Uni-CoT: Towards Unified Chain-of-Thought Reasoning Across Text and Vision [Version 0.1]

Luozheng Qin[1,*] Jia Gong[1,*] Yuqing Sun[1,*] Tianjiao Li[3]
Mengping Yang[1] Xiaomeng Yang[1] Chao Qu[4] Zhiyu Tan[1,2,#,†] Hao Li[1,2,†]
[1]Shanghai Academy of AI for Science [2]Fudan University
[3]Nanyang Technological University [4]INFTech
{qinluozheng, gongjia, sunyuqing}@sais.com.cn

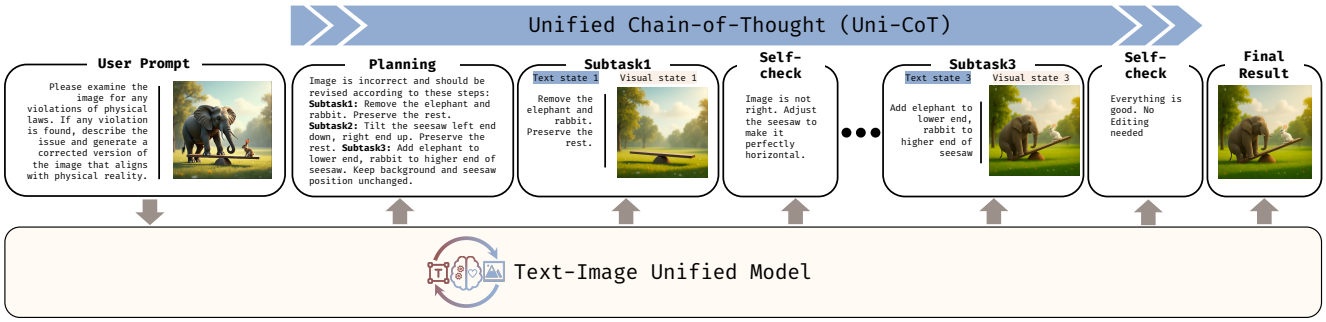

Figure 1. Interpretable multi-modal reasoning trajectory of Uni-CoT. Uni-CoT extends Chain-of-Thought (CoT) reasoning to the multi-modal domain, enabling a unified model to perform coherent, grounded, and step-by-step reasoning across text and images for complex multi-modal tasks.

Abstract

Chain-of-Thought (CoT) reasoning has been widely adopted to enhance Large Language Models (LLMs) by decomposing complex tasks into simpler, sequential subtasks. However, extending CoT to vision-language reasoning tasks remains challenging, as it often requires interpreting transitions of visual states to support reasoning. Existing methods often struggle with this due to limited capacity of modeling visual state transitions or incoherent visual trajectories caused by fragmented architectures.

To overcome these limitations, we propose Uni-CoT, a Unified Chain-of-Thought framework that enables coherent and grounded multimodal reasoning within a single unified model. The key idea is to leverage a model capable of both image understanding and generation to reason over visual content and model evolving visual states. However, empowering a unified model to achieve that is non-trivial, given the high computational cost and the burden of interleaved image-text training. To address this, Uni-CoT introduces a novel two-level reasoning paradigm: A Macro-Level CoT for high-level task planning, and A Micro-Level CoT for subtask execution. This hierarchical design significantly reduces the computational overhead. Furthermore, we introduce a structured training paradigm that combines interleaved image-text supervision for macro-level planning with multi-task auxiliary objectives for micro-level execution. Together, these innovations allow Uni-CoT to perform scalable, stable, and high-fidelity multi-modal reasoning over dynamic visual states. Furthermore, thanks to our design, all experiments can be efficiently completed using only 8 A100 GPUs with 80GB VRAM each. Experimental results on reasoning-driven image generation benchmark (WISE) and editing benchmarks (RISE and KRIS) indicates that Uni-CoT demonstrates state-of-the-art performance and strong generalization, establishing Uni-CoT as a promising solution for multi-modal reasoning. Project Page and Code: https://sais-fuxi.github.io/projects/uni-cot/

[*] Equal contribution, [#] project leader, [†] Corresponding authors.
Work in progress.

1. Introduction

Chain-of-Thought (CoT) reasoning [38] has emerged as a powerful paradigm for enhancing the performance of Large Language Models (LLMs) on complex tasks [5, 10, 23], by explicitly generating intermediate reasoning steps to enable step-by-step problem solving. Motivated by its success, recent works [24, 50, 53] have spent efforts to extend CoT to multi-modal settings, aiming to equip Multi-modal Large Language Models (MLLMs) with the capacity to handle vision-language reasoning tasks such as visual question answering [43], image editing with textual instructions [11], and embodied planning in interactive environments [25].

Prior works [9, 17, 41, 46, 49] explored utilizing reinforcement learning (RL) to enhance the text-based reasoning capabilities of MLLMs. While these approaches have shown promise, they often struggle in complex scenarios that require dynamic visual reasoning, such as geometric problems requiring diagram updates, visual puzzles demanding iterative image edits, or performing object rearrangement tasks involving spatial and causal understanding [4, 39]. A key reason of their failures possibly stems from the fundamental differences between textual and multi-modal reasoning: unlike purely text-based tasks, human multi-modal reasoning involves interpreting *dynamic visual state transitions* [1, 48], including object motion, spatial layouts, and visual causality, that naturally assist human reasoning but are inherently difficult to model through discrete symbolic representations alone.

Motivated by this, recent studies [12, 14–16, 31, 33] have explored programmatic visual manipulations, such as cropping, line plotting, and depth or segmentation estimation, to simulate visual transitions. While these methods capture certain dynamics within images, they fall short in modeling the structural changes critical for complex tasks like sequential image editing or goal-directed navigation, where visual states evolve significantly over time. To address these limitations, several approaches [13, 44, 52] adopt modular frameworks that combine MLLMs with image/video generators to achieve dynamic visual state transitions. However, the disjointed gradient flow between the reasoning and generation components in these works leads to fragmented reasoning trajectories and visually inconsistent or implausible visual state transitions, ultimately hindering the system's ability to perform coherent and grounded multi-modal reasoning.

To address these limitations, we propose **Uni-CoT**, a **Uni**fied **C**hain-**o**f-**T**hought framework, designed to support *structural visual transitions* for coherent multi-modal reasoning. Our core idea is to leverage a unified model [7], capable of both image understanding and generation, to support reasoning grounded in visual content and to model dynamic visual state transitions. This design is motivated by the intuition that using an unified model [3, 34, 35, 42] for both reasoning and generation naturally minimize the discrepancy between reasoning trajectories and visual transitions, thereby ensuring coherence in multi-modal inference. In addition, a unified architecture enables seamless implementation of end-to-end fine-tuning and reinforcement learning, facilitating higher coherence and performance ceilings for multi-modal reasoning tasks.

However, enabling the unified model to achieve multi-modal reasoning is not intuitive due to two significant challenges: (1) Computational Complexity: Unlike text-only reasoning within around 300 tokens per step, multi-modal reasoning jointly generate text and visual intermediate results, requiring around 10,000 tokens per step (see Sec. 2). This dramatic increase in sequence length significantly raises the computation and storage overhead of self-attention, limiting scalability and and even hindering the convergence of training. (2) Training Complexity: Interleaving image and text generation introduces significant optimization challenges due to mismatched learning dynamics and loss scales. Balancing cross-entropy for text with MSE for images often leads to instability, and reinforcement learning further amplifies this by requiring modality-aware rewards and careful optimization to maintain training stability.

To address computational complexity, we propose a novel two-level multi-modal reasoning framework comprising a *Macro-Level CoT* for high-level planning and a *Micro-Level CoT* for subtask execution. Specifically, the *Macro-Level CoT* performs global task planning by decomposing complex problems into simpler subtasks and aggregating their outcomes, while deliberately abstracting away execution detail of subtask. In contrast, the *Micro-Level CoT* focuses on solving individual subtasks while filtering out irrelevant information. To further simplify the reasoning process, we model micro-level reasoning as a Markov Decision Process (MDP), where each step depends only on the preceding state and current subtask instruction. With this hierarchical design, we decompose long reasoning trajectories to multiple short segments, thereby minimizing unnecessary token interactions and significantly reducing the computation complexity of self-attention layers.

Building upon this hierarchical design, the training of multi-modal reasoning can be naturally decomposed into two streamlined components: (1) Interleaved text-image supervision for the Macro-Level CoT branch, which guides the model in learning global planning strategies and final result synthesis from high-level trajectories; and (2) Multi-task auxiliary learning for the Micro-Level CoT branch, which supervises the MDP-style execution via a set of simplified objectives such as action generation, next-state prediction, and reward estimation. This decoupled training paradigm enables effective supervision at both global and local levels, facilitating efficient and scalable learning for

complex multi-modal reasoning tasks.

2. Preliminary: The Unifed Model - BAGEL

To enable unified chain-of-thought (CoT) reasoning across both image and text modalities, our framework builds upon **BAGEL** (Scala**b**le Gener**a**tive Co**g**nitive Mode**l**), a recently released open-source foundation model that supports joint vision-language understanding and generation. We briefly introduce the core elements of BAGEL in following sections.

Architecture. BAGEL adopts a unified, decoder-only transformer architecture featuring a Mixture-of-Transformer-Experts [21] design. Two experts are instantiated: one dedicated to *understanding* and the other to *generation*. Both experts operate over shared multi-modal token sequences through a unified self-attention mechanism, allowing flexible and lossless fusion across modalities without introducing task-specific bottlenecks.

Specifically, BAGEL integrates two modality-specific visual encoders:

- **Vision Transformer (ViT)** encoder for semantic-level understanding, which transforms an input image into approximately 4,900 tokens. It is initialized from a SigLIP2 [36] model and supports arbitrary aspect ratios via NaViT [6]-style positional encoding.
- **Variational Autoencoder (VAE)** from FLUX [20] for pixel-level generation, which encodes an image into a latent grid of 64×64 with 16 channels, resulting in 4,096 latent tokens.

The expert routing mechanism adopts hard gating: the understanding expert is activated for text and ViT tokens, while the generation expert exclusively processes VAE tokens. This dual-pathway design supports both fine-grained visual synthesis and high-level semantic grounding within a single, autoregressive modeling framework.

Training Objectives. BAGEL is jointly trained for multi-modal understanding and generation through two complementary loss functions:

- **Cross-Entropy Loss** for autoregressive token prediction over text tokens:

$$\mathcal{L}_{\text{CE}}^{\text{text}} = \sum_{i=1}^{C} x_i \log(\hat{x}_i), \tag{1}$$

  where x_i denotes the target token, $\hat{x}_i$ is the predicted token and C is the number of token classes, which may be textual or visual.
- **Mean Squared Error Loss.** For denoising-based generation under the Rectified Flow [22] paradigm, given a clean latent $\mathbf{x}_0$ and a Gaussian noise sample $\mathbf{x}_1$, we construct the noisy latent $\mathbf{x}_t$ via linear interpolation:

$$\mathbf{x}_t = (1-t)\cdot\mathbf{x}_0 + t\cdot\mathbf{x}_1, \quad t\in[0,1] \tag{2}$$

  The model is then trained to predict the velocity field using the Mean Squared Error (MSE) loss:

$$\mathcal{L}_{\text{MSE}}^{\text{image}} = \mathbb{E}\left[\|\mathbf{g}_\theta(\mathbf{x}_t \mid \mathbf{c}) - (\mathbf{x}_0 - \mathbf{x}_1)\|^2\right] \tag{3}$$

  Here, $\mathbf{g}_\theta(\mathbf{x}_t \mid \mathbf{c})$ denotes the velocity predicted by the model $\mathbf{g}$, conditioned on noisy latent $\mathbf{x}_t$ and context $\mathbf{c}$. Then the total training loss is a weighted combination of both objectives:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{CE}}\cdot\mathcal{L}_{\text{CE}}^{\text{text}} + \mathcal{L}_{\text{MSE}}^{\text{image}}, \tag{4}$$

  with empirically tuned weights $\lambda_{\text{CE}} = 0.25$ to ensure training stability.

Inference. At inference time, BAGEL operates autoregressively over interleaved multi-modal token sequences:

- Understanding tasks (e.g., visual question answering, multi-modal reasoning): BAGEL consumes ViT-encoded image tokens and text tokens, and outputs the next text token via standard next-token prediction.
- Generation tasks (e.g., text-to-image synthesis, image editing): BAGEL predicts VAE latent tokens conditioned on prompts or reference images and denoises them via Rectified Flow.

This unified inference mechanism enables BAGEL to handle diverse multi-modal tasks within a single autoregressive interface.

High Complexity for Multi-Modal Reasoning in BAGEL. As discussed above, while BAGEL provides a unified modeling backbone, it faces significant computational bottlenecks when applied to step-wise multi-modal reasoning. Unlike text-only Chain-of-Thought (CoT), where each reasoning step typically involves 512–1,024 tokens, multi-modal CoT often requires both image understanding and image generation within each step, introducing substantial overhead: Generating an image via VAE incurs approximately 4,096 tokens and Encoding an image via ViT for understanding introduces an additional 4,900 tokens. This results in nearly 9,000 visual tokens, in addition to around 1,000 text tokens, per reasoning step. As the number of steps increases, the overall training and inference costs become prohibitively expensive. Moreover, this token-intensive formulation imposes significant challenges on model optimization, often hindering convergence and limiting generalization, particularly in complex tasks that require long and compositional reasoning chains.

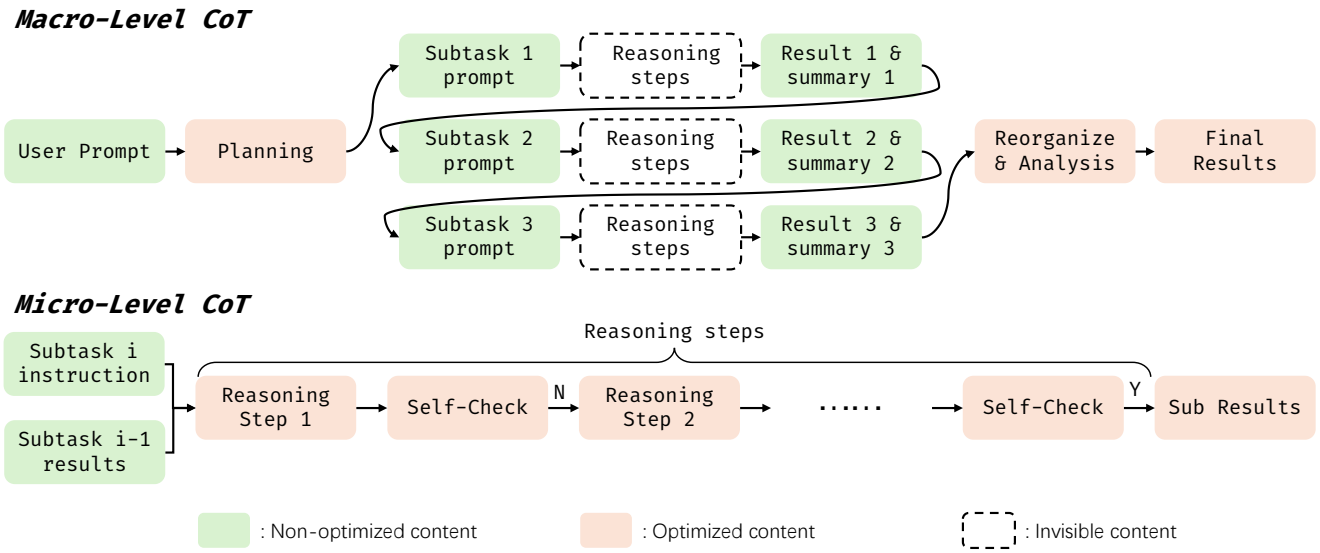

Figure 2. Overview of the Uni-CoT framework. Uni-CoT operates via two complementary branches: **Macro-Level CoT** decomposes a complex task into simpler subtasks and synthesizes their outcomes to derive the final answer, while keeping intra-subtask reasoning implicit to reduce learning and computational overhead; **Micro-Level CoT** formulates each subtask as a Markov Decision Process (MDP), where each reasoning step (a text-image pair) is a discrete node, and transitions depend solely on the previous step and current subtask prompt.

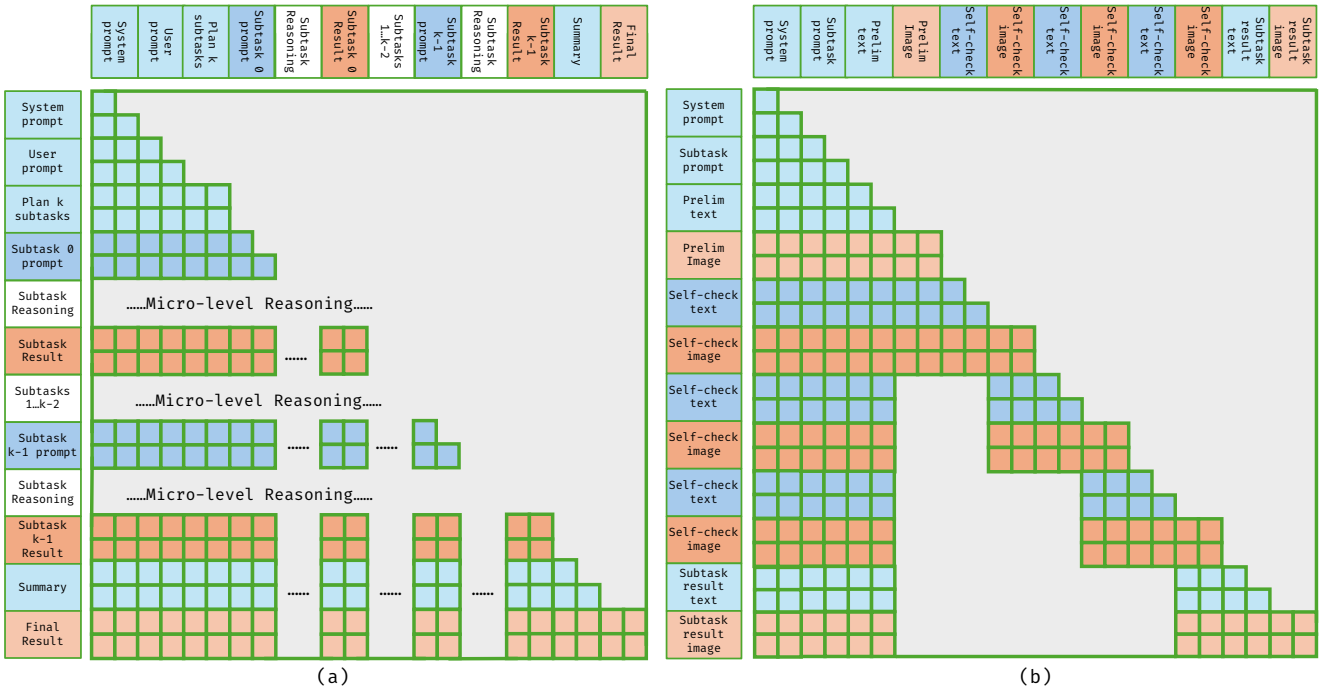

Figure 3. Uni-CoT masking strategy for hierarchical reasoning.(a) Macro masked attention scheme: Only the system prompt, high-level planning outputs, and final subtask outcomes are visible to the model. Intermediate reasoning traces—such as textual rationales or image edits produced during subtask execution—are fully masked. (b) Micro masked attention scheme: Each self-check (self-reflection) step can only attend to the immediate previous state (e.g., the last image-text pair) and the current subtask instruction.

3. Methodology

To address the inherent complexity of multi-modal reasoning, Uni-CoT adopts a hierarchical reasoning framework inspired by human cognitive strategies. Specifically, the reasoning process is divided into two levels: a **Macro-Level CoT**, which performs high-level task planning and summarization by decomposing complex problems into a sequence of simpler subtasks and aggregating their outcomes to synthesize the final result; and a **Micro-Level CoT**, which focuses on subtask execution through iterative, feedback-driven reasoning to ensure stable and reliable in-

termediate outputs. This two-tiered architecture promotes efficient reasoning, improved generalization, and enhanced interpretability across diverse multi-modal tasks.

We detail the planning strategies of Macro-Level CoT in Sec. 3.1 and the subtask execution mechanism of Micro-Level CoT in Sec. 3.2, followed by an explanation of the training procedures in Sec. 3.3.

3.1. Macro-Level CoT: Planning Strategies

Given a complex task, humans typically begin by outlining one or more abstract pathways toward the goal, rather than considering every low-level detail from the outset. Inspired by this cognitive behavior, a central function of the Macro-Level CoT is to formulate a global plan that decomposes the original task into a set of simpler, tractable subtasks.

Specifically, drawing inspiration from [8, 28, 45, 47], we propose three cognitively motivated planning mechanism:

1. **Sequential Decomposition Mechanism:** Humans often tackle complex tasks by addressing intermediate goals in a step-by-step manner. Analogously, we define a *Sequential Decomposition* strategy, wherein a task is split into a fixed sequence of subtasks. This structured progression simplifies the overall reasoning path and improves traceability.
2. **Parallel Decomposition Mechanism:** In collaborative scenarios, tasks are frequently divided into independent components that can be solved concurrently. Motivated by this, we propose a *Parallel Decomposition* strategy, which enables the model to reason over multiple subtasks in parallel, leveraging task modularity to improve efficiency.
3. **Implicit Planning via Progressive Refinement Mechanism:** In uncertain or dynamic environments, such as navigating a maze, human often forego rigid plans and instead refine their decisions iteratively. To emulate this behavior, we propose a *Progressive Refinement* strategy, wherein the model incrementally refines its plan and revise earlier steps if inconsistencies arise, thereby supporting adaptive and flexible reasoning.

These strategies empower the Macro-Level CoT to effectively address the "what-to-do" aspect of reasoning.

In addition to high-level task planning, the Macro-Level CoT is also responsible for summarizing and analyzing the outputs of individual subtasks and synthesizing them into a coherent final answer, as illustrated in Figure 2. Crucially, we treat the internal reasoning process within each subtask as implicit from the macro perspective, these fine-grained execution detail are abstracted away and handled entirely by the Micro-Level CoT. This deliberate abstraction allows the macro planner to operate from a global vantage point, focusing solely on the structural decomposition and overall reasoning trajectory without being entangled in the intricacies of subtask execution.

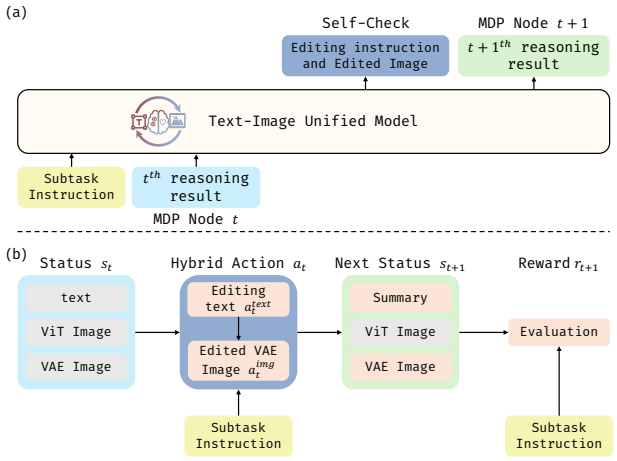

Figure 4. MDP-based reasoning architecture. (a) Overview of the sequential MDP process for multi-modal reasoning. (b) Architecture of a single MDP step $(s_t, a_t, s_{t+1}, r_{t+1})$. The transition from one state to the next is guided by the subtask instruction, with the learnable content highlighted in pink.

To enforce this abstraction during both training and inference, we introduce a macro masked-attention scheme, depicted in Figure 3 (a). This masking mechanism selectively reveals only the system prompt, macro-level planning outputs, and the final subtask outcomes. All intermediate reasoning traces, such as textual rationales and image modifications produced during micro-level subtask execution, are completely masked out. By restricting the visible context to only high-level information, this scheme encourages the model to reason at the task-structural level and prevents it from overfitting to low-level execution details. As a result, the Macro-Level CoT learns to maintain a top-down reasoning perspective, fostering modularity and scalability in multi-modal Chain-of-Thought reasoning.

3.2. Micro-Level CoT: Subtask Execution

Once a subtask is assigned by the Macro-Level planner, the Micro-Level CoT is responsible for its execution through iterative, feedback-driven reasoning. Since the overall reliability of the system depends on the consistency of each subtask's output, the core objective of the Micro-Level CoT is to ensure the generation of stable and high-quality results.

To support this, we introduce a **Self-Check** (**Self-Reflection**) mechanism to enhance both robustness and adaptability for subtask execution. Specifically, after attempting to complete a subtask, the model evaluates the quality of its output and determines whether revision is necessary. If logical inconsistencies or cross-modal mismatches are detected, the model revises its output and re-evaluates the result in a closed-loop feedback cycle. This process continues until the model deems the output satisfactory, thereby concluding the subtask execution. The self-

check mechanism ensures that generated outputs are aligned with the model's internal reasoning and multi-modal understanding, as illustrated in Figure 2.

Furthermore, to reduce the complexity of modeling the self-reflection mechanism, we formulate this reasoning process as a Markov Decision Process (MDP) [29], where each self-reflection step depends only on the result of the previous step and the given subtask instruction. As illustrated in Figure 4 (a), we define the current reasoning result (a multi-modal pair) as MDP node t, and the revised output after self-evaluation and editing as MDP node $t+1$. We assume that the transition from node t to $t+1$ is solely determined by the previous result and the fixed subtask instruction, which also aligns with human behavior during self-check. To enforce this locality, we introduce an *micro masked-attention scheme*, which restricts attention to short-range state transitions, as illustrated in Figure 3 (b). This design not only simplifies the learning objective but also enhances training stability and computational efficiency.

Formally, as illustrated in Figure 4, each element in our MDP step $(s_t, a_t, s_{t+1}, r_{t+1})$ is defined as follows:

- **State** s_t: The current reasoning state, consisting of the output from the previous step, including both textual and visual content;
- **Action** a_t: A hybrid operation that combines textual editing prompt generation with corresponding image editing;
- **Next state** s_{t+1}: The updated state, encompassing both the edited image and an aligned textual summary;
- **Reward** r_{t+1}: A textual judgment measuring alignment between the next state's result and the subtask objective.

With this formulation, the learning objectives of the Micro-Level CoT are formulated into two core competencies: (1) *subtask completion*, which involves both image editing and multi-modal understanding; and (2) *MDP modeling*, including hybrid action generation, next-state prediction, and reward estimation, highlighted in pink in Figure 4 (b).

3.3. Training Paradigm

The training of Uni-CoT is conducted in two sequential stages: Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL).

Supervised Fine-Tuning (SFT). As described in Sec. 3.1 and Sec. 3.2, our supervised fine-tuning stage comprises two complementary components: (1) interleaved multi-modal supervision for the Macro-Level CoT branch, and (2) multi-task learning for the Micro-Level CoT branch.

For the Macro-Level CoT, we follow BAGEL [7] to adopt a combined loss formulation that supervises both text and image generation. Specifically, we apply cross-entropy (CE) loss for textual output and mean squared error (MSE) loss for image generation:

$$\mathcal{L}_{\text{Macro}} = \lambda_{\text{CE}} \cdot \mathcal{L}_{\text{CE}}^{\text{text}} + \mathcal{L}_{\text{MSE}}^{\text{image}}, \tag{5}$$

Table 1. Details of auxiliary tasks for MDP modeling.

Objective	Data Structure	Loss
Text Action Generation	`[System Prompt, Subtask Prompt, Current Image and Text, "Editing Prompt"]`	Cross-Entropy Loss ($\mathcal{L}_{\text{CE}}^{\text{text}}$)
Image Action Generation	`[System Prompt, Subtask Prompt, Current Image and Text, Editing Prompt, "Edited Image"]`	MSE Loss ($\mathcal{L}_{\text{MSE}}^{\text{image}}$)
Next-State Prediction	`[System Prompt, Subtask Prompt, Edited Image, "Image Analysis"]`	Cross-Entropy Loss ($\mathcal{L}_{\text{CE}}^{\text{text}}$)
Reward Estimation	`[System Prompt, Subtask Prompt, Edited Image, Image Analysis, "Evaluation Data"]`	Cross-Entropy Loss ($\mathcal{L}_{\text{CE}}^{\text{text}}$)

where λ is a balancing coefficient that controls the relative importance of textual and visual losses.

For the Micro-Level CoT, subtask completion is supervised using interleaved multi-modal data, similarly adopting CE and MSE losses. Additionally, to support MDP-based modeling of the self-reflective reasoning process, we decompose learning into four auxiliary objectives and optimize model to learn them respectively as shown table. 1.

This multi-task supervision framework enables the Micro-Level CoT to learn both the mechanics of action execution and the reflective evaluation of outcomes within an MDP framework.

Reinforcement Learning (RL). To further enhance reasoning robustness and adaptability, we adopt reinforcement learning (RL) to optimize both Macro-Level and Micro-Level reasoning behaviors. Rewards are designed to reflect subtask completion quality, multi-modal consistency, and overall task success.

In this work, we employ a simplified yet effective RL strategy, Direct Preference Optimization (DPO) [30, 37], to align model outputs with human-preferred reasoning trajectories. We decouple the DPO training into two preference modeling stages:

- **Textual Preference Learning:** Encourages the model to prefer coherent and correct reasoning paths over suboptimal or inconsistent alternatives, using pairwise preference

annotations.

- **Visual Preference Learning:** Guides image editing behavior by optimizing toward preferred visual outputs, based on human or proxy feedback that captures semantic correctness, visual fidelity, and alignment with the instruction.

Further details on the reinforcement learning setup will be provided in a future version of this work.

4. Experiments

4.1. Implementation Details

Dataset Curation Process. Figure 6 illustrates the data pipeline we used for data curation. We collect interleaved text-image data for Macro-level and Micro-level reasoning paradigm respectively. We prepare text-to-image generation prompts from multiple datasets as seed prompts for prompt expansion.

For Macro-level reasoning data, we then enhance the prompts in the following two aspects: (1) We first check if the given prompts entail domain expertise knowledge or common-sense reasoning. If such reasoning or logical induction exist in prompts, we rewrite the prompts to explicitly state the result of reasoning deduction. For example, if the original prompt describes "a melting ice cream cone in desert sun", then the deduced rewritten prompt should elaborate on the object states under given condition, i.e. "dripping, puddle on hot sand" (2) We then enrich the generation prompts via adding auxiliary visual detail or attributes to concretely illustrate any abstract concepts or description, especially those regarding art styles, vibes, environment, etc. After prompt enhancement, we employ models such as GPT-4o or Qwen-plus to decompose the enhanced prompts into 2-3 subtasks such that complex image generation goal inferred by the enhanced prompts are broken into simpler and logically coherent sequential or parallel sub-goal. We then utilize models capable of image generation such as BAGEL-Think or GPT-4o, to perform image generation and editing following the subtask instruction. We also employ VLM models (GPT-4o) to evaluate on the results of subtask excecution and generate corresponding subtask refinement instruction. The textual detail of subtask planning, subtask evaluation and refinement, and the intermediate images generated in all subtask steps, are all collected as interleaved Macro-level data.

For Micro-level reasoning data, we directly generate preliminary images via BAGEL-Think. Next we perform several rounds of self-reflection. We repetitively evaluate on generated or edited images from the previous round using VLM models such as GPT-4o, outputting an assessment of image quality and instruction-following as well as an instruction on how should the image be refined in order to align with the intention of the original prompts better. We then employ image generation models (i.e. GPT-4o) capable of image editing to edit the images in prior rounds according to the refinement instruction. We collect the textual and visual output in each self-reflection loop as the interleaved text-image data for Micro-level reasoning SFT.

Training Details. Owing to our decomposition of interleaved image-text CoT via MDP modeling and high-quality data construction, the training of Uni-CoT is quite efficient, as all experiments can be accomplished on 8 NVIDIA A100 GPUs. Following common practices, we utilize FlashAttention, Fully Sharded Data Parallel (FSDP), and mixed-precision training for better computation efficiency. Throughout the training process, all model parameters are optimized using Adam optimizer with a constant learning rate of 2e-5. Additionally, we employ a linear warmup learning rate schedule, increasing the learning rate from zero over the first 200 steps. During each training step, we combine all data used for understanding and generation expert training, randomly sample from the combined dataset, and pack the samples into sequences with a target length of 32,768 tokens.

4.2. Experimental Setup

In this work, we primarily focus on image generation tasks within the Uni-CoT framework. Specifically, we evaluate Uni-CoT on two reasoning-centric generation tasks: reasoning-driven image generation and reasoning-driven image editing.

For the image generation task, following [7], we conduct experiments on a widely adopted benchmark, WISE [26], which are designed to assess the model's ability to generate coherent and faithful visual outputs conditioned on complex reasoning prompts. For the image editing task, we benchmark Uni-CoT on RISE [51] and KRIS [40], both of which target visual reasoning and editing under complex instructions. RISE [51] introduces the first benchmark for reasoning-informed visual editing, covering four key reasoning types: temporal, causal, spatial, and logical. In contrast, KRIS [40] is a diagnostic benchmark that categorizes editing tasks into three foundational knowledge types: factual, conceptual, and procedural.

4.3. Results for Reasoning-based Image Generation

Quantitative Results. Table 2 reports the evaluation results of Uni-CoT and other baselines on WISE, benchmarking their reasoning-based image generation ability across multiple knowledge domains. Among the evaluated models, Uni-CoT consistently achieves state-of-the-art results across all the evaluated domains, demonstrating significantly better reasoning-based image generation capability compared to open-source baselines.

Figure 5. Qualitative Results for Reliable Image Generation. Uni-CoT demonstrates impressive image generation capabilities on complex, abstract, and reasoning-intensive prompts. Notably, these results are achieved through joint image-text reasoning, where Uni-CoT iteratively evaluates the current visual state, provides textual instructions for modification, and then executes those modifications.

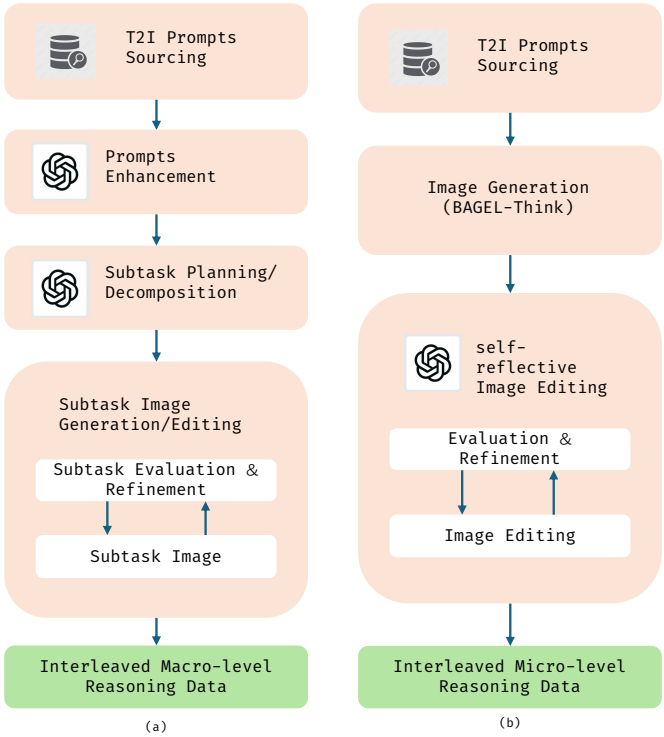

Figure 6. Data curation pipeline for hierarchical reasoning. We collect T2I prompts from various datasets, then for (a)Macro-level reasoning data pipeline: prompts are first enhanced via GPT-4o or Qwen3, then decomposed into several sequential or parallel subtasks. Then GPT-4o as well as Bagel-Think are used for subtask image generation, evaluation and refinement.(b)Micro-level reasoning data pipeline: we use Bagel-Think model to generate preliminary images based on collected T2I prompts. We then perform several rounds of self-reflection, using GPT-4o to first evaluate on current images and generate refinement instruction, then generate edited images conditioned on refinement instruction.

Table 2. Quantitative evaluation results on WISE [26]. Notably, the evaluation results are averaged over five independent runs to ensure statistical robustness. As can be drawn, Uni-CoT achieves the best performance among its open-source alternatives.

Model	Culture↑	Time↑	Space↑	Biology↑	Physics↑	Chemistry↑	Overall↑
Janus [3]	0.16	0.26	0.35	0.28	0.30	0.14	0.23
MetaQuery [27]	0.56	0.55	0.62	0.49	0.63	0.41	0.55
Bagel-Think [7]	**0.76**	0.69	0.75	0.65	0.75	0.58	0.70
Uni-CoT	**0.76**	**0.70**	**0.76**	**0.73**	**0.81**	**0.73**	**0.75**
GPT-4o [18]	*0.81*	*0.71*	*0.89*	*0.83*	*0.79*	*0.74*	*0.80*

Qualitative Analysis. As shown in Figure 5, Uni-CoT is capable of correcting semantically inaccurate generations via multi-round self-check. For prompts requires complex reasoning over commonsense and their visual appearance, the initial outputs of Uni-CoT may appear plausible but deviate from the intended semantics. However, by employing a multi-round self-check mechanism, Uni-CoT can iteratively re-evaluate and refine its generated outputs, ensuring that the final visual state in the multimodal reasoning trajectory achieves better alignment with the prompt and improved visual coherence.

4.4. Results for Reasoning-based Image Editing

Quantitative Results. We report KRIS and RISE benchmark results in Table 3 and Table 4. On KRIS benchmark, Uni-CoT outperforms all open-source baselines across perception, conceptual, and procedural categories. Remarkably, it also surpasses the closed-source Gemini 2.0 in overall score, highlighting its robust and interpretable editing capabilities under complex reasoning instructions. On RISE benchmark, Uni-CoT demonstrate comparable performance with Gemini 2.0 with respect to both overall performance on the four reasoning categories and the sub-dimension evaluation metrics of instruction reasoning, appearance consis-

Table 3. Quantitative comparisons on KRIS [40]. Uni-CoT achieves the top-1 performance among open-source models on the KRIS benchmark and even surpasses the commercial model Gemini 2.0 by 5.59 points in overall score,validating the robust and competitive image editing abilities owned by Uni-CoT.

Model	Perception				Conceptual Reasoning			Procedural Knowledge			Overall Score
	Attribute Perception	Spatial Perception	Temporal Perception	Average Score	Social Science	Natural Science	Average Score	Logical Reasoning	Instruction Decompose	Average Score	
Gemini 2.0 (Google) [19]	66.33	63.33	63.92	65.26	68.19	56.94	59.65	**54.13**	71.67	62.90	62.41
Step 3ϕ vision (StepFun) [32]	69.67	61.08	63.25	66.70	66.88	60.88	62.32	49.06	54.92	51.99	61.43
Doubao (ByteDance) [2]	70.92	59.17	40.58	63.30	65.50	61.19	62.23	47.75	60.58	54.17	60.70
BAGEL (ByteDance) [7]	64.27	62.42	42.45	60.26	55.40	56.01	55.86	52.54	50.56	51.69	56.21
BAGEL-Think (ByteDance) [7]	67.42	68.33	58.67	66.18	63.55	61.40	61.92	48.12	50.22	49.02	60.18
Uni-CoT	**72.76**	**72.87**	**67.10**	**71.85**	**70.81**	**66.00**	**67.16**	53.43	**73.93**	**63.68**	**68.00**
GPT-4o (OpenAI) [18]	*83.17*	*79.08*	*68.25*	*79.80*	*85.50*	*80.06*	*81.37*	*71.56*	*85.08*	*78.32*	*80.09*

Table 4. Quantitative comparisons on RISE [51]. Uni-CoT achieves competitive results compared to its baselines and show comparable performance with Gemini-2.0.

Model	Overall Performance (%)					Evaluation Sub-dimensions		
	Temporal	Causal	Spatial	Logical	Overall	Instruction Reasoning	Appearance Consistency	Visual Plausibility
GPT-4o (OpenAI) [18]	*34.1*	*32.2*	*37.0*	*10.6*	*28.9*	*62.8*	*80.2*	*94.9*
Gemini-2.0-Flash-exp (Google) [19]	8.2	15.5	**23.0**	**4.7**	**13.3**	48.9	68.2	82.7
Gemini-2.0-Flash-pre (Google) [19]	**10.6**	13.3	11.0	2.3	9.4	**49.9**	68.4	**84.9**
Uni-CoT	8.2	**18.9**	20.0	1.2	12.5	47.4	69.4	84.4
BAGEL-Think (ByteDance) [7]	5.9	17.8	21.0	1.2	11.9	45.9	**73.8**	80.1
BAGEL (ByteDance) [7]	2.4	5.6	14.0	1.2	6.1	36.5	53.5	73.0

tency and visual plausibility.

Qualitative Analysis. Figure 7 showcases step-by-step editing results on challenging instructions. Uni-CoT demonstrates strong fidelity to the original context while making precise visual modifications. As exhibited in this figure, our method achieves high prompt consistency, significant spatial accuracy, and plausible visual transitions—highlighting the effectiveness and interpretability of our CoT-guided editing strategy.

5. Conclusion

We present Uni-CoT, a unified Chain-of-Thought framework that enables coherent and grounded multimodal reasoning across vision and language within a single model. By introducing a two-level hierarchical reasoning architecture, comprising a Macro-Level CoT for high-level planning and a Micro-Level CoT for subtask execution modeled as an MDP, we significantly reduce computational complexity and improve reasoning efficiency. Our structured training paradigm further enables effective supervision and preference-based fine-tuning. Extensive experiments across reasoning-driven image generation and editing benchmarks demonstrate Uni-CoT's superiority in both performance and interpretability. We believe Uni-CoT offers a scalable foundation for future multi-modal reasoning systems.

6. Limitations and Path Forward

In the current implementation of the Uni-CoT framework, we have successfully integrated two core mechanisms that significantly enhance multimodal reasoning capabilities: the Sequential Decomposition Mechanism at the Macro-Level CoT and the Self-Reflection Mechanism at the Micro-Level CoT. These components have yielded notable improvements in reasoning-driven image generation and editing tasks. We will release the corresponding codebase shortly and warmly welcome the community to explore and build upon it.

Despite this progress, several challenges remain. First, we are actively developing more advanced strategies at the Macro-Level CoT, namely the Parallel Decomposition Mechanism and Implicit Planning via Progressive Refinement. While promising, these mechanisms introduce significantly more complex path and memory management, making it difficult to achieve stable training and robust generalization. This is particularly problematic in out-of-distribution scenarios, where reasoning trajectories often deviate from training-time patterns.

Second, our current focus has been on generation tasks, which are more tolerant of imprecise visual state transitions. In contrast, understanding tasks, such as adding auxiliary lines in geometric problems, require strict, step-by-step visual coherence and precise structural alignment. We have observed that directly applying preference learning in such

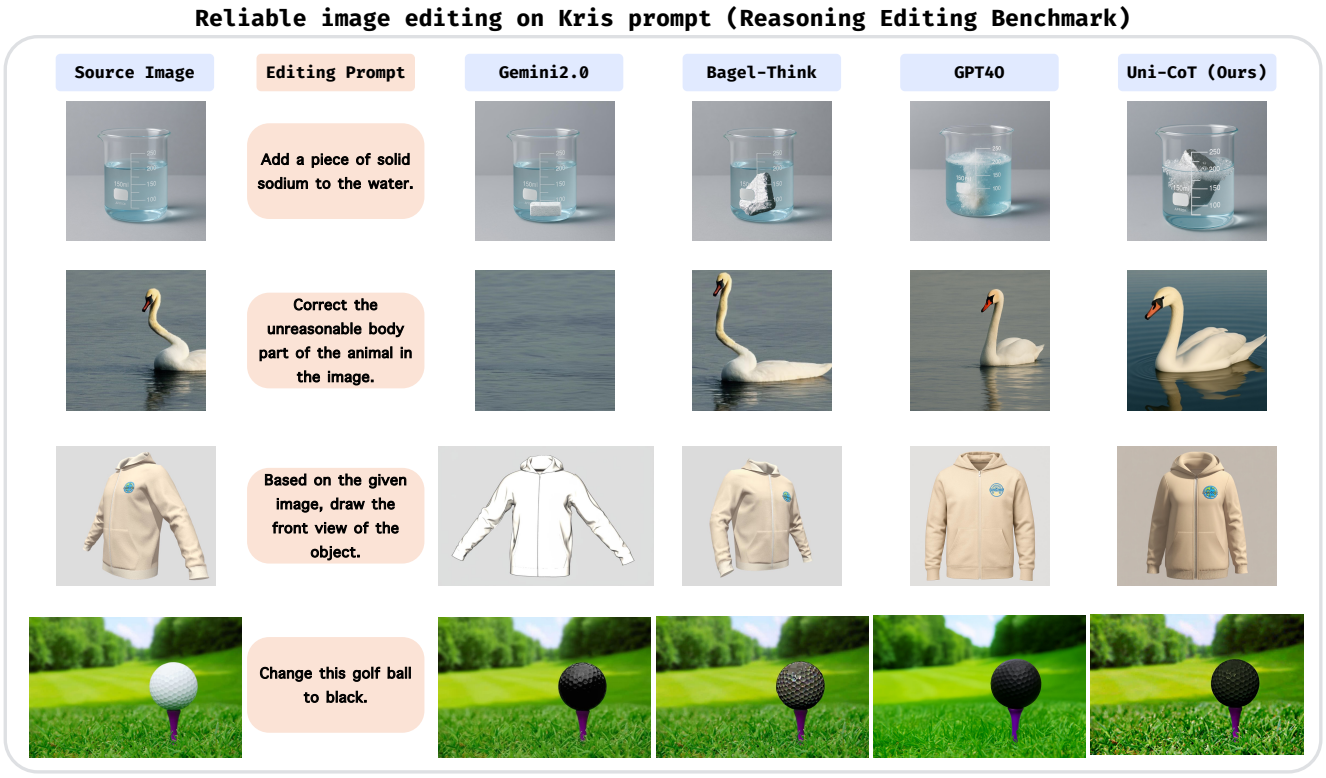

Figure 7. Qualitative Results for Reliable Image Editing. Uni-CoT demonstrates considerable image editing abilities, further supporting the effectiveness of its micro-level CoT reasoning. It can generate textual editing instructions and modify the current visual state accordingly.

cases leads to suboptimal results, largely due to its inability to enforce fine-grained visual consistency.

To address these limitations, we are exploring new strategies aimed at improving trajectory modeling, memory control, and visual transition accuracy. These include architectural improvements and learning paradigms specifically designed to support physically grounded and logically consistent visual reasoning. These enhancements are currently under development and will be featured in a future release.

We are encouraged by the progress thus far and look forward to sharing more comprehensive results and expanded capabilities in the next version of Uni-CoT.

7. Acknowledgment

We would like to thank Haoyu Pan, Zhengbo Zhang, Qi Lv, Jun Gao and Zhiheng Li for helpful discussion and feedback on this work. The computations in this research were performed using the CFFF platform of Fudan University.

References

[1] Renée Baillargeon. Infants' physical world. *Current directions in psychological science*, 13(3):89–94, 2004. 2

[2] ByteDance. Doubao: Bytedance's ai chat assistant. https://www.doubao.com/chat/, 2025. Accessed: 2025-05-08. 10

[3] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 2, 9

[4] Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23678–23686, 2025. 2

[5] Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations*. 2

[6] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023. 3

[7] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint*

arXiv:2505.14683, 2025. 2, 6, 7, 9, 10

[8] Qingxiu Dong, Li Dong, Yao Tang, Tianzhu Ye, Yutao Sun, Zhifang Sui, and Furu Wei. Reinforcement pre-training. *arXiv preprint arXiv:2506.08007*, 2025. 5

[9] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 2

[10] Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309, 2020. 2

[11] Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*, 2025. 2

[12] Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19520–19529, 2025. 2

[13] Litao Guo, Xinli Xu, Luozhou Wang, Jiantao Lin, Jinsong Zhou, Zixin Zhang, Bolan Su, and Ying-Cong Chen. Comfymind: Toward general-purpose generation via tree-based planning and reactive feedback. *arXiv preprint arXiv:2505.17908*, 2025. 2

[14] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14953–14962, 2023. 2

[15] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024.

[16] Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9590–9601, 2024. 2

[17] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 2

[18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 9, 10

[19] Kat Kampf and Nicole Brichtova. Experiment with gemini 2.0 flash native image generation, march 2025. *URL https://developers. googleblog. com/en/experiment-with-gemini-20-flash-native-image-generation/. Accessed*, pages 05–08, 2025. 10

[20] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 3

[21] Weixin Liang, LILI YU, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025. 3

[22] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*. 3

[23] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *The Eleventh International Conference on Learning Representations*. 2

[24] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024. 2

[25] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, 36:25081–25094, 2023. 2

[26] Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Kunpeng Ning, Chaoran Feng, Bin Zhu, and Li Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025. 7, 9

[27] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025. 9

[28] Du Phan, Matthew Douglas Hoffman, David Dohan, Sholto Douglas, Tuan Anh Le, Aaron Parisi, Pavel Sountsov, Charles Sutton, Sharad Vikram, and Rif A Saurous. Training chain-of-thought via latent-variable inference. *Advances in Neural Information Processing Systems*, 36:72819–72841, 2023. 5

[29] Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990. 6

[30] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 6

[31] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generaliz-

able r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025. 2

[32] stepfun. Step-3o-vision: Stepfun's ai chat assistant. https://www.stepfun.com/chats/new/, 2025. Accessed: 2025-07-24. 10

[33] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11888–11898, 2023. 2

[34] Zhiyu Tan, Hao Yang, Luozheng Qin, Jia Gong, Mengping Yang, and Hao Li. Omni-video: Democratizing unified video understanding and generation. *arXiv preprint arXiv:2507.06119*, 2025. 2

[35] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024. 2

[36] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 3

[37] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 6

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2

[39] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of llms: visualization-of-thought elicits spatial reasoning in large language models. *Advances in Neural Information Processing Systems*, 37:90277–90317, 2024. 2

[40] Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang Yu, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. Kris-bench: Benchmarking next-level intelligent image editing models. *arXiv preprint arXiv:2505.16707*, 2025. 7, 10

[41] Jiaer Xia, Yuhang Zang, Peng Gao, Yixuan Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*, 2025. 2

[42] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 2

[43] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024. 2

[44] Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18826–18836, 2025. 2

[45] Xinyu Yang, Yuwei An, Hongyi Liu, Tianqi Chen, and Beidi Chen. Multiverse: Your language models secretly decide how to parallelize and merge generation. *arXiv preprint arXiv:2506.09991*, 2025. 5

[46] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025. 2

[47] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. 5

[48] Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. Event perception: a mind-brain perspective. *Psychological bulletin*, 133(2):273, 2007. 2

[49] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025. 2

[50] Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, et al. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*. 2

[51] Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, et al. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *arXiv preprint arXiv:2504.02826*, 2025. 7, 10

[52] Zhonghan Zhao, Wenwei Zhang, Haian Huang, Kuikun Liu, Jianfei Gao, Gaoang Wang, and Kai Chen. Rig: Synergizing reasoning and imagination in end-to-end generalist policy. *arXiv preprint arXiv:2503.24388*, 2025. 2

[53] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023. 2