DO LATENT STATES REALLY REASON? MAKING IT GREAT AGAIN THROUGH INFORMATION CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

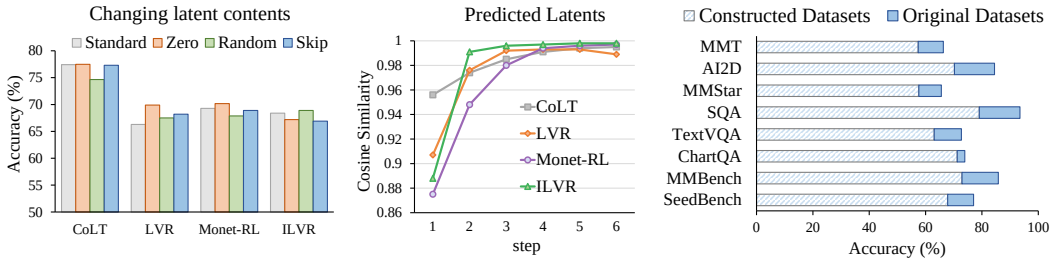
Textual chain-of-thought (CoT) improves multimodal reasoning but incurs substantial autoregressive decoding cost. Latent reasoning reduces this overhead by compressing intermediate reasoning into a fixed number of latent states. Despite recent progress, existing methods often implicitly assume that supervised latent states necessarily encode useful reasoning information. Our empirical results challenge this assumption, showing that latent states do not necessarily encode meaningful semantic information and that even skipping them during inference can have little impact on performance. These findings call into question whether current latent reasoning methods truly learn informative and interpretable latent representations. To address this, we propose ICLR (**I**nformation **C**ontrol for **L**atent **R**easoning), which explicitly controls information flow to make latent states both necessary for prediction and effective for reasoning. Specifically, we design a two-stage framework. In the first stage, we explicitly route all information through the latent states, making them necessary intermediates in the reasoning pathway. In the second stage, we further extract the visual evidence and textual semantics required for step-level reasoning and reinforce the latent states to encode this information, thereby enabling effective latent reasoning. Extensive experiments show that ICLR achieves state-of-the-art performance across multimodal benchmarks, surpassing existing latent reasoning baselines and latent visual reasoning approaches requiring auxiliary images and costly annotations. Compared to textual CoT methods, ICLR achieves an average performance gain of **3.6%** while delivering a **20.9**$\times$ generation speedup. Code will be released.

1 INTRODUCTION

Multimodal large language models (MLLMs) bridge visual perception and general reasoning across diverse tasks, often relying on textual chain-of-thought (CoT) to decompose complex problems and improve reasoning accuracy (Wang et al., 2025; Gao et al., 2024). However, textual CoT requires autoregressive token-by-token generation, incurring substantial decoding latency and computational cost. Moreover, textual decoding collapses diverse latent solution trajectories into a single natural-language realization, limiting alternative reasoning paths Hao et al. (2024). Latent reasoning has attracted increasing attention as an alternative to textual CoT, as it performs intermediate reasoning in a continuous representation space rather than through explicit text generation, thereby reducing decoding overhead while preserving reasoning capabilities (Zhu et al., 2025b; Yu et al., 2026).

Existing latent reasoning approaches can be broadly categorized into two groups: (i) Some methods introduce auxiliary visual supervision to align latent states with visual features encoded from auxiliary images, encouraging models to reason based on task-relevant visual information (Li et al., 2026a; Wang et al., 2026; Dong et al., 2026). (ii) Other methods align latent states with textual reasoning signals to preserve the structured semantics of intermediate reasoning while reducing reasoning costs (Hu et al., 2026; Tan et al., 2026; Wei et al., 2026). Despite achieving improvements across various tasks, these methods share a core assumption that injecting sufficient information into latent states can make them meaningful semantic reasoning variables. However, our empirical analysis challenges this assumption. As shown in Figure 1(a), we conduct intervention experiments by replacing latent states with information-free representations, such as zero or random vectors. The results show that model performance remains nearly unchanged across different benchmarks, indicating that these latent states may not contain the effective information required for prediction.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) (b) (c)

Figure 1: **Systematic analysis of latent visual reasoning.** (a) Changing latent contents has little impact on predictions or performance. (b) Latent representations remain similar across reasoning steps, lacking step-specific semantics. (c) Models (e.g., CoLT (Hu et al., 2026)) trained on reconstructed datasets (Viveiros et al., 2026) suffer substantial performance drops on other benchmarks.

Furthermore, we observe that different latent states exhibit highly similar semantic representations (Figure 1(b)), revealing a lack of step-level specificity corresponding to the reasoning process. These empirical findings suggest that existing latent reasoning methods may fail to learn sufficiently informative reasoning representations, raising a critical question: ***Do latent states really reason?***

Recent work has begun to investigate whether latent states encode useful semantics and to explore strategies for enhancing their semantic content (Liang & Pan, 2026; Li et al., 2026b). Viveiros et al. (Viveiros et al., 2026) show that cropping or zooming into local image regions introduces no fundamentally new information beyond the original image. They further improve latent visual reasoning through task-specific data reconstruction and intermediate visual supervision, while Zhang et al. Zhang et al. (2025) reactivate suppressed visual information in "silenced visual latents" via inference-time optimization. However, these approaches rely on task-specific reconstruction or additional supervision, thus incurring heavy reliance on specifically configured datasets and failng to generalize to broader backbones and data distributions. Moreover, we find that models trained on these reconstructed data exhibit degraded performance on other benchmarks, as shown in Figure 1(c), indicating task-specific overfitting and catastrophic forgetting. Therefore, the key challenge lies not in stronger augmentation or supervision, but in addressing two fundamental questions:

> *How to enable latent states to carry useful semantics, and what information each latent state should encode.*

To address this, we propose ICLR (**I**nformation **C**ontrol for **L**atent **R**easoning), a framework that enables multimodal models to effectively leverage latent states for reasoning through explicit information control. Our key insight is that reliable latent reasoning requires not only semantically meaningful latent states, but also actual reliance on these semantics during prediction. Building on this principle, we design a two-stage training strategy to ensure the functional necessity and semantic sufficiency of latent states: ***(i) Latent Dependency Pretraining.*** We first block direct access to the original visual input during answer generation, thereby explicitly routing all task-relevant information through the latent states. Unlike auxiliary objectives that only indirectly encourage latent utilization (Li et al., 2026a; Wang et al., 2026), this design makes latent states necessary intermediates in the reasoning pathway and lays the foundation for determining what information they should encode. **(ii) Dual-View Semantic Distillation.** To further strengthen the structured reasoning semantics encoded in the latent states, we restore access to visual information and perform step-level distillation from complementary visual and textual perspectives. For the k-th latent state h_k, the visual branch extracts a step-relevant evidence distribution over image patches conditioned on the input and CoT prefix, and uses it to supervise the visual attention of h_k. Meanwhile, the textual branch extracts the semantic importance distribution of the corresponding CoT step and aligns it with h_k. The visual signal determines where each latent state should attend, while the textual signal specifies what reasoning content it should encode, jointly grounding each latent state in visual evidence and step-specific semantics. In summary, our contributions are as follows:

- We systematically analyze existing latent reasoning methods and find that supervision-induced information in latent states does not ensure path necessity or step-level semantic effectiveness for prediction, providing insights for future research on latent reasoning.

- We propose **ICLR**, which explicitly controls information flow to establish both predictive necessity and semantic effectiveness of latent states. ICLR first employs Latent Dependency Pretraining to force semantic information through latent states, followed by Dual-View Semantic Distillation to learn input- and step-specific semantics from complementary visual and textual perspectives.
- ICLR achieves state-of-the-art performance across multimodal benchmarks, surpassing the current best latent reasoning baseline by **3.0** and achieving a **20.9**$\times$ generation speedup over textual CoT. Ablations further confirm its effectiveness in learning reasoning representations.

2 Related Works

Chain-of-Thought Reasoning in Multimodal Large Language Models. Chain-of-Thought (CoT) enhances large language model reasoning by explicitly generating intermediate steps and has been extended to multimodal tasks such as visual question answering (Qiu et al., 2024; Chen et al., 2024b), video understanding (Arnab et al., 2026), and embodied reasoning (Zhao et al., 2025; Bai et al., 2026). By integrating visual evidence with step-by-step linguistic reasoning, CoT substantially improves performance on challenging multimodal tasks. Recent work further advances this paradigm through structured reasoning, long-form CoT training, and reinforcement learning. For example, LLaVA-CoT (Xu et al., 2025) adopts structured multi-stage reasoning and test-time scaling, Multimodal-CoT (Zhang et al., 2023) separates rationale generation from answer inference, and GPT-5 (Singh et al., 2025a) and GLM-5 (Zeng et al., 2026) highlight the value of long-form CoT for capable reasoning agents. However, explicit CoT incurs substantial autoregressive decoding and computational overhead, while committing intermediate reasoning to discrete textual forms restricts exploration to a single trajectory. Therefore, reducing generation overhead while preserving reasoning quality has become an important research direction in multimodal reasoning.

Latent Visual Reasoning. Limitations of text-based reasoning have motivated growing interest in latent reasoning, which enables models to perform intermediate reasoning directly in latent space without explicitly generating intermediate discrete linguistic representations (Zhang et al., 2025; Yu et al., 2026; Hao et al., 2024). Recently, this paradigm has been extended to multimodal settings. LVR supervises latent states by reconstructing visual representations relevant to the given question (Li et al., 2026a); Monet combines teacher-state distillation with policy optimization for visual latents (Wang et al., 2026); and CoLT supervises latent representations by aligning them with textual CoT (Hu et al., 2026). Despite these differences, they generally assume that sufficiently supervised latent states will be effectively used for prediction. Recent studies have begun to challenge this assumption. CapImagine finds that input perturbations may have limited effects on latent states, while latent states may in turn have little impact on the final answer (Li et al., 2026b). Viveiros et al. (Viveiros et al., 2026) observe that latent states are prone to collapsing into narrow regions of the representation space and propose data reconstruction as a mitigation strategy. However, these studies primarily characterize failure modes and rely on auxiliary images for optimization. Our work differs in that we not only diagnose failures in latent visual reasoning but also explore a general solution applicable to existing latent reasoning paradigms and general-purpose training corpora.

3 Preliminaries

Latent Thought Representation. Let M_θ denote a multimodal large language model with parameters θ. Given a multimodal input consisting of an image $\mathbf{I}$ and a text query $\mathbf{q}$, the model first encodes the image to obtain visual tokens $\mathbf{V}_T = \mathrm{Encoder}(\mathbf{I})$, which are concatenated with the textual token embeddings $\mathbf{T}$ and jointly fed into the language model to form the initial multimodal prefix $\mathbf{U}^{(0)} = [\mathbf{V}_T; \mathbf{T}]$. In text-based Chain-of-Thought (CoT) reasoning, the model autoregressively generates an explicit textual reasoning sequence $\mathbf{r} = (r_1, r_2, \ldots, r_L)$ of length L before producing the final answer $\mathbf{y}$. For complex multimodal tasks, generating such long reasoning traces token by token introduces substantial decoding overhead and inference latency. Latent reasoning instead represents the intermediate reasoning process as a fixed-length trajectory $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_K) \in \mathbb{R}^{K\times d}$ in the continuous representation space, where K denotes the number of latent reasoning steps and d is the hidden dimension, typically with $K \ll L$. At each latent step k, the model directly uses the last-layer hidden state at the current reasoning position as the latent thought:

$$\mathbf{h}_k = M_\theta^{\text{last}}\left(\mathbf{U}^{(k-1)}\right) \in \mathbb{R}^d, \qquad \mathbf{U}^{(k)} = [\mathbf{U}^{(k-1)}; \mathbf{h}_k], \qquad k = 1, \ldots, K. \tag{1}$$

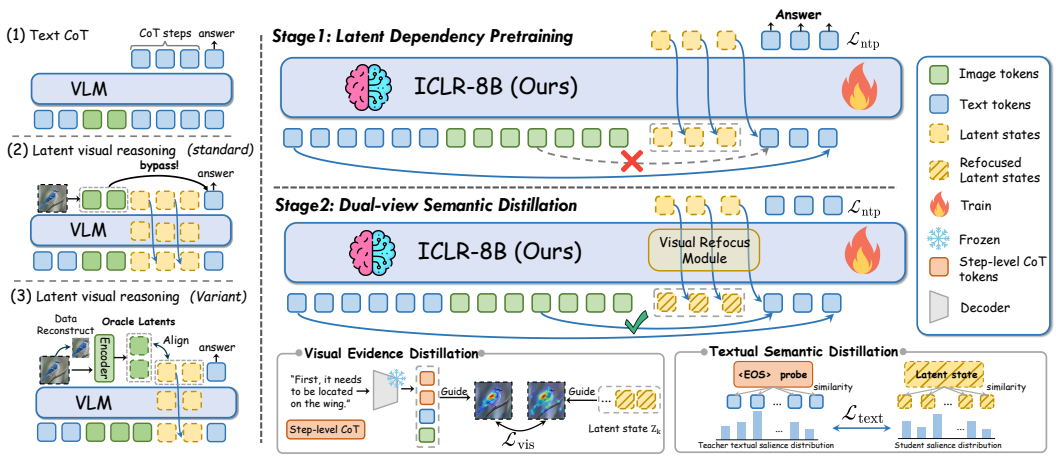


Figure 2: **Comparison of reasoning paradigms and overview of ICLR.** Left: text CoT incurs substantial reasoning overhead; standard latent visual reasoning may bypass latent states; and data-reconstruction variants suffer from limited applicability and generalization. Right: ICLR combines Latent Dependency Pretraining to enforce reliance on latent states with Dual-View Semantic Distillation to inject step-relevant visual evidence and CoT semantics, thereby establishing both the functional necessity and semantic sufficiency of latent states.

Each $\mathbf{h}_k$ is fed back as the continuous input for the next position, yielding the autoregressive latent trajectory $\mathbf{h}_1 \rightarrow \mathbf{h}_2 \rightarrow \cdots \rightarrow \mathbf{h}_K$. This allows intermediate reasoning to proceed entirely in the continuous hidden space, avoiding explicit decoding of lengthy textual CoT sequences and reducing intermediate generation overhead.

4 METHOD

In this section, we present ICLR, a latent reasoning framework that explicitly controls information flow to enable effective reasoning through latent states. We first provide an overview of the framework (Section 4.1), then describe our two-stage information control strategy (Section 4.2), which combines Latent Dependency Pretraining (Section 4.3) and Dual-View Semantic Distillation (Section 4.4). Finally, we present the complete training objective and inference procedure (Section 4.5).

4.1 OVERVIEW

As shown in Figure 2(left), we compare ICLR with text CoT, standard latent visual reasoning and its variants. Standard latent visual reasoning may bypass the latent states and directly generate the answer (Figure 1(a)). Variants using data reconstruction can strengthen the visual semantics of latent states, but lack general applicability and exhibit limited cross-benchmark generalization (Figure 1(c)). Thus, existing paradigms do not reliably ensure that latent states are effectively used for reasoning. To address this limitation, we propose ICLR, which improves latent reasoning through a two-stage information control strategy that ensures both the predictive necessity and semantic effectiveness of latent states, as shown in Figure 2(right). We describe the details below.

4.2 TWO-STAGE INFORMATION CONTROL STRATEGY

Effective latent reasoning requires latent states to be both functionally necessary for prediction and semantically informative at each reasoning step. Since semantic supervision alone does not ensure reliance on the latent pathway, while latent dependence alone does not guarantee meaningful step-specific semantics, we adopt a two-stage information-flow control strategy. Specifically, in Stage 1, we employ ***Latent Dependency Pretraining***, which restricts direct access to visual information during answer generation while preserving full visual access for latent construction, thereby forcing all information to flow through the latent pathway and increasing the variational lower bound on the conditional mutual information between the answer and the latent representation (Theory A.1). This makes the latent states necessary intermediates in the reasoning pathway. In Stage 2, we restore full

visual access and introduce ***Dual View Semantic Distillation***, which aligns latent states at each reasoning step with complementary visual-evidence and textual-semantic targets, thereby encouraging them to encode the task-relevant information required for prediction.

4.3 Latent Dependency Pretraining

In this stage, we explicitly restrict the direct visual pathway to answer generation, making the latent trajectory a necessary intermediary for transmitting task-relevant information. Specifically, the latent trajectory $\mathbf{h}_{1:K}$ is still autoregressively generated from the full multimodal prefix $\mathbf{U}^{(0)} = [\mathbf{V}_T; \mathbf{T}]$, allowing it to fully access the original visual input. In contrast, answer supervision is computed from a separately constructed vision-restricted input. We preserve the original text tokens and multimodal positional layout while zeroing the input embeddings at visual positions, thereby removing direct access to the original visual content:

$$\mathbf{U}^{(0)}_{\text{ans}} = [\mathbf{0}_{m\times d}; \mathbf{T}]. \tag{2}$$

Subsequently, the final latent state $\mathbf{h}_K$ is injected into the answer context, and the autoregressive loss is computed only over the answer tokens:

$$\mathcal{L}^{\text{dep}}_{\text{ans}} = -\sum_t \log p_\theta \left(y_t \mid \mathbf{U}^{(0)}_{\text{ans}}, \mathbf{h}_K, y_{<t} \right). \tag{3}$$

By controlling the information flow from the original visual input to answer generation, this design routes task-relevant information through the latent trajectory before prediction. From an information-theoretic perspective, it increases a variational lower bound on the conditional mutual information between the answer and the latent representation (Theory A.1), thereby making the latent states necessary for answer prediction. This stage establishes the functional necessity of the latent pathway and provides a foundation for learning step-specific semantics in Stage 2.

4.4 Dual-View Semantic Distillation

The preceding stage establishes the functional dependence of answer prediction on latent states. In this stage, we restore attention to visual tokens and further strengthen the latent states to encode information necessary for reasoning, enabling each state to capture step-specific reasoning semantics.

Latent Visual Refocusing. To enable each latent state to better focus on step-relevant visual information and support effective information encoding, we introduce a lightweight visual refocusing module R_ω after each latent step. Given the k-th latent state $\mathbf{h}_k$, the module maps it into a set of dynamic evidence slots, which aggregate step-relevant information through self-attention and cross-attention over the visual tokens $\mathbf{V}_T$. Let $\mathbf{c}_k$ and $\mathbf{A}_k$ denote the aggregated visual representation and the corresponding attention map, respectively. We then formulate the refocusing process as:

$$(\mathbf{c}_k, \mathbf{A}_k) = R_\omega(\mathbf{h}_k, \mathbf{V}_T), \qquad \mathbf{z}_k = \mathbf{h}_k + \sigma(g_\omega(\mathbf{h}_k))\, \mathbf{W}_\omega \mathbf{c}_k. \tag{4}$$

Here, $\mathbf{z}_k$ denotes the refocused latent state and is subsequently used for further latent reasoning and textual semantic distillation, while the visual attention $\mathbf{A}_k$ is used for visual evidence distillation.

Visual Evidence Distillation. From the visual perspective, we explicitly supervise *where each latent state should attend*. Specially, for the k-th reasoning step, we use the associated textual CoT prefix to guide the extraction of a patch-level teacher attention distribution $\mathbf{A}^{\text{tea}}_k$ from the original image and question, which specifies the visual evidence relevant to the current step. On the student side, the attention maps produced by the latent state are averaged across S evidence slots and normalized over valid visual tokens. We then align the resulting student attention with the teacher distribution using a confidence-weighted KL divergence:

$$\hat{\mathbf{A}}_k = \text{Normalize}\left(\frac{1}{S} \sum_{s=1}^{S} \mathbf{A}^{(s)}_k \right), \qquad \mathcal{L}_{\text{vis}} = \frac{\sum_{k=1}^{K} \gamma_k \, \text{KL}\left(\mathbf{A}^{\text{tea}}_k \| \hat{\mathbf{A}}_k \right)}{\sum_{k=1}^{K} \gamma_k}. \tag{5}$$

Here, γ_k is determined by the normalized entropy of $\mathbf{A}^{\text{tea}}_k$, such that steps with more concentrated visual evidence receive larger weights.

Textual Semantic Distillation. We further introduce textual semantic distillation as a complementary supervision signal to guide each latent state toward the salient reasoning content it should encode. Following CoLT (Hu et al., 2026), we reuse the backward auxiliary decoder D_ϕ together with a projection P_{back}. For the preceding reasoning span s_{k-1}, we append an <EOS> token and feed the resulting sequence into D_ϕ. Let $b_{k-1,j}$ denote the last-layer hidden representation of the j-th token produced by D_ϕ, and let b_{k-1}^{EOS} denote the hidden state at the <EOS> position, which serves as a span-level semantic summary. Based on the similarity between each token representation and the semantic summary, we derive a token-level salience distribution and aggregate the projected token representations into a semantic prototype:

$$\pi_{k-1,j} = \text{softmax}_j\left(\frac{\cos\big(\mathbf{b}_{k-1,j}, \mathbf{b}_{k-1}^{\text{EOS}}\big)}{\tau}\right), \qquad \boldsymbol{\mu}_{k-1} = \sum_j \pi_{k-1,j} P_{\text{back}}(\mathbf{b}_{k-1,j}), \tag{6}$$

where $\tau = 0.1$. Here, $\pi_{k-1,j}$ measures the relative salience of the j-th token to the overall span semantics, while $\boldsymbol{\mu}_{k-1}$ denotes the span-level semantic prototype obtained by salience-weighted aggregation. Subsequently, we use the refocused latent state $\mathbf{z}_k$ as a query over the same projected token representations to obtain the latent-induced salience distribution $\hat{\boldsymbol{\pi}}_{k-1}$. We then jointly align step-level semantics and token-level salience:

$$\mathcal{L}_{\text{text}} = \frac{1}{K-1}\sum_{k=2}^{K}\left[1 - \cos(\mathbf{z}_k, \text{sg}(\boldsymbol{\mu}_{k-1})) + \lambda\ \text{KL}(\boldsymbol{\pi}_{k-1} \| \hat{\boldsymbol{\pi}}_{k-1})\right], \tag{7}$$

where $\lambda = 0.25$, and $\text{sg}(\cdot)$ denotes the stop-gradient operator. Overall, Textual semantic distillation further aligns latent states with step-specific semantics, thereby encouraging them to encode semantic content essential for reasoning.

4.5 Training Objective

The training objective consists of two stages. In Stage 1, we optimize the latent dependency objective $\mathcal{L}_{\text{ans}}^{\text{dep}}$ in Eq. 3 to establish reliance on the latent reasoning pathway for answer prediction. In Stage 2, we restore full visual access and optimize the dual-view distillation objective:

$$\mathcal{L}_{\text{distill}} = \alpha\mathcal{L}_{\text{vis}} + \beta\mathcal{L}_{\text{text}}, \tag{8}$$

where α and β weight visual evidence and textual semantic distillation, respectively. We further retain the forward and backward decoding supervision from CoLT (Hu et al., 2026), $\mathcal{L}_{\text{fwd}}$ and $\mathcal{L}_{\text{bwd}}$, and optimize answer generation with the standard next-token prediction loss:

$$\mathcal{L}_{\text{ntp}} = -\sum_{t=1}^{|\mathbf{y}|}\log p_\theta\left(y_t \mid y_{<t}, \mathbf{h}_{1:K}, \mathbf{I}, \mathbf{q}\right). \tag{9}$$

Training procedure. We train the model via supervised fine-tuning on the image subset of One-Thinker (Feng et al., 2026), which provides high-quality CoT annotations for diverse visual reasoning tasks. For each sample, the <Think> and <Answer> markers delimit the reasoning trace and the final answer. We further partition each CoT sequence into K contiguous textual spans and use these spans as step-level semantic targets to supervise the corresponding latent states via an auxiliary decoder D_ϕ. During inference, we discard the external decoder and training losses, retaining only the latent reasoning and visual refocusing pathways. The model directly generates the final answer after K latent reasoning steps without decoding textual CoT, thereby reducing inference overhead.

5 Experiments

5.1 Experimental Setup

Benchmarks. We comprehensively evaluate our method on eight multimodal benchmarks, including SeedBench (Li et al., 2024), MMBench (Liu et al., 2024), ChartQA (Masry et al., 2022), TextVQA, ScienceQA (Saikh et al., 2022), MMStar (Chen et al., 2024a), AI2D (Kembhavi et al., 2016), and MMT-Bench (Liu et al., 2024), covering a broad range of visual understanding and

Table 1: Comprehensive comparison of closed- and open-source MLLMs, latent reasoning, and latent visual reasoning methods. 'SQA' is the abbreviation of the 'ScienceQA' benchmark. In this table, '†' denotes results reported in CoLT, while '‡' denotes our reproduced results.

Method	Size	SeedBench	MMBench	ChartQA	TextVQA	SQA	MMStar	AI2D	MMT	Avg.
Proprietary Models										
GPT-5-high†	-	-	83.8	86.2	-	97.4	76.4	89.7	77.2	-
Claude-Opus-4.1†	-	-	83.0	83.3	-	-	71.0	84.4	-	-
Gemini-2.5-Pro†	-	-	90.1	59.7	-	96.2	77.5	88.4	75.4	-
Open-Source MLLMs (7–8B scale)										
LLaVA-OneVision-1.5†	8B	77.3	84.1	86.5	-	95.0	67.7	84.2	-	-
InternVL3†	8B	-	83.4	86.6	80.2	-	68.2	85.2	65.0	-
Latent Reasoning										
CCoT†	8B	63.4	78.4	60.4	63.1	87.0	60.3	77.0	57.1	68.3
ICoT†	8B	64.7	79.1	62.1	65.6	87.6	61.2	77.8	57.9	69.5
CODI†	8B	68.6	80.2	61.6	67.2	88.4	62.5	79.0	59.1	70.8
SoftCoT†	8B	69.2	81.0	64.5	68.4	89.5	64.0	80.4	60.6	72.2
SIM-CoT†	8B	72.4	82.0	66.7	70.2	90.8	65.6	82.0	62.4	74.0
Latent Visual Reasoning										
Multimodal CoT†	8B	63.2	79.5	62.1	66.1	87.8	61.5	78.2	58.3	69.6
LaCoT†	7B	67.4	-	64.6	67.6	-	-	-	-	-
LVR†	7B	71.2	81.5	67.2	70.2	90.2	64.8	81.3	62.6	73.6
Qwen3-VL (Direct answer)†	8B	66.2	76.2	64.1	74.3	85.3	58.6	75.3	55.6	69.5
Qwen3-VL (Textual reasoning)†	8B	76.4	83.4	65.1	75.2	91.8	67.1	83.6	63.3	75.7
CoLT‡	8B	77.1	86.1	68.2	69.0	93.0	65.9	84.8	66.4	76.3
ICLR	8B	**78.4**	**87.5**	**78.8**	**74.6**	**94.9**	**66.8**	**85.6**	**67.7**	**79.3**
Δ vs. CoLT‡		+1.3	+1.4	+10.6	+5.6	+1.9	+0.9	+0.8	+1.3	+3.0

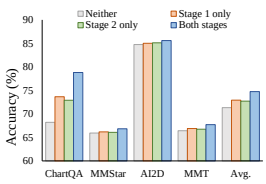
reasoning capabilities, such as general visual comprehension, chart and text reasoning, scientific question answering, and diagram understanding. Further details are provided in the Appendix A.3.

Baselines. We compare our method against state-of-the-art MLLM baselines, grouped into the following categories: (1) Proprietary MLLMs, including GPT-5-high (Singh et al., 2025b), Claude-Opus-4.1 (Anthropic, 2025), and Gemini-2.5-Pro (Comanici et al., 2025), serve as references for strong general-purpose multimodal models. (2) Open-source MLLMs, including LLaVA-OneVision-1.5 (An et al., 2025) and InternVL3 (Zhu et al., 2025a). (3) Latent reasoning methods, including CCoT (Cheng & Van Durme, 2024), ICoT (Deng et al., 2024), CODI (Shen et al., 2025), and so on, reduce reliance on explicit CoT by performing or supervising intermediate reasoning in continuous representation space. (4) Latent visual reasoning methods, including Multimodal CoT (Zhang et al., 2023), LaCoT (Sun et al., 2026), and LVR (Li et al., 2026a), further incorporate visual information into the latent reasoning process. (5) CoLT (Hu et al., 2026) provides fine-grained step-level supervision for multimodal latent states using textual CoT, without requiring auxiliary images or additional visual annotations. We also compare Qwen3-VL-8B under direct-answer and textual-reasoning settings to evaluate the gains of our method over standard reasoning paradigms.

Implementation details. We use Qwen3-VL-8B-Instruct as the backbone and Qwen3-0.6B as the auxiliary decoder D_ϕ. Both models belong to the Qwen3 family and share the same tokenizer and vocabulary (Bai et al., 2025). All models are trained for one epoch on the image subset of OneThinker and evaluated with identical prompts and configurations for fair comparison. Following CoLT (Hu et al., 2026), we set the number of latent reasoning steps to $K = 3$ and perform the two-stage supervised fine-tuning (SFT) procedure described in Section 4.5. During training, we use the AdamW optimizer with a cosine learning rate schedule. The learning rate and weight decay are set to 1×10^{-4} and 1×10^{-2}, with a batch size of 8. Additionally, we set $\tau = 0.1$ and $\lambda = 0.25$, with the distillation loss weights $\alpha = 0.2$ and $\beta = 0.2$. More details are provided in the Appendix A.2.

5.2 Main Results

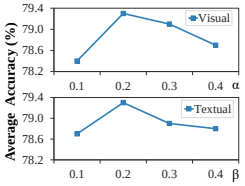
Comparison with State-of-the-Art Methods. Table 1 summarizes the results across eight multimodal benchmarks. ICLR achieves the highest average score of **79.3** and the best performance on six benchmarks, demonstrating strong and consistent performance across diverse visual understanding and reasoning tasks. Compared with direct answer and textual reasoning, ICLR improves the average score by **+9.8** (69.5→79.3) and **+3.6** (75.7→79.3), respectively, highlighting the benefits of reasoning and our latent reasoning design. Moreover, ICLR consistently outperforms CoLT, the strongest baseline in average performance, raising the average score from 76.3 to 79.3 (**+3.0**), with



(a)

Visual	Textual	Refocusing	ChartQA	MMStar	AI2D	MMT	Avg.
			71.9	66.0	84.4	66.4	72.2
		✓	73.8	66.7	85.1	65.9	72.9
✓		✓	**76.6**	66.3	84.5	66.6	73.5
	✓	✓	74.5	**66.5**	**85.2**	**67.2**	73.4
✓	✓	✓	**78.8**	**66.8**	**85.6**	**67.7**	**74.7**

(b)



(c)

Figure 3: **Ablation studies and hyperparameter analysis.** (a) Effects of Latent Dependency Pretraining (Stage 1) and Dual-View Semantic Distillation (Stage 2). (b) Effects of visual and textual distillation branches. (c) Sensitivity analysis of the visual and textual distillation weights α and β.

the largest gains on ChartQA (**+10.6**) and TextVQA (**+5.6**). Compared with existing latent reasoning methods, ICLR further surpasses the strongest baseline SIM-CoT (Wei et al., 2026) (79.3 vs. 74.0, **+5.3**), with particularly notable improvements on ChartQA (**+12.1**) and TextVQA (**+4.4**), suggesting that ICLR more effectively integrates visual and textual information for multimodal reasoning than methods that primarily perform implicit reasoning within the language model. Compared with latent visual reasoning methods specifically designed for visual information processing, ICLR also outperforms the strongest baseline, LVR (Li et al., 2026a) (79.3 vs. 73.6, **+5.7**), while avoiding the need for auxiliary image annotations. Overall, these results demonstrate that ICLR effectively handles diverse visual and textual information while maintaining general reasoning capability.

Inference Efficiency. A key advantage of latent reasoning over text-based CoT is that it avoids lengthy autoregressive reasoning generation. We compare inference efficiency in terms of both inference speed and token usage, with the results reported in Table 2. Since input encoding costs are relatively small compared with generation costs, the main efficiency gap arises during generation. On MMStar (Chen et al., 2024a) and MMT-Bench (Liu et al., 2024), text-based CoT requires 10.52 s and 12.25 s to generate approximately 177.5 and 187.1 reasoning tokens, whereas our method takes only 0.42 s and 0.67 s, yielding 24.9× and 18.3× generation speedups. Overall, ICLR achieves 11.1× and 10.2× end-to-end speedups while maintaining strong task performance, substantially reducing both inference latency and reasoning overhead.

Table 2: **Inference speed comparison on MMStar and MMT-Bench (per-sample average).** All methods use Qwen3-VL-8B on a single A100 GPU, with latency reported separately for input encoding and token generation (seconds).

Method	MMStar				MMT-Bench			
	CoT Tokens	Time Encode	Time Generate	Acc.	CoT Tokens	Time Encode	Time Generate	Acc.
Direct Answer	0	0.58	0.23	58.6	0	0.59	0.45	55.6
Text CoT	177.5	0.58	10.52	67.1	187.1	0.57	12.25	63.3
ICLR	**3** (↓59.2×)	**0.58**	**0.42** (↓24.9×)	**66.8**	**3** (↓62.4×)	**0.59**	**0.67** (↓18.3×)	**67.7**

5.3 Ablation Study

We conduct comprehensive ablation studies to assess the contribution of each component in ICLR. Component ablations are conducted on four representative benchmarks, while hyperparameter sensitivity is evaluated using the eight-benchmark average (Masry et al., 2022; Chen et al., 2024a; Kembhavi et al., 2016; Ying et al., 2024).

Effect of Two-Stage Training. Figure 3 evaluates the contributions of Latent Dependency Pretraining (Stage 1) and Dual-View Semantic Distillation (Stage 2). Stage 2 alone achieves an average score of 72.71, indicating that semantic refinement alone is insufficient to realize the full benefit of the latent reasoning pathway. Stage 1 alone reaches 72.93, supporting the effectiveness of restricting direct visual bypasses to encourage latent dependency. Combining both stages yields the best average score of 74.74, improving over Stage 2 alone by **+2.0** and Stage 1 alone by **+1.8**. These results highlight the complementarity of latent dependency and semantic refinement, and the importance of establishing dependency before refinement for effective latent visual reasoning.

Effect of Dual-View Semantic Distillation. Figure 3(b) validates the complementary roles of the two distillation signals. Removing Visual Evidence Distillation reduces performance by 1.3 points (74.7→73.4), showing that visual refocusing improves fine-grained visual information utilization by guiding latent states toward step-relevant evidence. Removing Textual Semantic Distillation causes a similar 1.2 point drop, demonstrating that CoT-based semantic supervision enhances stepwise reasoning semantics and semantic consistency. The full dual-view supervision achieves the

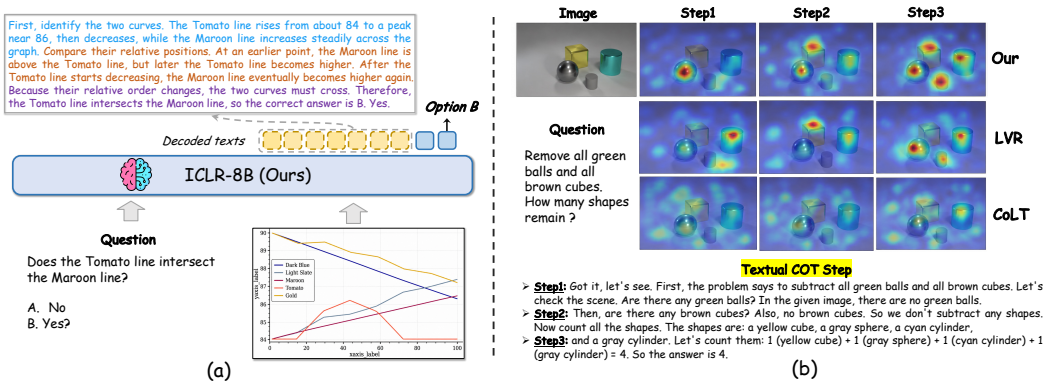


Figure 5: **Qualitative visualization of latent reasoning and step-level attention.** (a) Latent states are projected back into text using a forward decoder, with colors indicating reasoning content from different latent steps. (b) Compared with LVR and CoLT, ICLR progressively attends to step-relevant visual evidence and better aligns with the semantic content of the corresponding textual CoT.

best performance, confirming their complementary and non-redundant effects. Figure 3(c) further investigates the branch weights, where setting both coefficients to 0.2 achieves optimal performance.

Functional Necessity of Latent States. We test dependence on the latent states by zeroing, randomly replacing, or skipping latent states at inference, following the protocol in Section 1. These interventions induce only modest performance changes for the baselines (+1.7, -0.5, and -1.4), suggesting that they can largely bypass the latent states. In contrast, ICLR exhibits substantially larger performance drops (-4.4, -3.4, and -4.2), demonstrating markedly stronger dependence on latent reasoning. Figure 4 further visualizes ICLR's reasoning process under the Original Latent and Zeroed Latent conditions, showing that zeroing the latent states misdirects the model's reasoning focus. Together, these results demonstrate that ICLR routes task-relevant information through the latent trajectory, rather than directly exploiting the original multimodal input or bypassing the latent states for prediction.

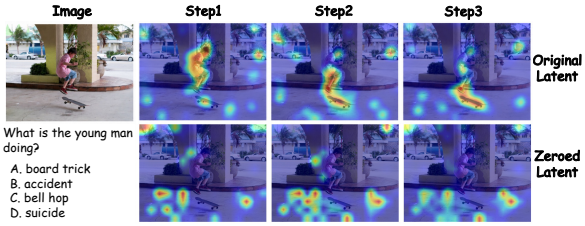


Figure 4: **Necessity of the latent reasoning pathway.**

5.4 Qualitative Analysis

Figure 5 qualitatively analyzes ICLR's reasoning process from the perspectives of latent reasoning and step-level attention. In Figure 5(a), we use Qwen3-0.6B as the forward decoder to project latent states back into text for interpretation. The decoded steps progressively identify the relevant curves, compare their trajectories, and derive the correct answer. In Figure 5(b), we further visualize step-level attention. Existing methods often exhibit diffuse attention or focus on regions irrelevant to the current reasoning step. In contrast, ICLR progressively attends to step-relevant visual evidence and remains better aligned with the semantic content of the corresponding textual CoT. Overall, these results demonstrate that ICLR effectively encodes and aligns step-level visual and textual semantics, forming a structured and interpretable multi-step reasoning trajectory.

6 Conclusion

In this work, we propose ICLR, a framework for learning effective latent reasoning representations through explicit information-flow control. We introduce a two-stage training pipeline—combining Latent Dependency Pretraining and Dual-View Semantic Distillation—that jointly establishes the functional necessity and semantic sufficiency of latent states. ICLR first routes all information through the latent pathway, making latent states indispensable intermediate representations, and then enriches latent reasoning with complementary visual evidence and textual semantics for step-specific reasoning. Extensive experiments show that ICLR outperforms existing latent reasoning methods and textual CoT while substantially reducing autoregressive decoding cost.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Anthropic. Claude opus 4.1, August 2025. URL https://www.anthropic.com/news/claude-opus-4-1.

Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Temporal chain of thought: Long-video understanding by thinking in frames. *Advances in Neural Information Processing Systems*, 38:143018–143046, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Shuanghao Bai, Jing Lyu, Wanqi Zhou, Zhe Li, Dakai Wang, Lei Xing, Xiaoguang Zhao, Pengwei Wang, Zhongyuan Wang, Cheng Chi, et al. Latent reasoning vla: Latent thinking and prediction for vision-language-action models. *arXiv preprint arXiv:2602.01166*, 2026.

David Barber and Felix Agakov. Information maximization in noisy channels: A variational approach. *Advances in Neural Information Processing Systems*, 16, 2003.

Hanzhu Chen, Lin Yang, Jie Wang, Junhao Yan, Zhe Wang, Xize Liang, and Jianye Hao. Latent-guided reasoning: Empowering small llms with large-model thinking. In *The Fourteenth International Conference on Learning Representations*, 2026.

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

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1254–1262, 2024b.

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

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

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

Shuai Dong, Siyuan Wang, Xingyu Liu, Chenglin Li, Haowen Hou, and Zhongyu Wei. Interleaved latent visual reasoning with selective perceptual modeling. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29316–29335, 2026.

Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, et al. Onethinker: All-in-one reasoning model for image and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5432–5443, 2026.

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

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

Lianyu Hu, Shengqian Qin, Zeqin Liao, Qing Guo, Liang Wan, Wei Feng, and Yang Liu. Colt: Teaching multi-modal models to think with chain of latent thoughts. In *European Conference on Computer Vision*, pp. 526–544. Springer, 2026.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 148076–148090, 2026a.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308. IEEE, 2024.

You Li, Chi Chen, Yanghao Li, Fanhu Zeng, Kaiyu Huang, Jinan Xu, and Maosong Sun. Imagination helps visual reasoning, but not yet in latent space. *arXiv preprint arXiv:2602.22766*, 2026b.

Jia Liang and Liangming Pan. Do latent-cot models think step-by-step? a mechanistic study on sequential reasoning tasks. *arXiv preprint arXiv:2602.00449*, 2026.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Chen Qiu, Zhiqiang Xie, Maofu Liu, and Huijun Hu. Explainable knowledge reasoning via thought chains for knowledge-based visual question answering. *Information Processing & Management*, 61(4):103726, 2024.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: a novel resource for question answering on scholarly articles: T. saikh et al. *International Journal on Digital Libraries*, 23(3):289–301, 2022.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025a.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025b.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

Guohao Sun, Hang Hua, Jian Wang, Jiebo Luo, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Latent chain-of-thought for visual reasoning. *Advances in neural information processing systems*, 38:103739–103762, 2026.

Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Ruihua Song, and Jian Luan. Think silently, think fast: Dynamic latent compression of llm reasoning chains. *Advances in Neural Information Processing Systems*, 38:4646–4668, 2026.

André G Viveiros, Nuno Gonçalves, André FT Martins, and Matthias Lindemann. What's holding back latent visual reasoning? *arXiv preprint arXiv:2605.18445*, 2026.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond image and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12030–12040, 2026.

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

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. Sim-cot: Supervised implicit chain-of-thought. In *International Conference on Learning Representations*, volume 2026, pp. 56721–56742, 2026.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2087–2098. IEEE, 2025.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

Xinlei Yu, Zhangquan Chen, Yongbo He, Tianyu Fu, Guanting Dong, Cheng Yang, Chengming Xu, Yue Ma, Xiaobin Hu, Zhe Cao, et al. The latent space: Foundation, evolution, mechanism, ability, and outlook. *arXiv preprint arXiv:2604.02029*, 2026.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Yuyi Zhang, Boyu Tang, Tianjie Ju, Sufeng Duan, and Gongshen Liu. Do latent tokens think? a causal and adversarial analysis of chain-of-continuous-thought. *arXiv preprint arXiv:2512.21711*, 2025.

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

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1702–1713. IEEE, 2025.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pp. 400–410, 2024.

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

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

A Appendix

A.1 Theoretical Analysis

A.1.1 Information-Theoretic Guarantee for Latent Dependency

Setup. Fix the model parameters and let (I, Q, Y) be an independent sample from the data distribution, where Y is a discrete answer sequence. Let $Z = \mathbf{h}_K$ denote the final Stage-1 latent state, and let $C = (Q, B)$ denote the non-latent context available to the restricted answer decoder. Here, B contains the retained visual-token count, positional layout, and other visible metadata induced by Eq. 2. If B is fixed or determined by Q, conditioning on C reduces to conditioning on Q. Here (I, Q, Y) are the random-variable counterparts of $(\mathbf{I}, \mathbf{q}, \mathbf{y})$ in the main text, and C encodes the information available in $\mathbf{U}^{(0)}_{\text{ans}}$. We retain the main-text decoder notation p_θ, with $p_\theta(Y_t \mid C, Z, Y_{<t})$ denoting the same restricted prediction as $p_\theta(y_t \mid \mathbf{U}^{(0)}_{\text{ans}}, \mathbf{h}_K, y_{<t})$ in Eq. 3.

Assume that $H(Y \mid C) < \infty$ and that the decoder defines a normalized distribution $p_\theta(Y \mid C, Z)$, positive on the data support, with finite expected negative log-likelihood. For variable-length answers, the sequence includes a termination token; equivalently, a fixed-length padded representation can be used. We use natural logarithms throughout and define the population expectation of the same loss in Eq. 3 as

$$\begin{aligned} \mathbb{E}[\mathcal{L}^{\text{dep}}_{\text{ans}}] &= \mathbb{E}[-\log p_\theta(Y \mid C, Z)] \\ &= \mathbb{E}\left[-\sum_t \log p_\theta(Y_t \mid C, Z, Y_{<t})\right]. \end{aligned} \tag{10}$$

The expectation is over fresh examples at fixed model parameters; it introduces no additional training objective. As in Eq. 3, the loss is summed over answer tokens, rather than divided by an example-dependent sequence length.

Proposition 1 (information certificate from dependency loss). *Under the setup above,*

$$\begin{aligned} I(Y; Z \mid C) &= H(Y \mid C) - \mathbb{E}[\mathcal{L}^{\text{dep}}_{\text{ans}}] + \mathcal{E}_{\text{dec}}, \\ \mathcal{E}_{\text{dec}} &= \mathbb{E}_{C,Z}\left[\mathrm{KL}\big(P_{Y|C,Z} \| p_\theta(\cdot \mid C, Z)\big)\right] \geq 0. \end{aligned} \tag{11}$$

Consequently,

$$I(Y; Z \mid C) \geq \max\{0, H(Y \mid C) - \mathbb{E}[\mathcal{L}^{\text{dep}}_{\text{ans}}]\}. \tag{12}$$

In particular, $\mathbb{E}[\mathcal{L}^{\text{dep}}_{\text{ans}}] \leq H(Y \mid C) - \delta$ for $\delta > 0$ certifies at least δ nats of answer-relevant information beyond the visible context.

Proof. For each (c, z), adding and subtracting the log of the true conditional probability gives the conditional cross-entropy decomposition

$$\begin{aligned} \mathbb{E}[-\log p_\theta(Y \mid c, z) \mid c, z] = H(P_{Y|c,z}) \\ + \mathrm{KL}\big(P_{Y|c,z} \| p_\theta(\cdot \mid c, z)\big). \end{aligned} \tag{13}$$

Averaging over (C, Z) yields $\mathbb{E}[\mathcal{L}^{\text{dep}}_{\text{ans}}] = H(Y \mid C, Z) + \mathcal{E}_{\text{dec}}$. Substitution into $I(Y; Z \mid C) = H(Y \mid C) - H(Y \mid C, Z)$ proves Eq. 11. Non-negativity of KL and conditional mutual information proves Eq. 12. The finiteness assumptions ensure these differences are well defined. □

This is an application of the variational information bound (Barber & Agakov, 2003); a closely related reconstruction-based formulation appears in Chen et al. (2026). Here the target is the answer, and the conditioning context explicitly accounts for the information retained by the restricted multimodal input.

Corollary 1 (cost of removing the informative pathway). *Suppose the restricted decoder has no input-dependent channel outside (C, Z). Replace Z by $\widetilde{Z}$ satisfying $\widetilde{Z} \perp Y \mid C$, without modifying C. For any normalized decoder $\widetilde{p}(Y \mid C, \widetilde{Z})$, its expected sequence log-loss ℓ_{rep} satisfies*

$$\ell_{\text{rep}} \geq H(Y \mid C), \qquad \ell_{\text{rep}} - \mathbb{E}[\mathcal{L}^{\text{dep}}_{\text{ans}}] \geq H(Y \mid C) - \mathbb{E}[\mathcal{L}^{\text{dep}}_{\text{ans}}]. \tag{14}$$

Proof. The conditional independence gives $P_{Y|C,\widetilde{Z}} = P_{Y|C}$. Applying the same cross-entropy decomposition to $\widetilde{p}$ gives $\ell_{\text{rep}} \geq H(Y \mid C, \widetilde{Z}) = H(Y \mid C)$. Subtracting $\mathbb{E}[\mathcal{L}_{\text{ans}}^{\text{dep}}]$ proves the second inequality. An infinite replacement loss also satisfies the inequalities. □

Implication. If

$$\mathbb{E}\big[\mathcal{L}_{\text{ans}}^{\text{dep}}\big] \leq H(Y \mid C) - \delta, \tag{15}$$

then

$$I(Y; Z \mid C) \geq \delta$$

and any replacement satisfying $\widetilde{Z} \perp Y \mid C$ incurs

$$\ell_{\text{rep}} - \mathbb{E}\big[\mathcal{L}_{\text{ans}}^{\text{dep}}\big] \geq \delta.$$

Thus, the restricted prediction objective provides a quantitative certificate of answer-relevant information transmitted through the latent pathway.

A.2 Implementation Details

We provide a detailed description of the implementation details in this section, with the key configurations summarized in Table 3.

Training data and latent-step construction. We follow the training data and preprocessing protocol of CoLT (Hu et al., 2026) and use the image reasoning subset curated from OneThinker. The complete training set contains approximately 122K samples, each consisting of an image, a question, an explicit chain of thought, and a final answer. The reasoning trace and final answer are delimited by the `<Think>` and `<Answer>` markers, respectively. We divide the training data into two disjoint subsets: approximately 33K samples are randomly drawn from the full training set for Latent Dependency Pretraining, and the remaining approximately 89K samples are used for Dual-View Semantic Distillation. This setup introduces no additional images, region annotations, or manually generated visual intermediates. Following CoLT, we preserve the original reasoning order and divide each chain of thought into $K = 3$ contiguous textual spans. Split points preferentially use punctuation or newline boundaries near equal-length positions; when no suitable natural boundary exists, we use the nearest equal-length position. An `<EOS>` token is appended to each span, which serves as the semantic supervision signal for the corresponding latent step.

Model and training configuration. We use Qwen3-VL-8B-Instruct as the multimodal backbone and Qwen3-0.6B as the auxiliary text decoder during training (Bai et al., 2025). The two models share the same tokenizer and vocabulary. We freeze the vision encoder and multimodal projector, while optimizing the language backbone, latent-state projections, and visual refocusing module. The auxiliary decoder is used only to produce supervision signals and is not retained at inference time. Training is implemented with PyTorch, Hugging Face Transformers, and LLaMA-Factory (Zheng et al., 2024), using bfloat16 and DeepSpeed ZeRO-3 on eight NVIDIA A100 GPUs with 80GB of memory each. Both stages use a fixed random seed of 42 and are trained for one epoch. Stage 2 is initialized from the model parameters obtained in Stage 1, while the optimizer and learning-rate scheduler are reinitialized.

Optimization and hyperparameters. Both stages use AdamW with a cosine learning-rate schedule. The learning rate is 1×10^{-4}, the weight decay is 1×10^{-2}, and the batch size is 8. We use $K = 3$ latent steps, set the textual salience temperature to $\tau = 0.1$, and set the token-salience KL coefficient within the textual loss to $\lambda = 0.25$. The visual and textual distillation weights are both set to 0.2, i.e., $\alpha = \beta = 0.2$.

Inference and evaluation. At inference time, we remove the visual teacher and the Qwen3-0.6B auxiliary decoder used for supervision, retaining only the Qwen3-VL backbone, the three latent reasoning steps, and the visual refocusing module. After the third latent step, the model directly generates the final answer without explicitly decoding the intermediate chain of thought. All models use the same input template, image preprocessing, and decoding configuration, and are scored with VLMEvalKit and the official evaluation protocol of each benchmark. We use deterministic greedy decoding by default to avoid sampling-induced differences in model comparisons.

Details of the Visual Refocusing Module. The visual refocusing module R_ω expands the current latent state into $S = 4$ dynamic evidence slots and applies two refocusing blocks, each performing

inter-slot self-attention followed by cross-attention over the original visual tokens. The module has a hidden dimension of 1536 and uses 12 attention heads. The slot representations are pooled and written back to the latent state through a gated residual connection, yielding $\mathbf{z}_k$.

Details of the Visual Evidence Distillation Branch. The visual branch uses a frozen Qwen3-VL-8B-Instruct teacher to extract a patch-level evidence distribution conditioned on the original image, the question, and the ground-truth CoT prefix up to the current step. We use eight fixed cross-layer attention heads: L23H4, L21H10, L26H20, L19H27, L23H5, L29H11, L25H10, and L24H13. For each attention head, we first aggregate the valid query positions for the current step using visual-mass query pooling, and independently compute and normalize its image-conditional patch attention map. We then average the eight distributions at the map level to form the teacher evidence distribution for that step. On the student side, the cross-attention maps from the four evidence slots are averaged and normalized over valid visual tokens. We then align the teacher and student evidence distributions using the entropy-based confidence γ_k and a confidence-weighted KL divergence, as defined in the main text.

Table 3: Implementation settings for the two training stages.

Configuration	**Stage 1**	**Stage 2**
Training data	Approx. 33K	Approx. 89K
Training epochs	1	1
Latent steps K	3	3
Optimizer	AdamW	AdamW
Learning-rate schedule	Cosine	Cosine
Learning rate	1×10^{-4}	1×10^{-4}
Weight decay	1×10^{-2}	1×10^{-2}
Batch size	8	8
Visual refocusing	Disabled	4 slots, 2 layers, 12 heads, hidden dim. 1536
Visual teacher aggregation	–	8 fixed cross-layer heads
Distillation parameters	–	$\tau = 0.1$, $\lambda = 0.25$, $\alpha = 0.2$, $\beta = 0.2$
Precision / parallelism	bfloat16 / DeepSpeed ZeRO-3	bfloat16 / DeepSpeed ZeRO-3

A.3 Benchmark Details

SEED-Bench. SEED-Bench (Li et al., 2024) is a large-scale benchmark designed to comprehensively evaluate the visual understanding and reasoning capabilities of multimodal large language models. The original benchmark contains approximately 19K human-annotated multiple-choice questions spanning 12 evaluation dimensions and covering diverse visual perception and reasoning skills. In our experiments, we focus on its image-based evaluation subset, which provides a broad assessment of general visual comprehension, spatial understanding, and multimodal reasoning.

MMBench. MMBench (Liu et al., 2024) is a comprehensive benchmark for evaluating fine-grained multimodal perception and reasoning capabilities. It contains approximately 3K multiple-choice questions covering 20 ability dimensions, hierarchically organized around perception and reasoning. In addition to diverse visual understanding tasks, MMBench adopts a carefully designed evaluation protocol to reduce answer-position bias and improve the robustness of model comparison. We use MMBench to assess the general multimodal understanding and reasoning ability of our model.

ChartQA. ChartQA (Masry et al., 2022) evaluates visual and logical reasoning over real-world charts. It contains 9.6K human-written questions and 23.1K questions generated from human-authored chart summaries. The questions require models not only to recognize visual elements and retrieve numerical values, but also to perform arithmetic operations, comparisons, and multi-step reasoning over structured chart information. Therefore, ChartQA provides a particularly suitable

testbed for evaluating whether latent reasoning can preserve and manipulate fine-grained visual evidence.

TextVQA. TextVQA (Singh et al., 2019) is designed to evaluate visual question answering that requires reading and reasoning over textual information embedded in natural images. The dataset contains 45,336 questions over 28,408 images. Solving these questions typically requires a model to jointly localize relevant textual regions, recognize their contents, and integrate them with other visual and linguistic cues. We use TextVQA to evaluate the ability of our latent reasoning framework to preserve fine-grained text-rich visual information.

ScienceQA. ScienceQA (Saikh et al., 2022) is a multimodal science question-answering benchmark containing 21,208 multiple-choice questions collected from elementary- and high-school science curricula. The questions cover diverse scientific domains and may involve image contexts, textual contexts, or both, with most examples additionally accompanied by lectures and explanations. ScienceQA requires the joint use of visual evidence, scientific knowledge, and multi-step reasoning, making it suitable for evaluating multimodal reasoning beyond direct visual recognition.

MMStar. MMStar (Chen et al., 2024a) is a vision-indispensable multimodal benchmark specifically designed to reduce the effects of text-only solvability and potential data leakage in existing evaluations. It contains 1,500 manually curated challenge samples covering six core capabilities and 18 fine-grained evaluation axes, with each question selected to require genuine visual information for successful prediction. This property makes MMStar particularly relevant to our study, as it provides a stringent evaluation of whether latent states effectively encode and utilize task-relevant visual evidence.

AI2D. AI2D (Kembhavi et al., 2016) focuses on understanding and reasoning over grade-school scientific diagrams. The dataset contains more than 5K diagrams together with approximately 15K question–answer pairs, covering scientific processes, spatial relations, object interactions, and diagrammatic structures. Unlike natural-image benchmarks, AI2D requires models to jointly interpret graphical elements, text labels, arrows, and their structural relationships, providing a complementary evaluation of structured visual reasoning.

MMT-Bench. MMT-Bench (Ying et al., 2024) is a large-scale multimodal multitask benchmark designed to comprehensively evaluate large vision-language models. It comprises 31,325 multiple-choice visual questions covering 32 meta-tasks and 162 subtasks, spanning capabilities such as visual recognition, localization, reasoning, planning, and expert-domain understanding. Its broad task diversity makes MMT-Bench particularly useful for assessing whether improvements in latent reasoning generalize across heterogeneous multimodal scenarios rather than being confined to a small set of task types.

A.4 TRAINING PROCEDURE

We provide pseudocode for the two-stage optimization procedure used to train ICLR in Algorithm 1.

Algorithm 1 ICLR Training Procedure

Require: Batch $\mathcal{B}$, MLLM M_θ, and auxiliary decoder D_ϕ
Require: Refocusing module R_ω, latent steps K, and evidence slots S
Require: Loss weights α, β, and λ, and temperature τ

Stage 1: Latent Dependency Pretraining
for each sample $(\mathbf{I}, \mathbf{q}, \mathbf{r}, \mathbf{y}) \in \mathcal{B}$ **do**
  Encode image: $\mathbf{V}_T \leftarrow \mathrm{Encoder}(\mathbf{I})$
  Initialize: $\mathbf{U}^{(0)} \leftarrow [\mathbf{V}_T; \mathbf{T}]$
  for $k = 1, \ldots, K$ **do**
    $\mathbf{h}_k \leftarrow M_\theta^{\text{last}}(\mathbf{U}^{(k-1)})$
    $\mathbf{U}^{(k)} \leftarrow [\mathbf{U}^{(k-1)}; \mathbf{h}_k]$
  end for
  Mask visual inputs for answer prediction: $\mathbf{U}_{\text{ans}}^{(0)} \leftarrow [\mathbf{0}_{m \times d}; \mathbf{T}]$
  Compute $\mathcal{L}_{\text{ans}}^{\text{dep}}$ using $\mathbf{h}_K$ and the restricted answer context
end for
Update trainable parameters with $\mathcal{L}_{\text{ans}}^{\text{dep}}$

Stage 2: Dual-View Semantic Distillation
for each sample $(\mathbf{I}, \mathbf{q}, \mathbf{r}, \mathbf{y}) \in \mathcal{B}$ **do**
  Restore standard visual visibility and split $\mathbf{r} \rightarrow \{\mathbf{s}_1, \ldots, \mathbf{s}_K\}$
  for $k = 1, \ldots, K$ **do**
    Generate latent state $\mathbf{h}_k$
    Refocus visual evidence: $(\mathbf{c}_k, \mathbf{A}_k) \leftarrow R_\omega(\mathbf{h}_k, \mathbf{V}_T)$
    Obtain refocused latent: $\mathbf{z}_k \leftarrow \mathbf{h}_k + \sigma(g_\omega(\mathbf{h}_k))\mathbf{W}_\omega \mathbf{c}_k$
    Extract step-conditioned teacher attention $\mathbf{A}_k^{\text{tea}}$
    Aggregate student evidence attention $\hat{\mathbf{A}}_k$ and confidence weight γ_k
  end for
  Compute visual distillation loss $\mathcal{L}_{\text{vis}}$
  for $k = 2, \ldots, K$ **do**
    Decode $[\mathbf{s}_{k-1}; \texttt{<EOS>}]$ with D_ϕ
    Compute token salience $\boldsymbol{\pi}_{k-1}$ and semantic prototype $\boldsymbol{\mu}_{k-1}$
    Obtain latent-induced salience $\hat{\boldsymbol{\pi}}_{k-1}$ from $\mathbf{z}_k$
  end for
  Compute textual semantic loss $\mathcal{L}_{\text{text}}$
  Compute $\mathcal{L}_{\text{fwd}}$, $\mathcal{L}_{\text{bwd}}$, and $\mathcal{L}_{\text{ntp}}$
end for
$\mathcal{L}_{\text{stage2}} \leftarrow \mathcal{L}_{\text{ntp}} + \mathcal{L}_{\text{fwd}} + \mathcal{L}_{\text{bwd}} + \alpha\mathcal{L}_{\text{vis}} + \beta\mathcal{L}_{\text{text}}$
Update trainable parameters using $\mathcal{L}_{\text{stage2}}$
Ensure: Trained M_θ and R_ω; remove D_ϕ during inference

A.5 Additional Qualitative Visualizations

Figure 6 complements Section 5.4 with six examples of step-specific latent semantics. The displayed trajectories progress from identifying relevant quantities or entities to expressing relations and forming an answer.

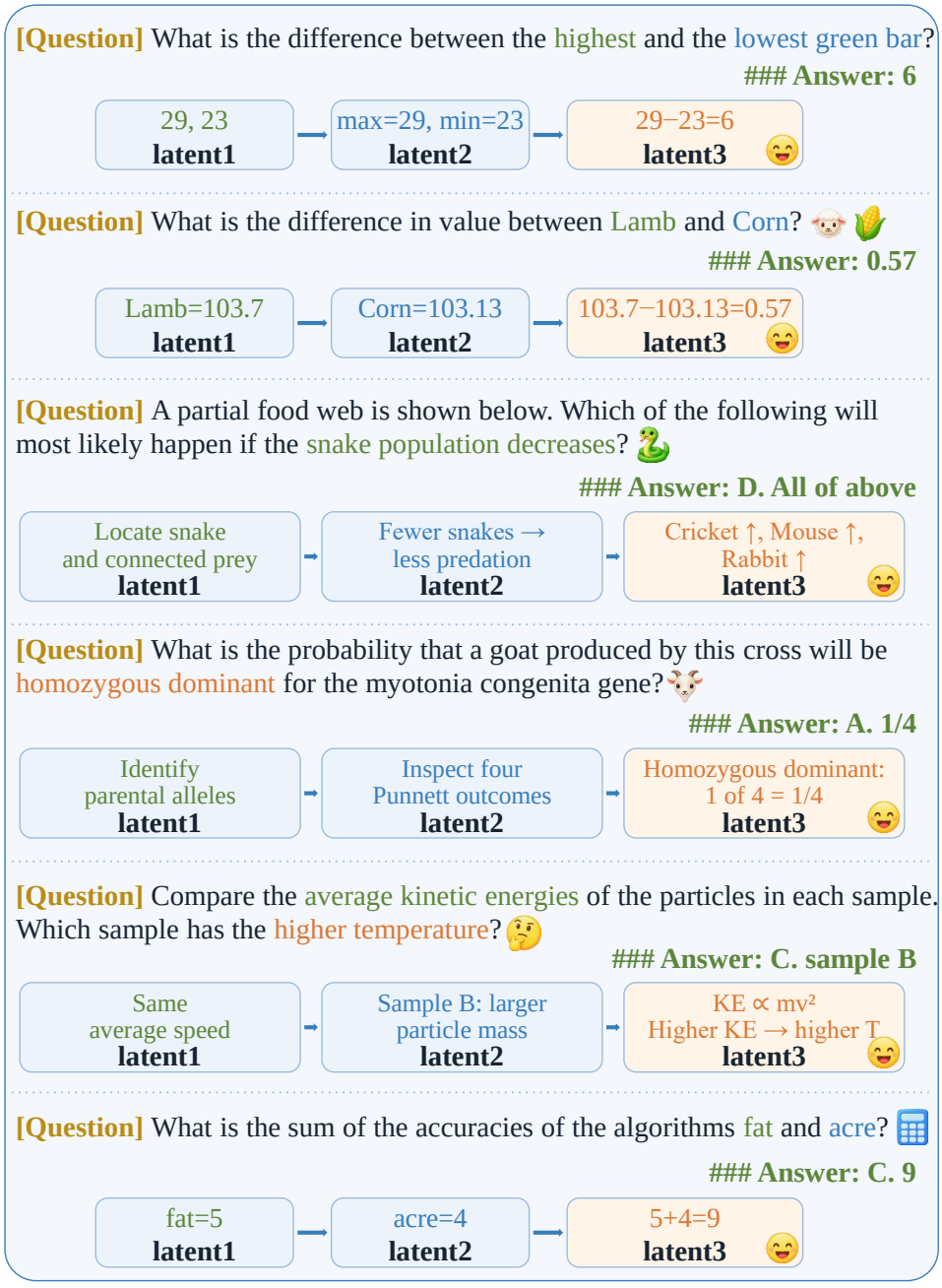
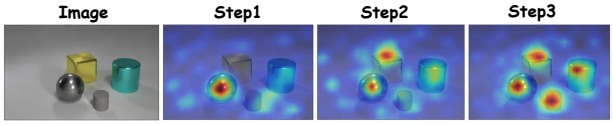


Figure 6: **Qualitative examples of step-wise latent reasoning.** Examples from ChartQA, AI2D, ScienceQA, and MMStar pair each question with textual interpretations of three successive latent states and the displayed answer. The trajectories illustrate numerical comparison, food-web reasoning, inheritance, kinetic-energy comparison, and addition. Colored text distinguishes step-associated content; arrows indicate the step order. These cases illustrate distinct semantic roles across steps, from identifying evidence to combining it into an answer.

A.5.1 Additional Step-Wise Visual Attention

Figure 7 extends the step-level attention analysis in Figure 5(b) with three additional examples from our model. For each example, the attention map at every latent step is paired with the corresponding textual CoT segment. Across anatomical diagrams, geographic maps, and relational graphs, the high-response regions evolve with the current reasoning target, progressing from evidence localization to relation assessment and answer formation.



Question: Remove all green balls and all brown cubes. How many shapes remain ?

Textual COT Step

- **Step1**: Got it, let's see. First, the problem says to subtract all green balls and all brown cubes. Let's check the scene. Are there any green balls? In the given image, there are no green balls.
- **Step2**: Then, are there any brown cubes? Also, no brown cubes. So we don't subtract any shapes. Now count all the shapes. The shapes are: a yellow cube, a gray sphere, a cyan cylinder,
- **Step3**: and a gray cylinder. Let's count them: 1 (yellow cube) + 1 (gray sphere) + 1 (cyan cylinder) + 1 (gray cylinder) = 4. So the answer is 4.

(a)

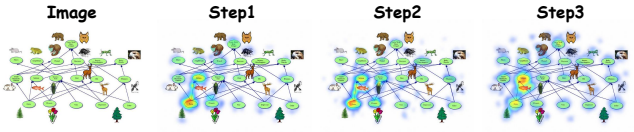


Question: Which of these states is farthest north?
A. West Virginia
B. Louisiana
C. Arizona
D. Oklahoma

Textual COT Step

- **Step1**: First, I need to determine the north direction from the map. The compass rose indicates that north is toward the top of the map.
- **Step2**: Now locate West Virginia, Louisiana, Arizona, and Oklahoma, and compare their positions along the north-south direction. West Virginia is positioned farther north than the other three states.
- **Step3**: Therefore, the state farthest north is West Virginia. The answer is A.

(b)



Question: In the diagram of the food web shown, if the number of ferns decrease, the supply of salmon will most likely?
A. decrease
B. can't tell
C. stay same
D. increase

Textual COT Step

- **Step1**: First, inspect the food-web connections involving ferns and salmon. The diagram shows that ferns serve as a food source connected to salmon.
- **Step2**: If the number of ferns decreases, this food source becomes less available. Although salmon may have other food sources, the reduction in ferns still reduces the food available to them.
- **Step3**: With less food available, the supply of salmon is most likely to decrease. Therefore, the answer is A.

(C)

Figure 7: **Additional examples of step-wise visual attention.** Each row shows the input image, attention maps for three successive latent steps, and the corresponding textual CoT. (a) Attention follows the airway and lung regions as the reasoning identifies the trachea as the connecting structure. (b) Attention moves from the map orientation and candidate regions to West Virginia as the northernmost option. (c) Attention traces relevant nodes and connections in the food web as the reasoning links fewer ferns to a decrease in the salmon supply. The examples exhibit step-conditioned alignment between visual evidence and textual reasoning across distinct visual formats.

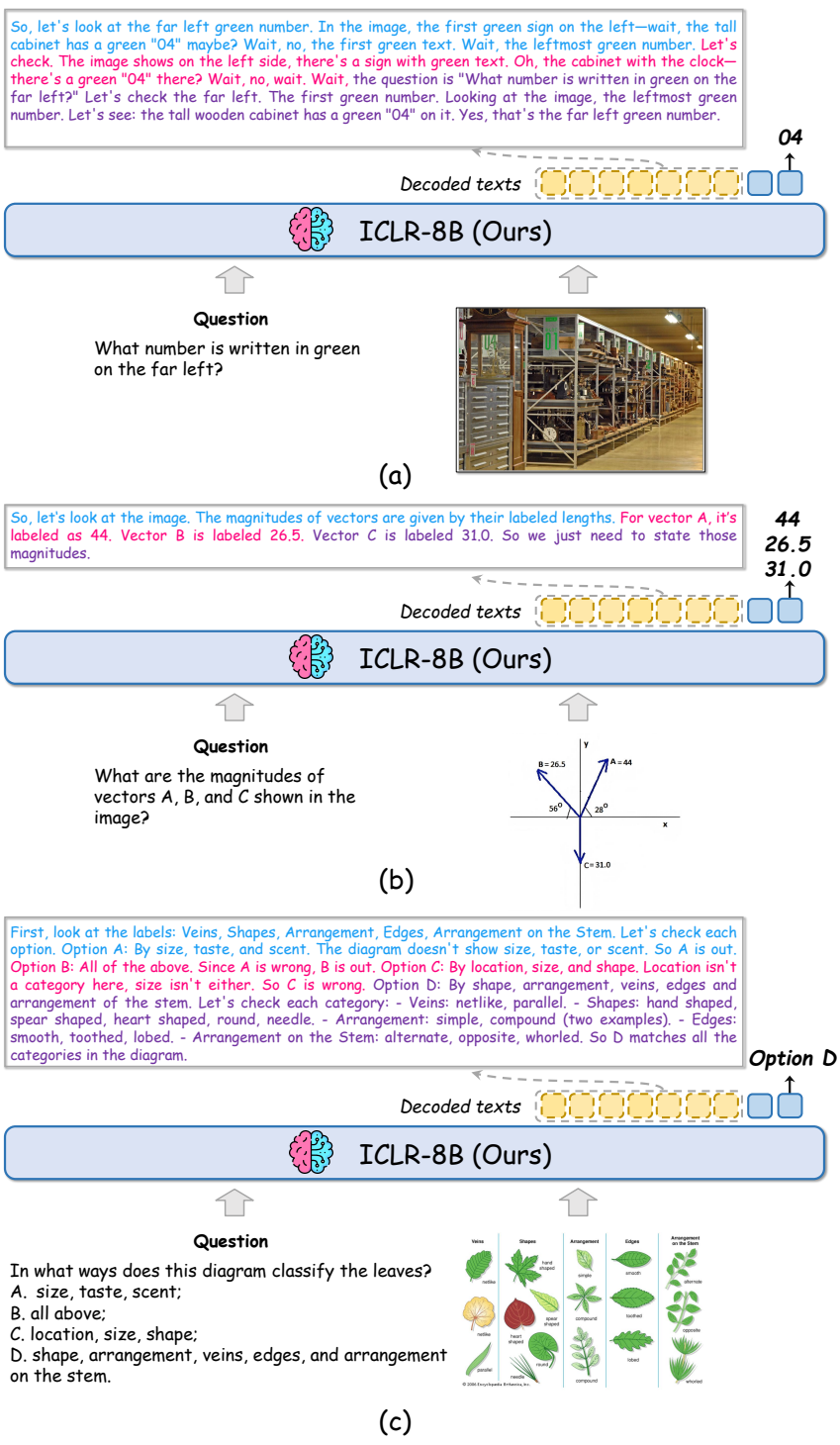


Figure 8: **Decoded latent content for visual question answering.** (a) The decoded text searches for the leftmost green number; the displayed output is 04. (b) The decoded text identifies vector magnitudes of 44, 26.5, and 31.0, matching the displayed output. (c) The decoded text matches the leaf-classification criteria to option D. Colored text distinguishes portions of the decoded passages, not spatial attention. Together, the examples illustrate decoded content associated with visual localization, label reading, and comparison of diagram categories.

A.5.2 DECODED LATENT CONTENT AND VISUAL INPUTS

Figure 8 presents three additional ICLR-8B examples, pairing the input image and question with decoded text and the final output. The decoded content refers to question-relevant objects, numerical labels, or diagram categories. These examples complement the compact trajectories in Figure 6 by showing longer textual interpretations of the task-relevant content.

A.6 ATTENTION-MASKING INTERVENTION

To test whether the model relies on high-attention visual evidence, we mask different attention regions and compare the resulting performance with the full model. Table 4 reports the absolute performance change under each masking strategy.

Table 4: Causal intervention results under attention-region masking strategies using the Qwen3-VL-8B-Instruct backbone. Δ denotes the absolute performance change relative to the full model without masking.

Masking strategy	**ChartQA**	Δ	**MMStar**	Δ	**AI2D**	Δ	**MMT**	Δ
Ours (Full)	**78.8**	–	**66.8**	–	**85.6**	–	**67.7**	–
Mask High-Attn	74.2	**–4.6**	62.5	**–4.3**	80.2	**–5.4**	63.6	**–4.1**
Mask Low-Attn	78.3	–0.5	66.9	+0.1	84.8	–0.8	67.4	–0.3
Random Mask	75.7	–3.1	63.5	–3.3	82.1	–3.5	64.6	–3.1
Mask All Steps	72.5	**–6.3**	60.8	**–6.0**	78.9	**–6.7**	61.8	**–5.9**