OctoT2I: A Self-Evolving Agentic Text-to-Image Router

Xu Jiang[1,*] Bin Chen[1,*] Gehui Li[1,*] Yule Duan[1,*] Ronggang Wang[1,2] Jian Zhang[1,2,†]
[1]School of Electronic and Computer Engineering, Peking University
[2]Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology, Shenzhen Graduate School, Peking University
{xjiang25,chenbin,ligehui921,ylduan25}@stu.pku.edu.cn zhangjian.sz@pku.edu.cn

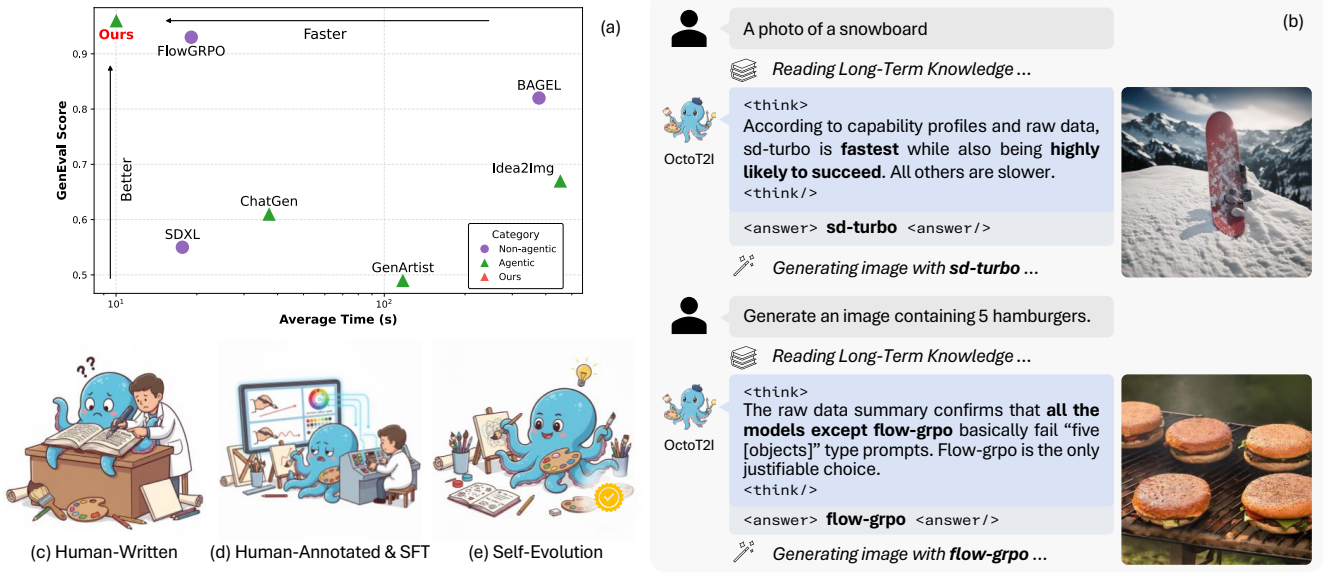

Figure 1. **The core advantages of OctoT2I in performance, decision-making, and knowledge acquisition.** **(a)** OctoT2I (Ours) achieves an exceptional performance-efficiency balance on GenEval. **(b)** This superior performance stems from its intelligent, evidence-based decisions, which route each user prompt to the most suitable T2I model. For example, it selects the most efficient tool (sd-turbo) for "a photo of a snowboard" while allocating the most justifiable tool (flow-grpo) for "Generate an image containing 5 hamburgers". **(c, d, e)** This intelligent decision-making capability is enabled by our novel **(e)** self-evolving mechanism, which overcomes the limitations of previous **(c)** handcrafted priors or **(d)** costly supervised fine-tuning (SFT) with human-annotated data.

Abstract

The explosive growth of Text-to-Image (T2I) models, from large-scale versions to lightweight, real-time ones, now faces diminishing marginal returns from single-model scaling. Agentic T2I methods emerged to alleviate this bottleneck by using multiple models. However, existing agentic T2I methods suffer from three key challenges: reliance on expensive handcrafted priors or human annotations, rigid single-path decision mechanisms, and a neglect of inference efficiency. To address these challenges, we introduce ***OctoT2I****, a novel agentic framework that reformulates the T2I task as a joint optimization of generation quality and inference efficiency. OctoT2I implements a stateful, multi-round routing strategy that adaptively selects the most suitable tool based on its knowledge and memory. This strategy is enabled by a knowledge base built from scratch by our novel Self-Evolving Mechanism. This mechanism, which requires no human supervision, first autonomously defines foundational Conceptual Dimensions (*e.g.*, style, color, count) and then intelligently explores their combinations via an itera-*

This work was supported in part by National Natural Science Foundation of China (62372016), National Science and Technology Major Project-Mobile Information Networks (2024ZD130060), Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology (2024B1212010006), Shenzhen Science and Technology Program (SYSPG20241211173440004) and Outstanding Talents Training Fund in Shenzhen.

*Equal contribution. †Corresponding author.

*tive "Propose–Solve–Evaluate–Learn" (PSEL) loop. The PSEL loop efficiently discovers each tool's capability frontier, driving continuous improvement without external guidance. Extensive experiments demonstrate that OctoT2I achieves competitive performance (**0.96**) on GenEval while delivering a **90.3%** inference speedup and a **56.6%** energy-efficiency gain over the leading baseline (Flow-GRPO), striking an exceptional balance between performance and efficiency. Code is available at* https://github.com/JaxJiang2642081986/OctoT2I.

1. Introduction

Text-to-Image (T2I) models translate textual language into corresponding images and thereby improve the efficiency of visual content creation. Architectural progress has been rapid: diffusion models [3, 4, 8, 34, 36, 51, 56, 58] now dominate, while autoregressive approaches [9, 30, 43, 52] continue to advance. In parallel, the field shows a clear split in scale. One line of work [11, 22, 24, 27, 49] focuses on building larger models to reach state-of-the-art performance, often with reinforcement learning during post-training; for example, reported parameter counts range from roughly $\sim$1B in early Stable Diffusion [36] variants and 0.6B in PixArt-α [7] to multi-billion scales such as 8B in Stable Diffusion 3 [11] and 24B in Playground v3 [26]. A second line of work [8, 24, 37, 51] emphasizes efficiency and yields lightweight models with much faster generation. Despite these gains, the marginal return from scaling a single model is diminishing, and the rapidly growing number of different models makes it difficult for non-professional users to select the most suitable tool for a given task.

To overcome the capability bottleneck of a single model, researchers have brought agents [6, 14, 23, 54, 60] into T2I. The core idea is to use an LLM as a controller that composes and schedules multiple off-the-shelf T2I models (referred to as "tools") so the system can tackle complex tasks that a single tool struggles to solve. Initial works [19, 45, 57, 60] provide preliminary evidence that these agentic methods can enhance the capability of T2I systems.

Despite early progress, the performance of agentic T2I methods remains constrained by how they acquire tool-usage knowledge. On the one hand, [5, 45] relies on human-written priors for decision making. Its performance, therefore, depends on the granularity and coverage of human-written tool descriptions, which consequently limit generalization and system performance. On the other hand, [19, 60] pursues better generalization by training LLMs on large, finely annotated datasets created by humans. This path is very expensive for both annotation and model training, exhibits limited scalability [40, 41], and is ultimately bounded by human supervision [17, 59]. Beyond knowledge, current systems are further limited by static and single-path decision mechanisms. Workflows such as [5, 19, 45] typically use only one specific T2I tool for a single round of generation. Idea2Img [57] enables multiple rounds of generation, but the generation tool used remains fixed. Finally, efficiency has been overlooked in previous work, leading to high inference latency and computational cost that harm user experience in interactive settings and hinder practical deployment and sustainability in the long term.

To this end, we reformulate agentic T2I as a joint optimization of generation quality and inference efficiency. We then introduce **OctoT2I**, a novel agentic framework designed to operationalize this new formulation. The core of this framework is a router agent capable of iterative decision-making, which relies on two key information sources: a Knowledge Module for storing long-term knowledge of tools, and a Memory Module for recording the execution history of the current task. In each decision round, the agent queries these two sources and decides which tool to execute. Subsequently, the Evaluation Module quantitatively scores the outcome, and this score is fed back into the memory as new experience to inform the next round of decisions, forming a complete "reason-act-reflect" loop.

Additionally, to alleviate the performance limitations and high costs in the knowledge acquisition of prior methods, we design a novel **Self-Evolving Mechanism** where the agent autonomously acquires knowledge from scratch. Specifically, the agent first autonomously defines a set of conceptual dimensions. Subsequently, the agent initiates an iterative "Propose–Solve–Evaluate–Learn" (PSEL) loop over conceptual dimension combinations, which is guided by an Exploration Space Pruning strategy for efficient and adaptive tool capability frontier search. In the "Propose" stage, it instantiates current dimension combination into concrete textual prompts. The agent then proceeds to the "Solve" and "Evaluate" stages by solving tasks with different tools and quantitatively assessing the generated images. Finally, in the "Learn" stage, it analyzes and summarizes the evaluation results to update its knowledge base. This data-driven process, built entirely on self-interaction, enables the agent to surpass the performance limitations of traditional, handcrafted priors.

❑ (1) We propose OctoT2I, a novel agentic router to operationalize the co-optimization of performance and efficiency. It implements a stateful, multi-round router that leverages knowledge and memory to adaptively select the most suitable tool for each prompt based on real-time feedback.

❑ (2) We introduce a self-evolving mechanism that enables the agent to acquire knowledge entirely through self-interaction, without external data or human supervision.

❑ (3) Extensive experiments show that OctoT2I achieves leading performance ($\mathbf{0.96}$) while delivering a $\mathbf{90.3}\%$ inference speedup and a $\mathbf{56.6}\%$ energy-efficiency gain over

the previous state of the art [27], striking an effective balance between performance and efficiency.

2. Related Work

2.1. Text-to-Image Generation

Recent advances in T2I generation have followed two prominent lines of development. One direction of research [11, 22, 24, 26, 27, 49] focuses on scaling up model capacity to push the performance frontier. This trend aims to improve image fidelity and the ability to handle complex textual instructions by increasing the number of parameters and the amount of training compute. The evolution of the Stable Diffusion series exemplifies this direction, with model sizes expanding from early variants [36] ($\sim$1B parameters) to the most recent Stable Diffusion 3 [11] (8B). Other large-scale models, such as FLUX [22], have further demonstrated strong compositional capabilities. Meanwhile, works like Flow-GRPO [27] and Qwen-Image [49] incorporate reinforcement learning in post-training stages to surpass the limitations of pretrained objectives. While scaling has proven effective in boosting performance, it also incurs significant computational costs, including high inference latency and substantial energy consumption.

In parallel, another direction of work [8, 24, 37, 51] emphasizes efficient T2I models that aim to reduce inference cost and latency. For example, SD-Turbo [37] compresses the multi-step diffusion process into a single step, enabling near real-time generation. SANA [8, 51] allows high-resolution image synthesis on consumer-grade GPUs. However, these lightweight approaches often trade off generation quality, especially when handling complex prompts.

The coexistence of these two lines has fostered a diverse T2I ecosystem. However, this diversity also raises a practical challenge: general users often lack the expertise to select the most appropriate model for their specific needs. This can result in suboptimal tool selection, unnecessary computational overhead, and diminished user experience. To the best of our knowledge, no prior work has attempted to integrate these two lines of work and consider the above challenges, which represents a key motivation for our work.

2.2. Agentic Systems

The rapid advancement of LLMs [1, 32] has catalyzed their transformation from mere language processors into agentic systems [6, 16, 20, 38, 50] capable of controlling how they accomplish tasks and tool usage. Pioneering works such as ToolLLM [35] and GPT4Tools [54] have demonstrated that LLMs can orchestrate external APIs and tools to solve complex tasks that extend beyond their intrinsic knowledge.

Recently, this agentic paradigm has been applied to T2I to overcome the limitations of single models. Early efforts [5, 19, 45, 57, 60] in this domain have explored several promising directions. Idea2Img [57]leverages an LLM for automated validation and prompt rewriting. DiffAgent [60] and GenArtist [45] have implemented agents for intelligent T2I model selection, while ChatGen [19] trained a model to handle conversational inputs and automatically assemble all components required for generation.

Despite the initial success of existing agentic T2I methods in managing multiple tools, their performance remains constrained due to knowledge acquisition and their static, single-path decision mechanisms. Furthermore, efficiency is overlooked, which hinders practical deployment. To solve these problems, our work introduces a self-evolving mechanism to enable the agent to autonomously construct tool knowledge, thereby surpassing the limitations of handcrafted priors. We also design a dynamic, multi-round routing framework to overcome the constraints of static decision-making. As the novel agentic T2I method to balance performance and efficiency, our approach offers significantly enhanced practical viability.

3. Methodology

3.1. Problem Formulation

To operationalize the co-optimization of generation quality and inference efficiency, we define routing as a constrained selection problem over a library of pretrained T2I models.

Setting. Let $\mathcal{P}$ be the space of prompts and $\mathcal{T} = \{\mathrm{t}_1, \ldots, \mathrm{t}_N\}$ a library of tools. Invoking t_i on a prompt $p \in \mathcal{P}$ returns an image $\mathbf{I} = \mathrm{t}_i(p)$. $\mathrm{q}(\mathbf{I}, p)$ denotes the ground-truth generation quality when t_i is run on prompt p, and $\mathrm{c}(\mathrm{t}_i)$ denotes the ground-truth computational cost of t_i.

Objective. Given a user-acceptable quality threshold θ, the ideal tool t^* is the cost-minimal tool that satisfies the quality requirement:

$$\mathrm{t}^*(p) = \arg\min_{\mathrm{t}_i \in \mathcal{T}} \ \mathrm{c}(\mathrm{t}_i) \ \text{s.t.} \ \mathrm{q}(\mathbf{I}, p) \geq \theta \,. \tag{1}$$

Challenges. While Eq. (1) defines the ideal objective, the agent lacks prior knowledge of the quality $\mathrm{q}(\cdot)$ and cost $\mathrm{c}(\cdot)$ functions for each tool. Consequently, we must estimate these quantities from prior data and execute routing decisions robustly in the presence of such estimation error.

3.2. OctoT2I System Design

System Overview. To solve the problem in Sec. 3.1, OctoT2I adopts a multi-round, dynamic routing mechanism. As shown in Fig. 2 (top), at each round r, the agent selects a tool t_r by conditioning on the user prompt p, the knowledge $\mathcal{K}$, and the prompt-specific working memory from the previous rounds, $\mathcal{M}_{r-1}$:

$$\mathrm{t}_r \ = \ \pi(p, \, \mathcal{K}, \, \mathcal{M}_{r-1}), \tag{2}$$

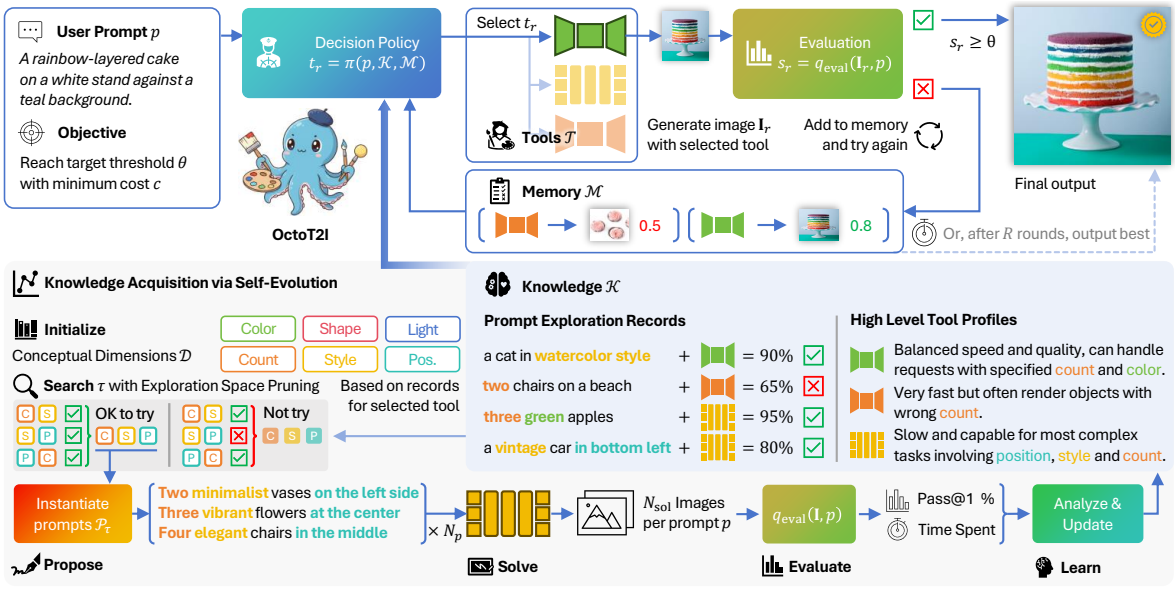

Figure 2. **The overall architecture of OctoT2I. (Top) Inference Workflow.** The agent takes a user prompt and executes a "reason-act-reflect" loop: The decision policy leverages its memory and knowledge modules to select an appropriate tool. This tool is then used to generate an image, which is subsequently evaluated to obtain a quality score. The workflow terminates when the score meets the quality threshold or the round limit is reached. **(Bottom) Self-Evolving Mechanism.** The knowledge module is autonomously and adaptively constructed from scratch. It includes hierarchical high-level tool profiles and exploration records. This construction is driven by the "Propose–Solve–Evaluate–Learn" (PSEL) loop, which is guided by the exploration space pruning strategy.

where $\pi(\cdot)$ is the decision policy. The agent then executes the chosen tool t_r on the prompt p to generate an image $\mathbf{I}_r$:

$$\mathbf{I}_r = \mathrm{t}_r(p). \tag{3}$$

Then the agent assesses the quality of $\mathbf{I}_r$, by calling the function q_{eval} to yield a score s_r:

$$s_r = \mathrm{q}_{\text{eval}}(\mathbf{I}_r, p). \tag{4}$$

The outcome of round r (*i.e.*, t_r, $\mathbf{I}_r$, s_r) is appended to the memory to form $\mathcal{M}_r$, which provides context for the next decision. The workflow terminates when the score meets the quality threshold θ or when the round limit R is reached, and the highest-scoring image is provided to the user.

Knowledge $\mathcal{K}$ serves as a long-term repository of prior knowledge about the toolset $\mathcal{T}$, providing the foundation for the reasoning process of decision policy π. This knowledge covers key performance characteristics such as each tool's core capabilities, applicable scenarios, and inference cost. Crucially, $\mathcal{K}$ is initially empty and is autonomously constructed, which is detailed in Sec. 3.3.

Memory $\mathcal{M}$ in round r, denoted $\mathcal{M}_r$, is a short-term working memory that is reinitialized for each new prompt. It consists of two components: a sequence of tuples $(\mathrm{t}_j, s_j, \mathbf{I}_j)$ recording the tool, score, and corresponding image for each step up to round r, and the best-so-far result $(\mathbf{I}_{\text{best}}, s_{\text{best}})$, which tracks the highest score encountered. The memory is defined as follows:

$$\mathcal{M}_r = \left(\{(\mathrm{t}_j, s_j, \mathbf{I}_j)\}_{j=1}^{r}, (\mathbf{I}_{\text{best}}, s_{\text{best}})\right). \tag{5}$$

The best-so-far result is updated after each evaluation as:

$$(\mathbf{I}_{\text{best}}, s_{\text{best}}) \leftarrow \begin{cases} (\mathbf{I}_r, s_r) & \text{if } s_r > s_{\text{best}} \\ (\mathbf{I}_{\text{best}}, s_{\text{best}}) & \text{otherwise} \end{cases}. \tag{6}$$

This memory, capturing both the execution history and the current optimum, provides the decision policy with the necessary information for adaptive routing.

Decision Policy π aims to solve the constrained optimization problem in Sec. 3.1. We transform this problem into an operational "filter-then-select" process. This policy is implemented by an LLM conditioned on an explicit chain-of-thought (CoT) template, $\mathbf{p}_{\text{decision}}$. At round r, this CoT prompt guides the LLM to approximate the reasoning based on the current state (*i.e.*, $p, \mathcal{K}, \mathcal{M}_{r-1}$). First, the LLM queries $\mathcal{K}$ and $\mathcal{M}_{r-1}$ to estimate the quality $\hat{\mathrm{q}}$ for each tool t_i, identifying a feasible set of tools whose estimated quality meets the threshold (*i.e.*, $\hat{\mathrm{q}}(\mathrm{t}_i(p), p) \geq \theta$). Second, the LLM retrieves the estimated cost $\hat{\mathrm{c}}$ from $\mathcal{K}$ for each valid tool and selects the one with the minimum estimated cost. This LLM-driven reasoning process can be expressed as:

$$\mathrm{t}_r = \pi(p, \mathcal{K}, \mathcal{M}_{r-1}) = \underset{\mathrm{t}_i \in \mathcal{T}, \hat{\mathrm{q}}(\mathrm{t}_i(p), p) \geq \theta}{\arg\min} \hat{\mathrm{c}}(\mathrm{t}_i). \tag{7}$$

Evaluation. As shown in Fig. 2 (top), the evaluation function q_{eval} provides a quantitative feedback score s_r on the quality of the generated image $\mathbf{I}_r$ given the prompt p, inspired by VQA-Score [25]. Specifically, we employ

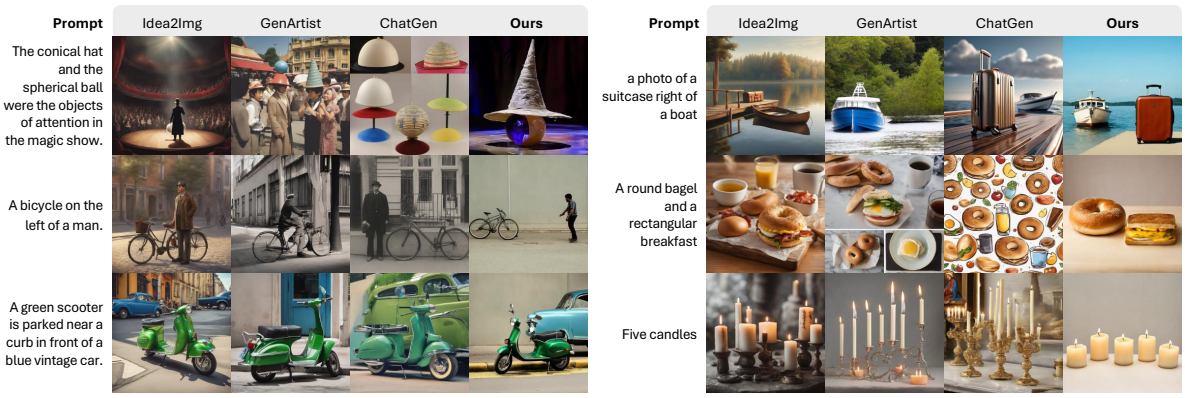

Figure 3. **Qualitative comparison** with agentic T2I methods.

Table 1. **Quantitative evaluation on GenEval benchmark.** Throughout this paper, the best and second-best results are marked in **bold red** and underlined blue, respectively. ↑ indicates higher is better. Obj.: Object; Attr.: Attribution.

Type	Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Non-agentic	SDXL-Lightning	0.98	0.69	0.36	0.85	0.09	0.18	0.52
	SDXL-Turbo	0.99	0.71	0.51	0.83	0.07	0.20	0.55
	Show-o	0.95	0.52	0.49	0.82	0.11	0.28	0.53
	JanusFlow	0.97	0.59	0.45	0.83	0.53	0.42	0.63
	Janus-Pro	0.99	0.89	0.59	0.90	0.79	0.66	0.80
	BAGEL	0.99	0.94	0.81	0.88	0.64	0.63	0.82
	FLUX.1-dev	0.99	0.85	0.74	0.79	0.21	0.48	0.68
	Flow-GRPO	0.99	**0.99**	0.88	**0.95**	0.95	**0.86**	0.93
Agentic	Idea2Img	0.98	0.81	0.68	0.93	0.25	0.41	0.67
	GenArtist	0.89	0.42	0.47	0.73	0.29	0.13	0.49
	ChatGen	0.93	0.49	0.23	0.75	0.12	0.14	0.44
	Ours	**1.00**	**0.99**	**0.95**	0.94	**1.00**	**0.86**	**0.96**

an MLLM to answer a binary "yes/no" question regarding prompt–image alignment. Instead of relying on the discrete textual response [51], we extract the raw logits for the "yes" and "no" tokens and softmax them into a continuous score s_r. This design yields a fine-grained feedback signal, providing the basis for both adaptive decision-making and the "Evaluate" stage of self-evolving mechanism in Sec. 3.3.

3.3. Knowledge Acquisition via Self-Evolution

As motivated previously, the major challenge in enabling the policy of Eq. (7) is how to acquire reliable tool knowledge $\mathcal{K}$. Rather than relying on handcrafted priors or costly human annotations, we introduce a novel **self-evolving mechanism** where the agent autonomously and adaptively constructs knowledge $\mathcal{K}$ from scratch.

Overview. As shown in Fig. 2 (bottom), the agent first initializes the exploration space by defining a foundational set of conceptual dimensions $\mathcal{D}$. Subsequently, for each tool $\mathrm{t}_i \in \mathcal{T}$, the agent initiates an iterative PSEL loop. Within this loop, the agent's exploration is guided by the exploration space pruning strategy, which leverages learned knowledge to decide which dimension combinations to explore next. By repeatedly executing this PSEL loop for each tool t_i, the agent progressively constructs the knowledge base $\mathcal{K}_i$.

Initialization. The self-evolution mechanism begins by establishing a structured basis for the vast and unstructured prompt space $\mathcal{P}$ to enable systematic exploration. Inspired by [12, 15, 47], the LLM is conditioned on the prompt $\mathbf{p}_{\text{define}}$ to generate $N_{\mathcal{D}}$ fundamental and orthogonal conceptual dimensions (*e.g.*, color, position, culture and counting), denoted as $\mathcal{D}$. These conceptual dimensions serve as building blocks to construct the complete exploration space, defined as the power set of $\mathcal{D}$ excluding the empty set:

$$\mathcal{C}_{\text{explore}} = 2^{\mathcal{D}} \setminus \{\emptyset\}. \tag{8}$$

The mechanism explores $\mathcal{C}_{\text{explore}}$ in a simple-to-complex order based on the number of conceptual dimensions in each combination τ, *i.e.*, from $|\tau| = 1$ up to $N_{\mathcal{D}}$.

Propose. Given a dimension combination $\tau \in \mathcal{C}_{\text{explore}}$, the agent acts as a task proposer and uses the prompt template $\mathbf{p}_{\text{propose}}$ to instantiate τ into N_p diverse and concrete textual prompts, denoted as $\mathcal{P}_\tau$. To ensure diversity, the agent maintains a history of previously proposed prompts

and generates only new prompts.

Solve. In the "Solve" stage, each prompt $p_\tau \in \mathcal{P}_\tau$ is executed with the current tool t_i. To ensure robustness, the agent performs N_{sol} independent runs per prompt, producing a set of candidate images $\{\mathbf{I}_{\tau,n}\}_{n=1}^{N_{\text{sol}}}$ for each p_τ.

Evaluate. In the "Evaluation" stage, the agent assesses the quality of the generated candidates $\{\mathbf{I}_{\tau,n}\}_{n=1}^{N_{\text{sol}}}$ for each exploration prompt $p_\tau \in \mathcal{P}_\tau$. For every image $I_{\tau,n}$, the evaluation function q_{eval} in Eq. (4) produces an alignment score $s_{\tau,n}$ with respect to p_τ. Considering the stochasticity of generation and our constrained-optimization objective in Eq. (1), we adopt the Pass@1 to estimate the probability that a tool's single-attempt output meets the quality threshold θ:

$$\text{Pass@1}(p_\tau, \mathrm{t}_i) = \frac{1}{N_{\text{sol}}} \sum_{n=1}^{N_{\text{sol}}} \mathbb{I}(s_{\tau,n} \geq \theta). \tag{9}$$

These evaluation results provide empirical evidence for the subsequent "Learn" stage, where the knowledge $\mathcal{K}$ is incrementally updated.

Learn. In the "Learn" stage, as shown in Fig. 2 (bottom), the agent structures the raw outcomes from the "Solve" and "Evaluate" stages into long-term knowledge along two complementary dimensions.

First, it consolidates fine-grained empirical evidence by recording explored prompts together with their Pass@1 success rates. These **prompt exploration records** provide data-driven, bottom-up references that the decision policy can reuse when encountering similar cases during inference.

Second, the agent abstracts these records into **high level tool profiles**, which summarize both the average inference cost and a semantic description of the tool's strengths and limitations. This high-level knowledge offers generalizable, top-down guidance for reasoning about novel prompts without direct precedents.

This multi-layered knowledge provides the decision policy with a comprehensive and complementary foundation. Additionally, it acts as the basis for more efficient exploration, detailed in the subsequent part.

Exploration Space Pruning. The vast size of the potential exploration space, C_{explore}, makes a brute-force search computationally infeasible. Moreover, our goal is not to serve as a static benchmark for comprehensive performance reporting, but rather as a dynamic explorer that efficiently discovers each tool's unique capability frontier.

To this end, our exploration space pruning strategy defines which tasks are worth exploring. As illustrated in Fig. 2 (bottom), it follows a recursive prerequisite principle: a complex dimension combination τ is only considered for exploration only if the agent, based on its existing knowledge, determines that the tool t_i has already mastered all simpler subtasks composing it ($\forall \tau' \subset \tau, \tau' \neq \emptyset$). This determination of whether a tool has mastered a given subtask is made by checking if the tool's historical average Pass@1 score on prompts proposed for τ' exceeds the threshold θ.

Through this mechanism, the agent dynamically constructs a personalized and dynamic exploration plan for each tool. The recursive checking process ensures that computational resources remain focused on tasks located at the capability boundary, thereby enabling the overall self-evolution to be both efficient and effective.

4. Experiment

4.1. Experimental Setting

Implementation Details. Toolset comprises five T2I models: Flow-GRPO [27], SDXL-Turbo [37], SD-Turbo [37], SANA1.5 [51], and SANA-Sprint [8]. We use NVILA-Lite-2B-Verifier [28] for evaluation module on the GenEval [12] and T2I-CompBench++ [15] and GPT-4o [18] on WISE [31]. The core controller of the agent is Qwen2-0.5B [42], obtained via policy distillation from GPT-4.1 [32]. R is 4, θ is 0.8, $N_{\mathcal{D}}$ is 7, N_p is 10, and N_{sol} is 5. All experiments are conducted on four NVIDIA RTX 3090 GPUs. More details on OctoT2I's workflow, function implementation, policy distillation and prompt templates are provided in the **Supplementary Material**.

Compared methods include SDXL-Lightning [24], SDXL-Turbo [37], Show-o [52], JanusFlow [30], Janus-Pro [9], BAGEL [10], FLUX.1-dev [22], SD1.5 [36], SD-Turbo [37], SANA Sprint [8], SDXL [34], SD3.5-M [11], DALLE 3 [2], SANA-1.5 [51], Flow-GRPO [27], Idea2Img [57], GenArtist [45] and ChatGen [19]. All compared methods use official default settings.

Evaluation Setups. Experiments are conducted on three widely-used benchmarks: GenEval [12], T2I-CompBench++ [15] and WISE [31], as well as on a user study with 30 random prompts from DiffusionDB [46]. We follow the default evaluation metrics for all benchmarks. Metrics reported as most aligned with human evaluations for T2I-CompBench++ are adopted. The calculation of CO_2e and kWh·PUE metrics follows [39].

4.2. Comparison with Other Methods

Quantitative Comparison. Tab. 1 and 2 demonstrate that our approach not only surpasses the strongest non-agentic method but also significantly outperforms all previous agentic frameworks. On GenEval, our overall score of 0.96 is substantially higher than that of other agentic methods. This advantage is further validated on T2ICompBench++, where our average score of 0.6618 leads all previous methods.

Efficiency Comparison OctoT2I delivers substantial improvements in inference efficiency, outperforming all compared agentic methods across all metrics as shown in Tab. 3. Our average inference time achieves a significant speedup over prior agentic methods (*e.g.*, approx. 45x

Table 2. **Quantitative evaluation results on T2ICompBench++.** ↑ indicates higher is better.

Method	Color	Shape	Texture	2D-Spatial	3D-Spatial	Numeracy	Non-Spatial	Complex	Average↑
Non-agentic Methods									
SD1.5	0.3591	0.3720	0.3961	0.0990	0.3139	0.4391	0.7866	0.3121	0.3847
SD-Turbo	0.5802	0.4485	0.5415	0.1345	0.3334	0.4979	0.8333	0.3506	0.4650
SANA Sprint	0.7400	0.4987	0.6278	0.3548	0.4050	0.5855	0.8867	0.3734	0.5590
SDXL	0.6111	0.4856	0.5201	0.1908	0.3558	0.5089	0.9067	0.3309	0.4887
SD3.5-M	0.7119	0.5339	0.7262	0.2868	0.3554	0.5700	0.8733	0.3418	0.5499
DALLE 3	0.7785	0.6205	0.7036	0.2865	0.3744	0.5296	0.9170	0.3771	0.5734
SANA-1.5	0.7735	0.5571	0.6744	0.3706	0.4100	0.6045	0.9233	0.3856	0.5874
Flow-GRPO	0.8119	0.6143	0.7194	0.5199	**0.4373**	0.6752	0.9133	0.3741	0.6332
Agentic Methods									
Idea2Img	0.6366	0.4596	0.5522	0.2799	0.3538	0.5240	0.8933	0.3490	0.5060
GenArtist	0.6089	0.4797	0.5241	0.1920	0.3357	0.5026	0.8733	0.3291	0.4807
ChatGen	0.4791	0.3628	0.3760	0.1562	0.2963	0.4374	0.7400	0.3157	0.3954
Ours	**0.8344**	**0.6270**	**0.7319**	**0.5272**	0.4334	**0.7508**	**0.9600**	**0.4300**	**0.6618**

Table 3. **Efficiency comparison with other competitive methods** on GenEval. ↓ indicates lower is better.

Metrics	BAGEL	Flow-GRPO	Idea2Img	ChatGen	GenArtist	Ours
Avg. Time (s)↓	377.39 (37.7×)	19.07 (1.90×)	453.22 (45.2×)	37.20 (3.71×)	117.29 (11.7×)	**10.02** (1.00×)
CO_2e (g)↓	11198.25 (20.0×)	878.72 (1.57×)	12033.28 (21.5×)	1258.47 (2.25×)	4403.95 (7.87×)	**559.50** (1.00×)
kWh · PUE↓	25.86 (20.0×)	2.02 (1.57×)	27.79 (21.6×)	2.91 (2.26×)	10.17 (7.89×)	**1.29** (1.00×)

Table 4. **Ablation on self-evolving mechanism on GenEval.**

Method	Counting	Position	Color Attri.	Overall↑
GPT Internal Knowledge	0.85	0.73	0.61	0.85
Hand-Crafted Prior	0.90	0.95	0.85	0.93
Self-Evolving Knowledge (Ours)	**0.95**	**1.00**	**0.86**	**0.96**

Table 5. **Ablation on decision policy (DP).**

Method	Color	Shape	Texture	2D-Spatial	Numeracy	Average↑
w/o DP	0.7016	0.5109	0.6342	0.3078	0.4815	0.5379
w/ DP (Ours)	**0.8344**	**0.6270**	**0.7319**	**0.5272**	**0.7508**	**0.6618**

Table 6. **Ablation on exploration space pruning (ESP) strategy.**

Method	Overall Score↑	Explored Prompts↓	Avg. Time (s)↓
w/o ESP	**0.96**	1270	6857.4
w/ ESP (Ours)	**0.96**	**370**	**2328.7**

faster than Idea2Img and 11x faster than GenArtist), which is critical for interactive use. Furthermore, our approach is the most computationally and environmentally efficient, recording the lowest carbon emissions (CO_2e) and energy consumption (kWh·PUE), leading to a more sustainable and environmentally responsible real-world deployment.

Qualitative Comparison. Fig. 3 visually demonstrates that OctoT2I, benefiting from the knowledge built by self-evolution, accurately analyzes prompts and routes them to the most suitable tool. For instance, when handling challenging spatial relationship prompts (*e.g.*, "A bicycle on the left of a man"), OctoT2I successfully generates the correct

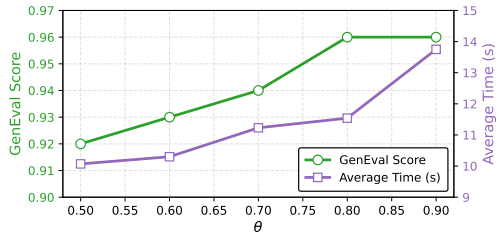

Figure 4. **Ablation study of** θ on performance (green, left axis) and inference time (purple, right axis) on the GenEval benchmark.

spatial layout. In contrast, Idea2Img, GenArtist, and ChatGen all fail to adhere to the spatial instruction. Furthermore, when faced with prompts requiring the combination of multiple unique objects (*e.g.*, "The conical hat and the spherical ball..."), OctoT2I correctly generates aesthetic images. GenArtist incorrectly merges the objects with a portrait, while ChatGen generates completely irrelevant objects.

4.3. Ablation Study

Effect of self-evolving mechanism. Tab. 4 shows a comparison of our method against two key ablation baselines. The first baseline, GPT Internal Knowledge, relies solely on the GPT-5's pretrained knowledge without access to any external knowledge base. The second baseline, Hand-Crafted Prior, initializes the knowledge base with manually designed descriptions written by research experts to mimic GenArtist [45]. Ours achieves gains of 0.11 over GPT Internal Knowledge and 0.03 over Hand-Crafted Prior, which validates that our self-evolving knowledge enables more evidence-based decision reasoning.

Effect of decision policy. Shown in Tab. 5, replacing the decision policy with a simple random tool selection policy leads to 0.23 relative performance drop on T2ICompBench++. It demonstrates that the knowledge-driven decision policy of OctoT2I is substantially more effective than an unguided multi-round trial and is the key

Table 7. **Generalization performance** on the Wise benchmark.

Method	Cultural	Time	Space	Biology	Physics	Chemistry	Average↑
SD3.5-M	0.43	0.50	0.52	0.41	0.53	0.33	0.45
SD3.5-L	0.44	0.50	0.58	0.44	0.52	0.31	0.46
BAGEL	0.44	**0.55**	**0.68**	0.44	0.60	0.39	0.52
Ours	**0.51**	**0.55**	0.67	**0.48**	**0.66**	**0.42**	**0.54**

Table 8. **Efficient New Tool Learning** on the Wise benchmark.

Method	Average↑	Δ Explored Prompts
Ours	0.54	0
Ours w/ 1 New Tool	0.61	60
Ours w/ 2 New Tools	**0.71**	130

Table 9. **Robustness to other LLMs** on the Wise benchmark.

Method	Cultural	Time	Space	Biology	Physics	Chemistry	Average↑
Ours	0.51	0.55	0.67	0.48	0.66	0.42	0.54
Ours w/ qwen	0.50	0.57	0.64	**0.53**	**0.70**	**0.48**	0.55
Ours w/ gpt5	**0.61**	**0.60**	**0.68**	0.52	0.68	0.46	**0.60**

driver of its leading performance.

Effect of Exploration Space Pruning strategy. Tab. 6 shows exploration space pruning strategy reduces the average number of explored prompts per tool by 70.9% and exploration time by 66.0% compared to the "w/o ESP" brute-force baseline, while achieving an identical overall score of 0.96. It proves our strategy enables a more efficient self-evolution that effectively discovers the capability frontiers.

Analysis of Quality Threshold θ. Fig. 4 shows that increasing the quality threshold θ improves performance at the cost of higher latency, with the most significant performance gain occurring up to $\theta = 0.8$. It is also worth noting that the parameter is adjustable, allowing users to flexibly control the trade-off based on their specific needs.

4.4. Generalization, Scalability, and Robustness

Generalization. Tab. 7 shows that OctoT2I achieves a competitive performance of 0.54, where the framework is directly adopted on the Wise benchmark. This demonstrates that OctoT2I is effective as a general-purpose framework that exhibits strong generalization capabilities.

Fast Learning and Integration of New Tools. Shown in Tab. 8, OctoT2I enables efficient, autonomous learning and seamless integration of new T2I tools—without requiring expert-designed handcrafted priors or costly training on new datasets, as demanded by prior methods [19, 45]. With just 60 exploration prompts, it incorporates Flux1.dev, lifting performance from 0.54 to 0.61. Adding gpt-image-1 further boosts the score to 0.71. This result demonstrates OctoT2I's exceptional scalability, transforming the traditionally high-cost, expert-dependent process of tool adaptation into a fully automated, low-overhead mechanism.

Robustness to Different LLMs. Tab. 9 demonstrates the LLM-agnostic robustness of OctoT2I. When the default LLM is replaced with Qwen3-VL-235B-A22B (w/ qwen) and GPT-5 (w/ gpt5), the performance improves to 0.55 and 0.60, respectively. This proves that the framework's performance gain is attributed to its universal hierarchical knowledge and routing architecture, rather than a specific LLM.

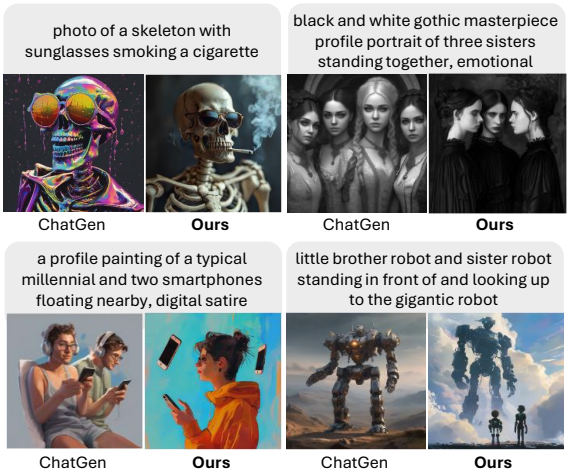

Figure 5. **Visual comparison** on real-world user prompts.

Table 10. **User study results and efficiency comparison** of two agentic methods with real user prompts.

Method	Total Votes ↑	Voting Rate (%)↑	Avg. Time(s) ↓
ChatGen	266	29.6	53.34
Ours	**634**	**70.4**	**18.45**

4.5. User Study on Real-World User Prompts

Tab. 10 shows the result of a user study involving 30 researchers and 30 real-world user prompts. Our method achieves a 70.4% voting rate, significantly outperforming ChatGen (29.6%). The visual comparison in Fig. 5 further demonstrates that OctoT2I produces better-aligned and more aesthetic visual results. Furthermore, OctoT2I (18.45s) is nearly 3× faster than ChatGen (53.34s), proving effectiveness in co-optimizing performance and efficiency.

5. Conclusion

This paper introduces **OctoT2I**, a novel agentic T2I framework that co-optimizes performance and efficiency, addressing key limitations of prior methods. We present two core innovations: (1) a **self-evolving mechanism** enabling autonomous knowledge acquisition via self-interaction, and (2) a **stateful, multi-round agentic router** that intelligently navigates the generation process using this knowledge. Experiments demonstrate that OctoT2I achieves leading performance while significantly reducing inference costs. Our work shows that self-evolving mechanisms and stateful routing represent a promising direction toward both more capable and efficient T2I systems. As future work, we plan to explore extending this T2I framework to other generative domains, such as image editing and 3D generation.

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 3

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 6

[3] Bin Chen, Gehui Li, Rongyuan Wu, Xindong Zhang, Jie Chen, Jian Zhang, and Lei Zhang. Adversarial diffusion compression for real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 28208–28220, 2025. 2

[4] Bin Chen, Weiqi Li, Shijie Zhao, Xuanyu Zhang, Junlin Li, Li Zhang, and Jian Zhang. Improved adversarial diffusion compression for real-world video super-resolution. *arXiv preprint arXiv:2603.00458*, 2026. 2

[5] Chieh-Yun Chen, Min Shi, Gong Zhang, and Humphrey Shi. T2i-copilot: A training-free multi-agent text-to-image system for enhanced prompt interpretation and interactive generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19396–19405, 2025. 2, 3

[6] Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F Karlsson, Jie Fu, and Yemin Shi. Autoagents: A framework for automatic agent generation. *arXiv preprint arXiv:2309.17288*, 2023. 2, 3

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion trans-former for photorealistic text-to-image synthesis. 2

[8] Junsong Chen, Shuchen Xue, Yuyang Zhao, Jincheng Yu, Sayak Paul, Junyu Chen, Han Cai, Song Han, and Enze Xie. Sana-sprint: One-step diffusion with continuous-time consistency distillation. *arXiv preprint arXiv:2503.09641*, 2025. 2, 3, 6

[9] 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, 6

[10] 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. 6

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2, 3, 6

[12] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023. 5, 6

[13] Rui Hao, Linmei Hu, Weijian Qi, Qingliu Wu, Yirui Zhang, and Liqiang Nie. Chatllm network: More brains, more intelligence. *arXiv preprint arXiv:2304.12998*, 2023.

[14] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023. 2

[15] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation . *IEEE Transactions on Pattern Analysis Machine Intelligence*, (01):1–17, 5555. 5, 6

[16] Qing Huang, Zhipei Xu, Xuanyu Zhang, and Jian Zhang. Unishield: An adaptive multi-agent framework for unified forgery image detection and localization. *arXiv preprint arXiv:2510.03161*, 2025. 3

[17] Edward Hughes, Michael Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul, and Tim Rocktaschel. Open-endedness is essential for artificial superhuman intelligence. *arXiv preprint arXiv:2406.04268*, 2024. 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. 6

[19] Chengyou Jia, Changliang Xia, Zhuohang Dang, Weijia Wu, Hangwei Qian, and Minnan Luo. Chatgen: Automatic text-to-image generation from freestyle chatting. *arXiv preprint arXiv:2411.17176*, 2024. 2, 3, 6, 8

[20] Xu Jiang, Gehui Li, Bin Chen, and Jian Zhang. Multi-agent image restoration. *arXiv preprint arXiv:2503.09403*, 2025. 3

[21] Sehoon Kim, Suhong Moon, Ryan Tabrizi, Nicholas Lee, Michael W Mahoney, Kurt Keutzer, and Amir Gholami. An llm compiler for parallel function calling. *arXiv preprint arXiv:2312.04511*, 2023.

[22] Black Forest Labs. Flux, 2024. 2, 3, 6

[23] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36: 51991–52008, 2023. 2

[24] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024. 2, 3, 6

[25] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. 4

[26] Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Chase Lambert, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*, 2024. 2, 3

[27] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. 2, 3, 6

[28] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. Nvila: Efficient frontier visual language models, 2024. 6

[29] Zheyuan Liu, Munan Ning, Qihui Zhang, Shuo Yang, Zhongrui Wang, Yiwei Yang, Xianzhe Xu, Yibing Song, Weihua Chen, Fan Wang, et al. Cot-lized diffusion: Let's reinforce t2i generation step-by-step. *arXiv preprint arXiv:2507.04451*, 2025.

[30] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7739–7751, 2025. 2, 6

[31] 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. 6

[32] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5):1, 2023. 3, 6

[33] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[34] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 6

[35] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023. 3

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 6

[37] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024. 2, 3, 6

[38] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[39] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13693–13696, 2020. 6

[40] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 2

[41] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*, 2024. 2

[42] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6

[43] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 2

[44] Z Wang, S Mao, W Wu, T Ge, F Wei, and H Ji. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona selfcollaboration. arxiv 2023. *arXiv preprint arXiv:2307.05300*, 2023.

[45] Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing. *Advances in Neural Information Processing Systems*, 37:128374–128395, 2024. 2, 3, 6, 7, 8

[46] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022. 6

[47] Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions?, 2025. 5

[48] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

[49] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 2, 3

[50] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023. 3

[51] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025. 2, 3, 5, 6

[52] 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, 6

[53] Hui Yang, Sifu Yue, and Yunzhong He. Auto-gpt for online decision making: Benchmarks and additional opinions. *arXiv preprint arXiv:2306.02224*, 2023.

[54] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[55] Shuzhou Yang, Xuanyu Zhang, Yinhuai Wang, Jiwen Yu, Yuhan Wang, and Jian Zhang. Diffl le: Diffusion-based domain calibration for weak supervised low-light image enhancement. *International Journal of Computer Vision*, 133 (5):2527–2546, 2025.

[56] Shuzhou Yang, Xiaoyu Li, Xiaodong Cun, Guangzhi Wang, Lingen Li, Ying Shan, and Jian Zhang. Gencompositor: Generative video compositing with diffusion transformer. In *The Fourteenth International Conference on Learning Representations*, 2026. 2

[57] Zhengyuan Yang, Jianfeng Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Idea2img: Iterative self-refinement with gpt-4v(ision) for automatic image design and generation. *arXiv preprint arXiv:2310.08541*, 2023. 2, 3, 6

[58] Jiexuan Zhang, Yiheng Du, Qian Wang, Weiqi Li, Yu Gu, and Jian Zhang. Alignedgen: Aligning style across generated images. *arXiv preprint arXiv:2509.17088*, 2025. 2

[59] Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*, 2025. 2

[60] Lirui Zhao, Yue Yang, Kaipeng Zhang, Wenqi Shao, Yuxin Zhang, Yu Qiao, Ping Luo, and Rongrong Ji. Diffagent: Fast and accurate text-to-image api selection with large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6390–6399, 2024. 2, 3

[61] Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, et al. Mindstorms in natural language-based societies of mind. *arXiv preprint arXiv:2305.17066*, 2023.