EndoCoT: Scaling Endogenous Chain-of-Thought Reasoning in Diffusion Models Supplementary Material

Xuanlang Dai[1,2], Yujie Zhou[1,3], Long Xing[1,4], Jiazi Bu[1,3], Xilin Wei[1,5], Yuhong Liu[1,3], Beichen Zhang[1,6], Kai Chen[1†], and Yuhang Zang[1†]

[1] Shanghai AI Laboratory, Shanghai, China
[2] Xi'an Jiaotong University, Xi'an, China
[3] Shanghai Jiao Tong University, Shanghai, China
[4] University of Science and Technology of China, Hefei, China
[5] Fudan University, Shanghai, China
[6] The Chinese University of Hong Kong, Hong Kong, China

Abstract. Recently, Multimodal Large Language Models (MLLMs) have been widely integrated into diffusion frameworks primarily as text encoders to tackle complex tasks such as spatial reasoning. However, this paradigm suffers from two critical limitations: (i) **MLLMs text encoder exhibit insufficient reasoning depth**. Single-step encoding fails to activate the Chain-of-Thought process, which is essential for MLLMs to provide accurate guidance for complex tasks. (ii) **The guidance remains invariant during the decoding process**. Invariant guidance during decoding prevents DiT from progressively decomposing complex instructions into actionable denoising steps, even with correct MLLM encodings. To this end, we propose Endogenous Chain-of-Thought (EndoCoT), a novel framework that facilitates structured visual reasoning in MLLMs by iteratively refining latent thought states through an iterative thought guidance module, and then bridges these states to the DiT's denoising process. Second, a terminal thought grounding module is applied to ensure the reasoning trajectory remains grounded in textual supervision by aligning the final state with ground-truth answers. With these two components, the MLLM text encoder delivers meticulously reasoned guidance, enabling the DiT to execute it progressively and ultimately solve complex tasks in a step-by-step manner. Extensive evaluations across diverse benchmarks (e.g., Maze, TSP, VSP, and Sudoku) achieve an average accuracy of 92.1%, outperforming the strongest baseline by 8.3 percentage points.

Keywords: Diffusion Models · Chain-of-Thought Reasoning

1 Introduction

Diffusion models [11, 15, 22, 23] have revolutionized visual generation, achieving unprecedented levels of photorealism and diversity. Recent efforts integrate Mul-

†Corresponding author.

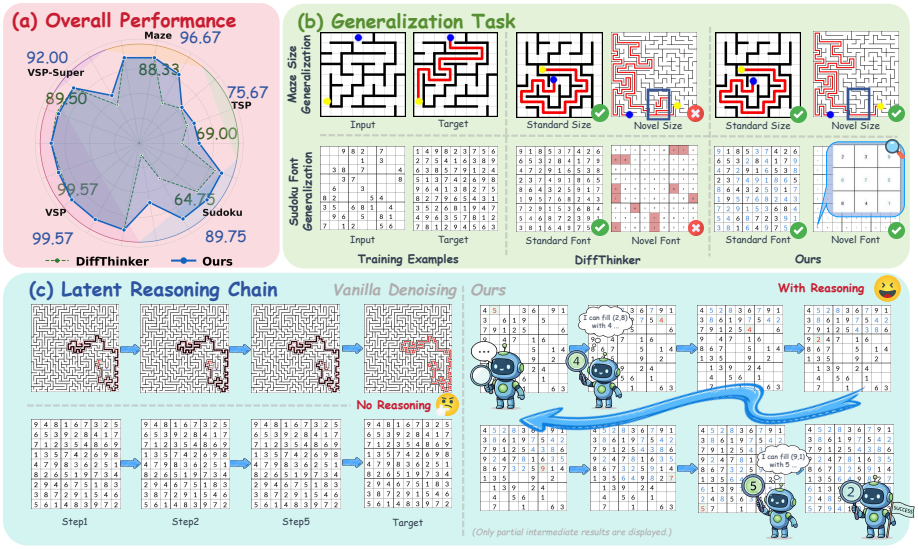

Fig. 1: EndoCoT enables endogenous chain-of-thought reasoning. (a) Radar plot showing EndoCoT outperforms baselines across all benchmarks. **(b)** On visual reasoning tasks requiring generalization (maze size, Sudoku font), previous work [10] fails on novel domains while EndoCoT consistently generalizes correctly. **(c)** Vanilla denoising (left) commits to solutions early without reasoning, while our approach (right) enables interpretable, step-by-step reasoning chains.

timodal Large Language Models (MLLMs) [3, 5, 16, 29] as text encoders to augment semantic understanding and logical alignment. Yet despite these advances, they struggle with tasks requiring *step-by-step logical reasoning*, such as solving mazes, planning Traveling Salesperson Problem (TSP) routes, or completing Sudoku puzzles, where solutions must adhere to strict sequential constraints and generalize to novel configurations.

Current paradigms relegate MLLMs to the role of *static conditional encoders*, computing text embeddings once at the beginning of generation.

This passive integration fails to exploit the dynamic reasoning potential of MLLMs. Even recent explicit attempts like DiffThinker [10] to "inject" reasoning into diffusion models result in superficial alignment rather than genuine cognitive processing. While previous work [10] reports improved metrics on specific benchmarks, they produce fragile solutions that fail catastrophically when generalized to novel domains, see Fig. 1(b).

Our analysis reveals why these failures occur: despite being built on powerful MLLMs, these models do not actually perform reasoning *during* generation.

Instead, as shown in in Fig. 1(c), they commit to their final solution within the first few denoising steps and merely refine visual quality thereafter. This suggests that without an *endogenous* mechanism, forced alignment results in fragile pattern matching rather than structured, iterative reasoning.

Motivated by these persistent failures in supervised fitting, we conduct a systematic empirical analysis (Sec. 2) that identifies two key bottlenecks: (1) *limited single-step reasoning*: MLLMs cannot encode all necessary logical constraints in a single forward pass; (2) *static-guidance failure*: Diffusion Transformers (DiTs) cannot maintain alignment with complex logical constraints when conditioned on static embeddings.

Chain-of-Thought (CoT) reasoning [28] has shown how to overcome similar limitations in MLLMs [2, 32] through iterative refinement, but diffusion models currently lack an analogous mechanism.

To this end, we propose *Endogenous* CoT (**EndoCoT**), a framework that enables diffusion models to perform structured reasoning by iteratively exploring the semantic latent space.

Specifically, EndoCoT encompasses two primary components: (1) **Iterative Thought Guidance**: iteratively updates latent states in the MLLM to create a CoT-like reasoning process and establishes correspondence with DiT's denoising process. (2) **Terminal Thought Grounding**: aligns the MLLM's final reasoning state with ground-truth answers, ensuring the reasoning trajectory remains grounded in textual supervision and preventing cumulative drift in the terminal output. Consequently, EndoCoT serves as a complete end-to-end pipeline, seamlessly integrating reasoning and generation within a unified diffusion framework. Experiments are conducted across four diverse reasoning tasks: Maze, TSP, Visual Spatial Planning (VSP), and Sudoku, demonstrating consistently superior performance over state-of-the-art baselines including DiffThinker [10] and Qwen3-VL-8B [3]. As summarized in Fig. 1(a), EndoCoT outperforms baselines across all benchmarks. Moreover, as task complexity scales, EndoCoT maintains exceptional accuracy: it achieves 90% on Maze-32 and 95% on Sudoku-35, outperforming the strongest baseline by 25% and 40% respectively. Unlike vanilla baselines that commit early and fail catastrophically, EndoCoT exhibits interpretable, step-by-step reasoning chains (see Fig. 1(c)). Our main contributions are summarized as follows:

1. To the best of our knowledge, we present an early attempt at enabling CoT-like reasoning in diffusion models via iterative latent state refinement, rather than pre-computing solutions in a single pass.
2. Through comprehensive layer-wise sensitivity and attention entropy analysis, we localize where reasoning arises in diffusion models and identify the two key bottlenecks that limit prior methods (Sec. 2).
3. EndoCoT achieves 25-40% gains over prior work on complex visual reasoning benchmarks, enables controllable inference-time scaling, and yields clearer and more controllable transformation trajectories in image editing tasks.

2 Analysis of Reasoning in Diffusion Models

To understand why current diffusion models fail at complex visual reasoning despite integrating powerful MLLMs, we first conduct a systematic empirical

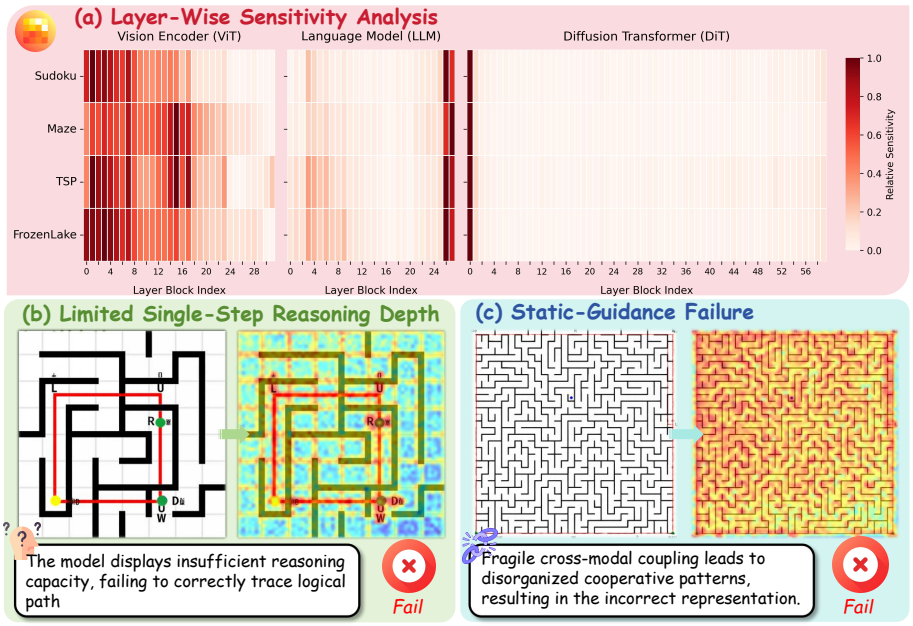

Fig. 2: (a) Layer-wise sensitivity across Vision Encoder, LLM, and DiT components (red: high sensitivity, white: low sensitivity). **(b) Limited single-step reasoning**: DiT performs spatial grounding but trajectory violates constraints. **(c) Static-guidance failure**: Dense topologies cause attention entropy to become diffuse.

analysis. These analyses directly motivate the design of our EndoCoT framework and reveal fundamental bottlenecks that limit performance. All experiments are conducted on Qwen-Image-Edit-2511 [29].

Layer-Wise Sensitivity Analysis. To investigate the internal dynamics of how reasoning capabilities emerge, we conduct a layer-wise sensitivity analysis on a representative image editing model. Specifically, we quantify the relative sensitivity by measuring the magnitude of activation responses across the distinct layer blocks of the model's three core components: the Vision Encoder, the Language Model, and the Diffusion Transformer.

As shown in Fig. 2(a), peak sensitivity is highly concentrated within the Vision Encoder and precisely at the architectural junction between the LLM (terminal layers) and the DiT (initial layers). This localized activation yields a critical insight: the heavy lifting of logical reasoning is predominantly handled by the MLLM. This observation strongly suggests that activating the MLLM during training is essential to fully unleash the model's visual reasoning potential. This finding directly informs our design choice to *jointly fine-tune* both the MLLM and DiT components in EndoCoT.

Limited Single-Step Reasoning. While the MLLM handles the bulk of logical reasoning, we observe that a single forward pass through the MLLM is insufficient

for complex tasks. As depicted in Fig.2(b), in lower-complexity scenarios (e.g., 8×8 mazes), the DiT successfully focuses on the generated trajectory. However, the generated paths systematically violate physical constraints (e.g., passing through walls). This indicates that while the DiT successfully executes spatial grounding effectively in simple tasks, the MLLM lacks the capacity to fully encode and enforce all logical constraints in a single pass. The MLLM needs to iteratively refine its understanding of the problem, rather than attempting to compute the complete solution in one shot. This motivates our design of *multi-round reasoning* in the latent space.

Static-Guidance Failure in Complex Dynamic DiT Decoding. Even if the MLLM could produce perfect reasoning in a single step, we observe a second critical limitation: the information coupling between the DiT and MLLM is fragile and static. To diagnose this, we analyze the cross-attention entropy between the generated spatial patches and the ground-truth reasoning tokens during the denoising process in Fig. 2(c). In high-complexity scenarios (e.g., 32×32 mazes), the attention entropy map becomes globally high and diffuse. This reveals a severe breakdown in the coupling between the DiT and the MLLM. When faced with dense spatial topologies, the DiT loses its ability to anchor spatial features to specific logical text tokens, causing the attention distribution to average out and resulting in a complete collapse of spatial grounding. The static, one-time injection of text embeddings is insufficient, and the conditioning signal needs to be *dynamically updated* throughout the reasoning process.

Summary. Our analysis across Fig. 2(a-c) highlights three key insights: (1) The MLLM has strong reasoning potential, but cannot fully exploit it in a single forward pass; (2) The DiT excels at spatial grounding, but requires dynamic, evolving conditioning to maintain alignment with complex logical constraints; (3) The current paradigm of static, one-time text injection is limited for multi-step reasoning tasks. This insight motivates our iterative latent reasoning mechanism, which we present in the next section.

3 Methodology

Building on our analysis in Sec. 2, we present ***Endo****genous* **C**hain-**o**f-**T**hought (**EndoCoT**), a framework that enables diffusion models to perform self-guided reasoning during generation. As shown in Fig. 3, our approach iteratively refines latent thought states in the MLLM to create a CoT reasoning process, aligns these reasoning states with explicit textual supervision, and uses progressive training to first build reasoning capability, then refine output quality.

We begin with necessary preliminaries on flow matching, then present our core methodology in three parts: (1) iterative thought guidance module, (2) terminal thought grounding, and (3) progressive training strategy. We conclude with the full inference algorithm.

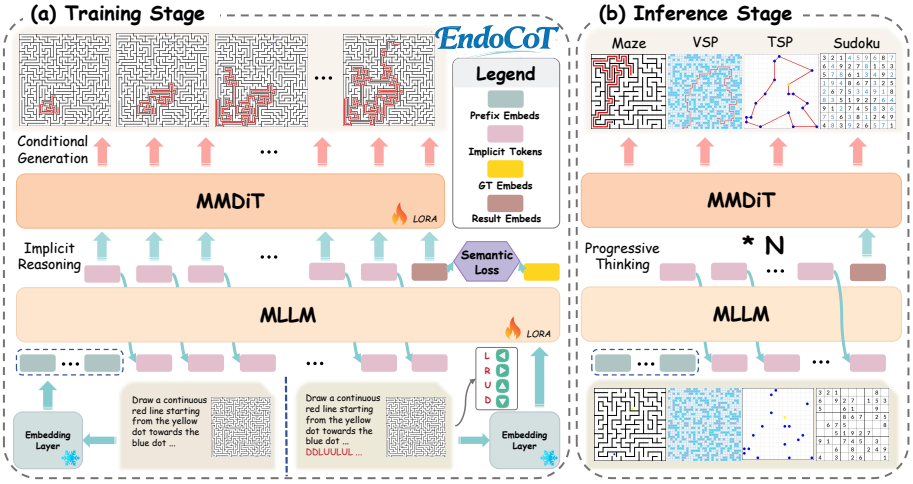

Fig. 3: Overview of EndoCoT. (a) Training: We propose a progressive two-stage training strategy: the first stage trains the model to fit both intermediate and final states at each reasoning step, capturing the full multi-step trajectory; the second stage freezes gradients on intermediate states and optimizes only the terminal state, refining generation quality while preserving learned reasoning dynamics (b) Inference: the model iteratively updates latent representations.

3.1 Preliminary

Flow Matching (FM). Flow Matching [15] provides a simplified framework for generative modeling by constructing a linear probability path between two distributions. Let $X_0 \sim \pi_0$ denote a clean data sample and $X_1 \sim \pi_1$ represent Gaussian noise from the prior. A linear trajectory X_t interpolates between X_0 and X_1 over a continuous time step $t \in [0, 1]$:

$$X_t = tX_1 + (1 - t)X_0. \tag{1}$$

By differentiating X_t with respect to time step t, we obtain the ground-truth vector field $u_t(X_t)$:

$$u_t(X_t) = \frac{dX_t}{dt} = X_1 - X_0. \tag{2}$$

A neural network $v_\theta(X_t, t, c)$ approximates this vector field, where c is an optional conditioning variable. The training objective $\mathcal{L}_{\text{FM}}$ is :

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t, X_0 \sim \pi_0, X_1 \sim \pi_1} \left\| v_\theta(X_t, t, c) - u_t(X_t) \right\|^2. \tag{3}$$

3.2 EndoCoT

Current diffusion models exhibit limited reasoning capabilities for complex, multi-step tasks. Recent analyses suggest these models possess reasoning potential

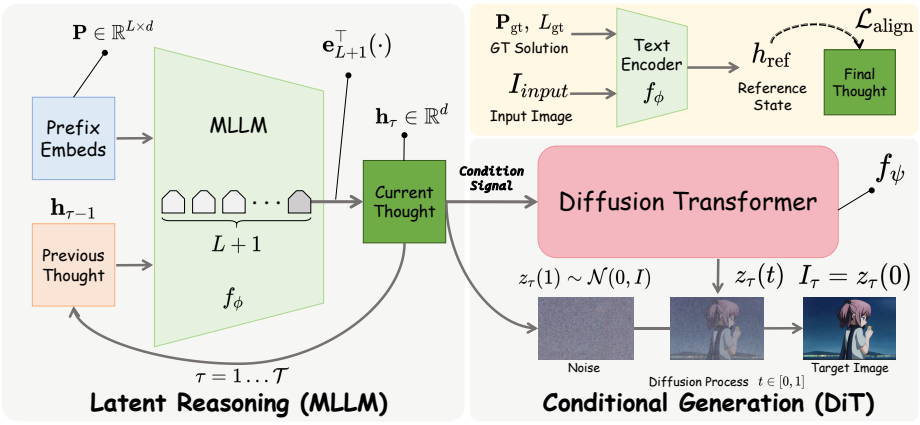

Fig. 4: Overview of notations and iterative thought guidance module. EndoCoT iteratively refines latent states $\mathbf{h}_\tau$ through the MLLM f_ϕ, then conditions the DiT f_ψ at each reasoning step τ to generate intermediate visual outputs $\mathbf{I}_\tau$.

matching their parameter scale. Architectural advances in text encoders [5, 29] provide the structural foundation to unlock this potential. We propose *Endogenous* Chain-of-Thought (**EndoCoT**), a framework enabling diffusion models to perform autonomous chain-of-thought reasoning.

Iterative Thought Guidance. We formulate the iterative thought guidance process as an iterative refinement of conditional latent states. Standard diffusion models generate an output by conditioning on static text embeddings. In contrast, our approach performs $\mathcal{T}$ reasoning steps, where each step refines both the visual output and the conditioning signal.

Intuitively, this process mimics human problem-solving: rather than attempting to generate the complete solution in one pass, we iteratively refine our understanding of the problem and proposed solution. Each reasoning step builds upon the previous one, allowing the model to explore solution spaces incrementally.

Let f_ϕ denote the MLLM (excluding embedding and projection layers) and f_ψ denote the Diffusion Transformer (DiT). A summary of the main notations is provided in Fig. 4. The reasoning process is formulated as a sequence of hidden state updates in the latent manifold $\mathbb{R}^d$. $\mathbf{P} \in \mathbb{R}^{L\times d}$ represents the prefix embeddings obtained by encoding the textual prompt and input image through their respective embedding layers. Given fixed prefix embeddings $\mathbf{P}$ (where L is the sequence length of the textual prompt), the τ-th reasoning step ($\tau \in \{1, \dots, \mathcal{T}\}$) updates the thought state $\mathbf{h}_\tau \in \mathbb{R}^d$ recursively:

$$\mathbf{h}_\tau = \mathbf{e}_{L+1}^{\top} f_\phi\left([\mathbf{P}; \mathbf{h}_{\tau-1}]\right), \quad \tau = 1, \dots, \mathcal{T}, \tag{4}$$

where $[\cdot;\cdot]$ denotes concatenation along the sequence dimension, and $\mathbf{e}_{L+1}$ is the one-hot basis vector used to extract the hidden state at the $(L+1)$-th sequence

position. Crucially, $\mathbf{h}_{\tau-1}$ directly serves as a high-dimensional input to the first layer of f_ϕ, bypassing the discrete embedding lookup table.

Conditional Flow Generation. Note that our reasoning steps $\mathcal{T}$ are distinct from the diffusion model's internal denoising timesteps. Each reasoning step τ involves a complete denoising trajectory from noise to image, conditioned on the current thought state $\mathbf{h}_\tau$. At each reasoning step τ, we condition the diffusion model on the current thought state $\mathbf{h}_\tau$ to generate an intermediate visual output $\mathbf{I}_\tau$ by solving the flow ODE:

$$\frac{d\mathbf{z}_\tau(t)}{dt} = v_\psi(\mathbf{z}_\tau(t), t, \mathbf{h}_\tau), \quad \mathbf{z}_\tau(1) \sim \mathcal{N}(\mathbf{0}, I), \quad \mathbf{I}_\tau = \mathbf{z}_\tau(0), \tag{5}$$

where $\mathbf{z}_\tau(t)$ denotes the latent state at flow time $t \in [0, 1]$. We optimize the model by supervising the generated output against a ground-truth intermediate target $\mathbf{I}^*_\tau$, using the Conditional Flow Matching loss based on Eq. 3:

$$\mathcal{L}_{\text{reasoning}} = \mathbb{E}_{\tau,t,\mathbf{z}_\tau(0),\mathbf{z}_\tau(1)} \left[\|(\mathbf{z}_\tau(0) - \mathbf{z}_\tau(1)) - v_\psi(\mathbf{z}_\tau(t), t, \mathbf{h}_\tau)\|^2 \right]. \tag{6}$$

The intermediate targets $\mathbf{I}^*_\tau$ can be obtained through sequential ground-truth decomposition (e.g., partial path segments for mazes).

Motivated by our analysis in Sec. 2, we apply LoRA fine-tuning to both f_ϕ and f_ψ. This joint adaptation enables collaborative reasoning across the conditioning and generation stages.

Terminal Thought Grounding. While the reasoning-enhanced framework enables iterative visual refinement, its reliance on purely visual supervision introduces two primary challenges: (1) a modality gap between visual targets and latent reasoning states, and (2) underutilization of explicit textual ground truth. Inspired by the success of explicit supervision in text-based latent reasoning, we introduce an auxiliary alignment objective that grounds the final reasoning state with textual supervision:

$$\mathbf{h}_{\text{ref}} = \mathbf{e}^\top_{L_{\text{gt}}+1} f_\phi \left([\mathbf{P}_{\text{gt}}, \mathbf{I}_{\text{input}}] \right), \tag{7}$$

where $\mathbf{P}_{\text{gt}} \in \mathbb{R}^{L_{\text{gt}} \times d}$ is the embedding of the ground-truth reasoning steps, and $\mathbf{I}_{\text{input}}$ denotes the task input (e.g., initial maze configuration). We align the final reasoning state $\mathbf{h}_\mathcal{T}$ with this reference using an L2 (Semantic) loss:

$$\mathcal{L}_{\text{align}} = \|\mathbf{h}_\mathcal{T} - \mathbf{h}_{\text{ref}}\|^2. \tag{8}$$

The overall training loss combines the flow matching loss with this alignment term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{FM}} + \mathbb{I}_{\{\tau=\mathcal{T}\}} \cdot \lambda_{\text{align}} \mathcal{L}_{\text{align}}, \tag{9}$$

where $\mathbb{I}_{\{\tau=\mathcal{T}\}}$ is the indicator function activating the alignment loss only at the final reasoning step, and λ_{align} balances visual generation quality and textual grounding. Empirically, we set $\lambda_{\text{align}} = 1$. As shown in our ablation study (Sec. 4.2), this term is critical for preventing drift in the latent reasoning states and ensuring the reasoning trajectory remains grounded.

Progressive Training Strategy Reasoning steps and final outputs serve distinct objectives: intermediate steps focus on exploring solution paths, while the final step emphasizes producing a correct answer. Directly optimizing both objectives simultaneously can lead to conflicting gradients. We therefore adopt a two-stage progressive training strategy.

Stage 1: Reasoning Development. As shown in Fig. 3 (a), we supervise all reasoning steps $\tau = 1, \ldots, \mathcal{T}$, enabling the model to learn step-by-step visual reasoning. The training objective is:

$$\mathcal{L}_{\text{stage1}} = \sum_{\tau=1}^{\mathcal{T}} \left(\mathcal{L}_{\text{FM}}^{\tau} + \mathbb{I}_{\{\tau=\mathcal{T}\}} \lambda_{\text{align}} \mathcal{L}_{\text{align}} \right), \tag{10}$$

where $\mathcal{L}_{\text{FM}}^{\tau}$ denotes the flow matching loss at reasoning step τ. By providing supervision at each intermediate step, we encourage the model to develop coherent, incremental reasoning trajectories.

Stage 2: Terminal Consolidation. Once the model has developed robust reasoning capabilities, we shift the training focus to solidifying the visual accuracy of the final output. As shown in Fig. 3 (b), while the intermediate reasoning process is preserved during the forward pass, gradients are computed exclusively for the final output:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{FM}}^{\mathcal{T}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}. \tag{11}$$

The intermediate steps $\tau < \mathcal{T}$ are executed in the forward pass only, serving as reasoning scaffolding without gradient propagation. To prevent the degradation of the previously learned reasoning chains, Stage 2 employs a short-cycle fine-tuning strategy with limited iterations.

Inference Process As shown in Fig. 3(b), EndoCoT does not require decoding intermediate visual states during inference. By specifying the number of reasoning steps $\mathcal{T}$, the model recursively updates its latent thought states to generate the final result.

4 Experiments

Model Setup. We build upon Qwen-Image-Edit-2511 [29] and apply LoRA fine-tuning [12] to efficiently adapt the model while preserving pre-trained knowledge. We use a LoRA rank of 32 for all targeted modules, a learning rate of 1×10^{-4}, and train for 5 epochs. Detailed hyperparameter and LoRA module configuration are provided in the Appendix.

4.1 Results on Visual Reasoning Tasks

Datasets. Following the evaluation protocol of [10], we select Maze navigation (Maze), Traveling Salesman Problem (TSP), Sudoku solving (Sudoku), Visual

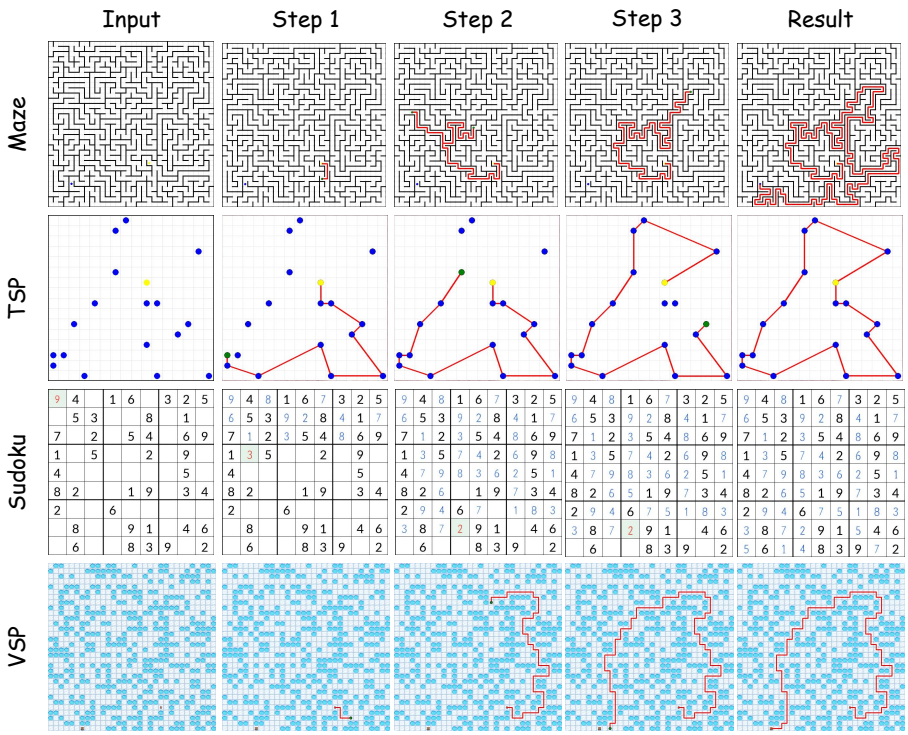

Fig. 5: Step-by-step reasoning process across four distinct tasks. Our model incrementally resolves complex visual reasoning tasks through intermediate reasoning steps. For each task (Maze, TSP, Sudoku, VSP), we show the initial input (leftmost), intermediate refinement steps, and the final optimal solution (rightmost).

Spatial Planning (VSP), and VSP-Pro with larger map sizes (up to 32×32) as evaluation benchmarks. Each task requires multi-step logical reasoning. We also follow [10] to construct the fine-tuning data of each task. All generated data are rendered at a resolution of 512×512. Detailed statistics and construction methods of fine-tuning data are provided in the Appendix.

Settings. We evaluate models under three distinct settings. *1) Zero-shot:* Vanilla baselines without any task-specific training. *2) Task-specific training:* Training and evaluation are performed on individual tasks separately (e.g., train on the Maze dataset, evaluate on Maze only), which is the default setting of DiffThinker [10]. *3) Unified training:* A single model is trained on the combined dataset of all tasks (Maze+TSP+Sudoku+VSP), and evaluated on all tasks simultaneously. The 'unified training' setting is more challenging as it requires the model to handle a wide range of problems.

Results of Task-Specific Training. As shown in Tab. 1, our model establishes a new state-of-the-art across all evaluated visual reasoning benchmarks under task-specific training. Our approach demonstrates exceptional robustness

Table 1: Evaluation results across visual reasoning tasks. Our model establishes a new state-of-the-art across both task-specific and unified training settings, demonstrating exceptional robustness to increasing task complexity.

Models	Maze			TSP			Sudoku				VSP						VSP-Super		Avg
	8	16	32	12	15	18	45	40	35	30	3	4	5	6	7	8	16	32	
▼ *Zero-Shot*																			
ThinkGen [13]	0	0	0	0	0	0	0	0	0	0	44	4	11	13	9	10	0	0	5.1
ChronoEdit [30]	1	1	26	0	0	0	0	0	0	0	60	20	12	7	16	13	1	1	8.8
Qwen3-VL-8B [32]	1	0	0	0	0	0	0	0	0	0	64	46	33	21	12	21	1	0	11.1
Qwen-Image-Edit-2511 [29]	0	0	0	0	0	0	0	0	0	0	50	55	44	16	23	23	0	0	11.7
▼ *Task-Specific Training*																			
Qwen3-VL-8B(SFT) [10]	53	37	0	59	60	43	30	17	2	0	99	96	98	96	92	86	61	8	58.56
Qwen3-VL-8B(GRPO) [10]	0	0	0	0	0	0	0	0	0	0	91	70	70	31	34	24	0	0	17.8
DiffThinker [10]	**100**	**100**	65	76	72	59	97	94	55	13	**100**	**100**	**100**	98	**100**	**100**	99	80	83.8
Ours	**100**	**100**	**90**	**77**	**77**	**73**	**100**	**100**	**95**	**64**	**100**	**100**	**100**	99	**100**	99	99	**85**	**92.1**
▼ *Unified Training*																			
DiffThinker [10]	**98**	**99**	**66**	**64**	49	34	94	45	26	37	**100**	**99**	**99**	99	**100**	97	97	**84**	77.1
Ours	97	98	52	**64**	**55**	**46**	**100**	**88**	**80**	**59**	**100**	98	98	**100**	**100**	**100**	**100**	80	**84.2**

to increasing complexity. At higher difficulty levels like Sudoku (scale 35) and Maze (scale 32), it achieves 95% and 90% accuracy respectively, significantly outperforming DiffThinker [10] (55% and 65%). On spatial planning tasks, our model maintains near-perfect scores across standard VSP levels and reaches 85% on the most challenging VSP-Super (scale 32), whereas generative baselines like ThinkGen [13] and ChronoEdit [30] fail completely. Fig. 5 shows the unique intermediate reasoning process of our model, which progressively refines solutions through multiple steps that are absent in prior approaches.

Results of Unified Training. The unified training results (single model trained on all tasks) show that our framework can learn transferable reasoning skills across tasks, though performance is slightly lower than task-specific training. As shown in Tab. 1, under unified training, our approach still achieves competitive performance compared to DiffThinker [10].

4.2 Ablation Study and Analysis

To validate our design choices, we conduct comprehensive ablation studies on the Maze benchmark, which serves as a representative testbed for long-horizon reasoning. We report both the accuracy and **path repetition rate** that measures the overlap ratio between the generated path and the ground truth path.

Effectiveness of the Semantic Loss. The auxiliary Semantic loss provides explicit textual supervision by aligning our continuous latent token representations with ground-truth reasoning steps. This constraint guides the implicit tokens toward semantically meaningful trajectories.

As shown in Tab. 2, removing this supervision signal (*w/o* semantic loss) leads to severe performance degradation. While the model retains basic routing ability on simple mazes, its logical consistency breaks down on harder instances: on the highly complex Maze-32 benchmark, accuracy drops from 90% to just

Table 2: Ablation Study on the Maze Benchmark. Removing the auxiliary semantic loss or using explicit tokens both lead to significant performance degradation compared to our default choice.

Models	Accuracy (ACC)			Path Repetition (%)		
	8	16	32	8	16	32
w/o semantic loss	39	42	14	93.44	92.23	67.24
Explicit token	34	8	0	81.47	33.35	0.08
Ours	**100**	**100**	**90**	**100.00**	**100.00**	**98.13**

Table 3: Performance comparison across different maze scales. Our method achieves near-perfect accuracy and correct path repetition rate, significantly surpassing the baselines.

Models	Accuracy			Path Repetition (%)		
	8	16	32	8	16	32
MLLM-Only	0	0	0	0.00	0.62	0.20
DiT-Only	57	43	18	89.75	90.80	80.96
Ours	**100**	**100**	**90**	**100.00**	**100.00**	**98.13**

Table 4: Inference-Time CoT Scaling on Maze Benchmarks. Comparison of Accuracy and Path Repetition Rate across different token budgets (τ). Performance consistently improves, particularly on challenging Maze-32.

Models	Accuracy			Path Repetition (%)			Time
	8	16	32	8	16	32	
DiffThinker [10]	100	97	56	100.00	100.00	73.74	15.72
Ours ($\tau = 2$)	100	72	11	100.00	89.15	45.26	16.02
Ours ($\tau = 5$)	100	94	27	100.00	98.49	63.90	16.54
Ours ($\tau = 10$)	100	99	49	100.00	99.93	82.33	17.52
Ours ($\tau = 20$)	100	100	74	100.00	100.00	96.47	19.27
Ours ($\tau = 50$)	**100**	**100**	**90**	**100.00**	**100.00**	**98.13**	24.81

14%, and path repetition rate falls from 98.13% to 67.24%. This confirms that the semantic loss acts as a crucial regularizer, preventing implicit token drift and enabling reliable long-horizon reasoning.

Implicit vs. Explicit Tokens. While continuous implicit tokens offer flexibility, one might reasonably consider explicit text generation as a more interpretable alternative for reasoning [9,21]. To investigate this design choice, we construct a variant that replaces implicit tokens with explicit autoregressive text generation, where the MLLM must produce discrete intermediate reasoning steps before generating the final visual output.

As reported in Tab. 2, explicit tokens lead to severe performance degradation. While the model achieves 34% accuracy on Maze-8, it fails on Maze-32 (0% accuracy). The failure case above reveals the cause: when planning long-horizon paths with a discrete vocabulary, the model becomes vulnerable to autoregressive error accumulation and mode collapse, degenerating into repetitive token loops.

Failure Case of Explicit Tokens

"⟨...⟩ <|im_start|>system\nDescribe the key features of the input image ⟨...⟩ Generate a new image that meets the user's requirements ⟨...⟩ Direction keys: D=Down, U=Up, R=Right, L=Left.<|im_end|>\n<|im_start|>assistant\n***Rencontre Rencontre Rencontre ⟨...⟩ Rencontre Rencontre***"

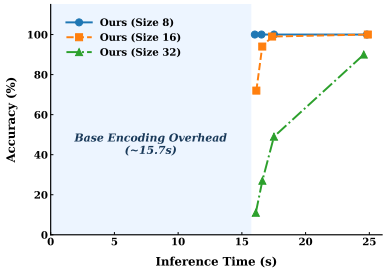

Fig. 6: The trade-off between model accuracy and overall inference time across different token budgets.

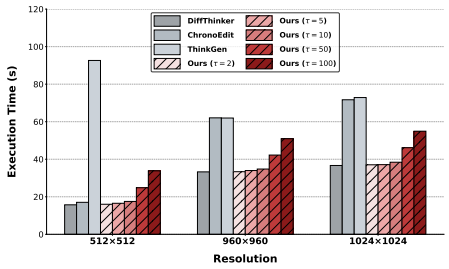

Fig. 7: Comparison of execution time against existing baselines across different resolutions and token budgets.

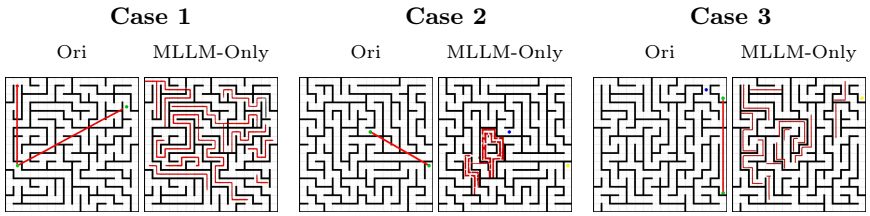

Fig. 8: Qualitative comparison of path reasoning. While MLLM-Only enables the model to understand the basic topological structure of the maze compared to the untuned baseline, it still lacks sufficient visual reasoning representations. Consequently, it exhibits erratic and wandering paths.

Inference-Time CoT Scaling. As shown in Tab. 4, our approach demonstrates a robust scaling law: dynamically increasing the number of implicit reasoning budgets τ smoothly and significantly elevates both accuracy and path repetition rate, particularly on the most challenging *Maze-32* benchmark. This indicates that our implicit tokens effectively refine complex reasoning steps through iterative exploration. Furthermore, true scalability requires predictable computational costs. As shown in Fig. 6, our model steadily trades inference time for higher accuracy.

Resolution Scaling. Beyond the token budget τ, our model demonstrates a crucial advantage in high-resolution tasks: *as spatial resolution increases, the relative computational cost of our method significantly decreases.* This efficiency gain stems fundamentally from the fact that our approach does not require repeating the computationally expensive DiT denoising steps. As shown in Fig. 7, our model maintains a stable and predictable inference latency.

Joint Training vs. MLLM-Only Baselines. Finally, we investigate the necessity of jointly optimizing both the MLLM and DiT components, rather than relying on either module in isolation. Our core premise is that spatial reasoning requires both high-level cognitive planning (from the MLLM) and low-level physical grounding (from the DiT).

Fig. 9: Progressive image editing results. Left: step-by-step object addition, where a stone lantern and a deer are sequentially introduced into the scene. **Right:** object transformation, where the deer is progressively modified into a sheep.

Quantitatively, Tab. 3 shows that decoupling components leads to severe performance degradation. Using only DiT limits the model's logical reasoning capacity, causing accuracy to drop to 18% on Maze-32. Interestingly, while we attribute the model's reasoning capability primarily to the MLLM (Section 2), relying solely on this module results in complete failure. This paradox indicates that while language models excel at abstract logic, they cannot directly map conceptual steps into spatial coordinates without the visual grounding provided by the DiT. This limitation is shown qualitatively in Fig. 8. The MLLM-only baseline produces erratic, wandering trajectories that frequently become trapped in dead ends, lacking global spatial awareness.

Results on Image Editing Tasks As shown in Fig. 9, our approach demonstrates a distinctive *progressive editing* capability driven by the internal reasoning trajectory. Given step-by-step editing instructions, the model iteratively plans and executes each modification during generation. By varying the reasoning step τ, we can control how many editing operations are carried out.

5 Related Work

Reasoning in Multimodal Large Language Models Chain-of-thought (CoT) and other test-time scaling strategies [4, 33, 36, 40] have proven effective in autoregressive large language models (LLMs). Recent works [1, 31] extend this paradigm to multimodal settings. OpenAI [17] introduces "Think with Images". Subsequent works [6, 18, 25, 27, 34] further extend this line of research to "Thinking with Video", using visual content as external evidence to support multi-step reasoning. Latent Sketchpad [35] proposes an interleaved autoregressive generation of text and visual latents. However, enabling reasoning in diffusion models (DMs) remains challenging. Since DMs typically rely on fixed-length text encoders [19, 20], their capacity to incorporate multi-step reasoning into generation is limited. Recent works [29] introduce vision-language models (VLMs) as encoders, providing richer semantic representations that facilitate reasoning-conditioned generation in diffusion models.

Reasoning in Diffusion Model Several works [13, 14] explore incorporating reasoning signals into DMs by injecting textual reasoning traces into the conditioning inputs. While those approaches enable reasoning-aware generation, the MMDiT is often treated as a conditional decoder. This over-reliance on VLMs

leads to a decoupled pipeline in which the MLLM mainly acts as a prompt enhancer. Recently, approaches [26, 30, 37] have leveraged video priors to perform complex editing, essentially treating logical state transitions as temporal sequences. However, these methods rely on the inherent smoothness of video models to "reason" through changes, which may not translate to discrete logical tasks. DiffThinker [10] takes a step toward internalizing reasoning by exploring the reasoning potential of MMDiT directly. Works like [7, 8] have attempted visual generation under the next-token autoregressive paradigm. Despite these advancements, existing frameworks still cannot perform genuine, iterative chain-of-thought reasoning within the diffusion process itself. In this work, we enable endogenous CoT reasoning through iterative latent state refinement, achieving robust logical planning and spatial grounding.

Latent Reasoning Latent reasoning has been validated in text-only domains, enabling multi-step reasoning in continuous latent spaces [9,24,38]. Latent-space reasoning compresses long reasoning chains and supports tree-structured exploration, improving inference efficiency and diversity [21, 39]. Motivated by this line of work, we train diffusion models end-to-end with latent tokens to enable test-time scaling within the diffusion process.

6 Conclusion

We presented **EndoCoT**, a novel framework that enables diffusion models to perform *endogenous* Chain-of-Thought reasoning. By iteratively refining latent thought states and grounding the final output in textual supervision, EndoCoT successfully bridges the gap between high-level logical planning and precise visual generation. Extensive experiments across diverse reasoning tasks demonstrate that this structured reasoning behavior is enabled by the synergistic coupling of the multimodal text encoder and the DiT backbone, rather than from standard denoising processes alone. While highly effective, our approach currently requires manual tuning for the optimal number of reasoning steps and relies on high-quality datasets with explicit intermediate supervision. Future work will focus on adaptive mechanisms for automatic reasoning depth control and extending the framework to broader general-purpose tasks.

Acknowledgements

This project is funded by Shanghai Artificial Intelligence Laboratory.

References

1. Agarwal, A., Sengupta, A., Chakraborty, T.: The art of scaling test-time compute for large language models. arXiv preprint arXiv:2512.02008 (2025) 14
2. Anthropic: Claude opus 4.6. Anthropic Official Website (February 2026), https://www.anthropic.com/news/claude-opus-4-6, release; Claude Opus 4.6 achieves new state-of-the-art performance in complex reasoning, coding, and long-context understanding 3
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631 2, 3
4. Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., et al.: Graph of thoughts: Solving elaborate problems with large language models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 17682–17690 (2024) 14
5. Black Forest Labs: Flux.2: Frontier visual intelligence. Black Forest Labs Official Website, https://blackforestlabs.ai/flux-2-frontier-visual-intelligence/, release; FLUX.2 provides significant improvements in visual quality, prompt adherence, and architectural efficiency 2, 7
6. Fei, H., Wu, S., Ji, W., Zhang, H., Zhang, M., Lee, M.L., Hsu, W.: Video-of-thought: Step-by-step video reasoning from perception to cognition. arXiv preprint arXiv:2501.03230 (2024) 14
7. Gao, Z., Shou, M.Z.: D-ar: Diffusion via autoregressive models. arXiv preprint arXiv:2505.23660 (2025) 15
8. Gu, J., Wang, Y., Zhang, Y., Zhang, Q., Zhang, D., Jaitly, N., Susskind, J., Zhai, S.: Dart: Denoising autoregressive transformer for scalable text-to-image generation. arXiv preprint arXiv:2410.08159 (2024) 15
9. 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) 12, 15
10. He, Z., Qu, X., Li, Y., Zhu, T., Huang, S., Cheng, Y.: Diffthinker: Towards generative multimodal reasoning with diffusion models. arXiv preprint arXiv:2512.24165 (2025) 2, 3, 9, 10, 11, 12, 15
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 1
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022) 9
13. Jiao, S., Lin, Y., Zhong, Y., She, Q., Zhou, W., Lan, X., Huang, Z., Yu, F., Yu, Y., Zhao, Y., et al.: Thinkgen: Generalized thinking for visual generation. arXiv preprint arXiv:2512.23568 (2025) 11, 14
14. Kou, S., Jin, J., Zhou, Z., Ma, Y., Wang, Y., Chen, Q., Jiang, P., Yang, X., Zhu, J., Yu, K., et al.: Think-then-generate: Reasoning-aware text-to-image diffusion with llm encoders. arXiv preprint arXiv:2601.10332 (2026) 14

15. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022) 1, 6
16. Liu, A.H., Khandelwal, K., Subramanian, S., Jouault, V., Rastogi, A., Sadé, A., Jeffares, A., Jiang, A., Cahill, A., Gavaudan, A., et al.: Ministral 3. arXiv preprint arXiv:2601.08584 (2026) 2
17. OpenAI: Thinking with images. OpenAI Official Website (April 2025), https://openai.com/index/thinking-with-images/, release; OpenAI o3 and o4-mini visual reasoning models can think with images in chain-of-thought 14
18. Qin, Y., Wei, B., Ge, J., Kallidromitis, K., Fu, S., Darrell, T., Wang, X.: Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. arXiv preprint arXiv:2511.19418 (2025) 14
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 14
20. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020) 14
21. Shen, Z., Yan, H., Zhang, L., Hu, Z., Du, Y., He, Y.: Codi: Compressing chain-of-thought into continuous space via self-distillation. arXiv preprint arXiv:2502.21074 (2025) 12, 15
22. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 1
23. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020) 1
24. Tan, W., Li, J., Ju, J., Luo, Z., Luan, J., Song, R.: Think silently, think fast: Dynamic latent compression of llm reasoning chains. arXiv preprint arXiv:2505.16552 (2025) 15
25. Tong, J., Mou, Y., Li, H., Li, M., Yang, Y., Zhang, M., Chen, Q., Liang, T., Hu, X., Zheng, Y., et al.: Thinking with video: Video generation as a promising multimodal reasoning paradigm. arXiv preprint arXiv:2511.04570 (2025) 14
26. Wang, M., Wang, R., Lin, J., Ji, R., Wiedemer, T., Gao, Q., Luo, D., Qian, Y., Huang, L., Hong, Z., Ge, J., Ma, Q., He, H., Zhou, Y., Guo, L., Mei, L., Li, J., Xing, H., Zhao, T., Yu, F., Xiao, W., Jiao, Y., Hou, J., Zhang, D., Xu, P., Zhong, B., Zhao, Z., Fang, G., Kitaoka, J., Xu, Y., Xu, H., Blacutt, K., Nguyen, T., Song, S., Sun, H., Wen, S., He, L., Wang, R., Wang, Y., Yang, M., Ma, Z., Millière, R., Shi, F., Vasconcelos, N., Khashabi, D., Yuille, A., Du, Y., Liu, Z., Li, B., Lin, D., Liu, Z., Kumar, V., Li, Y., Yang, L., Cai, Z., Deng, H.: A very big video reasoning suite (2026), https://arxiv.org/abs/2602.20159 15
27. Wang, S., Jin, J., Wang, X., Song, L., Fu, R., Wang, H., Ge, Z., Lu, Y., Cheng, X.: Video-thinker: Sparking" thinking with videos" via reinforcement learning. arXiv preprint arXiv:2510.23473 (2025) 14
28. 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) 3
29. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025) 2, 4, 7, 9, 11, 14

30. Wu, J.Z., Ren, X., Shen, T., Cao, T., He, K., Lu, Y., Gao, R., Xie, E., Lan, S., Alvarez, J.M., et al.: Chronoedit: Towards temporal reasoning for image editing and world simulation. arXiv preprint arXiv:2510.04290 (2025) 11, 15
31. Yan, Z., Li, X., He, Y., Yue, Z., Zeng, X., Wang, Y., Qiao, Y., Wang, L., Wang, Y.: Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. arXiv preprint arXiv:2509.21100 (2025) 14
32. 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) 3, 11
33. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., Narasimhan, K.: Tree of thoughts: Deliberate problem solving with large language models. Advances in neural information processing systems **36**, 11809–11822 (2023) 14
34. Zhang, H., Gu, X., Li, J., Ma, C., Bai, S., Zhang, C., Zhang, B., Zhou, Z., He, D., Tang, Y.: Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. arXiv preprint arXiv:2508.04416 (2025) 14
35. Zhang, H., Wu, W., Li, C., Shang, N., Xia, Y., Huang, Y., Zhang, Y., Dong, L., Zhang, Z., Wang, L., et al.: Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in mllms. arXiv preprint arXiv:2510.24514 (2025) 14
36. Zhang, Q., Lyu, F., Sun, Z., Wang, L., Zhang, W., Hua, W., Wu, H., Guo, Z., Wang, Y., Muennighoff, N., et al.: A survey on test-time scaling in large language models: What, how, where, and how well? arXiv preprint arXiv:2503.24235 (2025) 14
37. Zhang, Z., Chen, Z., Yang, Z., Yang, Y.: Are image-to-video models good zero-shot image editors? arXiv preprint arXiv:2511.19435 (2025) 15
38. Zhang, Z., He, X., Yan, W., Shen, A., Zhao, C., Wang, S., Shen, Y., Wang, X.E.: Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. arXiv preprint arXiv:2505.15778 (2025) 15
39. Zheng, Z., Lee, W.S.: Beyond imitation: Reinforcement learning for active latent planning (2026), https://arxiv.org/abs/2601.21598 15
40. Zheng, Z., Xie, Z., Wang, Z., Hooi, B.: Monte carlo tree search for comprehensive exploration in llm-based automatic heuristic design. arXiv preprint arXiv:2501.08603 (2025) 14