CoLT: Teaching Multi-Modal Models to Think with Chain of Latent Thoughts

Lianyu Hu[1], Shengqian Qin[2], Zeqin Liao[1], Qing Guo[3,4⋆], Liang Wan[5], Wei Feng[5], and Yang Liu[1]

[1] Nanyang Technological University, Singapore
[2] Shanghai Jiao Tong University, China
[3] NKIARI, Shenzhen Futian, China
[4] AAIS & VCIP, Nankai University, China
[5] Tianjin University, China

Abstract. Chain-of-thought (CoT) reasoning has enabled multi-modal large language models (MLLMs) to tackle complex visual reasoning tasks by generating explicit intermediate reasoning steps in natural language. However, this text-based reasoning paradigm is inherently slow at inference time with even thousands of tokens and fundamentally constrained by the expressiveness of natural language. In this paper, we propose CoLT (**C**hain **o**f **L**atent **T**houghts), a novel framework that teaches multi-modal models to reason through a chain of latent thought representations instead of verbose text tokens, which can perform thinking with as few as 3 steps. Naively forcing the model to think with latent states easily produces meaningless semantics and makes training unstable. To effectively regulate the latent reasoning process, we introduce a lightweight external decoder that provides step-level supervision for each latent reasoning step in two complementary directions: a *forward* mode that decodes latent thoughts into the textual reasoning of the next step, and a *backward* mode that aligns decoder hidden states with the model's latent thoughts given preceding textual context. We further incorporate *internal supervision* that encourages coherent step-by-step latent transitions. The decoder and internal supervision are removed during inference to maintain high efficiency of latent reasoning. Extensive experiments on eight benchmarks demonstrate that CoLT not only outperforms existing latent reasoning methods such as CODI and SIM-CoT, but also surpasses latent visual reasoning approaches that rely on auxiliary images with costly annotation requirements. Compared to text CoT methods, CoLT can notably reduce the inference time by 10.1× and text decoding time by 22.6×. Code is released at `https://github.com/hulianyuyy/CoLT`.

Keywords: Latent reasoning · Multi-modal models · Chain-of-thought

1 Introduction

Multi-modal large language models (MLLMs) have demonstrated remarkable capabilities in visual understanding and reasoning [3, 17–19, 25, 33, 38, 44, 66]. By

⋆ Corresponding author

integrating vision encoders with powerful language model backbones, these models can process both visual and textual information to solve complex tasks ranging from visual question answering to mathematical problem solving [31, 58]. A key ingredient underlying their success is chain-of-thought (CoT) reasoning [49], which instructs models to decompose problems into intermediate reasoning steps expressed in natural language before arriving at a final answer. Recent reasoning models such as DeepSeek-R1 [14] and QwQ [43] have further amplified this paradigm through long-form CoT, establishing it as the dominant reasoning strategy.

Despite its effectiveness, text-based CoT suffers from several fundamental limitations. First, generating lengthy reasoning chains in natural language significantly increases inference latency, as each reasoning token must be produced autoregressively. This becomes more severe as the thinking process typically consumes hundreds or even thousands of tokens [14, 19, 20], constituting the heaviest computing burden during inference. Second, the decoded textual expressions are inherently deterministic: once a reasoning step is committed to a specific natural language form, it collapses the rich distribution of potential solving trajectories into a single surface realization, eliminating alternative reasoning pathways. Third, errors made in early reasoning steps can propagate and compound through subsequent steps, a phenomenon that is particularly problematic in long reasoning chains [22, 47].

Recent work has begun to explore *latent reasoning* as an alternative, where the intermediate thinking process occurs in a continuous representation space rather than through explicit text generation. CoCoNut [15] pioneered this direction by training language models to reason in a continuous latent space. CODI [37] further explored compressing chain-of-thought into continuous representations via self-distillation. SemCoT [16] tried to align latent thoughts with textual CoT via a trained sentence transformer. However, these approaches are primarily designed for text-only language models and do not address the unique challenges of multi-modal reasoning. More recently, latent visual reasoning [22, 41] leverages auxiliary images to generate latent thought tokens, but introduces significant overhead by requiring additional image annotations and specialized encoders, resulting in high labelling costs and limited scalability.

In this paper, we propose CoLT (**C**hain **o**f **L**atent **T**houghts), a novel framework that enables multi-modal models to perform effective latent reasoning with principled supervision. A critical challenge is that naively forcing the model to perform latent thinking, *i.e.*, simply replacing text tokens with continuous vectors during reasoning, easily produces semantically meaningless latent representations and causes severe training instability, as the unconstrained continuous space provides no inductive bias for structured reasoning. To overcome this, we introduce fine-grained step-level supervision through a lightweight external decoder operating in two complementary modes:

- **Forward decoding supervision**: The decoder takes the preceding latent thoughts as input and is trained to decode the textual reasoning content of

the next step, ensuring that latent thoughts encode meaningful and progressive reasoning information.
- **Backward decoding supervision**: The decoder receives the preceding textual thoughts as input, and the generated hidden states are enforced to approach the latent thought representations, thereby anchoring the latent space to the semantic structure of explicit reasoning.

In addition to these external supervision signals, we introduce an **internal supervision** mechanism that directly encourages coherent transitions between consecutive latent steps by training the model to predict the next step's representation from the current one. The combination of forward, backward, and internal supervision provides a comprehensive training signal that effectively regularizes the latent reasoning process without requiring any auxiliary visual annotations.

Extensive experiments on eight multimodal benchmarks demonstrate that CoLT consistently outperforms existing latent reasoning methods, and surpasses latent visual reasoning approaches that rely on costly image-based supervision. Compared to text CoT methods, CoLT can notably reduce the inference time by 10.1× and text decoding time by 22.6×, achieving superior efficiency. Plentiful visualizations verify that CoLT can successfully encode meaningful thinking patterns within the latent thoughts.

2 Related Work

2.1 Multi-Modal Large Language Models

The rapid advancement of large language models has catalyzed the development of multi-modal large language models (MLLMs) that integrate visual perception with language understanding [1–3,7,12,13,19,28,51,63–66]. LLaVA [28] pioneered the visual instruction tuning paradigm, and its successors have significantly advanced the field: LLaVA-NeXT [27] improved reasoning and OCR capabilities, while LLaVA-OneVision [23] unified image, video, and multi-image understanding into a single framework. The InternVL series has pushed open-source performance, with InternVL 2.5 [7] achieving results competitive with commercial systems, and InternVL3 [66] further advancing capabilities via mixed-modality pre-training. In the Qwen-VL family, Qwen2-VL [45] introduced native dynamic resolution support, and Qwen3-VL [3] strengthened document parsing and long-video comprehension. Step3-VL [19] conducted a vision-centric exploration of MLLM design choices, providing insights into the role of visual representations. While these models have made substantial progress, their reasoning capabilities still heavily rely on text-based intermediate representations, which our work aims to address through latent reasoning.

2.2 Chain-of-Thought Reasoning

Chain-of-thought (CoT) prompting [49] has emerged as a powerful technique that enables language models to solve complex problems by generating intermediate reasoning steps. Li *et al.* [26] showed that CoT empowers transformers to

solve inherently serial problems that are otherwise intractable for constant-depth models. Recent work has dramatically scaled this paradigm: OpenAI o1 [20] demonstrated that training with long chains of thought yields substantial gains on scientific and mathematical benchmarks, while DeepSeek-R1 [14] showed that reinforcement learning can incentivize emergent reasoning capabilities. More recent works like OpenAI GPT-5 [39], GLM-5 [59] and Qwen2.5-xCoder [55] further verify that long-form CoT is the key protocol to build high-intelligence agents.

In the multi-modal domain, CoT reasoning has been extensively explored across diverse modalities. Image analysis [35, 52], video understanding [60] and embodied intelligence [56,62] have fully witnessed the power of CoT reasoning in promoting understanding performance via long-chain thoughts. Wang *et al*. [48] provided a comprehensive survey categorizing multimodal CoT methods. Despite strong performance, these explicit approaches incur significant computational costs due to lengthy text generation, often consuming thousands of tokens [5]. This motivates the exploration of more efficient alternatives that preserve reasoning quality while reducing the generation burden.

2.3 Latent Reasoning

The limitations of text-based reasoning have motivated a growing body of work on latent reasoning that operates in continuous representation spaces [6]. CoCoNut [15] introduced a pioneering framework that trains language models to reason in a continuous latent space, replacing discrete text tokens with continuous thought vectors, and demonstrated that the continuous space can encode multiple alternative reasoning paths. CODI [37] proposed compressing chain-of-thought into continuous representations via self-distillation. SoftCoT [53] introduced a small assistant model to generate continuous soft thought tokens, projected into the main LLM's embedding space to boost reasoning accuracy and efficiency. CoLaR [42] introduced dynamic latent compression that adaptively compresses reasoning chains based on problem difficulty. SIM-CoT [50] used an auxiliary decoder to provide step-level implicit supervision, which stabilizes latent reasoning states to avoid collapse and enhances interpretability.

In the multi-modal setting, latent visual reasoning [22, 46] has emerged as an approach that uses auxiliary images to generate latent thought tokens for reasoning. LVR [22] performed implicit visual inference by requiring the latent visual representation to approximate the encoded features of auxiliary images. MoNet [46] introduced a three-stage framework to gradually internalize auxiliary image features into the generation process of the backbone model. However, these approaches require auxiliary image annotations or generation models, resulting in high labelling costs and limited applicability to scenarios where visual intermediates are not readily available. Our proposed CoLT differs fundamentally by operating entirely in the language model's latent space, requiring no auxiliary visual annotations while providing comprehensive supervision through forward, backward, and internal signals.

3 Method

In this section, we present CoLT, our framework for teaching multi-modal models to reason through latent thought representations. We first provide an overview of the framework (Sec. 3.1), then describe the latent thought representation (Sec. 3.2), the external decoder supervision combining forward and backward decoding (Sec. 3.3), the internal step-level supervision (Sec. 3.4), and the complete training objective (Sec. 3.5).

3.1 Overview

Given a multi-modal input consisting of an image $\mathbf{v}$ and a text query $\mathbf{x}$, a standard MLLM first encodes the image through a vision encoder to obtain visual tokens $\mathbf{z}_v = \text{Encoder}(\mathbf{v})$, which are concatenated with the text tokens as input to the language model. In text-based CoT, the model generates explicit reasoning steps $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_L$ as natural language text, followed by the final answer $\mathbf{y}$:

$$P(\mathbf{y}|\mathbf{v}, \mathbf{x}) = \sum_{\mathbf{r}_{1:L}} P(\mathbf{y}|\mathbf{r}_{1:L}, \mathbf{v}, \mathbf{x}) \prod_{l=1}^{L} P(\mathbf{r}_l|\mathbf{r}_{<l}, \mathbf{v}, \mathbf{x}). \tag{1}$$

CoLT replaces the text reasoning steps with latent thought vectors $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_K \in \mathbb{R}^d$ with $K \ll L$, where d denotes the hidden dimension of the language model. The model directly produces latent representations at designated reasoning positions, bypassing the text generation bottleneck while preserving multi-path exploration in the latent space:

$$P(\mathbf{y}|\mathbf{v}, \mathbf{x}) = P(\mathbf{y}|\mathbf{h}_{1:K}, \mathbf{v}, \mathbf{x}), \tag{2}$$

where each $\mathbf{h}_k$ is obtained from the language model's hidden states at the corresponding latent thought position.

The key challenge in latent reasoning is ensuring that the latent thoughts encode meaningful and structured reasoning information, despite being unconstrained continuous vectors. The core contribution of CoLT is to introduce fine-grained *step-level supervision* that guides the latent thinking process. Specifically, we design two complementary supervision mechanisms:

- **External decoder supervision** (Sec. 3.3): We introduce a lightweight external decoder D_ϕ, a smaller language model from the same model family as the backbone that shares the same tokenizer and vocabulary. The decoder provides step-level supervision from two directions: *forward decoding* that decodes latent thoughts into the textual reasoning of the next step, and *backward decoding* that aligns decoder hidden states with the model's latent thoughts given preceding textual context.
- **Internal step-level supervision** (Sec. 3.4): A lightweight projection head that encourages coherent transitions between consecutive latent steps by predicting the next step's representation from the current one.

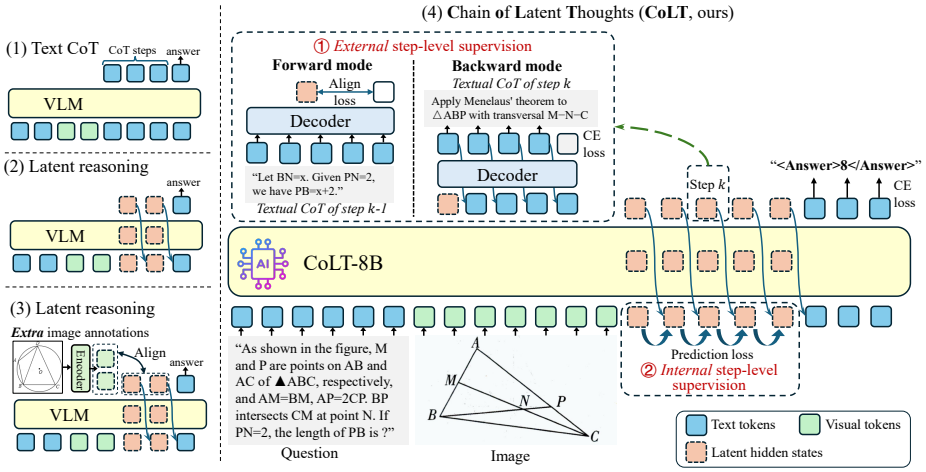


Fig. 1: Comparison of reasoning paradigms and overview of CoLT. Left: (1) text CoT generates verbose reasoning tokens before the answer; (2) latent reasoning replaces text with latent states but lacks explicit supervision; (3) latent visual reasoning aligns latent states with extra image annotations via an encoder. Right: CoLT generates latent thought vectors $\mathbf{h}_1, \ldots, \mathbf{h}_K$ regulated by two complementary mechanisms: external step-level supervision (a decoder D_ϕ operating in forward and backward modes) and internal step-to-step prediction.

Together, these signals comprehensively regulate the latent reasoning process, ensuring that each latent thought encodes progressively meaningful semantics. Figure 1 compares CoLT with existing reasoning paradigms (text CoT, latent reasoning, and latent visual reasoning) and illustrates our framework.

3.2 Latent Thought Representation

To enable structured latent reasoning, we divide the reasoning process into K discrete steps based on the logical structure of the chain-of-thought annotations. At each step, instead of generating explicit text tokens, the model directly produces a hidden state that serves as the latent thought vector. No special tokens are introduced into the vocabulary, and the latent thoughts exist purely as continuous hidden representations. The latent thought at step k is obtained from the last hidden layer of the language model:

$$\mathbf{h}_k = \mathrm{LM}^{\mathrm{last}}(\mathbf{z}_v, \mathbf{x}, \mathbf{h}_1, \ldots, \mathbf{h}_{k-1}) \in \mathbb{R}^d, \tag{3}$$

where $\mathrm{LM}^{\mathrm{last}}(\cdot)$ extracts the last-layer hidden state at the current reasoning position. Each latent thought $\mathbf{h}_k$ is directly fed back into the language model as the input embedding for the next position, allowing subsequent latent thoughts and the final answer to attend to all preceding latent representations through the standard self-attention mechanism. This design preserves the autoregressive

nature of the language model while enabling reasoning in the continuous latent space, avoiding any overhead from extending the model's vocabulary.

During training, we assume access to paired data where explicit text reasoning chains $\mathbf{r}_1, \ldots, \mathbf{r}_L$ are available (from existing CoT annotations or distilled from a teacher model). Each latent thought corresponds to $\frac{L}{K}$ consecutive textual tokens. The latent thought vectors are supervised to encode the same reasoning information through the mechanisms described below, and at inference time, the model uses only the latent thoughts without generating any text reasoning.

3.3 External Decoder Supervision

The external decoder D_ϕ provides step-level supervision from two complementary directions: forward decoding that ensures latent thoughts encode sufficient information to produce textual reasoning, and backward decoding that anchors the latent space to the semantic structure of explicit reasoning chains.

Forward decoding. The forward decoding supervision ensures that the latent thoughts encode sufficient information to reconstruct the explicit reasoning process. The external decoder D_ϕ receives the preceding latent thought $\mathbf{h}_k$ as input and is trained to decode the textual reasoning content of the next step $\mathbf{r}_{k+1}$:

$$\hat{\mathbf{r}}_{k+1} = D_\phi(\mathbf{h}_k), \tag{4}$$

where D_ϕ autoregressively generates the text tokens conditioned on the preceding latent thoughts. Since D_ϕ is a smaller language model from the same model family and shares the same vocabulary as the main backbone, it can naturally decode the latent thoughts into coherent textual reasoning. The forward decoding loss is defined as the negative log-likelihood of the ground-truth textual reasoning:

$$\mathcal{L}_{\text{fwd}} = -\frac{1}{K} \sum_{k=0}^{K-1} \sum_{t=1}^{|\mathbf{r}_{k+1}|} \log P_{D_\phi}(r_{k+1,t} | r_{k+1,<t}, \mathbf{h}_k), \tag{5}$$

where $r_{k+1,t}$ is the t-th token of the $(k+1)$-th reasoning step. This forward supervision provides a strong training signal: the latent thought at each step must capture enough information to enable accurate generation of the next reasoning step. The gradients from the decoder flow back through the latent thoughts to the main language model, encouraging the formation of informative latent representations.

Backward decoding. While forward decoding maps from latent to text, backward decoding provides the reverse constraint by anchoring the latent space to the semantic structure of text reasoning. The decoder D_ϕ receives the preceding textual reasoning steps $\mathbf{r}_{k-1}$ as input and produces a hidden state $\tilde{\mathbf{h}}_k$ trained to align with the model's latent thought:

$$\tilde{\mathbf{h}}_k = D_\phi^{\text{last}}(\mathbf{r}_{k-1}), \tag{6}$$

where D_ϕ^{last} extracts the last-layer hidden state of the decoder after processing the text input. Since D_ϕ belongs to the same model family and shares the same vocabulary as the main backbone, the decoder's text-conditioned hidden states naturally reside in a compatible representation space, enabling effective alignment without additional projection layers. The backward decoding loss enforces alignment between $\tilde{\mathbf{h}}_k$ and the corresponding latent thought $\mathbf{h}_k$:

$$\mathcal{L}_{\text{bwd}} = \frac{1}{K}\sum_{k=1}^{K}\left\| \frac{\tilde{\mathbf{h}}_k}{\|\tilde{\mathbf{h}}_k\|} - \frac{\text{sg}(\mathbf{h}_k)}{\|\text{sg}(\mathbf{h}_k)\|} \right\|^2, \tag{7}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator applied to prevent the latent thoughts from collapsing to match trivial decoder outputs. The normalization ensures that alignment is measured in direction rather than magnitude, providing more stable training dynamics. Together, forward and backward signals provide comprehensive decoder-based regulation: forward decoding ensures the latent thoughts *can produce* reasoning text, while backward decoding ensures they *align with* the semantic content of reasoning steps.

3.4 Internal Step-Level Supervision

Beyond the external decoder supervision, we introduce an internal supervision mechanism that directly regularizes the transitions between consecutive latent thought steps. The key intuition is that adjacent latent steps should have a predictable relationship, mirroring the logical progression in explicit reasoning chains. We implement this through a lightweight projection head $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$ (a two-layer MLP with GELU activation) that predicts the next step's latent thought from the current one:

$$\hat{\mathbf{h}}_{k+1} = f_\theta(\mathbf{h}_k), \tag{8}$$

with the internal supervision loss:

$$\mathcal{L}_{\text{int}} = \frac{1}{K-1}\sum_{k=1}^{K-1}\left(1 - \frac{\hat{\mathbf{h}}_{k+1}\cdot \text{sg}(\mathbf{h}_{k+1})}{\|\hat{\mathbf{h}}_{k+1}\|\cdot\|\text{sg}(\mathbf{h}_{k+1})\|}\right), \tag{9}$$

where we use cosine similarity with a stop-gradient on the target to prevent representational collapse. The internal supervision complements the external decoder signals: while the decoder-based losses anchor the latent space to the *content* of reasoning through text-latent correspondence, the internal prediction loss regularizes the *structure* of the latent chain through step-to-step consistency.

3.5 Training Objective

The complete training objective of CoLT combines the standard task loss with the three supervision signals:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha\mathcal{L}_{\text{fwd}} + \beta\mathcal{L}_{\text{bwd}} + \gamma\mathcal{L}_{\text{int}}, \tag{10}$$

where $\mathcal{L}_{\text{task}}$ is the standard next-token prediction loss for generating the final answer $\mathbf{y}$ conditioned on the latent thoughts:

$$\mathcal{L}_{\text{task}} = -\sum_{t=1}^{|\mathbf{y}|} \log P_{\text{LM}}(y_t | y_{<t}, \mathbf{h}_{1:K}, \mathbf{v}, \mathbf{x}), \tag{11}$$

and α, β, γ are hyperparameters controlling the relative importance of each supervision signal.

Training procedure. The training of CoLT proceeds in two stages. In the first stage (*supervised fine-tuning*), we train the external decoder and LLM of the backbone with the combined objective in Eq. (10) on CoT annotations from the OneThinker dataset [10], which provides high-quality textual reasoning annotations across diverse visual tasks. This stage teaches the model to produce structured latent thoughts regulated by the three supervision signals.

In the second stage (*reinforcement learning*), we apply Group Relative Policy Optimization (GRPO) [36] on 50K CoT samples, allowing the latent thoughts to update freely without the decoder and internal supervision constraints. By removing the auxiliary losses and optimizing solely through outcome-based rewards, the model is encouraged to explore latent reasoning strategies beyond what the supervised signals prescribe. During inference, the external decoder and projection head are discarded entirely, incurring no additional computational overhead.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate CoLT on eight diverse multi-modal benchmarks: SeedBench [24] for comprehensive visual understanding and reasoning; MMBench [29] covering multiple visual perception and reasoning abilities; ChartQA [34] for chart comprehension and data reasoning; TextVQA [40] targeting text-rich visual question answering; ScienceQA [32] for multi-modal science question answering; MMStar [4] with carefully curated vision-indispensable problems; AI2D [21] focusing on diagram understanding; and MMT-Bench [57] for comprehensive multitask evaluation.

Implementation details. We build CoLT on top of Qwen3-VL-8B-Instruct [3], and instantiate the external decoder D_ϕ as Qwen3-0.6B [54], a smaller model from the same Qwen3 family that shares the same tokenizer and vocabulary. The projection head f_θ is a 2-layer MLP with GELU activation. We set the number of latent thought steps $K = 3$. The loss weights are $\alpha = 0.2$, $\beta = 0.2$, and $\gamma = 0.2$. We dynamically split the textual CoT annotations into K splits during training to enhance the ability of model understanding. We train all models on the image subset of Onethinker dataset [11] with one epoch. All models are evaluated with the same prompt and same configurations to ensure fairness. We use the AdamW [30] optimizer with a cosine learning rate schedule and a batch size of 8. More details can be referred to our codebase.

Table 1: Comprehensive comparison across proprietary models, open-source MLLMs, latent reasoning methods and latent visual reasoning methods. 'SQA' is the abbreviation of 'ScienceQA' benchmark. †: methods built on the Qwen3-VL-8B backbone. Best non-proprietary results in **bold**.

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	**77.5**	**84.6**	**74.7**	**81.3**	**92.8**	**68.9**	**85.4**	**67.4**	**79.1**
Relative improvement		**+1.1**	**+1.2**	**+9.6**	**+6.1**	**+1.0**	**+1.8**	**+1.8**	**+4.1**	**+3.3**

4.2 Main Results

Table 1 presents a comprehensive comparison of CoLT against proprietary models, open-source MLLMs, latent reasoning methods, and latent visual reasoning approaches across all eight benchmarks.

Baseline comparison. Compared to direct answer (no reasoning), CoLT improves average performance by +9.6% (79.1 vs. 69.5), confirming that latent thoughts provide substantial reasoning capacity despite generating only 3 continuous vectors. Compared to textual reasoning with explicit chain-of-thought, CoLT achieves +3.4% higher average (79.1 vs. 75.7), with the largest gains on ChartQA (+9.6%) and TextVQA (+6.1%), while using significantly fewer tokens.

Comparison with latent reasoning methods. We compare against five latent reasoning methods, all built on the same Qwen3-VL-8B backbone for fair comparison. CCoT [8] compresses chain-of-thought into dense representations; ICoT [9] internalizes reasoning via distillation; CODI [37] distills CoT into latent thoughts; SoftCoT [53] uses soft continuous embeddings; and SIM-CoT [50] aligns implicit tokens with explicit reasoning. CoLT surpasses the strongest baseline SIM-CoT by +5.1% on average (79.1 vs. 74.0), with consistent gains including +5.1% on SeedBench, +8.0% on ChartQA, and +11.1% on TextVQA, demonstrating the effectiveness of three-way step-level supervision.

Comparison with latent visual reasoning. We further compare with approaches that incorporate auxiliary images for additional supervision: Multimodal CoT [61] decomposes reasoning into rationale generation and answer inference; LaCoT [41] performs visual reasoning through continuous latent rep-

Table 2: Ablation study on supervision signal combinations. ✓ indicates the corresponding loss is active during training.

$\mathcal{L}_{\text{fwd}}$	$\mathcal{L}_{\text{bwd}}$	$\mathcal{L}_{\text{int}}$	**SeedBench**	**MMStar**	**AI2D**	**MMT**	**Avg.**
			63.5	55.0	73.2	53.8	61.4
✓			68.2	59.5	77.6	57.8	65.8
	✓		69.5	60.8	78.8	59.5	67.2
		✓	65.8	57.2	75.5	56.0	63.6
✓	✓		72.0	63.5	81.5	62.0	69.8
✓		✓	70.5	61.5	80.0	60.2	68.1
	✓	✓	71.2	62.8	81.0	61.5	69.1
✓	✓	✓	**75.2**	**66.5**	**83.8**	**65.0**	**72.6**

resentations; and LVR [22] generates latent visual tokens as intermediate representations. CoLT outperforms the strongest baseline LVR by +5.5% on average (79.1 vs. 73.6), with notable gains on TextVQA (+11.1%) and SeedBench (+6.3%) while eliminating the need for auxiliary image annotations, validating our design of operating in the language model's latent space.

4.3 Ablation Study

We conduct comprehensive ablation studies to analyze the contribution of each component in CoLT. All ablations use the Qwen3-VL-8B backbone and report results on four representative benchmarks.

Effect of supervision signals. Table 2 presents a systematic ablation over all combinations of the three supervision signals. Several key findings emerge: (1) Among individual losses, backward decoding alone achieves the highest single-component accuracy (67.2%), outperforming forward decoding (65.8%) and internal supervision (63.6%). (2) Every pairwise combination substantially improves over its constituent single losses, with $\mathcal{L}_{\text{fwd}} + \mathcal{L}_{\text{bwd}}$ (69.8%) being the strongest pair. (3) Adding the third signal to any pair consistently yields further gains, with backward decoding contributing the largest marginal improvement (+4.5% when added to $\mathcal{L}_{\text{fwd}} + \mathcal{L}_{\text{int}}$), followed by forward (+3.5%) and internal (+2.8%). Overall, this confirms all three losses provide non-redundant regularization and complete three-way supervision is essential for optimal performance.

Effect of decoder architecture. We investigate the impact of the external decoder size in Tab. 3. All decoder variants are from the Qwen3 family and share the same tokenizer and vocabulary. Scaling the decoder from Qwen3-0.6B to Qwen3-8B yields only marginal improvements: Qwen3-0.6B performs only slightly worse than Qwen3-1.7B (−0.3%), and Qwen3-1.7B nearly matches Qwen3-4B (−0.2%). Further scaling to Qwen3-8B provides a negligible +0.1% gain. These results demonstrate that even a lightweight 0.6B decoder provides effective step-level supervision, and we use Qwen3-0.6B as the default for its best accuracy-efficiency trade-off.

Table 3: Ablation on external decoder architecture. All decoders are from the Qwen3 family and share the same vocabulary as the backbone.

Decoder	Params	SeedBench	MMStar	AI2D	MMT	Avg.
Qwen3-0.6B	0.6B	**75.2**	**66.5**	**83.8**	**65.0**	**72.6**
Qwen3-1.7B	1.7B	75.5	66.8	84.1	65.3	72.9
Qwen3-4B	4.0B	75.7	67.0	84.3	65.5	73.1
Qwen3-8B	8.0B	75.8	67.1	84.4	65.6	73.2

Table 4: Effect of the number of latent thought steps K on four benchmarks. Performance peaks at K=3 and gradually decreases with larger K.

K	SeedBench	MMStar	AI2D	MMT	Avg.
1	69.5	60.5	78.2	58.8	66.8
2	73.0	64.0	81.8	62.5	70.3
3	**75.2**	**66.5**	**83.8**	**65.0**	**72.6**
4	74.8	66.0	83.5	64.5	72.2
6	74.3	65.5	83.0	64.0	71.7
8	73.8	65.0	82.6	63.5	71.2

Effect of number of latent steps. Table 4 presents the performance of CoLT as a function of the number of latent thought steps K. Performance increases rapidly from K=1 (66.8%) through K=2 (70.3%) to K=3 (72.6%), beyond which it gradually decreases: K=4 (72.2%), K=6 (71.7%), and K=8 (71.2%). This suggests that K=3 provides the optimal trade-off, where too few steps limit reasoning capacity while excessive steps introduce redundancy that slightly degrades performance. We use K=3 as the default.

Generalization across latent steps. Beyond the default setting where training and testing use the same K, we investigate cross-K generalization in Tab. 5. The diagonal entries (matching K) consistently achieve the best performance, confirming that the model learns step-specific reasoning patterns. Notably, CoLT exhibits strong robustness to K mismatch: models trained with K=3 and tested at K=6 incur only minor drops (0.7% on SeedBench, 0.7% on MMStar, 0.7% on MMT), and testing at K=8 similarly retains most of the performance with drops of merely 1.2%. Even the largest mismatch (training with K=1, testing at K=8) causes at most 2.0% degradation, confirming K=3 as a practical default that generalizes well across different test-time step counts.

4.4 Robustness to Input Noise

Compared to explicit text decoding, latent reasoning can preserve the probability of multi-path reasoning within latent spaces and thus can perform more robustly to input noise. To assess the robustness of latent reasoning under noisy inputs, we evaluate CoLT against Direct Answer and Text CoT under two categories

Table 5: Cross-K generalization: performance (%) on individual benchmarks when training and testing with different numbers of latent steps. Best result in **bold**.

(a) SeedBench

Train K	Test K = 1	3	6	8
1	**69.5**	68.8	68.0	67.5
3	70.0	**75.2**	74.5	74.0
6	69.0	74.0	**74.3**	74.0
8	68.5	73.5	73.6	**73.8**

(b) MMStar

Train K	Test K = 1	3	6	8
1	**60.5**	59.8	59.2	58.8
3	61.0	**66.5**	65.8	65.3
6	60.0	65.2	**65.5**	65.2
8	59.5	64.8	64.8	**65.0**

(c) MMT

Train K	Test K = 1	3	6	8
1	**58.8**	58.0	57.5	57.0
3	59.5	**65.0**	64.3	63.8
6	58.5	63.8	**64.0**	63.8
8	58.0	63.2	63.3	**63.5**

Table 6: Robustness to input noise. Per-benchmark accuracy degradation (%) under visual noise and text noise on SeedBench, MMStar, and MMT.

Noise	Level	SeedBench Direct Answer	SeedBench Text CoT	SeedBench CoLT	MMStar Direct Answer	MMStar Text CoT	MMStar CoLT	MMT Direct Answer	MMT Text CoT	MMT CoLT
Visual Noise										
Gauss. blur	σ=1	−7.5	−5.2	**−2.8**	−8.8	−6.0	**−3.4**	−8.2	−5.6	**−3.0**
	σ=2	−11.8	−8.0	**−4.8**	−13.0	−9.0	**−5.5**	−12.2	−8.5	**−5.0**
	σ=3	−15.8	−11.0	**−6.8**	−17.5	−12.5	**−8.0**	−16.2	−11.5	**−7.2**
Occlusion	10%	−8.2	−5.6	**−3.0**	−9.2	−6.4	**−3.6**	−8.5	−5.8	**−3.2**
	20%	−12.5	−8.5	**−5.2**	−14.0	−9.8	**−6.0**	−13.0	−9.0	**−5.5**
	30%	−17.0	−12.2	**−7.8**	−19.0	−13.8	**−9.0**	−17.5	−12.5	**−8.2**
Text Noise										
Spell. err.	5%	−7.8	−5.5	**−3.2**	−7.2	−5.0	**−2.8**	−8.0	−5.8	**−3.4**
	10%	−12.2	−8.2	**−5.0**	−11.5	−7.8	**−4.5**	−12.5	−8.5	**−5.2**
Key del.	10%	−10.5	−7.5	**−4.2**	−9.5	−6.8	**−3.8**	−10.8	−7.8	**−4.5**
	20%	−15.5	−11.2	**−6.8**	−14.2	−10.0	**−6.2**	−16.0	−11.5	**−7.2**

of perturbations. For visual noise, we apply Gaussian blur ($\sigma \in \{1, 2, 3\}$) and random occlusion (masking 10%/20%/30% of image patches). For text noise, we introduce character-level spelling errors (5%/10% of tokens) and key information deletion (10%/20% of question tokens). Table 6 reports per-benchmark accuracy degradation on the SeedBench, MMStar, and MMT benchmarks.

CoLT consistently exhibits the lowest performance degradation across all noise conditions and benchmarks. Under visual noise, MMStar shows the highest sensitivity due to its reliance on vision-indispensable problems: at the most severe occlusion level (30%), Direct Answer degrades by 19.0%, Text CoT by 13.8%, and CoLT by only 9.0%. In contrast, SeedBench is comparatively more resilient, with CoLT degrading by at most 7.8%. For text noise, MMT is the most affected due to its diverse multitask evaluation, while MMStar shows the smallest drops given its focus on vision-critical problems. Overall, CoLT's degradation ranges from 2.8% to 9.0% across all conditions, substantially lower than Direct Answer (7.2% to 19.0%) and Text CoT (5.0% to 13.8%). We attribute this robustness to the continuous nature of latent representations: whereas discrete text tokens propagate local input errors through the sequential decoding process,

Table 7: Inference speed comparison on MMStar and SeedBench (per-sample average). All methods use the same Qwen3-VL-8B backbone on a single H200 GPU. Time is decomposed into input encoding and token generation (in seconds).

Method	MMStar				MMT-Bench			
	CoT Tokens	Time		Acc.	CoT Tokens	Time		Acc.
		Encode	Generate			Encode	Generate	
Direct answer	0	0.45	0.14	58.6	0	0.46	0.15	55.6
Text CoT	142.1	0.47	7.24	67.1	138.5	0.48	7.38	63.3
CoLT	3	0.44	0.32	**68.9**	3	0.45	0.33	**67.4**
	(↓ **47.4**×)		(↓ **22.6**×)		(↓ **46.2**×)		(↓ **22.4**×)	

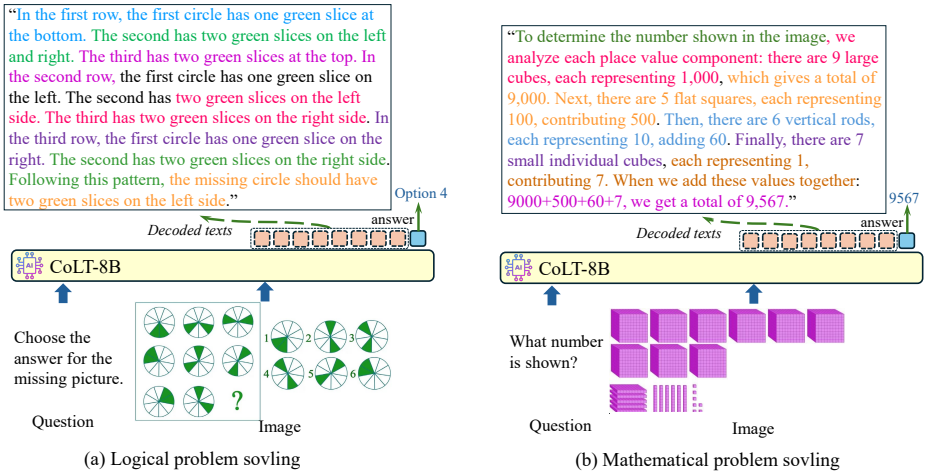
continuous latent vectors distribute information across all dimensions, enabling more graceful degradation under noise. Furthermore, the three-way supervision in CoLT encourages latent states to capture abstract reasoning patterns rather than surface-level input features, making them inherently less sensitive to input perturbations.

4.5 Inference Speed Comparison

A key advantage of latent reasoning over text-based CoT is the reduced inference time. Table 7 decomposes the per-sample inference time on MMStar and MMT-Bench into input encoding and token generation. The encoding phase is roughly constant across methods (∼0.44–0.48 s), as it processes the same inputs regardless of reasoning strategy. The critical difference lies in generation: text CoT requires 7.24 s (MMStar) and 7.38 s (MMT-Bench) to produce ∼142 and ∼139 reasoning tokens respectively, whereas CoLT completes generation in 0.32 s and 0.33 s using only 3 latent vectors, achieving 22.6× and 22.4× reduction in generation time. Overall, CoLT delivers 10.1× (MMStar) and 10.1× (MMT-Bench) end-to-end speedup while maintaining higher accuracy.

4.6 Qualitative Analysis

Figure 2 presents two MathVista examples where we use the forward decoder to project the latent thoughts of CoLT back into text for interpretation. In the logical pattern-matching problem (a), the decoded text reveals that the latent steps systematically describe the visual pattern row by row, identifying the position and number of green slices in each circle, and then infer the missing pattern to arrive at the correct answer. In the mathematical problem (b), the decoded text shows that the latent steps decompose the base-10 block representation into individual place-value components and compute the final sum. The color-coded segments demonstrate that each latent step encodes a distinct and meaningful portion of the reasoning chain, confirming that CoLT's latent representations capture structured, interpretable thought patterns. The decoded content also maintains logical coherence across consecutive latent steps, meaning that our internal supervision enforces meaningful step-to-step transitions.



(a) Logical problem sovling (b) Mathematical problem sovling

Fig. 2: Two qualitative examples from MathVista with decoded latent thoughts. We use the forward decoder to project latent states back into text. Color-coded segments indicate reasoning content decoded from different latent steps.

5 Conclusion

We presented CoLT (Chain of Latent Thoughts), a novel framework that teaches multi-modal models to reason through latent thought representations instead of explicit text-based CoT. The core innovation is introducing three complementary step-level supervisions: forward decoding, backward decoding, and internal supervision. Extensive experiments on eight benchmarks demonstrate that CoLT consistently outperforms existing latent reasoning methods and achieves up to 22.6× speedup over text CoT. Promising directions include adaptive latent steps by problem difficulty, and exploring hybrid latent-text reasoning frameworks.

Acknowledgements

Our work is supported by National Natural Science Foundation of China (Grant No. U2574216), and Emerging Frontiers Cultivation Program of Tianjin University Interdisciplinary Center. This research is also supported by the Fundamental Research Funds for the Central Universities (No. XXX-63263254); by the National Research Foundation Singapore and the Cyber Security Agency under the National Cybersecurity R&D Programme (NCRP25-P04-TAICeN); and is part of the IN-CYPHER programme and is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. Any opinions, findings and conclusions, or recommendations expressed in these materials are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore, Cyber Security Agency of Singapore, Singapore.

References

1. Bai, J., Lei, Y., Shi, Q., Feng, A., Xin, Y., Zhao, Z., Shen, F., Yu, K., Li, J.: Threshold-guided optimization for visual generative models. arXiv preprint arXiv:2605.04653 (2026)
2. Bai, J., Li, Y., Zhu, Y., Xin, Y., Shi, Q., Feng, A., Liu, X., Tao, M., Xue, J., Li, X., et al.: Prism: Efficient test-time scaling via hierarchical search and self-verification for discrete diffusion language models. arXiv preprint arXiv:2602.01842 (2026)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? Advances in Neural Information Processing Systems **37**, 27056–27087 (2024)
5. Chen, Q., Qin, L., Liu, J., Peng, D., Guan, J., Wang, P., Hu, M., Zhou, Y., Gao, T., Che, W.: Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. arXiv preprint arXiv:2503.09567 (2025)
6. Chen, X., Zhao, A., Xia, H., Lu, X., Wang, H., Chen, Y., Zhang, W., Wang, J., Li, W., Shen, X.: Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. arXiv preprint arXiv:2505.16782 (2025)
7. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
8. Cheng, J., Van Durme, B.: Compressed chain of thought: Efficient reasoning through dense representations. arXiv preprint arXiv:2412.13171 (2024)
9. Deng, Y., Prasad, K., Fernandez, R., Smolensky, P., Chaudhary, V., Shieber, S.: Implicit chain of thought reasoning via knowledge distillation. arXiv preprint arXiv:2311.01460 (2023)
10. Feng, K., Zhang, M., Li, H., Fan, K., Chen, S., Jiang, Y., Zheng, D., Sun, P., Zhang, Y., Sun, H., et al.: Onethinker: All-in-one reasoning model for image and video. arXiv preprint arXiv:2512.03043 (2025)
11. Feng, K., Zhang, M., Li, H., Fan, K., Chen, S., Jiang, Y., Zheng, D., Sun, P., Zhang, Y., Sun, H., 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)
12. Fu, R., Li, Y., Zhang, Z., Wu, J., Liu, Y., Cao, S., Zeng, Y., Zhang, Y., Du, X., Fong, S.: Neurosymactive: Differentiable neural-symbolic reasoning with active exploration for knowledge graph question answering. arXiv preprint arXiv:2602.15353 (2026)
13. Fu, R., Wang, Y., Xu, T., Liu, Y., Tang, W., Wu, W., Ma, X., Fong, S.: S-path-rag: Semantic-aware shortest-path retrieval augmented generation for multi-hop knowledge graph question answering. In: Proceedings of the ACM Web Conference 2026. pp. 4057–4068 (2026)
14. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
15. Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., Tian, Y.: Training large language models to reason in a continuous latent space. arXiv preprint arXiv:2412.06769 (2024)

16. He, Y., Zheng, W., Zhu, Y., Zheng, Z., Su, L., Vasudevan, S., Guo, Q., Hong, L., Li, J.: Semcot: Accelerating chain-of-thought reasoning through semantically-aligned implicit tokens. Advances in Neural Information Processing Systems (2025)
17. Hu, L., Gao, L., Shang, F., Wan, L., Feng, W.: illava: An image is worth fewer than 1/3 input tokens in large multimodal models. International Conference on Learning Representation (2026)
18. Hu, L., Ma, X., Liao, Z., Liu, Y.: Tvi-cot: Text-visual interleaved chain-of-thought reasoning for multimodal understanding. International Conference on Machine Learning (2026)
19. Huang, A., Yao, C., Han, C., Wan, F., Guo, H., Lv, H., Zhou, H., Wang, J., Zhou, J., Sun, J., et al.: Step3-vl-10b technical report. arXiv preprint arXiv:2601.09668 (2026)
20. Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al.: Openai o1 system card. arXiv preprint arXiv:2412.16720 (2024)
21. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European conference on computer vision. pp. 235–251. Springer (2016)
22. Li, B., Sun, X., Liu, J., Wang, Z., Wu, J., Yu, X., Chen, H., Barsoum, E., Chen, M., Liu, Z.: Latent visual reasoning. International Conference on Learning Representation (2026)
23. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
24. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench: Benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13299–13308 (2024)
25. Li, J., Huang, Y., Wu, M., Zhang, B., Ji, X., Zhang, C.: Clip-sp: Vision-language model with adaptive prompting for scene parsing. Computational Visual Media **10**(4), 741–752 (2024)
26. Li, Z., Liu, H., Zhou, D., Ma, T.: Chain of thought empowers transformers to solve inherently serial problems. In: The Twelfth International Conference on Learning Representations (2024)
27. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llavanext: Improved reasoning, ocr, and world knowledge (2024)
28. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
29. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
30. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
31. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
32. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems **35**, 2507–2521 (2022)

33. Lyu, K., Yuan, Z., He, J., Yan, Q., Su, X., Hu, N., Liu, Y., Hao, C., Qin, S., Hu, L., et al.: Photocraft: Agentic reasoning with hierarchical self-evolving memory for deep image search. arXiv preprint arXiv:2606.03099 (2026)
34. Masry, A., Tan, J.Q., Joty, S., Hoque, E., 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)
35. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems **37**, 8612–8642 (2024)
36. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
37. Shen, Z., Yan, H., Zhang, L., Hu, Z., Du, Y., He, Y.: 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)
38. Shi, H., Liu, W., Li, Z., Fang, X., Meng, X., Peng, W., Zhong, H., Liu, M., Wang, Y.: Intelligent robot systems: a survey from the perspective of visual intelligence. Visual Intelligence **4**(1), 14 (2026)
39. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2026)
40. Singh, A., Natarjan, V., Shah, M., Jiang, Y., Chen, X., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8317–8326 (2019)
41. Sun, G., Hua, H., Wang, J., Luo, J., Dianat, S., Rabbani, M., Rao, R., Tao, Z.: Latent chain-of-thought for visual reasoning. arXiv preprint arXiv:2510.23925 (2025)
42. Tan, W., Li, J., Ju, J., Luo, Z., Song, R., Luan, J.: Think silently, think fast: Dynamic latent compression of llm reasoning chains. arXiv preprint arXiv:2505.16552 (2025)
43. Team, Q.: Qwq: Reflect deeply on the boundaries of the unknown (2024)
44. Wang, C., Peng, H.Y., Liu, Y.T., Gu, J., Hu, S.M.: Diffusion models for 3d generation: A survey. Computational Visual Media **11**(1), 1–28 (2025)
45. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
46. Wang, Q., Shi, Y., Wang, Y., Zhang, Y., Wan, P., Gai, K., Ying, X., Wang, Y.: Monet: Reasoning in latent visual space beyond images and language. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2026)
47. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. International Conference on Learning Representation (2023)
48. Wang, Y., Wu, S., Zhang, Y., Yan, S., Liu, Z., Luo, J., Fei, H.: Multimodal chain-of-thought reasoning: A comprehensive survey. arXiv preprint arXiv:2503.12605 (2025)
49. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)

50. Wei, X., Liu, X., Zang, Y., Dong, X., Cao, Y., Wang, J., Qiu, X., Lin, D.: Simcot: Supervised implicit chain-of-thought. International Conference on Learning Representation (2026)
51. Xiao, J., Wu, Z., Lin, H., Chen, Y., Liu, Y., Zhao, X., Wang, Z., He, Z.: Not just what's there: Enabling clip to comprehend negated visual descriptions without finetuning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 10978–10986 (2026)
52. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2087–2098 (2025)
53. Xu, Y., Guo, X., Zeng, Z., Miao, C.: Softcot: Soft chain-of-thought for efficient reasoning with llms. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 23336–23351 (2025)
54. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
55. Yang, J., Zhang, W., Miao, Y., Quan, S., Wu, Z., Peng, Q., Yang, L., Liu, T., Cui, Z., Hui, B., et al.: Qwen2. 5-xcoder: Multi-agent collaboration for multilingual code instruction tuning. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 13121–13131 (2025)
56. Ye, A., Zhang, Z., Wang, B., Wang, X., Zhang, D., Zhu, Z.: Vla-r1: Enhancing reasoning in vision-language-action models. arXiv preprint arXiv:2510.01623 (2025)
57. Ying, K., Meng, F., Wang, J., Li, Z., Lin, H., Yang, Y., Zhang, H., Zhang, W., Lin, Y., Liu, S., et al.: Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. arXiv preprint arXiv:2404.16006 (2024)
58. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9556–9567 (2024)
59. Zeng, A., Lv, X., Hou, Z., Du, Z., Zheng, Q., Chen, B., Yin, D., Ge, C., Xie, C., Wang, C., et al.: Glm-5: from vibe coding to agentic engineering. arXiv preprint arXiv:2602.15763 (2026)
60. Zhang, S., Hao, X., Tang, Y., Zhang, L., Wang, P., Wang, Z., Ma, H., Zhang, S.: Video-cot: A comprehensive dataset for spatiotemporal understanding of videos based on chain-of-thought. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 12745–12752 (2025)
61. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2023)
62. Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al.: Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1702–1713 (2025)
63. Zhu, C., Lin, Y., Chen, S., Wang, Y., Lin, J.: Medeyes: Learning dynamic visual focus for medical progressive diagnosis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 13916–13924 (2026)
64. Zhu, C., Lin, Y., Shao, J., Lin, J., Wang, Y.: Pathology-aware prototype evolution via llm-driven semantic disambiguation for multicenter diabetic retinopathy diagnosis. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 9196–9205 (2025)

65. Zhu, C., Zeng, J., Jiang, J., Lin, J., Wang, Y.: Medsynapse-v: Bridging visual perception and clinical intuition via latent memory evolution (2026), https://arxiv.org/abs/2604.26283
66. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)