Stabilizing Latent Visual Reasoning with Multi-Step Visual-Thought Alignment

Anonymous ACL submission

Abstract

Vision-language models (VLMs) demonstrate strong capabilities in visual question answering (VQA), yet their reasoning remains confined to the text modality, limiting their alignment with inherently visual tasks. Latent visual reasoning emerges as a promising alternative, offering more efficient and flexible inference. However, existing approaches focus primarily on modeling latent visual cues, overlooking the importance of aligning latent reasoning with standard text decoding. In this work, we propose a supervised multi-step latent reasoning framework that scales implicit reasoning to VLMs. Our method introduces step-level supervision on latent hidden states via a freezing LLM decoder during training, bridging the modality gap and enriching semantic diversity. The LLM decoder is removed at inference to maintain native decoding efficiency. Experiments show that our latent visual reasoning approach matches the performance of explicit CoT finetuning on Qwen3-VL, while significantly reducing token usage.

1 Introduction

Vision-Language Models (VLMs) have achieved remarkable success in multimodal understanding and reasoning, driven by large-scale pretraining and advancements in supervised fine-tuning. By aligning visual encoders with Large Language Models (LLMs), these systems can process complex visual inputs and generate coherent textual responses(Wang et al., 2023; Chen et al., 2023; Li et al., 2024). This capability has been further amplified by Chain-of-Thought (CoT) finetuning and Reinforcement Learning (RL)(Zhai et al., 2024), which encourage models to decompose complex queries into intermediate reasoning steps. Such paradigms have become the standard for handling complex tasks with multi-hop reasoning an establish a robust baseline for multimodal intelligence.

Despite these advancements, the reasoning process in current VLMs remains text-centric. Standard CoT makes the model to verbalize every intermediate thought into discrete text tokens, we describe it as **thinking about images**. This externalization introduces a modality gap: rich, high-dimensional visual cues are inevitably compressed and lost during the textual reasoning process, leading to hallucinations or insufficient visual usage(Sun et al., 2025; Qin et al., 2025). To address this, recent works have explored **thinking with images** via tool-use or interleaved multimodal CoT generation, where models invoke external tools to sketch, crop, or zoom the visual input (Hu et al., 2024; Qiao et al., 2025; Zhang et al., 2025a; Zheng et al., 2025; Fan et al., 2025; Su et al., 2025b; Chern et al., 2025; Wang et al., 2025). However, these methods often suffer from significant inefficiencies due to external calls; besides, text-driven control pipelines may prevent the model from performing fully native multimodal reasoning.

Latent reasoning, originally emerging in Large Language Models (LLMs), offers a promising alternative by shifting the locus of reasoning from the discrete text token space to the continuous embedding space (Hao et al., 2024a; Shen et al., 2025b). A line of work internalizes or compresses explicit CoT into latent trajectories to reduce generation overhead (Deng et al., 2024; Cheng and Durme, 2024; Su et al., 2025a). Other efforts study latent compression and soft thinking to trade off reasoning depth and efficiency (Tan et al., 2025b; Zhang et al., 2025b). Separately, recurrent-depth and looped-transformer approaches scale test-time compute through latent iteration (Geiping et al., 2025; Saunshi et al., 2025; Shen et al., 2025a). Instead of decoding logits into text, models generate a sequence of continuous hidden states which we named latent thoughts. They feed back autoregressively into the model as inputs for subsequent reasoning steps. Because these states remain in the

high-dimensional continuous space, they retain a significantly higher bandwidth of information compared to discrete text tokens. This allows the model to maintain a rich, uncollapsed representation of its working memory, facilitating more complex logic manipulation without the computational burden of autoregressive text generation.

While latent reasoning has shown promise in LLMs, its extension to the multimodal domain remains underexplored. Existing attempts at latent visual reasoning typically focus on explicitly modeling latent visual cues(Ma et al., 2025; Yang et al., 2025). Recent work often employing reconstruction objectives to force latent states to represent visual features instead of pure thoughts. We argue that this approach artificially decouples visual and textual semantics within the reasoning process. In the deep layers of a well-aligned VLM, visual and textual representations are already fused into a unified semantic space. Forcefully separating them or over-emphasizing visual reconstruction neglects the abstract logical nature of reasoning. The goal of a reasoning model should be to optimize the trajectory of thought toward a solution, not to act as an image compressor.

In this work, we propose a novel framework for Step-level Supervised Latent Visual Reasoning that embraces the unified nature of high-dimensional semantics in latent thoughts. We claim that the optimal reasoning state is modality-agnostic when operating at the deepest layers of the LLMs, leading to a subconscious reaction that forms before explicit decoding to discrete text tokens. To harness this, we introduce a mechanism where the model generates continuous thought vectors supervised by an step-level decoder. This Explainable Decoder (a freezing LLM) acts as a probe, ensuring that the latent thoughts encode valid, interpretable reasoning steps aligned with explicit CoT. Critically, this decoder is used only during training to align the latent trajectory with ground-truth reasoning. At inference, decoder is removed to ensure the model reasoning purely in the efficient, high-bandwidth latent space.

Our contributions are threefold:

1. We introduce a supervised latent visual reasoning framework that eliminates the reliance on external tool-uss, enabling fully internalized multimodal reasoning.

2. We propose a novel step-level visual-thought alignment mechanism via an explainable decoder, which effectively grounds continuous latent states in interpretable logic without enforcing visual-text-latent modality decoupling.

3. We demonstrate that our approach matches explicit CoT methods on benchmarks such as ScienceQA and VSR while significantly reducing inference latency and token consumption, validating the efficiency of latent visual thinking.

2 Related Work

2.1 Multimodal LLM

Modern VLMs typically bridge the gap between vision and language by connecting a frozen visual encoder to an LLM via a learnable interface. Architectures such as Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023) employ cross-attention mechanisms (e.g., Perceiver Resampler or Q-Former) to query visual features, while LLaVA (Liu et al., 2023b) utilizes a simplified linear projection to map image patches directly into the textual embedding space. While these architectures provide a foundation for perception, their reasoning capabilities are heavily augmented by Chain-of-Thought (CoT) prompting (Wei et al., 2022; Bi et al., 2025; Tian et al., 2025; Huang et al., 2025), which decomposes tasks into intermediate textual steps. Recent works have further refined this via Instruction Tuning and Reinforcement Learning from Human Feedback (RLHF) (Sun et al., 2023), yet these methods fundamentally rely on discrete text generation, constraining the reasoning process to the limitations of the vocabulary and sequence length.

2.2 Reasoning in Latent Space

Latent reasoning aims to bypass the constraints of discrete token generation by operating in continuous space. In the language domain, Coconut(Hao et al., 2024a) proposed treating the final-layer hidden states of an LLM as continuous thoughts, looping them back as inputs to facilitate internal reasoning without decoding. CODI(Shen et al., 2025b) and SIM-CoT(Wei et al., 2025) further advanced this by compressing explicit CoT into latent trajectories via self-distillation. In the multimodal domain, research is nascent. CoCoVa (Ma et al., 2025) attempts to extend this by iteratively refining joint vision-language representations through

a latent Q-Former loop, allowing the model to revisit visual features. However, unlike our approach which emphasizes semantic alignment of the reasoning trajectory, these methods often focus on mechanism design for feature fusion rather than explicitly supervising the interpretability and logical progression of the latent states.

2.3 Thinking with Images: From Static Perception to Active Interaction

To overcome the passive nature of standard visual encoding, several frameworks have introduced mechanisms to actively interact with visual inputs, termed *thinking with images*. Visual Sketchpad(Hu et al., 2024) introduced Visual Sketchpad, allowing VLMs to generate auxiliary visual markers (lines, boxes) to aid geometric reasoning. V-Thinker(Qiao et al., 2025) and Thyme(Zhang et al., 2025a) integrated code-based tools, enabling models to perform operations like cropping and zooming to attend to fine-grained details. Deepeyes(Zheng et al., 2025) further optimized this using Reinforcement Learning to learn policies for selecting pixel-level operations. While these methods enhance visual grounding, they rely on external, non-differentiable tools and discrete control tokens. Our work diverges from this paradigm by internalizing the "active" reasoning process into the differentiable latent space, avoiding the overhead of external tool invocation.

3 Methodology

We propose a Supervised Latent Visual Reasoning framework that scales implicit reasoning to Vision-Language Models (VLMs). Our method enables reasoning in a continuous high-dimensional space while strictly enforcing semantic interpretability through a frozen auxiliary supervisor.

3.1 Preliminaries

Let $\mathcal{V}$ denote the vocabulary and $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d}$ be the token embedding matrix. We define the multimodal reasoning task on a dataset of triplets $(\mathbf{X}_v, \mathbf{X}_q, \mathbf{A})$, where $\mathbf{X}_v$ represents the input image, $\mathbf{X}_q$ is the textual query, and $\mathbf{A}$ is the ground-truth answer. To facilitate complex reasoning, we leverage ground-truth Chain-of-Thought (CoT) rationales $\mathbf{C} = \{s_1, s_2, \ldots, s_K\}$, where each step s_k consists of a sequence of discrete tokens.

We adopt Qwen3-VL (Bai et al., 2025) as our reasoning backbone $\mathcal{M}_\theta$. A critical feature of this architecture is the DeepStack mechanism (Meng et al., 2024), which injects multi-scale visual tokens directly into the deep layers of the LLM via interleaved cross-attention. Unlike shallow-fusion VLMs where visual information fades over long contexts, DeepStack ensures that the hidden states remain visually grounded throughout the reasoning process. This property is vital for our framework, as it prevents the latent reasoning trajectory from decoupling from the visual input.

3.2 Latent Visual Reasoning Process

As illustrated in Figure 1, our framework replaces the discrete tokens of $\mathbf{C}$ with a sequence of continuous latent thoughts $\mathbf{Z} = \{z_1, \ldots, z_K\}$, where each $z_k \in \mathbb{R}^d$ acts as a compressed semantic representation of the reasoning step s_k. The inference process proceeds in two phases: an implicit reasoning phase followed by an explicit generation phase.

Implicit Reasoning Phase. The model iterates for K steps in a recurrent manner without decoding discrete tokens. Let $\mathbf{H}^{(0)}$ be the initial embedding sequence of the input $(\mathbf{X}_v, \mathbf{X}_q)$. At each reasoning step $k \in \{1, \ldots, K\}$, the model computes the hidden states:

$$\mathbf{H}^{(k)} = \mathcal{M}_\theta(\mathbf{H}^{(k-1)} \oplus z_{k-1}) \tag{1}$$

where $\oplus$ denotes concatenation along the sequence dimension. The latent thought token z_k is defined as the last-layer hidden state at the final position of the current sequence:

$$z_k = \text{Last}(\mathbf{H}^{(k)}) \in \mathbb{R}^d \tag{2}$$

Crucially, z_k is not projected to the vocabulary $\mathcal{V}$. Instead, it is directly appended to the input context for the subsequent step. This establishes a continuous gradient flow, preserving high-bandwidth semantic information that is typically lost during discrete sampling.

Explicit Generation Phase. Upon completion of K latent steps, the accumulated context $\mathbf{H}^{(K)} = [\text{Embed}(\mathbf{X}_v, \mathbf{X}_q); z_1; \ldots; z_K]$ is used to generate the final answer $\mathbf{A}$. The model switches to standard autoregressive decoding:

$$p_\theta(\mathbf{A} \mid \mathbf{X}_v, \mathbf{X}_q, \mathbf{Z}) = \prod_{t=1}^{M} p_\theta(a_t \mid \mathbf{X}_v, \mathbf{X}_q, \mathbf{Z}, a_{<t}) \tag{3}$$

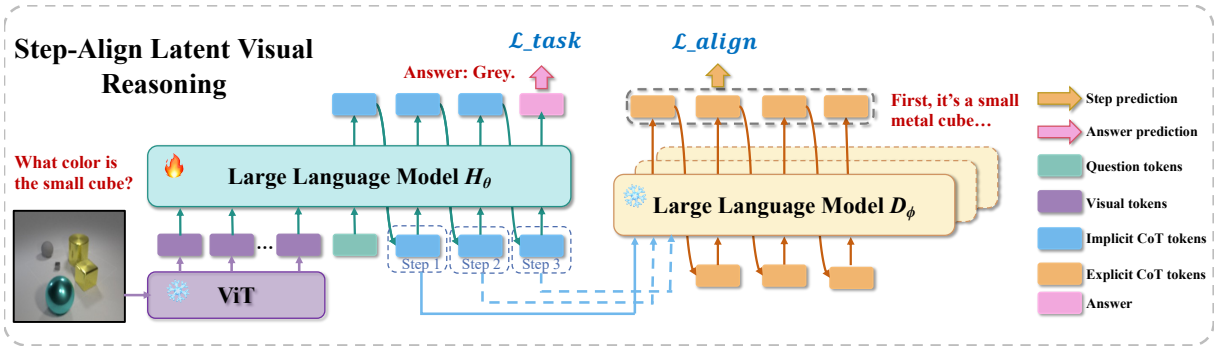


Figure 1: **Overview of the Supervised Latent Visual Reasoning framework.** The main backbone H_θ (left) performs implicit reasoning by generating continuous latent thoughts (blue tokens), which are recurrently fed into the next step. During training, a frozen LLM decoder D_ϕ (right) aligns these latent thoughts with explicit reasoning steps via $\mathcal{L}_{\text{align}}$. At inference, D_ϕ is removed.

3.3 Step-Level Alignment via Frozen Decoding

A fundamental challenge in latent reasoning is the risk of representation collapse, where continuous vectors $\mathbf{Z}$ degenerate into uninterpretable noise lacking logical structure. To mitigate this, we propose step-level visual-thought alignment, which supervises the latent space using an explainable decoder, denoted as $\mathcal{D}_\phi$.

Frozen LLM Backbone as Semantic Anchor. We instantiate $\mathcal{D}_\phi$ as a frozen replica of the reasoning LLM backbone $\mathcal{M}_\theta$. Specifically, $\mathcal{D}_\phi$ is initialized with the pre-trained weights of the base model, and its parameters are kept fixed ($\phi = \text{const}$) throughout training. This design choice is pivotal: by using a frozen copy of the backbone, we establish a static semantic anchor. Unlike a trainable decoder which might co-adapt with the encoder to interpret degenerate signals, the frozen decoder enforces strict compatibility. It requires the reasoning model $\mathcal{M}_\theta$ to generate latent thoughts z_k that align naturally with the pre-existing linguistic manifold of the LLM.

Step-level Reconstruction. During training, $\mathcal{D}_\phi$ takes the latent thought z_k as a soft prompt and reconstructs the corresponding explicit reasoning step s_k. We minimize the negative log-likelihood of the text step conditioned on the latent vector:

$$\begin{aligned}\mathcal{L}_{\text{align}}^{(k)} &= -\log p_\phi(s_k \mid z_k) \\ &= -\sum_{j=1}^{L_k} \log p_\phi(y_{k,j} \mid z_k, y_{k,<j}) \end{aligned} \tag{4}$$

This objective forces the latent vector z_k to compress the complete semantic content of s_k. Note that $\mathcal{D}_\phi$ is utilized exclusively during training to provide supervision signals; it is discarded during inference to ensure efficiency.

3.4 Training Objective

The framework is trained end-to-end using a composite objective function. We strictly enforce alignment at every latent step while optimizing for the final answer accuracy. The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda \sum_{k=1}^{K} \mathcal{L}_{\text{align}}^{(k)} \tag{5}$$

Here, $\mathcal{L}_{\text{task}} = -\sum_{t=1}^{M} \log p_\theta(a_t \mid \mathbf{X}_v, \mathbf{X}_q, \mathbf{Z}, a_{<t})$ represents the cross-entropy loss for the final answer generation. The hyperparameter λ balances the strength of the semantic constraint. By optimizing this objective, $\mathcal{M}_\theta$ learns to synthesize latent thoughts that are simultaneously effective for problem-solving via $\mathcal{L}_{\text{task}}$ and linguistically valid via $\mathcal{L}_{\text{align}}$.

4 Experiments

In this section, we empirically evaluate the efficacy of our proposed Supervised Latent Visual Reasoning framework. Our experimental design aims to validate three core hypotheses. We present a comprehensive comparison against baselines across diverse benchmarks.

4.1 Experimental Setup

Models and Training Our framework is instantiated using the **Qwen3-VL-4B-Instruct** backbone. We implement parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA) on the reasoning backbone, with a rank of 16 and alpha of 32. Optimization is performed using AdamW with a fixed

Table 1: **Main Results.** Comparison of accuracy (%) on In-Domain and Out-of-Domain (OOD) tasks. The left block includes standard benchmarks and the in-domain splits of Reason-RFT. The right block demonstrates generalization capabilities on the OOD splits of Reason-RFT.

Method	**In-Domain Performance**					**Out-of-Domain Performance**		
	SciQA	**M3CoT**	**VSR**	**Structure**	**Spatial**	**Structure**	**Spatial-L**	**Spatial-R**
No-CoT	78.68	59.70	62.88	20.61	22.72	30.63	7.69	7.54
Coconut	**92.46**	55.09	67.59	-	-	-	-	-
SFT-CoT	84.18	55.31	74.14	23.90	49.47	**57.63**	30.79	29.64
Ours	88.10	**60.83**	**82.80**	**38.66**	**55.81**	35.88	**41.75**	**42.53**

Table 2: **Efficiency comparison.** We report average generated tokens and inference time (s) across In-Domain (ID) and Out-of-Domain (OOD) splits. The speedup metric compares our method against the SFT-CoT baseline.

Method	Metric	SciQA	M3CoT	VSR	Structure		Spatial		
					ID	OOD	ID	OOD-L	OOD-R
No-CoT	Avg. Tokens	281.5	748.1	146.2	27.6	137.7	353.0	730.7	755.1
	Time (s)	7.77	20.38	8.05	3.01	5.23	8.56	15.36	16.12
Coconut	Avg. Tokens	7.18	71.9	4.1	-	-	-	-	-
	Time (s)	3.72	10.99	5.49	-	-	-	-	-
SFT-CoT	Avg. Tokens	768.0	3370.1	123.8	157.2	180.7	327.1	381.8	396.3
	Time (s)	16.31	34.51	7.32	7.60	8.83	13.67	15.66	16.42
Ours	Avg. Tokens	6.7	98.3	1.0	3.0	3.19	29.86	35.13	33.55
	Time (s)	3.45	10.50	3.13	2.80	3.23	4.15	3.91	3.88
Speedup (vs. SFT-CoT)		4.73×	3.29×	2.34×	2.71×	2.73×	3.29×	4.01×	4.23×

learning rate of 5×10^{-5}. To ensure training stability, we employ a curriculum learning strategy: the model is initially supervised with explicit Chain-of-Thought data to establish logical grounding, before gradually transitioning to the latent reasoning objective. We adopt a dynamic latent budget strategy, assigning $K = 10$ for logic-heavy tasks like M3CoT and $K = 5$ for concise tasks such as ScienceQA and VSR. We conducted all experiments on 8 NVIDIA A100 (40G) GPUs.

Datasets and Benchmarks We evaluate our framework across a comprehensive suite of multimodal reasoning benchmarks to assess both scientific synthesis and fine-grained visual logic. Our evaluation includes **ScienceQA** (Lu et al., 2022) for multimodal scientific knowledge and the complex split of **M3CoT** (Chen et al., 2024) for multi-step deductive reasoning. To scrutinize the model's ability to maintain spatial and visual details often lost in text, we explicitly incorporate the relation-reasoning subset of **VSR** (Shao et al., 2024; Liu et al., 2023a). Furthermore, we report performance on the in-domain and out-of-domain splits of **Reason-RFT** (Tan et al., 2025a) to verify generalization capabilities across distribution shifts.

Baselines We compare our method against three representative paradigms. Direct Generation (No-CoT) serves as a lower-bound baseline where the model directly predicts answers without intermediate reasoning. Explicit CoT represents the standard approach utilizing discrete text tokens, offering high accuracy but suffering from high latency. We also include Coconut (Hao et al., 2024b) as a latent reasoning baseline that lacks our proposed semantic thought alignment, serving as an ablation study for our supervised objective.

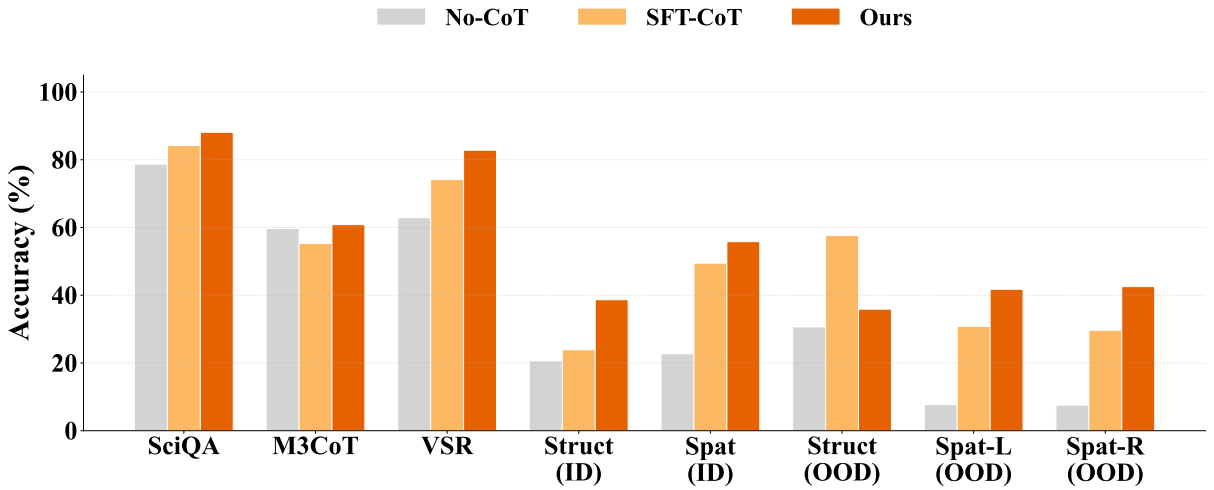


Figure 2: **Main Results on Reasoning Benchmarks.** Comparison of accuracy across standard benchmarks (SciQA, M3CoT, VSR) and Reason-RFT splits (Structure, Spatial). Our method (Orange) consistently outperforms baselines, particularly in OOD scenarios where Coconut often struggles. Visualizing Table 1.

4.2 Experimental Results

Performance Analysis. Table 1 summarizes the performance across all benchmarks. Our proposed framework consistently outperforms the Direct Generation baseline, confirming that internalizing reasoning steps is crucial for complex multimodal tasks. On ScienceQA, we achieve a substantial improvement of 9.42% over the no-reasoning baseline. Most notably, on the VSR benchmark, our method demonstrates a remarkable gain, surpassing the Direct Generation baseline by nearly 20% (from 62.88% to 82.80%) and significantly outperforming standard Explicit CoT (74.14%). This result supports our hypothesis that continuous latent tokens preserve high-bandwidth visual information such as spatial coordinates and fine-grained features, are often compressed or lost during the translation to discrete text. Furthermore, in the M3CoT deductive reasoning task, our method achieves the highest accuracy of 60.83%, indicating that the semantic thoughts alignment mechanism effectively regularizes the latent space for complex logical operations.

Computational Efficiency. Table 2 highlights the computational advantages of shifting reasoning to the latent space. By bypassing the autoregressive generation of intermediate text tokens, we drastically reduce the sequence length required for inference. On ScienceQA, the average token count drops from 281.5 in the No-CoT baseline to just 6.7 in our method, translating to a 2.25× wall-clock speedup. Similarly, on the computationally intensive M3CoT dataset, we observe a 1.94× speedup compared to Explicit CoT. These results verify that our framework breaks the dependency between reasoning depth and generation latency, offering a sustainable path for scaling inference-time compute.

Table 3: **Effect of Latent Thought Budget (K).** We scale the latent reasoning steps on M3CoT. Performance peaks at $K = 10$, balancing accuracy and token consumption.

Budget	Accuracy (%)	Avg. Tokens
2	54.44	67.20
4	56.08	73.00
6	56.61	72.60
8	56.17	78.40
10	**60.83**	98.30
12	60.05	87.03
14	58.24	83.90

4.3 Ablation Study

We investigate the scaling laws of latent reasoning by varying the latent token budget K on the M3CoT dataset. As shown in Table 3 and visualized in Figure 4, there is a clear positive correlation between the capacity of the latent thought process and task performance. Accuracy improves from 54.44% with minimal reasoning ($K = 2$) to a peak of 60.83% at $K = 10$. This trajectory suggests that allocating more thinking time and reasoning step in the latent space allows the model to perform deeper verification and integration of multimodal evidence. While setting $K = 10$ marginally increases the in-

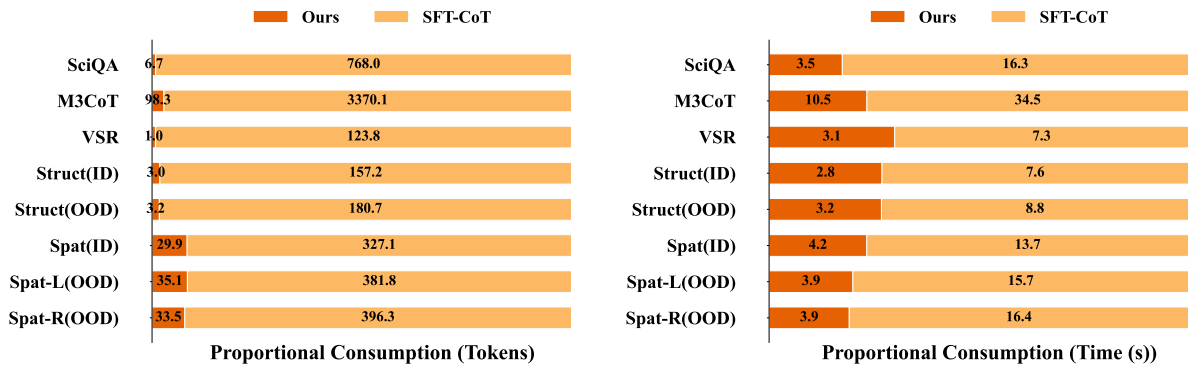


Figure 3: **Efficiency Comparison.** Side-by-side comparison from Table 2 of proportional resource consumption. **Left:** Average Generated Tokens. **Right:** Inference Time (s). The segments show the raw values; a smaller segment for Ours (Orange) indicates higher efficiency compared to SFT-CoT (Yellow). Our method significantly reduces overhead across most tasks in both metrics.

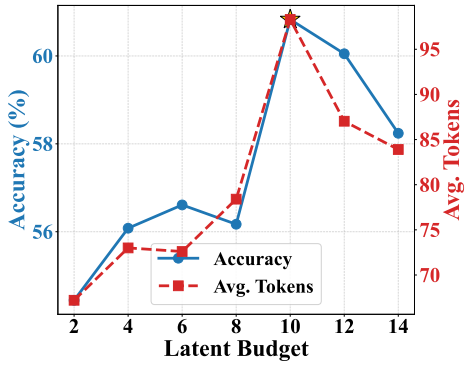


Figure 4: **Effect of Latent Thought Budget.** We scale the latent reasoning steps on M3CoT. Performance peaks at $K = 10$, effectively balancing reasoning accuracy with token consumption. Visualizing Table 3.

ference cost to 98.3 tokens, this investment yields a significant +6.39% accuracy gain, justifying the trade-off for complex reasoning tasks.

5 Analysis

Existing latent reasoning approaches often suffer from latent instability, where unsupervised continuous thoughts drift away from the language manifold, leading to semantic collapse. In this section, we scrutinize our framework to verify whether the proposed Semantic Thought Alignment effectively mitigates these risks. We conduct our analysis from two perspectives: the geometric topology of the latent space and the semantic interpretability of the decoded thoughts.

5.1 Geometric Structure and Modality Bridge

To validate that our latent thoughts maintain semantic validity under the supervision of the frozen decoder, we visualize the geometric relationship between modality-specific representations. We randomly sample 2,000 vectors from three distinct sources across the test sets of VSR, M3CoT, and Reason-RFT: Visual Tokens extracted by the vision encoder, Text Tokens from the pre-trained LLM embeddings, and our generated Latent Tokens ($\mathbf{Z}$). We project these high-dimensional representations into a two-dimensional space using t-SNE.

As illustrated in Figure 5, the visualization reveals a consistent topological pattern across all three datasets. Unlike unsupervised approaches where latent vectors often collapse into a singularity or drift into undefined regions, our latent tokens form a structured and dispersed manifold. Crucially, the Latent Token cluster is situated proximal to the Text Token manifold, partially overlapping with it. This geometric proximity serves as strong empirical evidence for our alignment hypothesis: the frozen LLM decoder acts as a semantic anchor, constraining the latent thoughts to reside within the valid linguistic activation space of the LLM.

Furthermore, the Latent cluster acts as a topological bridge, extending from the Text manifold towards the Visual manifold. This suggests that the continuous reasoning process successfully performs modality fusion. The latent vectors absorb high-bandwidth visual information, represented by the drift towards the visual cluster while retaining the discrete logical structure required by the language model. This bridging effect explains the model's superior performance on visual-heavy tasks like VSR, as the latent space captures fine-grained visual gradients that are typically lost when forcing a collapse to purely discrete text tokens.

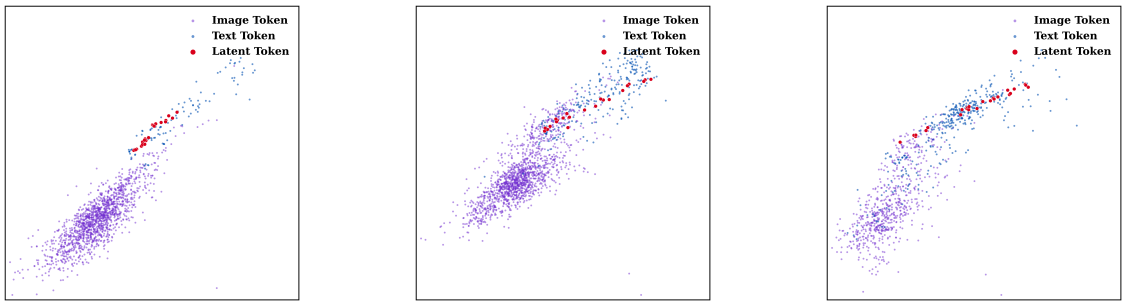


Figure 5: **Visualization of Token Distributions.** We visualize the latents and explicit tokens on VSR, M3CoT and Reason-RFT, respectively. In all three datasets, the latent tokens form a compact cluster structurally positioned between the visual and textual manifolds. This topological bridging and connecting effect confirms that the latent thoughts remain grounded in the linguistic space.

5.2 Semantic Decodability and Step-wise Logic

A prevalent failure mode in latent reasoning is semantic homogenization, where continuous latent thoughts degenerate into repetitive numerical patterns devoid of meaningful logic. To prove the internal reasoning process of our model, we utilize the frozen LLM decoder $\mathcal{D}_\phi$ to conduct decoding process. Although $\mathcal{D}_\phi$ is discarded during inference, it allows us to project the subconscious latent stream $\mathbf{Z} = \{z_1, \ldots, z_K\}$ back into the interpretable discrete text token space for qualitative analysis. Detailed cases are shown in Appendix A.

Temporal Evolution of Reasoning. Upon decoding the latent sequences generated by our model, we observe a coherent temporal evolution that mirrors the cognitive process of explicit Chain-of-Thought. As visualized in the qualitative examples in Appendix A, the reasoning trajectory exhibits a clear shift from perception to deduction. The early latent states typically decode into explicit visual grounding statements, capturing fine-grained attributes and spatial relationships extracted from the image. As the reasoning progresses to deeper steps, the decoded content transitions towards hypothesis verification and logical synthesis, aggregating the previously attended visual evidence to derive the final conclusion. This structured transition confirms that our frozen supervision effectively prevents the latent space from collapsing into noise, forcing the vectors to encode a semantic trajectory that is isomorphic to explicit human reasoning.

Richness of Latent Semantics. Furthermore, the decoding results reveal that the latent thoughts retain high-bandwidth visual information that is often compressed in standard text generation. We find that the decoded sequences frequently describe subtle visual cues, such as texture, fine-grained object states, or complex spatial configurations. These are pivotal for correct answering but might be omitted in a concise text-only CoT. This supports that the latent space operates as a unified modality space, allowing the model to think and reasoning using a dense representation that fuses visual features with linguistic logic. The ability of the frozen decoder to reconstruct these details verifies that z_k serves as a valid semantic interface between the visual encoder and the language backbone.

6 Conclusion

In this work, we presented a novel framework for Supervised Latent Visual Reasoning, bridging the gap between implicit continuous thought and explicit multimodal logic. By introducing a frozen explainable decoder as a semantic supervisor, we successfully stabilized the latent reasoning trajectory, ensuring that high-dimensional thought vectors remain interpretable and grounded in visual evidence. Our experiments on multiple benchmarks demonstrate that this approach not only achieves competitive accuracy compared to explicit Chain-of-Thought methods but also drastically reduces inference latency. These findings suggest that the future of efficient multimodal intelligence lies in shifting reasoning from the surface level of discrete token generation to the deep semantic space of latent representations, offering a robust path toward scalable reasoning in VLMs.

Limitations

While our framework improves efficiency and reasoning stability, several limitations remain. First, our method relies on high-quality Chain-of-

Thought annotations for step-level alignment during training; extracting such supervision for open-world visual tasks remains a challenge. Second, although the frozen decoder allows for post-hoc interpretation, the reasoning process during inference operates fundamentally as a black-box, making real-time user intervention or error correction less straightforward compared to explicit text generation. Finally, the current implementation employs a fixed budget for latent reasoning steps; developing adaptive mechanisms to dynamically determine the necessary reasoning depth based on instance complexity is a promising direction for future research.

References

Jean-Baptiste Alayrac and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Jinhe Bi, Danqi Yan, Yifan Wang, Wenke Huang, Haokun Chen, Guancheng Wan, Mang Ye, Xun Xiao, Hinrich Schuetze, Volker Tresp, and Yunpu Ma. 2025. Cot-kinetics: A theoretical modeling assessing lrm reasoning process.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024. M^3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *Preprint*, arXiv:2405.16473.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Preprint*, arXiv:2312.14238.

Jeffrey Cheng and Benjamin Van Durme. 2024. Compressed chain of thought: Efficient reasoning through dense representations. *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. *Preprint*, arXiv:2505.22525.

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

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. 2025. Grit: Teaching mllms to think with images. *Preprint*, arXiv:2505.15879.

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

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

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

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. *arXiv preprint arXiv:2406.09403*.

Xingyue Huang, Rishabh, Gregor Franke, Ziyi Yang, Jiamu Bai, Weijie Bai, Jinhe Bi, Zifeng Ding, Yiqun Duan, Chengyu Fan, Wendong Fan, Xin Gao, Ruohao Guo, Yuan He, Zhuangzhuang He, Xianglong Hu, Neil Johnson, Bowen Li, Fangru Lin, and 27 others. 2025. Loong: Synthesize long chain-of-thoughts at scale through verifiers. *Preprint*, arXiv:2509.03059.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *Preprint*, arXiv:2407.07895.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. 2023a. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*.

Haotian Liu, Chunyuan Li, Pengchuan Zhang, Xiangning Chen, Jianwei Yang, and 1 others. 2023b. Llava: Large language and vision assistant. *Preprint*, arXiv:2304.08485.

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 *Advances in Neural Information Processing Systems*, volume 35.

Jizheng Ma, Xiaofei Zhou, Yanlong Song, and Han Yan. 2025. Cocova: Chain of continuous vision-language thought for latent space reasoning.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. 2024. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. In *Advances in Neural Information Processing Systems*, volume 37, pages 23464–23487.

Runqi Qiao, Qiuna Tan, Minghan Yang, Guanting Dong, and 1 others. 2025. V-thinker: Interactive thinking with images. *arXiv preprint arXiv:2511.04460*.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. 2025. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *Preprint*, arXiv:2511.19418.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. 2025. Reasoning with latent thoughts: On the power of looped transformers. *Preprint*, arXiv:2502.17416.

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 *Advances in Neural Information Processing Systems*, volume 37, pages 8612–8642. Curran Associates, Inc.

Xuan Shen, Yizhou Wang, Xiangxi Shi, Yanzhi Wang, Pu Zhao, and Jiuxiang Gu. 2025a. Efficient reasoning with hidden thinking. *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.

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

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

Nan Sun, Zhenyu Zhang, Xixun Lin, Kun Wang, Yanmin Shang, Naibin Gu, Shuohuan Wang, Yu Sun, Hua Wu, Haifeng Wang, and Yanan Cao. 2025. V-iti: Mitigating hallucinations in multimodal large language models via visual inference-time intervention. *Preprint*, arXiv:2512.03542.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning large multimodal models with factually augmented rlhf.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. 2025a. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. *Preprint*, arXiv:2503.20752.

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

Yijun Tian, Shaoyu Chen, Zhichao Xu, Yawei Wang, Jinhe Bi, Peng Han, and Wei Wang. 2025. Reinforcement mid-training. *Preprint*, arXiv:2509.24375.

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

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. 2025. Sim-cot: Supervised implicit chain-of-thought. *Preprint*, arXiv:2509.20317.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. 2025. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *Preprint*, arXiv:2506.17218.

Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and Sergey Levine. 2024. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *Preprint*, arXiv:2405.10292.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, and 1 others. 2025a. 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. 2025b. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *Preprint*, arXiv:2505.15778.

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. *Preprint*, arXiv:2505.14362.

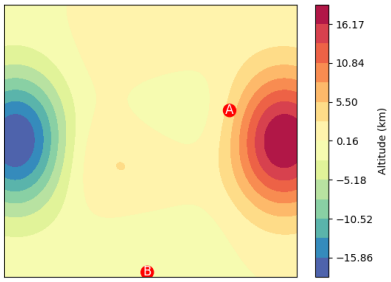
A Qualitative Analysis of Decoded Latent Thoughts

In this section, we present the full decoding results from our frozen auxiliary decoder $\mathcal{D}_\phi$ across five diverse reasoning scenarios from the M3CoT dataset. These cases cover fine-grained visual perception, abstract geometric reasoning, deductive logic, and complex spatial/mathematical estimation. The tables below show the exact text reconstructed from the continuous latent tokens $\mathbf{Z} = \{z_1, \ldots, z_{10}\}$ generated by our model during the implicit reasoning phase.

Question: What can be inferred about the child in the picture?
GT Answer: C (The child enjoyed her birthday cake and is 1 year old)

Step	Decoded Thought $\mathcal{D}_\phi(z_k)$
z_2	option B is the correct answer. B
z_3	that she enjoyed eating the cake. Step2 : Therefore, the correct answer is C. Step3 : Therefore, option C is the correct answer. C
z_4	her mouth, indicating that she has finished eating her cake. Step3 : Therefore, it can be inferred that the child enjoyed her birthday cake and is 1 year old. Step4 : Therefore, the correct answer is option C. C
z_6	on her face indicates that she enjoyed eating the cake. Step3 : Therefore, the correct answer is C, which states that the child enjoyed her birthday cake and is 1 year old. Step4 : Therefore, the answer is C. C
z_7	's face smeared with frosting suggests that she enjoyed her cake and smeared the frosting on her face. Step3 : Therefore, option C is the correct answer. \answer\ C
z_8	on her face suggests that she enjoyed her cake enough to smear it on her face. Step2 : The evidence suggests that the child is 1 year old and has a light brown tattoo on her right arm. Step3 : The evidence does not provide any information about the child's taste in candy or whether she had a birthday cake. Step4 : Therefore, the correct answer is C, that the child enjoyed her birthday cake and is 1 year old. C
z_9	, option C is the correct answer. C

Figure 6: **Case 1: Fine-grained Visual Perception.** The model identifies details like "smeared with frosting" (z_7) and "1 year old" (z_4, z_6) to infer the child's status.



Input Image

Question: Point out the spot on the picture where the altitude is at its lowest.
GT Answer: B (Point B)

Step	Decoded Thought $\mathcal{D}_\phi(z_k)$
z_1	the point B. - Examine the corresponding color and compare it to the legend. - We can make a rough estimation that the altitude is roughly 0.2 kilometers. Step2 : Determine the altitude of point A. - Search for the contour lines that intersect or come close to this specific point. - Examine the corresponding color and compare it to the legend. - We can make a rough estimation that the altitude is roughly 5.2 kilometers. Step3 : With a clear conscience, we can state that 0.2 < 5.2. Consequently, Point B's altitude is the lowest. Step4 : So choice "(B) Point B" is the correct answer. B
z_2	proximity to the point B. - Analyze the associated color and match it with the legend. - We can roughly gauge that the altitude stands at approximately 0.4 kilometers. Step2 : Since 0.4 < 5.6, it is clear that Point B holds the record for the lowest altitude. Step3 : Therefore, choice "B.Point B" should be selected. B
z_3	to the point B. - Examine the corresponding color and compare it to the legend. - We can make a rough estimation that the altitude is roughly 0.1 kilometers. Step3 : Now, let's compare the altitudes of the two points: - The altitude of point A is approximately 4.4 kilometers. - The altitude of point B is approximately 0.1 kilometers. Step4 : Thus, Point B has the lowest altitude. Step5 : Hence, the solution is choice "(B) Point B". B
z_4	through or closely approach this location. - Read the corresponding color, and compare it with the color in the legend. - We can make an approximate estimate that the elevation is around 5.1 kilometers. Step2 : Determine the altitude of point B. - Now, seek out the contour lines that intersect with it or pass nearby. - Analyze the associated color and match it with the legend. - We can roughly gauge that the altitude stands at approximately -0.1 kilometers. Step3 : - Due to the hierarchy of numbers, -0.1 < 5.1. - Thus, point B's position at the lowest altitude. Step4 : Therefore, the answer is choice "(B) Point B" B
z_5	go through or closely approach this location. - Read the corresponding color, and compare it with the color in the legend. - We can make an approximate estimate that the elevation is around 0.4 kilometers. Step2 : Determine the altitude of point B. - Examine the contour lines that either go through or closely approach this location. - Examine the corresponding color and compare it to the legend. - We can make a rough estimation that the altitude is roughly 1.7 kilometers. Step3 : - Due to the hierarchy of numbers, 0.4 < 1.7. - Hence, it is clear that Point B's position at the lowest altitude. Step4 : Therefore, the solution is choice "(B) Point B". \answer\ B
z_9	contour lines that either go through or closely approach this location. - Observe the corresponding color and cross-reference it with the legend. - We can make a rough estimation that the altitude is roughly 0.6 kilometers. Step2 : Ascertain the elevation of point B. - Now, seek out the contour lines that intersect with it or pass nearby. - Read the corresponding color, and compare it with the color in the legend. - We can make a rough estimation that the altitude is roughly 0.8 kilometers. Step3 : Because 0.6 < 0.8. Hence, Point B has the lowest altitude. Step4 : For these reasons, choice "(B) Point B" is the correct answer. \answer\ B

Figure 7: **Case 2: Quantitative Spatial Analysis.** The model reads the legend and estimates numerical altitude values to perform comparisons ($0.2 < 5.2$).

Input Image

Question: What can be inferred from the image?
GT Answer: D (The store caters to customers of all ages)

Step	Decoded Thought $\mathcal{D}_\phi(z_k)$
z_2	Step2 : Option B is incorrect because the store also sells toys and other items that may not be suitable for children. Step3 : Option C is incorrect because there is no evidence that the step stool is used by employees, only that it is used by customers to reach higher shelves. Step4 : Option D is correct because the store sells books, toys, and other items that may be suitable for customers of all ages. Conclusion: Therefore, option D is the correct answer. |answer| D
z_3	and caters to customers of all ages. Therefore, option D is the correct answer. Step2 : The step stool could be used by employees to reach higher shelves, but this is not directly related to the age range of customers. Therefore, option C is incorrect. Step3 : There is no direct evidence to suggest that the store is exclusively for children, so option B is incorrect. Step4 : The store sells more than just books, as evidenced by the stuffed animals in a basket, so option A is incorrect. Therefore, option D is the correct answer. |answer| D
z_4	correct answer. |answer| D
z_6	in a basket on the floor may not be for sale, but could be used for decoration or as a way to engage with younger customers. Step2 : Therefore, option D is the correct answer. D
z_8	caters to customers of all ages. Step2 : Therefore, the correct answer is D. Step3 : The other options are incorrect as there is no evidence to suggest that the store only sells books or is exclusively for children. Therefore, option D is the correct answer. D

Figure 8: **Case 3: Deductive Logic and Exclusion.** The latent thoughts demonstrate a rigorous elimination process, explicitly ruling out Options A, B, and C based on visual evidence.