting in a multimodal latent representation $[h_{t-1}^{\text{latent}}, z_{t-1}^{\text{selected}}]$. The input to the model at step t thus consists of all prior hidden states and their selected visual features, along with any preceding question embeddings, which can be written as $E_t = [e_1, \ldots, e_N, h_1^{\text{latent}}, z_1^{\text{selected}}, \ldots, h_{t-1}^{\text{latent}}, z_{t-1}^{\text{selected}}]$. The model fuses these multimodal representations to obtain $e_t^{\text{fused}} = f(E_t)$, which is projected through the output head to yield the next-token distribution $\mathcal{M}(x_{t+1} \mid E_t) = \text{softmax}(W \cdot e_t^{\text{fused}})$. This design allows the model to perform step-wise multimodal latent reasoning without generating intermediate reasoning sequences.

3.2 Training Procedure

The objective of IVT-LR is to enable multimodal reasoning within the latent space. Inspired by Deng et al. (2024), we adopt a multi-stage training strategy to progressively boost the model's reasoning capability. In the preprocessing stage, each reasoning trajectory is segmented into up to N steps, followed by the final answer. At stage 0, as shown in Figure 3, the model is trained with standard CoT supervision, where all reasoning steps are explicitly generated to strengthen symbolic reasoning ability. Afterwards, the latent reasoning steps are progressively introduced within the N stages: at each stage, one additional explicit reasoning step is replaced by a latent reasoning step, denoted by the special token <*latent*>, beginning with the first step. In this way, the model learns to progressively substitute explicit reasoning using

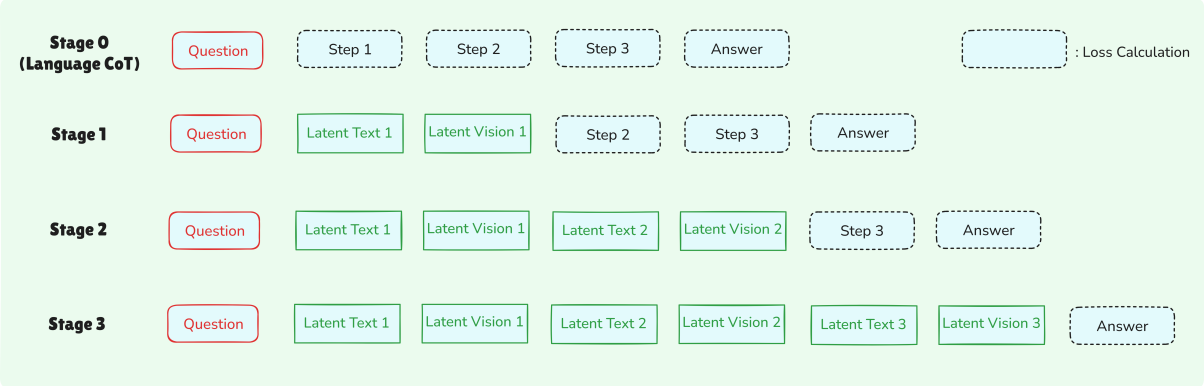


Figure 3: Overview of the Multi-Stage Progressive Training Strategy used for IVT-LR. The strategy begins with full explicit CoT and then gradually substitutes one explicit reasoning step with latent text and latent vision. Training loss is calculated exclusively over the remaining explicit steps and the final answer.

latent textual and visual representations while still being supervised on the final answer.

Training is optimized using negative log-likelihood (NLL) loss, with supervision applied only to reasoning steps and the final answer. Latent reasoning steps and question tokens are masked out. This design ensures that the supervision signal is placed only on the reasoning steps and the final answer. By avoiding excessive alignment between latent representations and explicit rationales, the model learns to internalize reasoning trajectories in latent space with essential image features, while still being driven toward correct final predictions. Compared to single-step fine-tuning, the proposed multi-stage training introduces additional stages but reuses the same backbone and training data, resulting in a moderate increase in training cost that remains practical and acceptable.

3.3 Inference Process

Since all rationales in training have been segmented into a certain number of steps, at inference time, the same number of <*latent*> tokens are appended after the question and image inputs. This setup ensures that reasoning is fully conducted in latent space and no explicit reasoning steps are produced before the final answer.

To evaluate the intermediate models at stage n, inference uses n latent tokens, yielding mixed explicit-latent reasoning consistent with the training stage. Importantly, latent text and latent vision co-exist only during the latent reasoning phase, where visual evidence is integrated into the hidden trajectory. Outside this phase, the model operates in a purely linguistic generation mode.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation. We evaluate our method on two widely used multimodal reasoning benchmarks: M^3CoT (Chen et al., 2024) and ScienceQA (Lu et al., 2022). M^3CoT is a large-scale benchmark focusing on multimodal chain-of-thought reasoning, where models must combine both visual and textual inputs to perform multi-step reasoning. ScienceQA is a diverse dataset covering natural science, language science, and social science, with many questions accompanied by diagrams or images. To further verify the generalizability and robustness of IVT-LR, we provide additional experimental results in Appendix C. We evaluate using exact-match answer accuracy, along with the average number of autoregressive steps and the average response time per question. These metrics capture both correctness and reasoning efficiency.

Baselines and Implementation Details. We compare IVT-LR against six representative methods, including text-only reasoning: CCoT (Mitra et al., 2024); vision-text involved reasoning: Chain-of-Focus (Zhang et al., 2025), SCAFFOLD (Lei et al., 2025), ICoT (Gao et al., 2025), Multimodal-CoT (Zhang et al., 2024); and No-CoT that directly predicts answers without generating intermediate steps. Detailed implementation configurations and comprehensive comparison settings are provided in Appendix B.

For fair comparison, we evaluate IVT-LR and all baselines with Qwen2-VL-7B (Wang et al., 2024a) and Chameleon-7B (Team, 2024) back-

Backbone	Methods	M^3CoT			ScienceQA		
		Acc.(%) ↑	# AR Steps ↓	Avg. Time(s) ↓	Acc.(%) ↑	# AR Steps ↓	Avg. Time(s) ↓
Qwen2-VL	No-CoT	45.4	-	-	64.4	-	-
	Multimodal CoT(Zhang et al., 2024)	42.5	106.3	3.10	58.3	83.9	2.44
	CCoT(Mitra et al., 2024)	44.1	177.2	5.31	63.8	164.0	5.23
	ICoT(Gao et al., 2025)	46.0	96.5	2.86	65.4	77.4	2.28
	SCAFFOLD(Lei et al., 2025)	44.9	170.8	5.14	62.5	162.3	4.91
	Chain-of-Focus(Zhang et al., 2025)	64.3	185.7	2.63	91.2	162.3	2.09
	IVT-LR	**71.8**	**10.0**	**0.65**	**94.6**	**11.0**	**0.67**
Chameleon	No-CoT	28.4	-	-	48.5	-	-
	Multimodal CoT(Zhang et al., 2024)	30.6	110.5	3.62	50.7	98.7	3.33
	CCoT(Mitra et al., 2024)	31.4	168.4	5.35	51.3	174.2	5.39
	ICoT(Gao et al., 2025)	32.3	110.9	5.43	53.4	92.4	4.62
	SCAFFOLD(Lei et al., 2025)	31.1	194.3	6.12	47.5	160.6	6.03
	Chain-of-Focus(Zhang et al., 2025)	36.5	739.4	3.09	61.2	717.1	2.56
	IVT-LR	**41.8**	**10.0**	**1.13**	**64.0**	**11.0**	**1.14**

Table 1: Comparison of IVT-LR with various multimodal reasoning baselines on the M^3CoT and ScienceQA benchmarks. The reported metrics include: Answer Accuracy (Acc.), Average number of Autoregressive Steps (# AR Steps), and Average Generation Time (Avg. Time). Experiments are conducted using backbone models: Qwen2-VL-7B and Chameleon-7B.

bones. Training details are in Appendix A.

4.2 Main Results

The results on M^3CoT and ScienceQA are summarized in Table 1. Analyzing these outcomes, we draw the following key observations:

Multimodal Reasoning Accuracy. IVT-LR achieves the highest accuracy on both the M^3CoT and ScienceQA benchmarks, consistently outperforming all baselines across both backbones. Compared to the strongest baseline, Chain-of-Focus, IVT-LR yields improvements of 5% (Chameleon backbone) to 7.5% (Qwen2-VL backbone) on M^3CoT. Similar gains are observed on the ScienceQA benchmark. Beyond this, IVT-LR surpasses other methods by margins of 10% to 25%, depending on the backbone and task. These results demonstrate that IVT-LR enables more effective cross-modal interaction in the latent space, leading to stronger multimodal reasoning capability on complex tasks.

Reasoning Efficiency. Beyond accuracy, a critical advantage of IVT-LR is its significantly enhanced inference efficiency, which is quantified by fewer autoregressive steps and lower inference latency compared to baselines. 1) Fewer autoregressive steps. Across both backbones, IVT-LR achieves at least a $\mathbf{9}\times$ reduction in the number of autoregressive steps required for generation compared to most baselines. This efficiency is achieved by conducting reasoning in the latent space, where each latent reasoning step corresponds to a single autoregressive step. 2) Lower Inference Latency. With the Qwen model, IVT-LR achieves an average inference time of approximately 0.66s, making it 3 to 8 times faster than all other baselines. A similar trend of significant speedup holds true across the Chameleon backbone. While No-CoT achieves the absolute lowest latency by completely sacrificing deep reasoning (around 0.35s), IVT-LR delivers state-of-the-art accuracy at an inference speed only marginally longer than the minimal No-CoT, demonstrating superior efficiency in the high-accuracy setting.

In summary, IVT-LR demonstrates both superior accuracy and improved reasoning efficiency in VQA tasks. By performing multi-step reasoning in latent space, the model not only achieves the highest accuracy among all baselines but also significantly reduces the number of autoregressive steps and achieves a substantially lower inference latency. These results highlight the effectiveness of latent reasoning in combining textual and visual info. Additional evaluations on other reasoning benchmarks and general VQA datasets are provided in Appendix C to further demonstrate the robustness and generalizability of IVT-LR.

4.3 Ablation Study

To understand the contributions of IVT-LR's key components, we conduct ablation experiments on visual reasoning tasks, including the effects of latent text, latent vision, and attention map selection strategies.

Latent Text. As shown in Table 2, removing latent text(w/o latent text) leads to a noticeable

Methods	M^3CoT	ScienceQA
IVT-LR	**71.83**	**94.6**
w/o latent text	52.20 (-19.63)	84.7 (-9.8)
w/o latent vision	46.64 (-25.19)	82.3 (-11.8)
w/o the whole latent part	58.02 (-13.81)	86.4 (-7.7)

Table 2: Accuracy comparison of IVT-LR on Qwen2-VL, showing the performance impact with and without its core latent components (*latent text* and/or *latent vision*). Values in parentheses indicate performance drop relative to full IVT-LR.

Selection Strategy	**M^3CoT (Acc. %)**
First Layer	71.70
Middle Layer	71.53
Last Layer	72.13
Average (All Layers)	71.83
Weighted Sum	**72.39**

Table 3: Ablation study of attention map selection strategies on M^3CoT. The Weighted Sum approach emphasizes deeper layers to capture more abstract semantic features.

drop in accuracy on both M^3CoT and ScienceQA. This demonstrates that latent text plays a crucial role in model performance: it provides a compact, continuous representation of intermediate reasoning states. This allows the model to internalize multi-step reasoning trajectories directly in the latent space, avoiding biases introduced by language-based alignment. Furthermore, operating in continuous hidden spaces, it effectively mitigates the amplification of errors typical in discrete, step-by-step textual reasoning.

Latent Vision. Table 2 also shows that removing latent vision (w/o latent vision) also results in decreased performance. This indicates that incorporating the most informative visual cues is vital for precise multimodal reasoning. Without this mechanism, the model cannot focus on the critical regions of the image, reducing the effectiveness of each reasoning step. Besides, based on attention-driven integration, latent vision ensures that the latent space receives rich, contextually relevant visual information. It also mitigates interference from irrelevant image regions, leading to more accurate and robust reasoning.

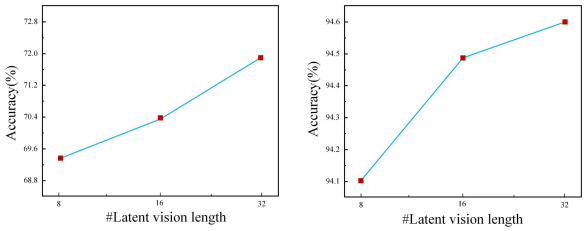
Attention Selection Strategies. To study the effectiveness of different attention map selection strategies, we compare five strategies on Qwen2-VL using the M^3CoT benchmark. These include three single-layer options—**First Layer**, **Middle Layer**, and **Last Layer**—and two multi-layer combinations—**Average** (simple mean across all layers) and **Weighted Sum** (assigning linearly increasing weights to deeper layers).

The results in Table 3 yield three critical insights into multimodal feature alignment. 1) Semantic hierarchy matters. The superior performance of deeper layers suggests that abstract, high-level features in VLMs align more effectively with the information-dense latent text space than low-level visual cues. 2) Information dilution is a risk. Simple averaging proves suboptimal, as shallow-layer noise tends to interfere with deep-layer semantic signals. 3) Weighted aggregation is optimal. Prioritizing deeper layers while retaining earlier structural cues achieves the best balance, underscoring the necessity of layer-aware modeling for robust reasoning.

4.4 In-depth Analysis

Length of Latent vision. We investigate the impact of varying the latent vision length per step.As shown in Figure 4, accuracy steadily increases with this length, indicating that longer latent vision sequences provide richer visual cues necessary for complex reasoning. Since the latent vision is formed by adaptively selecting visual embeddings from the image, increasing the length of these selections allows the model to gradually approach full-image utilization (e.g., 32 embeddings over three steps roughly cover the whole image in Qwen2-VL). This ensures that essential visual details, often required for global comprehension, are not omitted. Moreover, because embeddings are selected step-by-step across the latent reasoning stages, the process achieves targeted and cumulative coverage: each round complements the previous ones, enabling the model to integrate both localized critical features and broader global context in a structured, effective manner.

Stages of Latent Reasoning. We evaluate models with 1, 2, and 3 latent reasoning steps to study the effect of progressively replacing explicit reasoning. As shown in Table 4, accuracy improves as more reasoning steps are conducted in latent space, showing that latent representations provide a more robust reasoning mechanism than explicit language. This is because latent states avoid errors from language alignment and allow smoother integration



(a) M^{3}CoT (b) ScienceQA

Figure 4: Accuracy comparison of IVT-LR on the length of *latent vision* per reasoning step across two reasoning benchmarks: (a) M^{3}CoT and (b) ScienceQA.

Latent Stage	Science	Commonsense	Mathematics	Total
1	56.66%	64.40%	38.59%	56.30%
2	61.71%	70.11%	43.57%	61.48%
3	**70.90%**	**79.78%**	**63.07%**	**71.83%**

Table 4: Accuracy on M^{3}CoT across different latent reasoning stages. Results are shown both overall and broken down by domain.

with image embeddings.

Domain-wise results show that science and mathematics benefit most from additional latent tokens, highlighting that structured reasoning tasks are particularly suited for latent-space inference. The accuracy in commonsense also improves, but with smaller gains, since it often relies less on multi-step deduction. Together, these findings confirm that latent reasoning scales effectively with task complexity, supporting both efficiency and accuracy.

Attention Shift over Step-wise Embeddings. To further investigate the internal mechanisms of IVT-LR, we analyze how the model allocates its attention to image embeddings under our method and explicit reasoning with selected image embeddings. We use **Attention Ratio** and **Attention Focus** as metrics to analyze the model's focus.

(1) Attention Ratio:

$$R = \frac{\sum_{j \in \mathcal{I}} \text{Attn}(E_j)}{\sum_{i \in \mathcal{T}} \text{Attn}(E_i)}, \tag{1}$$

where $\mathcal{I}$ denotes the visual reasoning part, specifically the set of selected image embeddings, and $\mathcal{T}$ denotes the text tokens or the latent text part. This ratio reflects the relative allocation of attention between visual and textual information.

(2) Attention Focus (Inverse Entropy):

$$H = -\sum_k p_k \log p_k, \quad p_k = \frac{\text{Attn}(E_k)}{\sum_m \text{Attn}(E_m)},$$

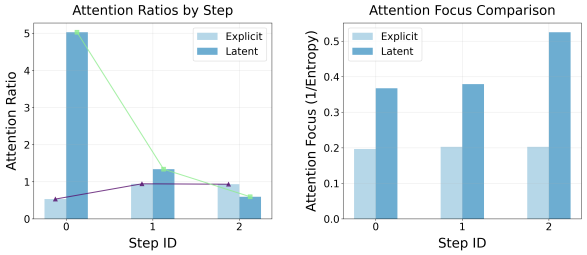


Figure 5: Attention analysis comparison between explicit and latent reasoning approaches. Left: attention ratios of visual part to textual part across reasoning steps. Right: attention focus measured by inverse entropy.

$$F = \frac{1}{H + \epsilon}, \quad \epsilon \ll 1, \tag{2}$$

where F is the Attention Focus. Higher F indicates more concentrated attention, while lower F reflects dispersed focus.

The result is shown in Figure 5. We found significant differences in model behavior between latent and explicit multimodal reasoning modes:

(1) Dynamic Attention Ratio: A Core of Visio-Linguistic Perception. In the latent reasoning mode, the attention ratio exhibits a clear downward trend across reasoning steps. Initially, the model focuses predominantly on latent vision, but over subsequent steps, attention gradually shifts to its latent text for deeper textual reasoning. This dynamic adjustment demonstrates the model's ability to prioritize the most informative visual cues and adaptively reallocate focus, reflecting enhanced visio-linguistic perception. In contrast, under explicit reasoning, the attention ratio remains largely unchanged and consistently below 1, indicating persistent focus on textual tokens. This suggests that, with interference from abundant textual information, explicit reasoning struggles to effectively filter and leverage critical visual features.

(2) Rising Attention Focus: A Hallmark of Efficient Reasoning. Beyond changes in attention ratio, our analysis of attention focus also reveals important insights. In latent reasoning, attention focus shows a progressively increasing trend, showing that the model's attention becomes increasingly concentrated over reasoning steps. This suggests that at each step, the model effectively filters and refines multimodal information, gradually converging on the most critical and relevant cues—a pattern reminiscent of human problem-solving, where distractions are progressively eliminated and attention is concentrated on core evidence. In contrast, under explicit reasoning, attention focus is not only

markedly lower than in implicit reasoning but also exhibits little change across steps. This indicates that explicit reasoning distributes attention more diffusely and lacks clear direction, processing substantial amounts of redundant or less relevant information, which reduces reasoning efficiency and limits the effective extraction of key visual-textual information.

5 Conclusion and Future Work

In this work, we present IVT-LR, the first vision-language reasoning framework that performs multimodal latent reasoning. IVT-LR utilizes latent text and latent vision to internalize complex reasoning trajectories, thereby realizing comprehensive multimodal latent reasoning. This approach effectively mitigates the attention dilution problem present in existing methods that rely on explicit textual reasoning and full-image processing. On VQA and other visual reasoning tasks, IVT-LR significantly outperforms multiple strong baselines, achieving new state-of-the-art results in both reasoning accuracy and efficiency. Our findings demonstrate the potential of interleaved vision-text reasoning in latent space, offering a promising paradigm for building more efficient and perceptive vision-language models and inspiring future research on multimodal reasoning strategies.

Future work could explore more dynamic ways of visual latent reasoning, such as adaptively determining the optimal number of latent steps based on the complexity of the question, rather than relying on a fixed stage number. Furthermore, this approach is highly promising for extending its application beyond pure reasoning to broader sequential multimodal tasks, including planning and complex decision-making in dynamic environments.

Limitations

While IVT-LR achieves substantial success in advancing both the accuracy and efficiency of multimodal reasoning, there are a few limitations that need to be addressed. 1) The adaptive selection of latent vision inevitably introduces a small, fixed amount of additional tokens per step. However, these tokens are processed internally, not generated externally, which ensures our final inference speed remains the absolute best in the high-accuracy setting. 2) IVT-LR requires a specialized multi-stage training curriculum, making it inherently more complex than simple prompt-based methods. However, this complexity is a justifiable investment: it is the direct catalyst for the massive gains in both accuracy and efficiency, and the required training resources and time investment remain modest and acceptable relative to the scale of the performance benefits achieved across existing training paradigms.

Ethics Statement

The datasets employed in our research were publicly released and constructed via human interaction in English. This process ensures that user privacy is fully protected, with no inclusion of personal information in the data. All scientific artifacts utilized are publicly accessible for academic purposes under permissive licenses, and their application in this paper adheres to their intended use. Therefore, we believe that our research work meets the ethical standards of the conference.

Acknowledgements

The work described in this paper was supported by Research Grants Council of Hong Kong (PolyU/15207122, PolyU/15209724, PolyU/15213323, PolyU/15205325) and PolyU internal grants (BDWP).

References

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221.

Jeffrey Cheng and Benjamin Van Durme. 2024. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*.

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

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

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.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2024. Think before you speak: Training language models with pause tokens. In *The Twelfth International Conference on Learning Representations*.

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*.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.

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. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 139348–139379.

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. 2025. Scaffolding coordinates to promote vision-language coordination in large multimodal models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2886–2903.

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*.

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

Dairu Liu, Ziyue Wang, Minyuan Ruan, Fuwen Luo, Chi Chen, Peng Li, and Yang Liu. 2025. Visual abstract thinking empowers multimodal reasoning. *arXiv preprint arXiv:2505.20164*.

Pan Lu, Swaroop Mishra, Tony 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. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 2507–2521.

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.

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

OpenAI. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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*.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 8612–8642.

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

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xinyi Wang, Lucas Caccia, Oleksiy Ostapenko, Xingdi Yuan, William Yang Wang, and Alessandro Sordoni. 2024b. Guiding language model reasoning with planning tokens. In *First Conference on Language Modeling*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 24824–24837.

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

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*.

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

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 5168–5191.

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

A Training Details

A.1 Training Setup

In IVT-LR training, we use a stage number N of four, a batch size of four, and the Adam optimizer with a learning rate set to 4×10^{-5} and $\beta_1 = 0.9$. All experiments are run on four NVIDIA A6000 GPUs (48GB VRAM each).

A.2 Rationale for $N = 4$ Training Stages

In the IVT-LR training, we set the number of stages (N) to 4, which corresponds to three core reasoning steps ($\mathbf{N - 1 = 3}$). This design choice is not arbitrary; it is motivated by a statistical analysis of the native rationale lengths in the target datasets (M^3CoT and ScienceQA).

We first examined the distribution of rationale steps (segmented by sentence) in the two datasets. As shown in Figure 6, the median number of rationale steps for both datasets is around 10. Each subtask in the reasoning process usually requires about two to three sentences to complete a causal inference. Thus, a full rationale can be naturally divided into three major reasoning steps, each corresponding to a distinct subtask. Therefore, setting $N = 4$ (three reasoning steps) provides a balanced and interpretable abstraction of the overall reasoning process.

Moreover, the statistical analysis shows that over 70% of the samples in both datasets contain more than three reasoning steps, and their distributions are highly dispersed. This high dispersion mandates merging adjacent steps for standardization and enhanced computational efficiency. Critically, we simultaneously retain a portion of the original one- and two-step samples. This strategy is essential to preserve the model's short reasoning ability and bolster generalization across varying reasoning depths, ensuring robust performance regardless of the input's complexity.

Method	Accuracy (%) ↑	#AR Steps ↓	Avg. Time (s) ↓
ICoT	53.7	78.3	2.30
SCAFFOLD	54.6	121.2	4.54
Chain-of-Focus	79.3	156.9	2.67
IVT-LR	**81.1**	**11.0**	**0.66**

Table 5: Performance of IVT-LR and baselines on A-OKVQA.

A.3 Examples of Merged Rationales

Table 7 and 8 illustrates examples of consolidated rationales with step indices.

B Baselines Details and Comparison Settings

For clarity of comparison, we explicitly outline the training or evaluation setup for each method. IVT-LR and Multimodal-CoT are trained via in-domain fine-tuning on the specific reasoning datasets used for evaluation, ensuring task-specific supervision. Chain-of-Focus is trained on a constructed dataset (MM-CoF) using a two-stage procedure of supervised fine-tuning followed by reinforcement learning before evaluation, representing a cross-domain fine-tuning setting. SCAFFOLD, CCoT, and ICoT are evaluated in 1-shot prompting settings without dataset-specific fine-tuning: SCAFFOLD employs positional scaffolding prompts, CCoT uses compositional scene graph based prompting, and ICoT uses interleaved-modal reasoning rationales. These distinctions clarify which methods rely on explicit training data and which operate with minimal or no fine-tuning, helping readers fairly interpret the performance differences presented in the main results.

C Extended Experiments and Generalizability Analysis

C.1 Justification for Selecting M3CoT and ScienceQA

We initially chose M3CoT and ScienceQA because, similar to many Chain-of-Thought approaches, our

method requires explicit reasoning rationales for supervision. These two datasets are currently the most comprehensive, commonly adopted, and highest-quality datasets in the multimodal reasoning domain that provide the necessary training rationales. Other large VQA datasets often lack such step-by-step reasoning annotations.

C.2 Broader Validation

To further support the generality of IVT-LR, we conducted supplementary experiments on additional reasoning datasets, general VQA benchmarks, and an alternative backbone model.

C.2.1 New Reasoning Dataset: A-OKVQA

We evaluated IVT-LR by training Qwen2-VL on the less commonly used A-OKVQA dataset, which provides reasoning rationales in the training set. The results are summarized in Table 5.

IVT-LR maintains superior accuracy and efficiency on A-OKVQA, demonstrating robustness across different reasoning data distributions.

C.2.2 Additional Backbone: Qwen2.5-VL-7B

To further validate IVT-LR's effectiveness, we evaluate it along with several baselines on M^3CoT and ScienceQA using the Qwen2.5-VL-7B backbone. The results are summarized in Table 6.

Summary. Across all experiments, including different datasets and multiple backbone models, IVT-LR consistently demonstrates superior accuracy and significantly improved inference efficiency. These results confirm that our method is robust, generalizable, and effective for multimodal latent reasoning in diverse reasoning scenarios and model architectures.

Backbone	Methods	M^3CoT			ScienceQA		
		Acc.(%) ↑	# AR Steps ↓	Avg. Time(s) ↓	Acc.(%) ↑	# AR Steps ↓	Avg. Time(s) ↓
Qwen2.5-VL-7B	ICoT	48.2	98.1	2.90	68.3	76.9	2.15
	SCAFFOLD	46.4	152.9	4.93	63.7	146.4	4.82
	Chain-of-Focus	66.6	166.3	2.55	92.3	160.2	1.98
	IVT-LR	**75.0**	**10.0**	**0.67**	**94.9**	**11.0**	**0.68**

Table 6: Comparison of IVT-LR with baselines on M^3CoT and ScienceQA using the Qwen2.5-VL-7B backbone.

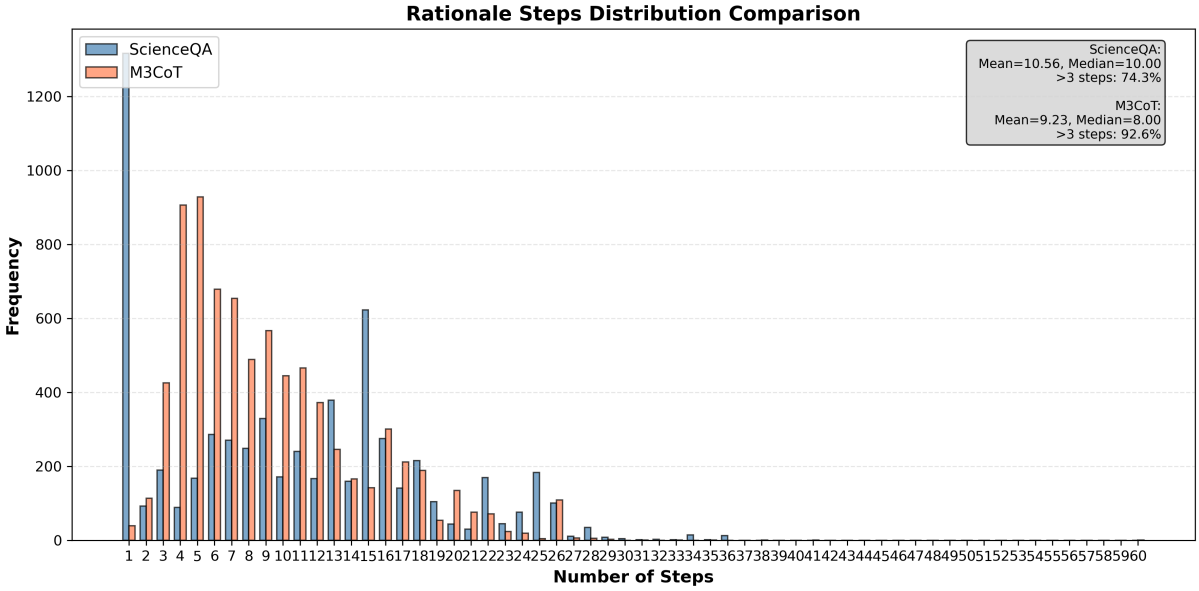


Figure 6: Distribution of native rationale steps across the M3CoT and ScienceQA datasets.

Question	What is the purpose of the hairdryer in the adult's hand?
Rationale	**Step 1:** According to the picture, the hair dryer in the adult's hand is not pointed at the hair. This suggests it has other uses besides drying hair. **Step 2:** There is a light ball in the air suggests that the air from the hairdryer is holding the ball up. Combined with the dancing little boy in the picture, this shows that the hairdryer is being used to entertain the little boy. **Step 3:** Therefore, (C) "Entertaining the little boy" is the right answer.

Table 7: An example of consolidated rationales for clarity.

Question	What is the purpose of the metal pylon on the street near the brick apartment building?
Rationale	**Step 1:** The metal pylon on the street indicates that cars are not allowed to drive in the pedestrian area. This inference is derived from the fact that the building next to it is an apartment building which suggests a residential area with high pedestrian traffic. **Step 2:** Additionally, the presence of thick white stripes across the street indicates a pedestrian crosswalk. Therefore, it can be concluded that the metal pylon is placed to prevent any intrusion from cars into the area reserved for pedestrians. **Step 3:** Option B is the correct answer.

Table 8: Another example of consolidated rationales for clarity.