Generation Navigator: A State-Aware Agentic Framework for Image Generation

Jinming Liu[1,2,*], **Ruoyu Feng**[3,*], **Yuqi Wang**[3], **Wenjun Zeng**[2], **Xin Jin**[2,†]
[1]Shanghai Jiao Tong University [2]Eastern Institute of Technology, Ningbo, China [3]Independent

Abstract

Despite rapid advances in text-to-image generation, faithfully realizing user intent remains challenging, often requiring manual multi-turn trial and error. To automate this process, existing systems rely on either simple prompt rewriting or closed-loop agents driven by hand-crafted rules, rather than learning to adapt actions to the evolving generation process. In this paper, we reformulate image generation as a state-conditioned action-making problem and propose Generation Navigator, a multi-turn T2I agent that learns to dynamically steer the generation trajectory and output the next action. However, training this agent via reinforcement learning introduces a critical credit assignment challenge: naively rewarding a trajectory based solely on a single state assigns equal credit to all actions in the rollout, ignores the quality dynamics across turns, and fails to distinguish actions that improve the trajectory from those that degrade it or waste turns without progress. We resolve this with PRE-GRPO (Peak-Retention-Efficiency Group Relative Policy Optimization), a trajectory-level reinforcement learning objective that explicitly rewards discovering a high-quality image (Peak), avoiding subsequent quality degradation across turns (Retention), and minimizing unnecessary turns (Efficiency). Experiments show substantial improvements across benchmarks, reaching a WISE score of 0.90 and 79.06% reasoning accuracy on T2I-ReasonBench.

1 Introduction

Text-to-image (T2I) generation models have made remarkable progress [Rombach et al., 2022, Saharia et al., 2022, Esser et al., 2024]. Yet satisfying user intent remains challenging in a single pass, especially when the intended image involves spatial layouts, commonsense reasoning, fine-grained style, or ambiguous details. In practice, users address this through trial and error: generating an image, inspecting the result, editing it or restarting from scratch depending on its current quality. This suggests that such T2I generation should not be treated merely as a one-shot prompt-to-image mapping, but as a sequential action-making process: after each generation, choosing the next action given the current result.

However, existing methods usually model such T2I generation merely as a prompt rewriting or fixed-workflow task. Prompt-centric methods [Lian et al., 2024, Yang et al., 2024, Wang et al., 2025] improve initial prompts but operate only before generation, leaving no recourse even though the resulting image is partially correct or structurally flawed. Recent closed-loop or agentic systems [Jiang et al., 2026, Wan et al., 2025] introduce reviewer (e.g., multi-modal large language models) feedback or self-critique, but still rely on fixed workflows or hand-crafted editing rules rather than learning a state-conditioned action policy, which differs from real user behavior.

This raises a natural question: does dynamic state-conditioned action making actually outperform fixed workflows? To answer this, we conduct a pilot study that quantifies the benefit of dynamic action.

*Equal contribution.
†Corresponding author: Xin Jin (jinxin@eitech.edu.cn).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

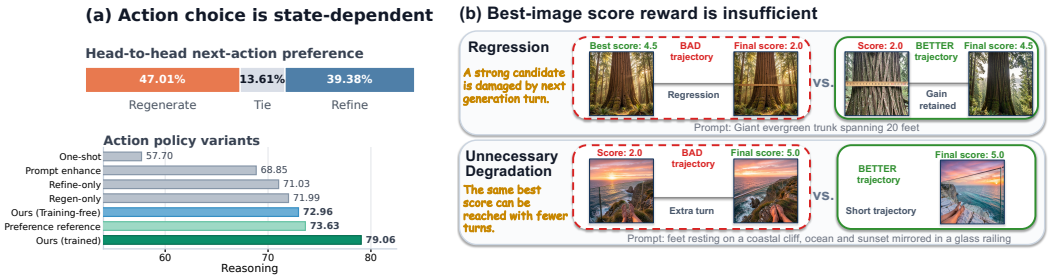


Figure 1: **The necessity of dynamic state-conditioned actions and trajectory-level rewards.** (a) Action choice is state-dependent: a preliminary study on T2I-ReasonBench [Sun et al., 2025] shows that different actions have respective advantages, and dynamic action making is superior to fixed workflows. (b) RL training relying only on a best-image-score reward struggles to distinguish regressive trajectories or unnecessary degradation.

Starting from an initial image produced by a single-pass generator, we construct an empirical action-preference reference. Mimicking real user behavior-where one either refines an acceptable draft or revises the prompt and retries—the reference at each turn executes two candidate actions, REFINE (editing the current image) and REGENERATE (producing a new image from a revised prompt), scores each outcome with the benchmark evaluator, and proceeds along the higher-scoring branch into the next turn (detailed settings are in the Appendix B). Results in Fig. 1(a) show that REGENERATE wins 47.01% of cases, REFINE wins 39.38%, and 13.61% are ties. At the system level, fixed refine-only and regenerate-only workflows improve over one-shot generation but still underperform state-conditioned action making (training-free preference reference and our method). These results suggest that **the optimal action is state-dependent and fixed workflows are suboptimal.**

To learn dynamic state-conditioned actions, reinforcement learning (RL) is a natural tool. However, it raises a second bottleneck: credit assignment. Vanilla RL methods [Shao et al., 2024] reward only a single state within the trajectory—typically the final or the best-generated image—while ignoring the quality dynamics across the full generation trajectory. More concretely, rewarding only the final image discards useful mid-trajectory signal, making learning inefficient; rewarding only the best image credits all actions equally, whether they improve, degrade, or waste turns, leading to suboptimal trajectories (Fig. 1(b)). Thus, **a good reward should capture the trajectory-level quality:** whether a high-quality image is discovered, whether quality is improved turn-by-turn, and whether the trajectory is efficient without unnecessary turns.

Building on these observations, we propose Generation Navigator, a multi-turn T2I agent that learns to make state-conditioned actions from experience. At each turn, the navigator observes the original prompt and interaction history, then outputs a structured action consisting of an action choice $d_t \in \mathcal{A} = \{\text{STOP}, \text{REFINE}, \text{REGENERATE}\}$ and a revised prompt p_t. STOP preserves a satisfactory image rather than risking unnecessary edits; REFINE edits and improves a partially correct draft in place; and REGENERATE abandons a structurally flawed image and starts over with a revised prompt. This turns multi-turn generation from a fixed workflow into a learnable agent problem.

To train this navigator, we introduce PRE-GRPO (Peak-Retention-Efficiency Group Relative Policy Optimization), which assigns credit to an entire multi-turn trajectory and decomposes the trajectory reward into three complementary terms: (i) Peak—discovering a high-quality image, (ii) Retention—avoiding subsequent quality degradation across turns, and (iii) Efficiency—reaching the peak in as few turns as possible.

Our contributions are summarized as follows:

- We analyze the limitations of fixed T2I workflows and formulate T2I generation as a state-conditioned action-making problem.
- We propose Generation Navigator, a T2I agent framework, together with a trajectory data pipeline. To optimize the agent, we further introduce PRE-GRPO, a trajectory-level RL objective that jointly reduces score regression across turns and penalizes unnecessary turns.
- Extensive experiments on demanding benchmarks like T2I-ReasonBench and WISE demonstrate that our framework substantially outperforms existing SOTA methods.

2 Related work

2.1 Text-to-image generation

Modern T2I systems built on diffusion models, rectified flow, and autoregressive multimodal generation have improved image fidelity and prompt following [Rombach et al., 2022, Saharia et al., 2022, Podell et al., 2024, Esser et al., 2024]. Recent work further integrates language reasoning and multimodal understanding into generation pipelines [Chen et al., 2025b, Deng et al., 2025, Qin et al., 2026]. Despite these advances, single-pass generation frequently fails to faithfully satisfy challenging user intents.

2.2 Prompt rewriting-based image generation

Prompt-level methods leverage LLMs to optimize user prompts before generation. Early and concurrent approaches parse prompts into structured layouts, regional sub-prompts, or richer descriptions [Lian et al., 2024, Yang et al., 2024, Kou et al., 2026], while learned rewriters such as PromptEnhancer [Wang et al., 2025] train chain-of-thought prompt expansion with image-text alignment rewards. These methods improve intent interpretation, but they mainly operate only before generation. Our method further considers the action after generation and forms a closed loop.

2.3 Workflow-driven closed-loop generation

Recent systems go beyond prompt rewriting by organizing generation as multi-turn workflows. Proactive T2I agents ask clarification questions under intent uncertainty [Hahn et al., 2025]; other agent methods coordinate rewriting, self-verification, iterative improvement, and pairwise best-candidate tracking [Wan et al., 2025, Zhou et al., 2026, Chu et al., 2025, Chen et al., 2025a]; dialogue-oriented systems such as DialogGen and Talk2Image address multi-turn user interaction, modality switching, intention drift, and dialogue-state consistency [Huang et al., 2025, Ma et al., 2025], while Agentic Retoucher, JarvisEvo, and AgentComp study specialized repair or alignment signals [Shen et al., 2026, Lin et al., 2025, Zarei et al., 2025]. However, most of these methods either adopt manually designed training-free workflows or employ fixed-workflow approaches optimized by vanilla reinforcement learning. They overlook the optimization of the T2I agent's entire trajectory. To address this issue, we propose a state-conditioned agent and utilize PRE-GRPO to enhance the trajectory-level learning capability of the T2I agent.

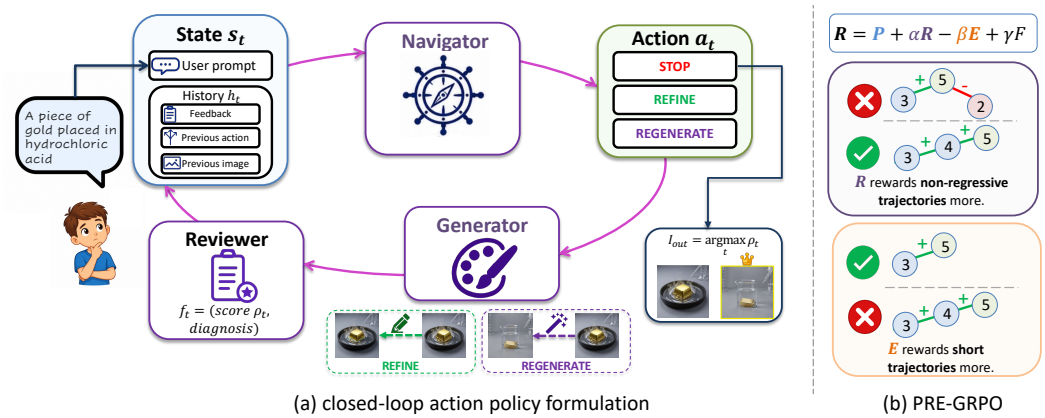
3 Generation Navigator: A State-Aware T2I Agent

As discussed in Section 1, prompt-rewriting and fixed-workflow closed-loop systems typically yield suboptimal results because they apply the same actions regardless of the current generation state. A partially correct image may benefit from targeted refinement; a structurally flawed image is better abandoned and regenerated from scratch; and a satisfactory image should be preserved rather than exposed to unnecessary edits. The core problem is therefore not merely how to rewrite the prompt, but which structured action to output given the current state. To this end, we propose Generation Navigator, a closed-loop multi-turn T2I agent that learns to make state-conditioned actions.

3.1 State-Conditioned Action Policy

To form a closed-loop pipeline, Generation Navigator consists of three components: **navigator**, **generator**, and **reviewer**, as shown in Fig. 2(a). The navigator is the only learned action-making component, implemented as a multimodal LLM. The generator and reviewer serve as environment interfaces: the generator executes a generation action (text-to-image or image-to-image editing), and the reviewer evaluates the resulting image and returns textual critique together with a scalar score.

At each turn, the navigator observes the current state and outputs a structured action $a_t = (d_t, p_t)$, where $d_t \in \mathcal{A} = \{\textsc{Stop}, \textsc{Refine}, \textsc{Regenerate}\}$ is the discrete action choice and p_t is a revised prompt tailored to that choice. The initial turn and subsequent turns serve different roles. At $t{=}1$, no image or feedback is available; the navigator acts as a prompt rewriter, transforming p_{orig} into an enhanced prompt p_1 that is sent to the generator for text-to-image generation. From $t{\geq}2$ onward, the navigator observes the full state, consisting of the original prompt p_{orig} and the interaction history



(a) closed-loop action policy formulation (b) PRE-GRPO

Figure 2: **Overview of Generation Navigator. (a)** At each turn, the navigator observes the prompt and interaction history, then outputs a structured action with one of three action choices—STOP, REFINE, or REGENERATE—along with a revised prompt. Until the turn budget is exhausted, or STOP is selected, the final output is selected as the highest-scored candidate across the entire trajectory. **(b)** PRE-GRPO decomposes the trajectory reward into (i) discovering a high-quality image (P), (ii) avoiding subsequent quality degradation across turns (R), (iii) turn efficiency (E), and format correctness (F). These terms jointly encourage efficient, non-regressive generation trajectories.

h_t, whose latest entry contains the generated image I_{t-1} and reviewer feedback f_{t-1}. It outputs a state-conditioned action: REFINE edits the current image guided by p_t; REGENERATE produces a new image from scratch using p_t; and STOP terminates exploration. It can be formulated as:

$$a_t \sim \pi_\theta(\cdot \mid s_t), \quad s_t = \begin{cases} (p_{\text{orig}},\ \varnothing) & t = 1 \\ (p_{\text{orig}},\ h_t) & t \geq 2 \end{cases} \tag{1}$$

where $h_t = \{(a_k, I_k, f_k)\}_{k=1}^{t-1}$ is the accumulated interaction history before turn t, containing all previous actions, generated images, and reviewer feedback. At $t{=}1$ the history is empty and the navigator sees only the original prompt.

We record the generated candidates in a rollout as

$$\tau = (s_1, a_1, I_1, f_1,\ \ldots,\ s_T, a_T, I_T, f_T), \quad T \leq T_{\max}. \tag{2}$$

The rollout stops when the navigator selects STOP at the next decision state or when the maximum turn budget $T_{\max}$ is exhausted. Importantly, termination does not commit the system to the final generated image; the delivered output is always the highest-scored candidate across the entire trajectory:

$$I_{\text{out}} = I_{t^\star}, \quad t^\star = \arg\max_{1 \leq t \leq T} \rho_t. \tag{3}$$

The learning objective is to find a policy that maximizes expected trajectory return:

$$\pi^\star = \arg\max_{\pi_\theta}\ \mathbb{E}_{\tau \sim \pi_\theta}\big[R(\tau)\big]. \tag{4}$$

where $R(\tau)$ is a scalar reward assigned to the entire trajectory τ. However, since the delivered image is the best candidate across the trajectory, a reward based solely on $\rho_{t^\star}$ assigns equal credit to every action in the rollout—including later turns that degrade the image or waste turns after a strong candidate has already been found. A good trajectory reward should therefore capture not only whether the agent discovers a high-quality image, but also whether it retains that quality toward the end and reaches it without unnecessary turns. We present such a reward in Section 3.2.

3.2 Learning the Navigator

As shown in Fig. 1(a), the training-free (TF) agent already improves over fixed-workflow agents by conditioning its actions on the current state. However, a gap remains compared to the preference reference, indicating that actions still leave room for improvement in both action selection and prompt

rewriting. To close this gap, we adopt a two-stage training procedure. The first stage is supervised fine-tuning: we construct a trajectory data pipeline yielding 103K trajectories and use them to teach the navigator the interaction format and the semantics of each action. The second stage applies reinforcement learning to optimize trajectory-level preferences. Since a vanilla RL reward cannot distinguish trajectories that regress after finding a strong candidate from those that retain quality, nor penalize unnecessary turns that waste turns without improvement, we propose PRE-GRPO to explicitly reward efficient, non-regressive trajectories.

3.2.1 Action trajectory construction

Simple prompt-image pairs are insufficient for learning an action policy because they do not contain intermediate states, structured actions, or outcomes. We therefore construct full trajectory data. We first score the prompts in the prompt pool from multiple dimensions, including Text/symbolic reasoning, Cardinality/counting, Linguistic reasoning, Domain knowledge, Attribute binding, Counter-intuitive reasoning, and Spatial/structural reasoning. Based on the scores, we select complex samples and rewrite/enhance part of the simple samples. Afterwards, the heuristic explorer performs tree-based reasoning over the actions of REFINE and REGENERATE. Meanwhile, an external multimodal reviewer scores each generated image and selects the optimal branch for subsequent reasoning. Each recorded branch stores the current state, selected action, revised prompt, generated image, reviewer feedback, and subsequent scores. We further filter out regression trajectories and those with low final scores. Finally, we obtain a filtered dataset of 103K trajectories. More details can be found in the Appendix C.

3.2.2 Supervised fine-tuning for cold start

After trajectory construction, we convert the filtered trajectories into multi-turn conversational format and perform supervised fine-tuning. The goal of SFT is to turn a general-purpose multimodal language model into a functional generation navigator before online policy optimization—i.e., cold-start initialization. Through standard next-token cross-entropy training, the model learns the concrete semantics of our defined action-choice space (STOP, REFINE, REGENERATE), acquires the ability to produce the correct format, and establishes the behavioral foundation required for subsequent on-policy reinforcement learning.

3.2.3 PRE-GRPO for trajectory-level preference optimization

However, because the SFT objective maximizes the likelihood of reference actions, it is not directly optimizing for better trajectories or higher-quality generated images. The learned policy inevitably inherits biases from the constructed data, and cannot perform trajectory-level credit assignment. We therefore further optimize the navigator with reinforcement learning.

Vanilla GRPO with a best image score reward conflates unnecessary turns and regression cases as shown in Fig. 1(b), because it assigns the same scalar credit to all actions in a rollout regardless of what happens after the best turn. We address this with PRE-GRPO, as shown in Fig. 2(b) and Equation 6, which decomposes trajectory quality into three complementary components.

For each training prompt x, the current navigator samples a group of K complete rollouts $G(x) = \{\tau_i\}_{i=1}^{K}$. Let $\rho_{i,t}$ denote the reviewer score at turn t in rollout τ_i, let T_i denote the number of generated candidates in rollout τ_i, and let $\hat{\rho}_{i,t} = \rho_{i,t}/\rho_{\max}$ be the normalized score. We compute three trajectory-level statistics:

$$P_i = \max_{1 \le t \le T_i} \hat{\rho}_{i,t}, \quad R_i = \hat{\rho}_{i,T_i}, \quad E_i = \frac{T_i - 1}{T_{\max} - 1}, \tag{5}$$

where P_i measures how good the best image found in the trajectory is, R_i measures whether the trajectory ends near that best image rather than degrading afterward, and E_i measures how many turns were consumed to get there.

$$R(\tau_i) = \underbrace{P_i}_{\textbf{P}\text{eak discovery}} + \alpha \cdot \underbrace{R_i}_{\text{final score }\textbf{R}\text{etention}} - \beta \cdot \underbrace{E_i}_{\text{turn }\textbf{E}\text{fficiency}} + \gamma \cdot \underbrace{F_i}_{\text{format correctness}}, \tag{6}$$

where F_i is the **format correctness term** that encourages well-formed actions and valid prompts. The **peak discovery term** rewards the policy for discovering a strong candidate anywhere in the

trajectory. The **retention term** discourages regression degradation: if the agent finds a good image but then continues into worse states, R_i drops well below P_i and the reward decreases. The **efficiency term** penalizes unnecessary turns, especially when the current candidate is already satisfactory. In particular, the first two terms can be rewritten as:

$$P_i + \alpha R_i = (1 + \alpha)P_i - \alpha(P_i - R_i). \tag{7}$$

The gap $P_i - R_i$ directly measures how much quality was lost between the best image and the last generated candidate. The combined objective penalizes this gap while preserving the incentive to find a high-quality image in the first place.

We then normalize returns within the rollout group:

$$A_i = \frac{R(\tau_i) - \text{mean}_{\tau_j \in \mathcal{G}(x)} R(\tau_j)}{\text{std}_{\tau_j \in \mathcal{G}(x)} R(\tau_j) + \epsilon}. \tag{8}$$

Comparing trajectories under the same prompt removes sensitivity to prompt difficulty and reviewer-score scale. We then apply the standard GRPO clipped objective [Shao et al., 2024] using A_i as the trajectory-level advantage.

Through this objective, PRE-GRPO moves credit assignment from the individual best image to entire trajectories. Rather than rewarding or penalizing a single generation turn in isolation, it evaluates the full trajectory the navigator takes—whether it found a strong candidate, whether it improve quality turn by turn, and whether it did so efficiently. This trajectory-level signal allows the policy to learn not just what action to take at each state, but how to plan a sequence of actions that leads to a high-quality outcome through a short, non-regressive trajectory.

4 Experiments

4.1 Setup and evaluation protocol

Base configuration. Unless otherwise specified, we instantiate Generation Navigator with Qwen3-VL-8B-Instruct as the navigator [Bai et al., 2025], Doubao-Seed1.5 as the reviewer [Guo et al., 2025], and FLUX.2-Klein-9B as the omni-generator [Black Forest Labs, 2026]. The same generator is used for both REFINE and REGENERATE actions.

Training data and hyperparameters. Our trajectory dataset contains 103K structured multi-turn trajectories constructed by the proposed data pipeline. We conducted one epoch of SFT and trained the PRE-GRPO stage for 300 optimization steps. The maximum turn budget is set to three turns. In Eq. 6, we empirically set $\alpha = 0.25$, $\beta = 0.025$, and $\gamma = 0.1$ for all main experiments. More experiments related to hyperparameters are provided in the Appendix E.

Benchmarks and comparisons. We evaluate on three benchmarks. T2I-ReasonBench [Sun et al., 2025] measures compositional reasoning across four categories and reports both reasoning accuracy and image quality. WISE [Niu et al., 2025] targets knowledge-intensive generation across six domains. GenEval [Ghosh et al., 2023] evaluates whether generated images correctly satisfy the compositional semantic requirements of prompts. Noting that these methods typically evaluate results using manually designed questions or additional detection models. This is entirely decoupled from and different from how our reviewer scores image quality. We compare Generation Navigator against a diverse set of baselines: proprietary one-shot systems such as GPT-4o [OpenAI, 2024], Gemini-2.0 [Google DeepMind, 2024], and SeedDream4.0 [Seedream Team, 2025]; open-source generators including Qwen-Image [Wu et al., 2025] and FLUX.2-Klein-9B [Black Forest Labs, 2026]; and agent baselines including GenAgent [Jiang et al., 2026], Think-then-Generate [Kou et al., 2026], IRG [Huang et al., 2026], and StruVis [Lyu et al., 2026].

4.2 Main benchmark results

Tables 1 and 2 summarize the main comparison. Across both benchmarks, our method outperforms the comparison methods. On T2I-ReasonBench, Generation Navigator reaches 79.06% overall reasoning accuracy. It significantly improves over the FLUX.2-Klein-9B one-shot generator [Black Forest Labs, 2026], is comparable to GPT-4o [OpenAI, 2024], and is superior to recent reasoning

Table 1: **Performance on T2I-ReasonBench.** Generation Navigator is comparable to some proprietary models.

Type	Method	Idiom		Textual		Entity		Scientific		Overall	
		Acc.	Qual.	Acc.	Qual.	Acc.	Qual.	Acc.	Qual.	Acc.	Qual.
Proprietary	Gemini-2.0 [Google DeepMind, 2024]	52.40	87.80	73.00	83.30	67.00	94.30	66.70	89.30	64.80	88.70
Proprietary	GPT-4o [OpenAI, 2024]	75.70	94.50	86.90	97.60	77.50	96.60	74.70	94.30	78.70	95.80
Open-source	SD-3.5-large [Esser et al., 2024]	35.60	85.30	62.20	75.40	46.60	92.60	52.90	84.50	49.30	84.40
Open-source	Bagel w/CoT [Deng et al., 2025]	44.60	84.30	44.00	73.70	52.40	91.60	57.70	88.30	49.70	84.50
Open-source	HunyuanImage-3.0 [Cao et al., 2025]	25.40	80.20	54.20	80.90	52.30	92.20	56.80	84.40	47.20	84.40
Open-source	UniCoT [Qin et al., 2026]	49.00	84.20	58.10	92.30	73.50	92.90	51.90	71.70	58.10	85.30
Open-source	Qwen-Image [Wu et al., 2025]	48.10	84.30	66.50	85.80	57.10	84.70	59.50	85.30	57.80	87.50
Open-source	FLUX.2-Klein-9B [Black Forest Labs, 2026]	55.35	93.83	43.21	77.02	57.78	84.29	74.44	76.12	57.70	82.82
Agent	Think-then-Generate [Kou et al., 2026]	58.50	90.60	75.20	89.50	68.80	95.20	72.90	93.50	68.30	92.20
Agent	StruVis [Lyu et al., 2026]	70.15	92.99	75.22	84.75	76.74	97.33	72.18	92.12	73.57	91.80
Agent	Generation Navigator	74.96	94.79	89.71	94.62	73.90	97.08	77.66	92.83	79.06	94.83

Table 2: **Performance on the knowledge-intensive WISE benchmark (judged by GPT-4o).** Generation Navigator achieves superior results across all sub-tasks compared with other baselines.

Type	Method	Cultural	Time	Space	Biology	Physics	Chemistry	Overall
Proprietary	SeedDream4.0 [Seedream Team, 2025]	0.84	0.78	0.90	0.85	0.83	0.66	0.82
Proprietary	GPT-4o [OpenAI, 2024]	0.81	0.71	0.89	0.83	0.79	0.74	0.80
Open-source	SD-3.5-large [Esser et al., 2024]	0.44	0.50	0.58	0.44	0.52	0.31	0.46
Open-source	Bagel w/CoT [Deng et al., 2025]	0.76	0.69	0.75	0.65	0.75	0.58	0.70
Open-source	HunyuanImage-3.0 [Cao et al., 2025]	0.58	0.57	0.72	0.56	0.68	0.35	0.58
Open-source	UniCoT [Qin et al., 2026]	0.76	0.70	0.76	0.73	0.81	0.73	0.75
Open-source	Qwen-Image [Wu et al., 2025]	0.62	0.63	0.78	0.55	0.67	0.35	0.61
Open-source	FLUX.2-Klein-9B [Black Forest Labs, 2026]	0.58	0.65	0.75	0.63	0.65	0.42	0.61
Agent	GenAgent [Jiang et al., 2026]	0.78	0.67	0.78	0.72	0.71	0.55	0.72
Agent	Think-then-Generate [Kou et al., 2026]	0.80	0.74	0.83	0.81	0.85	0.66	0.79
Agent	IRG [Huang et al., 2026]	0.78	0.72	0.76	0.81	0.82	0.78	0.77
Agent	Generation Navigator	0.93	0.87	0.96	0.86	0.87	0.82	0.90

or agentic baselines such as StruVis [Lyu et al., 2026]. On WISE, Generation Navigator obtains an overall score of 0.90. It exceeds proprietary models such as SeedDream4.0 [Seedream Team, 2025], and outperforms closed-loop baselines such as Think-then-Generate [Kou et al., 2026].

4.3 Ablation on Action Training

We first analyze the performance gains brought by each component. As shown in Fig. 3, introducing the agent framework, along with SFT and reinforcement learning, yields consistent performance improvements on both T2I-ReasonBench and WISE. Furthermore, our PRE-GRPO achieves more prominent gains compared with the vanilla GRPO algorithm. However, on GenEval, both TF Agent and SFT fall below the one-shot baseline. We found that GenEval prompts are characteristically short, templated phrases such as "a photo of a cat and a dog" that specify only basic object presence or spatial relations. This input style differs from what TF Agent and SFT are adapted to: the former's hand-crafted decision rules are designed for complex scenarios, while the latter is trained on multi-

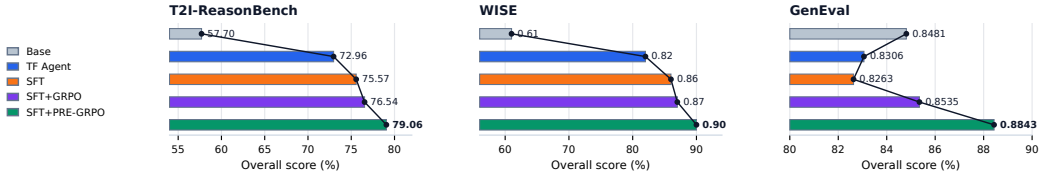


Figure 3: **Training-stage ablation reveals progressive gains on complex tasks and learned restraint on simple tasks.** For complex reasoning and knowledge-intensive tasks (T2I-ReasonBench and WISE), multi-turn action making and policy optimization yield steady improvements. For simple tasks (GenEval), TF Agent and SFT degrade below the one-shot baseline due to the narrow action distribution caused by hand-crafted decisions and trajectory cloning. PRE-GRPO fixes this with trajectory-level rewards by pushing the policy toward suitable action choices on both simple and complex prompts and exhibits strong robustness and adaptability.

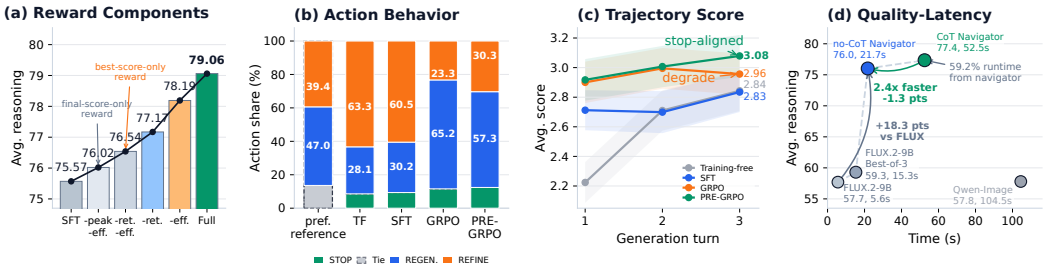


Figure 4: **PRE-GRPO action analysis and quality–latency trade-off.** (a) **Reward components:** PRE-GRPO outperforms final-score-only and best-score-only reward variants, and both the retention and efficiency terms contribute to the final performance. Here, "-peak", "-ret." and "-eff." denote reward variants w/o the peak, retention, and efficiency terms, respectively. (b) **Action behavior:** PRE-GRPO shifts the action distribution closer to the empirical preference reference, balancing decisive regenerations and targeted refinements. (c) **Trajectory score:** the score of PRE-GRPO increases monotonically at each turn, yielding superior trajectories. (d) **Quality–latency trade-off:** Our method can achieve a quality-latency balance by removing CoT.

turn trajectories dominated by hard prompts—both resulting in a narrow action distribution. In contrast, PRE-GRPO optimizes the entire trajectory with trajectory-level rewards, enabling the policy to adaptively calibrate its behavior to different prompts, exhibiting strong robustness and adaptability.

4.4 PRE-GRPO analysis

Reward ablation. As shown in Fig. 4(a), PRE-GRPO achieves substantial performance gains compared with GRPO with a final-score-only reward and GRPO with a best-score-only reward. Meanwhile, removing either the retention term or the efficiency term in Equation 6 leads to performance degradation. This verifies that the two terms can effectively facilitate action-policy learning.

Action behavior. Fig. 4(b) shows how each training stage reshapes the action distribution. TF Agent and SFT both deviate significantly compared to the preference reference, heavily favoring edits (REFINE accounts for over 60% of actions), yet this bias does not translate into better performance. Vanilla GRPO overcorrects in the opposite direction by using 65.15% REGENERATE, becoming overly aggressive about restarting. PRE-GRPO achieves a more balanced distribution that is closest to the preference reference.

Trajectory score. Fig. 4(c) evaluates the average score of each turn. Vanilla GRPO finds a strong intermediate candidate but then regresses to a worse final state, because best-score-only rewards credit all actions equally and cannot distinguish those that improve quality from those that later degrade it. In contrast, PRE-GRPO maintains a steadily improving trajectory.

4.5 Robustness and transfer

We further test whether the learned state-conditioned action policy depends on a particular navigator backbone, generator, or reviewer. Fig. 5 summarizes these component-level variations.

Navigator backbone. Fig. 5(a) shows that trajectory-level training improves state-conditioned action-making capability across navigator backbones and scales. The gains are not limited to the largest navigator: Qwen3-VL-4B improves from 69.59% to 77.53%, while Qwen3-VL-8B improves from 72.96% to 79.06% [Bai et al., 2025]. This indicates that our training paradigm is effective across different MLLMs, demonstrating the generalization ability of our method.

Generator transfer. Fig. 5(b) evaluates whether the learned navigator transfers when the underlying image generator changes. Specifically ,we directly transfer the navigator trained on FLUX.2-9B to different generators for evaluation. The same navigator can improve both Qwen-Image [Wu et al., 2025] and FLUX-family generators [Black Forest Labs, 2026], with the largest relative benefit on

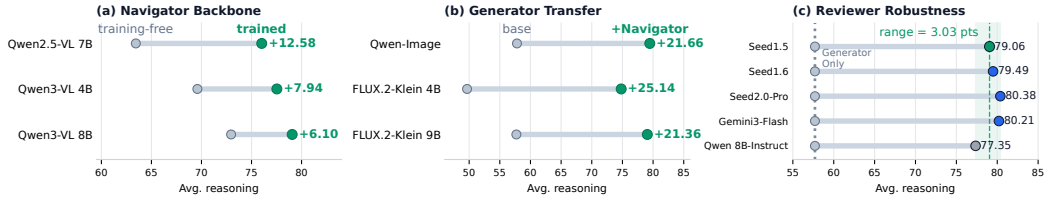


Figure 5: **Robustness and transfer across system components.** The learned state-conditioned action policy shows consistent transfer across navigator, generator, and reviewer choices. (a) **Navigator backbone:** training improves state-conditioned action-making performance across 4B, 7B and 8B navigators, indicating the robustness of our training method across different navigators. (b) **Generator transfer:** The navigator trained on Flux.2-9B can transfer to different underlying generators, which demonstrates that our method does not overfit to a single generator model. (c) **Reviewer robustness:** when we replace the reviewer used during training with a different one in inference, the performance only fluctuates within a range of 3 points, preserving a large gain over one-shot generation. This verifies that our method is not affected by the bias of a single reviewer.

the weaker FLUX.2-Klein-4B generator (49.66% to 74.80%). This demonstrates that our method effectively learns how to perform action-making, rather than overfitting to the pattern of a single generator.

Reviewer robustness and bias analysis. We further investigate whether the navigator's competence is merely a proxy for the reviewer's capability. To test this, we fix the navigator (trained exclusively with Seed1.5 feedback) and swap the inference-time reviewer across four alternatives spanning different architectures and capability tiers: Seed1.6, Seed2.0-Pro, Gemini3-Flash, and Qwen3-VL-8B-Instruct Bai et al. [2025]. **In this case, the reward model for training, the reviewer for inference, and the benchmark evaluator are completely distinct models.** As Fig. 5(c) shows, performance remains within a narrow band of merely 3.03 points (77.35%–80.38%). Notably, even replacing the proprietary Seed1.5 with an entirely different architecture (Gemini3-Flash) yields 80.21% — actually exceeding the training-time reviewer's result. This confirms that our navigator is highly robust and generalizable, capable of making effective action choices based on diverse feedback and current state. This action-making skill is therefore architecture-agnostic: the navigator learns how to act, not what a specific reviewer prefers.

4.6 Efficiency

Agentic generation introduces additional latency, so Fig. 4(d) reports the quality-latency trade-off on T2I-ReasonBench by using Qwen3-VL-8B-Instruct as a reviewer for runtime measurement (results of WISE are in Appendix J). The full-CoT Navigator reaches 77.4% accuracy at 52.5s per example. However, we found that 60% of the overhead stems from the Navigator because of its long Chain-of-Thought (CoT). For a faster alternative, we train a no-CoT Navigator that achieves 76.0% accuracy with an inference latency of 21.7 seconds per sample. It is 2.4x faster than the full-CoT Navigator while still substantially outperforming both the FLUX.2-Klein-9B generator-only baseline and its Best-of-3 variant. This demonstrates a practical quality-latency balance. Appendix I further shows that our method gains over the baseline even with fewer max turns, allowing users to flexibly balance quality and latency.

5 Conclusion

We presented Generation Navigator, an agentic framework motivated by the observation that the optimal next action in multi-turn T2I generation is inherently state-dependent—yet existing methods apply fixed workflows regardless of the current image quality. Our framework recasts this process as a learnable, state-conditioned action-making problem, replacing hand-designed heuristics with a data-driven policy. To address trajectory-level credit assignment, we introduced PRE-GRPO, which jointly rewards peak quality, score retention, and turn efficiency. Experiments across multiple benchmarks confirm substantial gains over existing one-shot and agentic baselines.

Acknowledgments

This work was supported by Grants of NSFC 62672260, 62302246, ZJNSFC LQ23F010008, Ningbo 2023D-001-01 & 2023Z237 & 2024Z284 & 2024Z289 & 2023CX050011 & 2025Z038 & 2025Z028 & 2025Z059, and supported by High Performance Computing Center at Eastern Instituteof Technology and Ningbo Institute of Digital Twin.

References

Shuai Bai et al. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Black Forest Labs. FLUX.2 [klein] 9B Base. https://huggingface.co/black-forest-labs/FLUX.2-klein-base-9B, 2026. Model card.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

Siyu Cao et al. HunyuanImage 3.0 technical report, 2025. URL https://arxiv.org/abs/2509.23951.

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 (ICCV)*, pages 19396–19405, October 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/html/Chen_T2I-Copilot_A_Training-Free_Multi-Agent_Text-to-Image_System_for_Enhanced_Prompt_Interpretation_ICCV_2025_paper.html.

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, 2025b. URL https://arxiv.org/abs/2501.17811.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Meng Chu, Senqiao Yang, Haoxuan Che, Suiyun Zhang, Xichen Zhang, Shaozuo Yu, Haokun Gui, Zhefan Rao, Dandan Tu, Rui Liu, and Jiaya Jia. VisionDirector: Vision-language guided closed-loop refinement for generative image synthesis, 2025. URL https://arxiv.org/abs/2512.19243.

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, 2025. URL https://arxiv.org/abs/2505.14683.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12606–12633. PMLR, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 52132–52152, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Abstract-Datasets_and_Benchmarks.html.

Google DeepMind. Introducing Gemini 2.0: Our new AI model for the agentic era. https://blog.google/innovation-and-ai/models-and-research/google-deepmind/google-gemini-ai-update-december-2024/, 2024. Official Google blog post.

Dong Guo et al. Seed1.5-VL technical report, 2025. URL https://arxiv.org/abs/2505.07062.

Meera Hahn, Wenjun Zeng, Nithish Kannen, Rich Galt, Kartikeya Badola, Been Kim, and Zi Wang. Proactive agents for multi-turn text-to-image generation under uncertainty. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 21591–21628. PMLR, 2025. URL https://proceedings.mlr.press/v267/hahn25a.html.

Minbin Huang, Yanxin Long, Xinchi Deng, Ruihang Chu, Jiangfeng Xiong, Xiaodan Liang, Hong Cheng, Qinglin Lu, and Wei Liu. DialogGen: Multi-modal interactive dialogue system with multi-turn text-image generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 411–426, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.25. URL https://aclanthology.org/2025.findings-naacl.25/.

Wenxuan Huang, Shuang Chen, Zheyong Xie, Shaosheng Cao, Shixiang Tang, Yufan Shen, Qingyu Yin, Wenbo Hu, Xiaoman Wang, Yuntian Tang, Junbo Qiao, Hangyu Guo, Yao Hu, Zhenfei Yin, Philip Torr, Yu Cheng, Wanli Ouyang, and Shaohui Lin. Interleaving reasoning for better text-to-image generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lLNNzBQPas.

Kaixun Jiang, Yuzheng Wang, Junjie Zhou, Pandeng Li, Zhihang Liu, Chen-Wei Xie, Zhaoyu Chen, Yun Zheng, and Wenqiang Zhang. Genagent: Scaling text-to-image generation via agentic multimodal reasoning, 2026. URL https://arxiv.org/abs/2601.18543.

Siqi Kou, Jiachun Jin, Zetong Zhou, Ye Ma, Yugang Wang, Quan Chen, Peng Jiang, Xiao Yang, Jun Zhu, Kai Yu, and Zhijie Deng. Think-then-generate: Reasoning-aware text-to-image diffusion with LLM encoders, 2026. URL https://arxiv.org/abs/2601.10332.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=hFALpTb4fR.

Yunlong Lin, Linqing Wang, Kunjie Lin, Zixu Lin, Kaixiong Gong, Wenbo Li, Bin Lin, Zhenxi Li, Shiyi Zhang, Yuyang Peng, Wenxun Dai, Xinghao Ding, Chunyu Wang, and Qinglin Lu. JarvisEvo: Towards a self-evolving photo editing agent with synergistic editor-evaluator optimization, 2025. URL https://arxiv.org/abs/2511.23002.

Yuanhuiyi Lyu, Kaiyu Lei, Ziqiao Weng, Xu Zheng, Lutao Jiang, Teng Li, Yangfu Li, Ziyuan Huang, Linfeng Zhang, and Xuming Hu. StruVis: Enhancing reasoning-based text-to-image generation via thinking with structured vision, 2026. URL https://arxiv.org/abs/2603.06032.

Shichao Ma, Yunhe Guo, Jiahao Su, Qihe Huang, Zhengyang Zhou, and Yang Wang. Talk2Image: A multi-agent system for multi-turn image generation and editing, 2025. URL https://arxiv.org/abs/2508.06916.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, and Li Yuan. WISE: A world knowledge-informed semantic evaluation for text-to-image generation, 2025. URL https://arxiv.org/abs/2503.07265.

OpenAI. GPT-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

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. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Luozheng Qin, Jia Gong, Yuqing Sun, Tianjiao Li, Haoyu Pan, Mengping Yang, Xiaomeng Yang, Chao Qu, Zhiyu Tan, and Hao Li. Uni-CoT: Towards unified chain-of-thought reasoning across text and vision. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5nevWRoNjn.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

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 (CVPR)*, pages 10684–10695, June 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html.

Seedream Team. Seedream 4.0: Toward next-generation multimodal image generation, 2025. URL https://arxiv.org/abs/2509.20427.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Shaocheng Shen, Jianfeng Liang, Chunlei Cai, Cong Geng, Huiyu Duan, Xiaoyun Zhang, Qiang Hu, and Guangtao Zhai. Agentic retoucher for text-to-image generation, 2026. URL https://arxiv.org/abs/2601.02046.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, and Xihui Liu. T2I-ReasonBench: Benchmarking reasoning-informed text-to-image generation, 2025. URL https://arxiv.org/abs/2508.17472.

Xingchen Wan, Han Zhou, Ruoxi Sun, Hootan Nakhost, Ke Jiang, Rajarishi Sinha, and Sercan O. Arik. Maestro: Self-improving text-to-image generation via agent orchestration, 2025. URL https://arxiv.org/abs/2509.10704.

Linqing Wang, Ximing Xing, Yiji Cheng, Zhiyuan Zhao, Donghao Li, Tiankai Hang, Jiale Tao, Qixun Wang, Ruihuang Li, Comi Chen, Xin Li, Mingrui Wu, Xinchi Deng, Shuyang Gu, Chunyu Wang, and Qinglin Lu. Promptenhancer: A simple approach to enhance text-to-image models via chain-of-thought prompt rewriting, 2025. URL https://arxiv.org/abs/2509.04545.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788, 2020.

Chenfei Wu et al. Qwen-Image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 56704–56721. PMLR, 2024. URL https://proceedings.mlr.press/v235/yang24ai.html.

Arman Zarei, Jiacheng Pan, Matthew Gwilliam, Soheil Feizi, and Zhenheng Yang. AgentComp: From agentic reasoning to compositional mastery in text-to-image models, 2025. URL https://arxiv.org/abs/2512.09081.

Sashuai Zhou, Qiang Zhou, Jijin Hu, Hanqing Yang, Yue Cao, Junpeng Ma, Yinchao Ma, Jun Song, Tiezheng Ge, Cheng Yu, Bo Zheng, and Zhou Zhao. Unified thinker: A general reasoning modular core for image generation, 2026. URL https://arxiv.org/abs/2601.03127.

A Appendix Roadmap

The appendix provides reproducibility details, metric definitions, qualitative evidence, and a human validation study that complement the main paper:

1. **Pilot study details.** Appendix B expands the setup behind the introductory pilot study, including the three-turn workflow and the fixed-action versus preference-reference comparison.
2. **Data pipeline and distribution.** Appendix C describes the construction of the 103K structured trajectories, including prompt-pool scoring, targeted prompt augmentation, branch-and-select exploration, and trajectory filtering. This section supports the training-data claims in Section 4.1 and clarifies how action-trajectory supervision differs from ordinary prompt-image data.
3. **Agent prompts.** Appendix D gives the complete reviewer and navigator prompt templates.
4. **Hyperparameter studies.** Appendix E analyzes the reward weights α and β in Eq. 6, the turn budget $T_{\max}$, and representative reward-scale examples. These studies explain why the default PRE-GRPO configuration balances candidate discovery, terminal retention, and turn efficiency.
5. **Controlled sampling-budget comparison.** Appendix F compares one-shot generation, best-of-3 selection, prompt enhancement, fixed-workflow agents, and trained action-policy variants under a single T2I-ReasonBench view. This isolates the net contribution of state-conditioned action making from gains caused by extra sampling budget.
6. **Best-score versus final-score selection.** Appendix G ablates the interaction between trajectory-level reward choice and inference-time output selection, explaining why Generation Navigator returns the highest-scored candidate along the trajectory.
7. **Average generation turns.** Appendix H reports the average number of generation turns on GenEval as an indirect signal of action calibration on simple compositional prompts.
8. **Quality–latency trade-off.** Appendix I studies how the maximum turn budget affects accuracy and latency for the no-CoT Navigator.
9. **WISE latency analysis.** Appendix J compares the quality–latency trade-off on WISE against generator-only and agent baselines.
10. **Qualitative visualizations.** Appendix K presents representative multi-turn cases across textual image design, counting, spatial relations, scientific reasoning, and two-object generation. The examples expose the actual prompts, reviewer feedback, actions, and selected images behind the aggregate results.
11. **Human evaluation.** Appendix L reports a pairwise human study comparing human preferences with reviewer-induced preferences, providing an empirical check that the automatic reviewer is a useful signal for PRE-GRPO trajectory optimization.
12. **Limitations and future directions.** Appendix M discusses computational trade-offs, and the use of reviewers as environment signals.

B Pilot Study Details

Pilot study details (Section 1). We construct a three-turn workflow on T2I-ReasonBench to compare action strategies. In the first turn, each prompt is rewritten by Doubao-Seed1.5 and then fed to FLUX.2-Klein-9B for initial image generation. In subsequent turns, three action strategies diverge: (i) *Refine-only* applies image-to-image editing on the current image at every turn; (ii) *Regenerate-only* uses the MLLM (Doubao-Seed1.5) to rewrite the prompt and performs text-to-image generation from scratch at every turn; (iii) *Preference reference* executes both actions in parallel at each turn, scores both outcomes with the benchmark evaluator, and proceeds along the higher-scoring branch into the next turn.

The preference reference thus provides an empirical upper bound on greedy per-turn action selection without any learned policy. Results in Fig. 1(a) show that neither Refine nor Regenerate dominates—Regenerate wins 47.01% of head-to-head comparisons, Refine wins 39.38%, and 13.61% are ties—confirming that the optimal action is state-dependent and motivating a learned action policy.

C Data Pipeline and Distribution

This section details how we construct the 103K structured multi-turn trajectories used for navigator training. We construct this data to support the subsequent cold-start training of SFT.

C.1 Prompt Pool and Complexity Scoring

The first stage starts from 923,274 candidate prompts. Each prompt is scored along seven complexity dimensions: spatial and structural composition, attribute binding, cardinality and counting, counter-intuitiveness, text and symbolic content, domain knowledge, and linguistic reasoning. We use these scores to identify prompts that are likely to require nontrivial state-conditioned actions during generation. As shown in Fig. 6, the raw pool is skewed toward Easy/Medium/Hard prompts, while Expert cases and score-5 tails are rare. In particular, we sample complex seeds from Hard/Expert prompts and from prompts whose score is at least 4 in any complexity dimension. We also sample Easy/Medium prompts as rewriting seeds for the augmentation stage, so that the final data contain both intrinsically difficult prompts and simpler prompts transformed into different scene cases. This endows our samples with diversity.

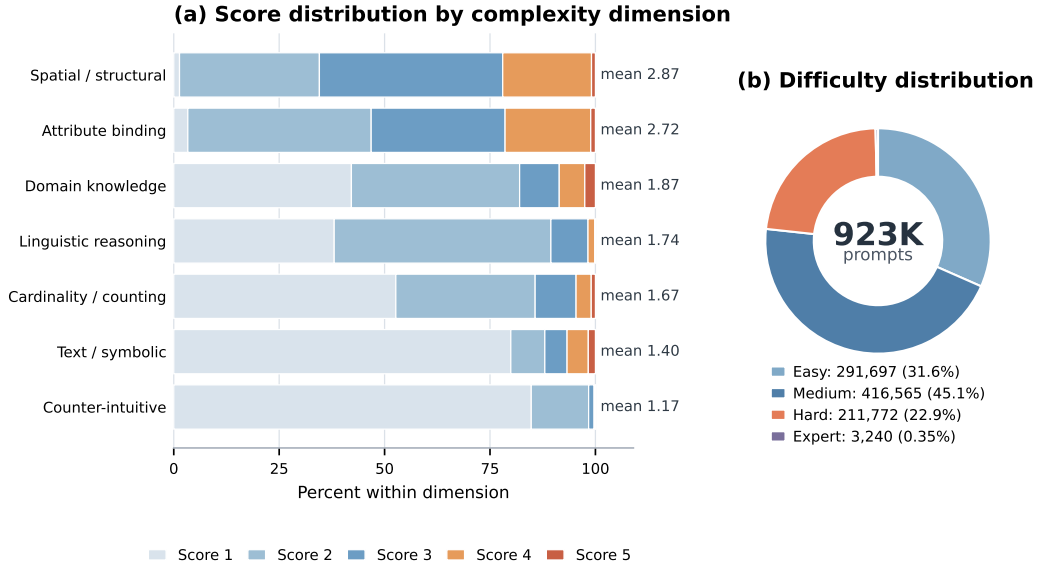


Figure 6: Prompt-pool distribution used for data construction. (a) Per-dimension score distribution over the 923K scored prompt pool, with mean score annotated on the right. (b) Overall difficulty distribution.

C.2 Prompt Augmentation

In the second stage, we employ an LLM to rewrite Easy/Medium samples to enrich domains with insufficient hard samples. Meanwhile, we augment idiom and scientific samples that are difficult to obtain in daily scenarios, ensuring broader coverage of our dataset. The rewriting prompt is presented as follows:

Idiom prompt generation rules. The rule template used for idiom prompt generation is shown below.

```
Track-specific rules for idiom_interpretation:
- Produce one compact everyday sentence that uses an idiom or metaphor in context.
- Usually 8-18 words, with at most one short setup clause before the idiom.
- Do not explain the figurative meaning.
- Keep it concrete enough that a later model can visualize it.
- The sentence may borrow a motif from the seed, but it should read like ordinary language.
```

Scientific prompt generation rules. The rule template used for scientific prompt generation is shown below.

```
Track-specific rules for scientific_reasoning:
- Convert the seed into a short scientific prompt.
- Usually keep prompts within 2-22 words.
- Do not describe a generic short scene.
- Keep explanations implicit. Do not explain why something happens.
- Before finalizing each scientific prompt, silently check: is this a setup or comparison, not a revealed result or
      decorative scene?
- If the seed is scenic, surreal, or decorative, discard that surface framing and keep only one motif that can become a
      benchmark-like scientific prompt.
```

C.3 Branch-and-select Trajectory Construction

The third stage constructs multi-turn action trajectories with a branch-and-select procedure. Algorithm 1 summarizes the implementation used to construct trajectory logs. At each turn, the explorer expands multiple candidate actions, scores their generated images, and continues from the best branch.

Algorithm 1 Branch-and-select trajectory construction

Input: request x, navigator N, generator G, reviewer R, branch size K, budget $T_{\max}$, threshold ρ_{thr}
$\mathcal{P}_1 \leftarrow \emptyset$, $\mathcal{H}_1 \leftarrow \emptyset$, $\rho_{\text{best}} \leftarrow -\infty$
for $t = 1$ **to** $T_{\max}$ **do**
  $\mathcal{A}_t \leftarrow N(x, \mathcal{P}_t)$, where $\mathcal{A}_t = \{a_t^1, \ldots, a_t^K\}$
  for $k = 1$ **to** K **do**
    $I_t^k \leftarrow G(a_t^k)$
    $(s_{\text{visual}}^k, s_{\text{instruction}}^k, c_t^k) \leftarrow R(x, I_t^k)$
    $\rho_t^k \leftarrow 0.3\, s_{\text{visual}}^k + 0.7\, s_{\text{instruction}}^k$
  end for
  $k^\star \leftarrow \arg\max_{1 \le k \le K} \rho_t^k$
  $\mathcal{H}_{t+1} \leftarrow \mathcal{H}_t \cup \{(a_t^k, I_t^k, \rho_t^k, c_t^k)\}_{k=1}^K$
  if $\rho_t^{k^\star} \ge \rho_{\text{thr}}$ **or** $t = T_{\max}$ **or** $\rho_t^{k^\star} \le \rho_{\text{best}}$ **then**
    return $\mathcal{H}_{t+1}$
  end if
  $\rho_{\text{best}} \leftarrow \max(\rho_{\text{best}}, \rho_t^{k^\star})$
  $\mathcal{P}_{t+1} \leftarrow \mathcal{P}_t \cup \{(a_t^{k^\star}, I_t^{k^\star}, \rho_t^{k^\star}, c_t^{k^\star})\}$
end for

The selected branch defines the next state for continued exploration. Note that while $\mathcal{H}_{t+1}$ logs the full tree of candidate actions and critiques for dataset construction, the navigator N is conditioned only on the active selected trajectory $\mathcal{P}_t$ during rollout. In our implementation, the stopping threshold is empirically set to $\rho_{\text{thr}} = 4.5$ on the 0–5 reviewer-score scale.

C.4 Trajectory Filtering

The heuristic explorer is used to generate broad action coverage, not as an oracle policy to be cloned directly. Before converting rollouts into SFT trajectories, we apply two filtering criteria: (1) we retain only trajectories whose best reviewer score exceeds 4.5 on the 0–5 scale, and (2) within each trajectory, we keep only the branches that show strict monotonic score improvement across turns. Branches that degrade or plateau are discarded. After filtering, the final dataset contains 103K structured multi-turn trajectories with state s_t, feedback $f_t = (c_t, \rho_t)$, target action $a_t = (d_t, p_t)$, and trajectory-level scores.

Note that the SFT model does not imitate the heuristic explorer wholesale; it only sees the successful trajectories that survived filtering. PRE-GRPO then further refines action preference at the trajectory level.

C.5 Training Data Contamination Analysis

Since our training trajectories are collected from an internal prompt pool, we verify that no benchmark prompt has leaked into the training data. We adopt a multi-granularity protocol combining semantic and lexical metrics:

1. **Embedding cosine similarity.** We encode each benchmark prompt and each training prompt with the `sentence-transformers/all-MiniLM-L6-v2` encoder [Reimers and Gurevych, 2019, Wang et al., 2020] and report the maximum cosine similarity between each benchmark prompt and its nearest neighbor in the training pool. This captures paraphrase-level semantic overlap.
2. **5-gram Jaccard index.** Measures symmetric lexical overlap at the 5-gram level.
3. **5-gram containment.** Measures the fraction of a benchmark prompt's 5-grams that appear in its nearest training prompt, capturing asymmetric substring inclusion.
4. **8-gram containment.** Following the contamination protocol used in PaLM [Chowdhery et al., 2023], we flag a benchmark sample as potentially contaminated if $\geq$70% of its 8-grams are contained in any training prompt.
5. **13-gram collision.** Following GPT-3 [Brown et al., 2020], we check for exact 13-gram matches between benchmark and training prompts.

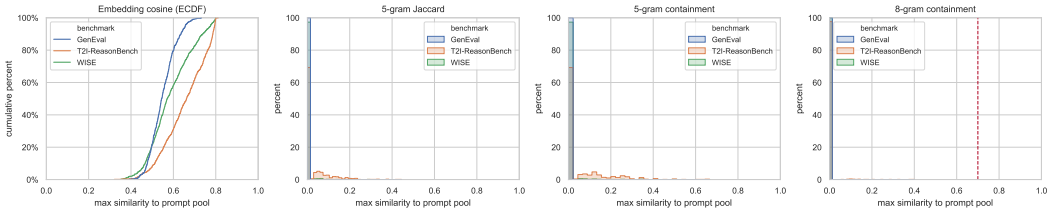


Figure 7: Training data contamination analysis. From left to right: embedding cosine similarity (ECDF), 5-gram Jaccard (histogram), 5-gram containment (histogram), and 8-gram containment (histogram) with the 70% threshold from the PaLM contamination protocol [Chowdhery et al., 2023] shown as a dashed line. No benchmark prompt exceeds the flagging threshold under any lexical metric.

Fig. 7 reports the results for the first four metrics across all three evaluation benchmarks (GenEval, T2I-ReasonBench, WISE). The leftmost panel plots the embedding cosine ECDF. Nearest-neighbor similarities stay below the 0.8 semantic near-duplicate threshold. T2I-ReasonBench shows relatively higher values because its prompts frequently describe common natural phenomena in generic terms; the lexical metrics confirm these overlaps are not surface-level copies. Under both 5-gram Jaccard and 5-gram containment, nearly all benchmark prompts have near-zero similarity to the training pool. In the 8-gram containment panel, no prompt approaches the 70% threshold of the PaLM contamination protocol [Chowdhery et al., 2023]. Under the 13-gram collision protocol of GPT-3 [Brown et al., 2020], we find zero exact matches.

We conclude that our results are not inflated by benchmark contamination.

D Agent Prompts

This section provides the complete prompt templates used by the reviewer and navigator. We use longer, more instructional prompts during data construction to make heuristic rollouts stable and interpretable. After trajectory data are built, the trained navigator is driven by a shorter prompt during SFT, RL, and inference, reducing token overhead while preserving the same action semantics.

Reviewer prompt. The reviewer prompt used during data construction and inference is shown below. It returns a textual diagnosis and two scalar scores. The scalar reviewer score used for ranking candidates is $\rho = 0.3 \cdot \text{Visual Quality} + 0.7 \cdot \text{Instruction Comprehension}$.

```
**Role**: You are a strict **Image Quality Critic and QA Specialist**.
```

```
**Task**:
1. Analyze the **Current Image** strictly against the **User Request**. Keep the thinking process concise (within 512
      tokens).
2. Provide a detailed diagnosis of flaws.

**Inputs**:
- **User Request**: {user_request}
- **Current Image**: (Visual Input)

**Evaluation Criteria**:
**1. Aesthetic Quality**:
**Aesthetic & Technical Quality Scoring Rules (0.0-5.0)**:
Evaluate the overall aesthetic appeal of the image and provide a score:

Assess the image for technical flaws (artifacts, anatomy distortions, blurriness) and aesthetic appeal (lighting,
      composition).
- **5 (Excellent):** Flawless. Professional lighting, perfect anatomy (hands/faces), sharp details, high aesthetic value.
- **4 (Good):** High quality, but may have negligible artifacts (e.g., slightly odd background texture) or average
      lighting.
- **3 (Acceptable):** Noticeable flaws (e.g., slightly distorted hands, weird physics, over-saturation) but the image is
      coherent.
- **2 (Poor):** Obvious artifacts, severe anatomy issues (extra fingers, melted faces), low resolution, or blurry.
- **1 (Unusable):** Complete noise, broken geometry, or unrecognizable visual garbage.

**2. Fulfillment (Crucial)**:
**Completeness Fulfillment (Coverage) Scoring Rules (0.0-5.0)**:
Evaluate the completeness of the image based on the requirements and provide a score:

How accurately does the image reflect the elements, attributes, style, and spatial relationships described in the prompt?
- **5 (Perfect):** All elements, modifiers, and styles are present and accurate. No missing objects.
- **4 (High):** Key elements are correct. Very minor details (e.g., a small background object or exact shade of color)
      might be slightly off, but the main intent is perfectly captured.
- **3 (Moderate):** Main subject is present, but some modifiers (e.g., "holding a red cup" becomes "holding a blue cup")
      or styles are incorrect.
- **2 (Low):** Significant deviations. Key subjects missing or incorrectly interacting. The image vaguely resembles the
      prompt but fails on specifics.
- **1 (Failure):** Completely irrelevant to the prompt or hallucinates entirely different subjects.

**Output Format**:
Provide a valid JSON object.
‘‘‘json
{
    "evaluation": {
        "Instruction_Comprehension": [0-5 float],
        "Visual_Quality": [0-5 float]
    },
    "diagnosis": "Short objective description of flaws. Be specific. E.g., 'The lighting is flat', 'The cat is missing the
       hat', 'The style is 3D instead of anime'."
}
‘‘‘
```

Navigator prompt for data construction. During data construction, the navigator prompt is intentionally verbose so that the heuristic agent produces diverse but interpretable rollouts.

```
**Role**: You are an expert **Art Director and Prompt Strategist**.

Keep the thinking process concise (within 384 words).

**Context**: A Critic has evaluated the previous generation and found issues. You have final authority to choose the next
      action.

**Inputs**:
- **Original User Request**: {user_request}
- **Current Prompt Used**: {current_prompt}
- **Critic's Visual Quality Score**: {visual_score}/5
- **Critic's Instruction Comprehension Score**: {instruction_score}/5
- **Critic's Diagnosis**: "{diagnosis}"
- **Visual Context**: (See attached image)

**Task 1: CHOOSE the Action**
Analyze the Diagnosis and Image to choose ONE action:
- **STOP**: The image is excellent and meets the User Request perfectly.
- **REGENERATE (T2I)**: Major failures. Examples: Wrong subject, garbled composition, wrong art style, severe anatomy
      issues. Needs a fresh start.
- **REFINE (I2I)**: The foundation is solid, but specific details need editing. Examples: Wrong clothing color, missing
      specific item, bad hands, unwanted object.

**Task 2: EXECUTE the Prompt Writing**

### Scenario A: If REGENERATE (T2I)
**Strategy Shift**: The previous prompt FAILED. Do NOT just copy it or make minor tweaks.
1. **Rephrase**: Describe the subject using different adjectives or synonyms.
2. **Reorder**: Move the missing or distorted elements mentioned in the ‘diagnosis‘ to the **very beginning** of the
      prompt.
3. **Simplify vs. Enrich**:
   - If diagnosis says "garbled/messy" -> **Simplify** details, focus on main subject.
   - If diagnosis says "boring/wrong style" -> **Enrich** with style modifiers (e.g., "cinematic lighting", "concept art").
```

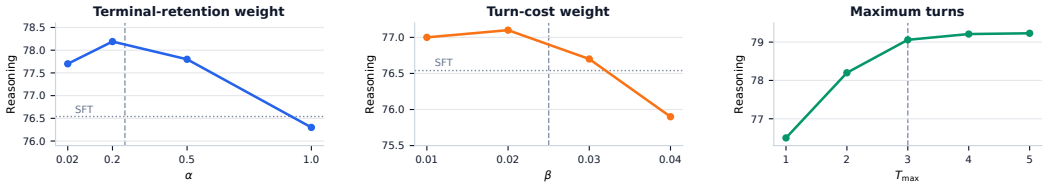


Figure 8: **Hyperparameter studies.** We sweep the terminal retention weight α, turn-cost weight β, and maximum turn budget $T_{\max}$. Dashed vertical lines mark the default setting used in the main experiments. The dotted horizontal baseline in the first two panels denotes the corresponding SFT baseline. Both α and β consistently outperform the baseline over a wide range of values, demonstrating that our method is robust to hyperparameters. Within this broad interval, both α and β can effectively impose path constraints.

```
### Scenario B: If REFINE (I2I)
Write a specific **Editing Instruction**.
1. **Format**: Use "Add [obj]", "Remove [obj]", "Change [obj] to [obj]", or "Make [obj] [action]".
2. **Specificity**: Explicitly state the target. AVOID "it/him/her". Use "the panda", "the red car".
3. **Single Focus**: Target **ONLY ONE** specific area mentioned in the `diagnosis`. Do not fix everything at once.
4. **Avoid Anatomy**: Do NOT try to fix eyes/gaze via I2I. Use REGENERATE for that.
5. **Example**: "Make the tall warrior wear a red cape." (NOT "Change clothes").

**Task 3: Output**
Only provide a valid JSON object.
```json
{
    "decision": "STOP" | "REGENERATE" | "REFINE",
    "reasoning": "Explain why you chose this action based on the diagnosis.",
    "revised_prompt": "String OR null. If STOP, null. If REGENERATE, the full new T2I prompt. If REFINE, the specific I2I
      instruction."
}
```
```

Navigator prompt for training and inference. After collecting trajectory data, we replace the long construction prompt with the following compact prompt:

```
You are an Art Director and Prompt Strategist. Based on diagnosis and current result, respond with strict JSON using keys
      decision and revised_prompt. The decision field must be STOP, REGENERATE, or REFINE.
```

The data-construction prompt and inference prompt differ in length and explicit guidance, but they share the same action space and output contract: stop, regenerate with a full revised T2I prompt, or refine with a localized I2I instruction.

E Hyperparameter Studies

We study three practical hyperparameters that affect action-policy learning: the terminal-retention weight α, the turn-cost weight β in Eq. 6, and the maximum turn budget $T_{\max}$. The format correctness weight γ is kept fixed because it mainly guards structured output validity rather than the trajectory-quality trade-off. To reduce sweep cost, these ablations use 100-step PRE-GRPO runs while varying each α or β factor at a time; the main model reported in Section 4 is trained for 300 steps.

Reward weights. Fig. 8 shows that α remains strong from 0.02 to 0.5, while larger values such as 1.0 reduce performance. In the useful range, the score-retention term acts as a stabilizer for stop alignment without disturbing peak discovery. When α becomes too large, the policy can overemphasize score retention and lose some exploratory benefit. The turn-cost weight β behaves as a weak efficiency regularizer: a small penalty is sufficient to prefer shorter trajectories when quality is comparable, while an overly large penalty suppresses useful exploration.

Scale intuition. The default weights are intended to regularize action selection rather than to override candidate discovery. For readability, let $y = [\rho_1, \ldots, \rho_T]$ be a raw reviewer-score sequence and ignore the fixed format correctness term. With the default $T_{\max} = 3$, $\alpha = 0.25$, and $\beta = 0.025$,

Eq. 6 reduces to the following normalized comparison:

$$\widetilde{R}(y) = \frac{\max_t \rho_t}{5} + 0.25 \cdot \frac{\rho_T}{5} - 0.025 \cdot \frac{T-1}{2}.$$

Each extra turn costs $0.025/(T_{\max}-1) = 0.0125$ in normalized reward under $T_{\max} = 3$; using all three turns instead of one therefore costs 0.025, equivalent to roughly 0.10 raw reviewer-score points when peak and terminal scores move together. This scale yields the following representative preferences, where numbers in parentheses are $\widetilde{R}$ values:

$$\text{(i)} \quad [4.80]\ (1.2000) > [4.50, 4.80]\ (1.1875),$$
$$\text{(ii)} \quad [3.00, 4.00, 4.80]\ (1.1750) > [3.00, 4.80, 4.00]\ (1.1350).$$

These comparisons show that the reward prefers a shorter trajectory at equal quality and, for trajectories with the same peak and length, prefers ending near the discovered peak. Thus, α and β regularize the state-conditioned action policy toward stable and economical trajectories without turning the objective into terminal-only or length-only optimization.

The same calculation explains why larger weights hurt. If β is increased to 0.05, the length penalty can become too aggressive:

$$\widetilde{R}_{\beta=0.05}([4.83]) = 1.2075 \ > \ \widetilde{R}_{\beta=0.05}([4.0, 4.6, 5.0]) = 1.2000,$$

so the reward would reject a trajectory that eventually reaches a substantially better candidate. Similarly, if α is made much larger, terminal retention can make lower-peak stable trajectories overtake trajectories that discover a stronger peak:

$$\widetilde{R}_{\alpha=4.0}([3.00, 4.50, 4.70]) = 4.6750 \ > \ \widetilde{R}_{\alpha=4.0}([3.00, 4.90, 4.64]) = 4.6670.$$

This matches the sweep trend in Fig. 8: overly large terminal or length weights make the trajectory objective less faithful to high-quality candidate discovery.

Turn budget. Increasing $T_{\max}$ from one to three turns yields a clear gain, confirming that user intents benefit from multi-turn correction. The performance gains are substantial in the first three turns. Considering both efficiency and performance, we therefore use $T_{\max} = 3$ in the main experiments to balance quality and inference cost.

F Controlled Sampling-Budget Comparison

We further analyze the performance gains brought by each individual operation. Fig. 9 separates the gain from extra sampling budget, prompt enhancement, fixed-workflow multi-turn execution, state-conditioned action making, and policy optimization on T2I-ReasonBench. The best-of-3 controls with benchmark selection improve only modestly over their corresponding one-shot settings: FLUX improves from 57.8 to 59.3, while prompt enhancement improves from 68.9 to 69.4. In contrast, fixed-workflow agents, training-free state-conditioned action making, and learned action-policy variants produce progressively larger gains, with PRE-GRPO reaching 79.06. This controlled view indicates that the final improvement is not explained by sampling budget alone; the main gains come from state-conditioned action making and trajectory-level training.

G Best-Score vs. Final-Score Selection

In the main paper, we use the highest-scoring image along the generation trajectory as the delivered output rather than simply returning the final generated image. To examine this output rule, we compare final-output selection with best-score selection at inference. We also include two reward variants during training, a final-step-only reward and a best-step reward, to check whether the same selection trend holds under different training signals. Results on T2I-ReasonBench are shown in Fig. 10.

Across both reward variants, best-score selection outperforms final-output selection. This indicates that, in a multi-turn generation trajectory, the last image is not always the strongest candidate according to the evaluation score, so preserving the best-scored candidate is a simple and more reliable output rule. The combination of best-step reward and best-score selection yields the strongest result (76.5%). We therefore adopt best-score trajectory selection as the default output strategy, including in the final PRE-GRPO setting.

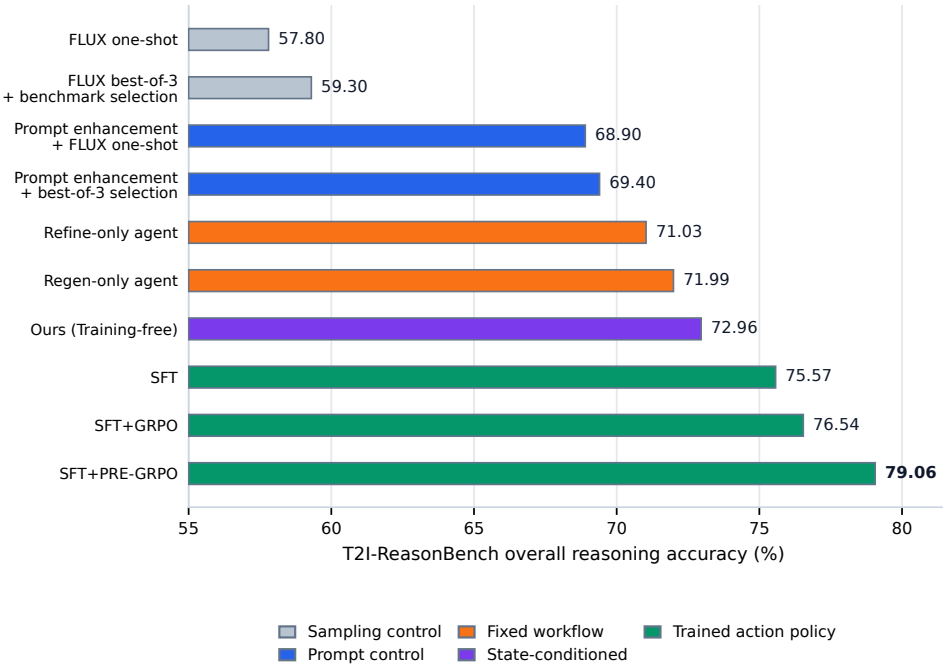


Figure 9: **Controlled sampling-budget comparison on T2I-ReasonBench.** We compare one-shot generation, best-of-3 controls with benchmark selection, prompt-enhanced controls, fixed-workflow agents, training-free state-conditioned action making, and trained action-policy variants. The small gains from best-of-3 selection isolate the effect of additional sampling, while the larger gains from state-conditioned and trained action policies show the net contribution of action-policy learning.

H Average Generation Turns on GenEval

To further examine the simple-prompt behavior discussed in Section 4.3, we report the average number of generation turns consumed by each method on the GenEval benchmark (Table 3). This statistic is not meant to explain performance by turn count alone, since our final output selector keeps the highest-scored candidate across the trajectory. Instead, it serves as an indirect signal of how often each method proposes additional actions on prompts where the initial candidate is often already competitive.

Table 3: Average generation turns on GenEval. PRE-GRPO uses fewer turns while achieving the best score, indicating better-calibrated action choices on simple compositional tasks.

Method	Avg. Turns	GenEval score
TF Agent	1.95	0.831
SFT	1.97	0.826
SFT + GRPO	1.89	0.854
SFT + PRE-GRPO	**1.67**	**0.884**

GenEval prompts are compositionally simple (e.g., “a red apple and a green bench”), meaning that a competent generator can often produce an adequate candidate with few turns. In this regime, the main risk is not merely using more turns, but choosing worse and unnecessary actions may hurt the candidate pool. A training-free agent can be overly sensitive to this mistake, while SFT imitates demonstrated actions without trajectory-level reward feedback.

Table 3 is consistent with this interpretation. TF Agent and SFT consume nearly two full turns on average (1.95 and 1.97), yet both underperform the one-shot GenEval baseline reported in Fig. 3. Vanilla GRPO reduces the average only slightly (1.89) and recovers part of the score. PRE-GRPO

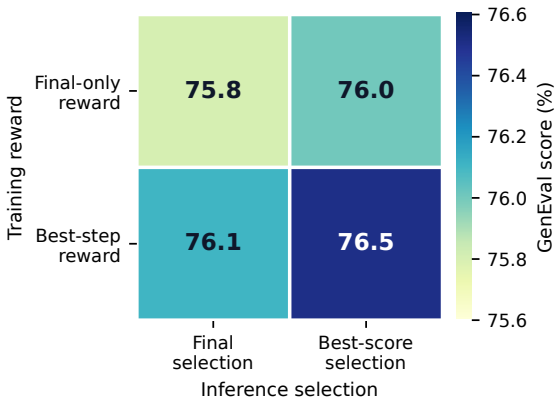


Figure 10: **Best-score vs. final-score selection on T2I-ReasonBench.** We ablate the training reward (final-step only vs. best-step) and the inference selector (final output vs. best-scored trajectory candidate). Best-score selection consistently improves over final-output selection under both reward choices.

lowers the average to 1.67 while achieving the best GenEval score (0.8843), suggesting that its trajectory-level objective discourages low-value actions and yields better-calibrated action choices on simple prompts.

I Effect of Turn Budget on Quality and Latency

We measure the quality-latency trade-off under different maximum turns. We vary the maximum turn budget from 1 to 4 using the no-CoT Navigator with Qwen3-VL-8B-Instruct as the reviewer and evaluate on T2I-ReasonBench. Results are visualized in Fig. 11.

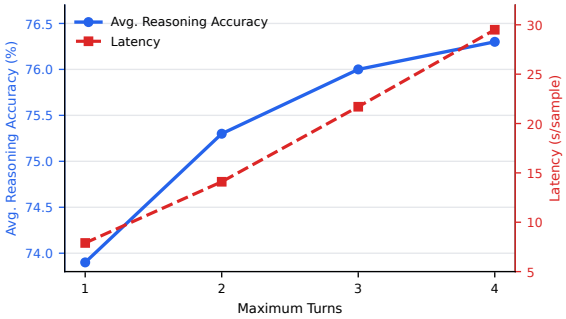


Figure 11: Quality–latency trade-off as the turn budget increases from 1 to 4. Users can adjust the maximum turn adaptively according to their computational budget.

Latency increases linearly. Each additional turn introduces one extra generator inference and one additional critic model invocation, making the cost variation highly predictable. When the turn number T is set to 3, the system can retain most of the benefits brought by multi-turn interaction while keeping the latency within a reasonable range. Therefore, setting $T = 3$ as the default number of turns is a reasonable choice. Meanwhile, even when T is small, the method already reaches performance well above the one-shot FLUX.2-9B baseline. This means users can achieve satisfactory performance at a lower cost in practical applications.

J Latency Analysis on WISE

Complementing the quality–latency analysis in Section 4.6, we further compare different methods on the WISE benchmark [Niu et al., 2025]. Fig. 12 plots each method by its WISE overall score and per-sample latency. The base model FLUX.2-9B [Black Forest Labs, 2026] finishes in 5.7 seconds but only reaches a WISE score of 0.61. Qwen-Image [Wu et al., 2025] obtains the same WISE score of 0.61 while requiring 101.2 seconds. The agent-based method IRG [Huang et al., 2026] improves quality to 0.77, but its latency increases to 104.8 seconds, about 18.4× slower than FLUX.2-9B, due to complex LLM reasoning and full-image regeneration within its closed loop.

Generation Navigator achieves a WISE score of 0.90 with CoT reasoning at 53.8 seconds. Removing CoT reduces the latency to 22.1 seconds while retaining a high WISE score of 0.88. Compared with IRG, the no-CoT Navigator is 4.7× faster and improves the WISE score by 0.11. This shows that Generation Navigator provides a stronger quality–latency trade-off than existing agent methods on knowledge-intensive generation.

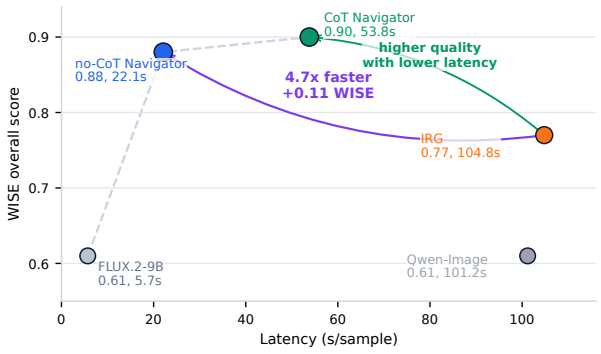


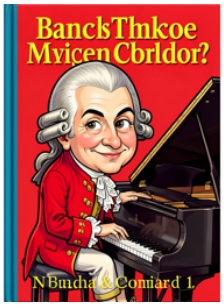
Figure 12: **Quality–latency trade-off on WISE.** Generation Navigator achieves substantially higher WISE scores than IRG while using much lower latency, especially in the no-CoT setting.

K Qualitative Visualizations

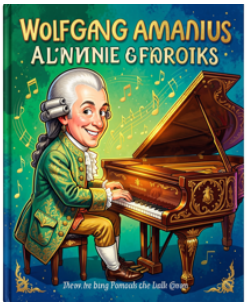
We provide representative qualitative cases covering textual image design, counting, spatial-relation, and two-object prompts. The top row shows the one-shot baseline and the navigator trajectory as vector-labeled image cards. The baseline is the generator output from the original prompt. For navigator images, labels indicate the action that produced the image, followed by the reviewer score and next action.

K.1 Case 1: Textual Image Design

Original prompt. Design a colorful book cover featuring a caricature of Mozart playing the piano.



One-shot baseline
generator only



Turn 1: initial T2I
score 3.30
next: REGENERATE



Turn 2: by REGENERATE
score 4.88
next: STOP

Navigator turn details.

Turn 1: initial T2I; next action REGENERATE (reviewer score 3.30)

Generation prompt. A vibrant, colorful book cover features a whimsical caricature of Wolfgang Amadeus Mozart playing a grand piano. The cover bursts with bold, saturated hues—emerald, sapphire, and gold—creating a lively backdrop. Mozart, in a playful caricature style, wears an exaggerated powdered wig, a detailed 18th - century coat with intricate patterns, and plays a grand piano with ornate carvings and glowing keys. Musical notes swirl around him, enhancing the musical theme. The design blends classical artistry with humorous exaggeration, ideal for a music - themed book, capturing Mozart's genius in a fun, eye - catching, artistically rich composition.
Reviewer diagnosis. All text content on this book cover is nonsensical AI-generated gibberish. Mozart's name is incorrectly rendered as garbled nonsense text instead of the correct name, and the small text at the bottom of the cover is also meaningless unreadable garbage, which fails to meet the requirement of a proper functional book cover. The core visual elements (colorful style, caricature of Mozart playing piano) are correctly implemented, the overall visual aesthetic is good with no major technical defects.

Turn 2: generated by REGENERATE; next action STOP (reviewer score 4.88)

Generation prompt. Vibrant, colorful book cover (emerald, sapphire, gold hues) features a whimsical caricature of Wolfgang Amadeus Mozart (exaggerated powdered wig, detailed 18th - century coat) playing a grand piano (ornate carvings, glowing keys). Musical notes swirl around. Top text: "WOLFGANG AMADEUS MOZART" in bold golden 18th - century - inspired font. Bottom text: "A Musical Journey" (legible) in 18th - century - style. Design blends classical artistry with humorous exaggeration, ideal for a music - themed book, capturing Mozart's genius in a fun, eye - catching, artistically rich composition.
Reviewer diagnosis. There are no significant flaws. The image fully meets all specified requirements: it is a properly formatted colorful book cover, correctly features a clear caricature of Wolfgang Amadeus Mozart playing a grand piano, with high technical quality, harmonious color matching, good layered lighting, sharp clear details, and cohesive subject-background integration.

K.2 Case 2: Spatial Relation

Original prompt. a photo of a bus above a boat.



One-shot baseline
generator only



Turn 1: initial T2I
score 3.30
next: REGENERATE



Turn 2: by REGENERATE
score 4.82
next: REGENERATE



Turn 3: by REGENERATE
score 4.85
next: STOP

Navigator turn details.

Turn 1: initial T2I; next action REGENERATE (reviewer score 3.30)

Generation prompt. A photorealistic image depicts a vibrant red city bus (with large windows, sleek modern design) floating above a small, weathered wooden rowboat. The bus hovers mid - air, positioned directly above the boat, which rests on calm, rippled water that reflects the bus's underside. The background features a clear sky with soft clouds, enhancing the surreal, dreamlike atmosphere. The scene contrasts the bus's sleek design with the boat's rustic texture, ideal for artistic photography, capturing the unexpected bus - over - boat juxtaposition.
Reviewer diagnosis. The bus is unrealistically levitating in mid-air above the empty boat, rather than being supported on the boat as expected for a bus above a boat, creating a physically impossible, unnatural arrangement that does not match the implicit expectation of a plausible photo.

Turn 2: generated by REGENERATE; next action REGENERATE (reviewer score 4.82)

Generation prompt. A photorealistic image depicts a vibrant red city bus (sleek modern design, large windows) resting on top of a small, weathered wooden rowboat. The bus is positioned directly atop the boat, which floats on calm, rippled water that reflects the bus's underside and the boat's weathered surface. The background features a clear sky with soft clouds, enhancing the surreal, dreamlike atmosphere. The scene contrasts the bus's sleek design with the boat's rustic texture, ideal for artistic photography, capturing the bus - supported - by - boat juxtaposition (not floating mid - air).
Reviewer diagnosis. The image correctly meets the core requirement of positioning a bus above a boat, with high overall visual quality. The only flaw is that the small wooden rowboat is physically implausible to support the large size and weight of the full-sized passenger bus, which violates real-world physical consistency.

Turn 3: generated by REGENERATE; next action STOP (reviewer score 4.85)

Generation prompt. A photorealistic image depicts a vibrant red city bus (sleek modern design, large windows) resting on a large, sturdy, weathered wooden boat (replacing the small rowboat) that can support the bus's weight. The bus is positioned atop the larger, robust vessel, which floats on calm, rippled water reflecting the bus's underside and the boat's weathered surface. The background features a clear sky with soft clouds, enhancing the surreal, dreamlike atmosphere. The scene contrasts the bus's sleek design with the boat's rugged, sturdy texture, ideal for artistic photography, capturing the bus - supported - by - larger - boat juxtaposition (not floating mid - air).
Reviewer diagnosis. The image fully meets the user's requirement of showing a bus positioned above a boat, with no critical flaws. It has natural lighting, accurate color reproduction, clear subject details, and harmonious composition with the background. There are only very minor shortcomings: it lacks ultra-high resolution micro-texture details, which does not affect the overall presentation and compliance with the request.

K.3 Case 3: Spatial Relation with Refinement

Original prompt. a photo of a frisbee right of a motorcycle.



One-shot baseline
generator only



Turn 1: initial T2I
score 3.40
next: REFINE



Turn 2: by REFINE
score 4.70
next: REGENERATE



Turn 3: by REGENERATE
score 4.79
next: STOP

Navigator turn details.

Turn 1: initial T2I; next action REFINE (reviewer score 3.40)

Generation prompt. A photorealistic photo depicts a sleek motorcycle (metallic silver, black accents, chrome wheels) on the left, with a vibrant circular frisbee (bright yellow, smooth surface) positioned to its right. The backdrop is a neutral outdoor setting (e.g., paved road or grassy field), lit by natural sunlight. The motorcycle is stationary, while the frisbee rests or floats, creating a clear left - right spatial arrangement. Details include the motorcycle's streamlined body, handlebars, and the frisbee's glossy texture, rendered with sharp, realistic detail for a high - quality image.
Reviewer diagnosis. While the frisbee is correctly positioned to the right of the motorcycle as requested, the frisbee is unnaturally floating in mid-air above the road instead of being placed naturally in the scene. The frisbee is obviously artificially added, poorly integrated with the environment, and looks inconsistent with the realistic style of the background and motorcycle, creating a jarring visual effect.

Turn 2: generated by REFINE; next action REGENERATE (reviewer score 4.70)

Generation prompt. Make the bright yellow frisbee lie on the paved road to the right of the motorcycle, ensuring it appears naturally placed and integrated with the road surface and the outdoor setting.
Reviewer diagnosis. The image correctly fulfills the core request: the frisbee is placed to the right of the motorcycle, and both required objects are present. There are minor AI generation flaws: the motorcycle incorrectly has three side mirrors instead of the standard two, and the model text on the motorcycle body is garbled unreadable gibberish. The rest of the image has good technical quality, natural lighting and clear details.

Turn 3: generated by REGENERATE; next action STOP (reviewer score 4.79)

Generation prompt. A photo of a silver motorcycle with two standard side mirrors and clear legible model text, parked on a paved road. A bright yellow frisbee lies naturally on the road to the right of the motorcycle. The background features a grassy field with distant trees under a clear sky, with natural lighting and sharp details. Photorealistic style, ensuring the motorcycle's mirror count and text are accurate.
Reviewer diagnosis. The image correctly meets the requirement: the yellow frisbee is placed to the right of the motorcycle, as requested. Only a minor subtle flaw exists: the text branding on the motorcycle's body is garbled/distorted, which is a small AI generation artifact. The image has good overall technical quality, clear details, harmonious natural colors and proper lighting.

K.4 Case 4: Counting with Refinement and Regeneration

Original prompt. a photo of three benches.



One-shot baseline
generator only



Turn 1: initial T2I
score 2.60
next: REFINE



Turn 2: by REFINE
score 2.69
next: REGENERATE



Turn 3: by REGENERATE
score 4.85
selected at budget end

Navigator turn details.

Turn 1: initial T2I; next action REFINE (reviewer score 2.60)

Generation prompt. Three weathered brown wooden benches with metal armrests and legs are arranged in a park - like setting: one in the foreground, one in the mid - ground, and one in the background, creating a sense of depth. Lush green grass, trees with autumnal foliage, and a clear blue sky form the backdrop. Each bench has a slatted seat and backrest, with subtle wood wear and slightly rusted metal, suggesting regular use. Styled as a high - resolution, naturalistic photograph, it is ideal for showcasing outdoor furniture in a peaceful environment.
Reviewer diagnosis. The user requested a photo of three benches, but the current image only contains two park benches, failing to meet the required quantity of benches, which is the key flaw. The image itself has good overall visual quality, with no obvious technical or aesthetic defects.

Turn 2: generated by REFINE; next action REGENERATE (reviewer score 2.69)

Generation prompt. Add a third weathered brown wooden bench with metal armrests and legs in the background, positioned behind the mid - ground bench to maintain the depth arrangement, matching the style and weathered appearance of the existing two benches.
Reviewer diagnosis. The image only contains 2 park benches; it fails to meet the user's explicit requirement of three benches, with the third bench completely missing from the scene. There are no other significant technical or aesthetic flaws to the image itself.

Turn 3: generated by REGENERATE; selected at budget end (reviewer score 4.85)

Generation prompt. Add a third weathered brown wooden park bench with metal armrests and legs, positioned behind the mid - ground bench (the right - hand bench) to form a three - bench depth arrangement, matching the weathered appearance and style of the existing two benches, set in a park with autumn - hued trees and green grass, under natural daylight.
Reviewer diagnosis. The rightmost (third) bench is partially cropped by the image boundary, with only half of its structure visible in the frame. There are no other significant technical, color, lighting, or composition flaws; the image is otherwise sharp, well-exposed, and harmoniously composed.
Termination. The next suggested action was a REFINE to fix cropping, but the turn budget was exhausted; the output selector keeps this best-scored candidate.

K.5 Case 5: Scientific Reasoning

Original prompt. A cup of icy coffee in warm humid air.



One-shot baseline
generator only



Turn 1: initial T2I
score 3.10
next: REGENERATE



Turn 2: by REGENERATE
score 4.74
next: REFINE



Turn 3: by REFINE
score 4.91
next: STOP

Navigator turn details.

Turn 1: initial T2I; next action REGENERATE (reviewer score 3.10)

Generation prompt. A clear glass cup of iced coffee—filled with dark coffee and ice cubes—rests in a warm, humid atmosphere. Condensation droplets glisten on the cup's exterior, formed as the cool drink interacts with the moist, sunlit air. The air around the cup is heavy with humidity, creating a hazy, tropical ambiance (e.g., a sunlit patio) to emphasize the heat. The scene, rendered in a photorealistic style, captures the contrast between the cold, refreshing coffee and the warm, humid surroundings, with detailed condensation and atmospheric moisture.
Reviewer diagnosis. The image has a core content error that conflicts with the request: visible rising steam is incorrectly added to the iced coffee. Cold icy coffee cannot produce visible steam in warm humid air; steam is a characteristic of hot beverages, not iced coffee. While the condensation droplets on the glass correctly reflect the humid environment, the erroneous steam directly contradicts the icy-coffee requirement.

Turn 2: generated by REGENERATE; next action REFINE (reviewer score 4.74)

Generation prompt. A clear glass cup of iced coffee—filled with dark coffee and ice cubes—rests on a sunlit wooden patio table in warm, humid air. Condensation droplets glisten on the cup's exterior, formed as the cool drink interacts with the moist, sunlit air. The air around the cup is heavy with humidity, creating a hazy, tropical ambiance (lush green foliage in the background) to emphasize the heat. The scene, rendered in a photorealistic style, captures the contrast between the cold, refreshing coffee and the warm, humid surroundings, with detailed condensation and atmospheric moisture—no visible steam (since iced coffee does not produce steam in warm air). The table shows water droplets from condensation, and the background features sunlit greenery to enhance the tropical, humid setting.
Reviewer diagnosis. There are no significant flaws that violate the request. Only minor softening of the out-of-focus background foliage remains; the image accurately presents icy coffee in warm humid air, with proper condensation on the glass that demonstrates the physical effect of a cold beverage in a humid warm environment.

Turn 3: generated by REFINE; next action STOP (reviewer score 4.91)

Generation prompt. Make the background foliage more sharply defined and vibrant to reduce the softening effect, enhancing the tropical, humid ambiance.
Reviewer diagnosis. There are no significant flaws. Only a minor imperfection remains: a small number of liquid stains on the wooden table look slightly artificial, but this does not affect the overall presentation. The image correctly conveys the effect of iced coffee in warm humid air, with accurate condensation on the glass and an appropriate warm natural environment.

K.6 Case 6: Spatial Relation

Original prompt. a photo of a book above a laptop.



One-shot baseline
generator only



Turn 1: initial T2I
score 2.75
next: REGENERATE



Turn 2: by REGENERATE
score 4.41
next: REGENERATE



Turn 3: by REGENERATE
score 4.91
next: STOP

Navigator turn details.

Turn 1: initial T2I; next action REGENERATE (reviewer score 2.75)

Generation prompt. A photorealistic image depicts a brown hardcover book positioned vertically above a sleek silver laptop, creating a stacked arrangement. The laptop, open with its screen facing upward, rests on a wooden desk, its keyboard and trackpad visible. The book, with a textured cover and "BOOK" embossed on the spine, sits atop the laptop's lid, slightly tilted to suggest balance. The desk features subtle details: a pen and a coaster, adding realism. Natural lighting highlights the book's texture and the laptop's metallic sheen, rendered in a detailed, photorealistic style to emphasize the spatial relationship between the two objects.
Reviewer diagnosis. The image incorrectly includes two stacked laptops (an open base laptop with an extra closed laptop stacked on top of it) instead of the requested single laptop with the book positioned above it, which deviates from the original user requirement. The image itself has good overall aesthetic and technical quality, only the core element composition does not match the request.

Turn 2: generated by REGENERATE; next action REGENERATE (reviewer score 4.41)

Generation prompt. A photorealistic image depicts a brown hardcover book positioned vertically above a single sleek silver laptop, with the laptop open (screen facing upward) on a wooden desk. The laptop, with its keyboard and trackpad visible, rests on the desk. The book, featuring a textured cover and "BOOK" embossed on the spine, sits atop the laptop's lid, slightly tilted to suggest balance. The desk includes a pen and a coaster, adding realism. Natural lighting highlights the book's texture and the laptop's metallic sheen, rendered in a detailed, photorealistic style to emphasize the spatial relationship between the book and the single laptop.
Reviewer diagnosis. The book is unrealistically levitating/floating unsupported above the laptop keyboard, creating a physically impossible, unnatural arrangement which is an AI generation artifact. Aside from this issue, the image has good overall technical quality, harmonious colors and clear details.

Turn 3: generated by REGENERATE; next action STOP (reviewer score 4.91)

Generation prompt. A photorealistic image depicts a brown hardcover book resting (not floating) on the lid of a single sleek silver laptop, with the laptop open (screen facing upward) on a wooden desk. The laptop, with its keyboard and trackpad visible, rests on the desk. The book, featuring a textured cover and "BOOK" embossed on the spine, sits atop the laptop's lid balanced naturally, no floating on the desk. The desk includes a pen and a coaster, adding realism. Natural lighting highlights the book's texture and the laptop's metallic sheen, rendered in a detailed, photorealistic style to emphasize the correct spatial relationship between the book and the single laptop.
Reviewer diagnosis. No significant flaws. The image correctly meets all requirements: it shows a book placed above a laptop as requested, has clear sharp details, natural soft lighting, harmonious natural colors, and good integration between the subject and the background.

K.7 Case 7: Two-Object Repair

Original prompt. a photo of a toothbrush and a bench.



One-shot baseline
generator only



Turn 1: initial T2I
score 3.30
next: REGENERATE



Turn 2: by REGENERATE
score 3.30
next: REGENERATE



Turn 3: by REGENERATE
score 4.85
next: STOP

Navigator turn details.

Turn 1: initial T2I; next action REGENERATE (reviewer score 3.30)

Generation prompt. A photorealistic image features a toothbrush and a bench in a serene outdoor park setting. The bench, crafted from weathered wooden planks with metal armrests, rests on a grassy lawn dotted with fallen leaves and small shrubs. The toothbrush, with a white handle and blue bristles, lies on the bench's seat, positioned centrally, its bristles slightly angled. Soft natural light casts gentle shadows, highlighting the contrast between the everyday toothbrush and the rustic, weathered bench, creating a calm, everyday-scene aesthetic.
Reviewer diagnosis. The request asks for one toothbrush and a bench, but this image contains two toothbrushes. One toothbrush is incorrectly generated, being abnormally embedded into the left metal armrest of the bench, which is physically implausible and is a generation artifact.

Turn 2: generated by REGENERATE; next action REGENERATE (reviewer score 3.30)

Generation prompt. A photorealistic image features a single toothbrush and a bench in a serene outdoor park setting. The bench, crafted from weathered wooden planks with metal armrests, rests on a grassy lawn dotted with fallen leaves and small shrubs. The toothbrush, with a white handle and blue bristles, lies on the bench's seat not embedded in the armrest, positioned centrally, its bristles slightly angled. Soft natural light casts gentle shadows, highlighting the contrast between the everyday toothbrush and the rustic, weathered bench, creating a calm, everyday-scene aesthetic.
Reviewer diagnosis. The toothbrush is incorrectly generated as a distorted, impossible malformed object: it has a single brush head connected to two separate full handles, which does not match the normal structure of a standard toothbrush requested. The bench is correctly presented, the background and overall image technical quality are good.

Turn 3: generated by REGENERATE; next action STOP (reviewer score 4.85)

Generation prompt. A photorealistic image features a standard toothbrush (single white handle, single blue-bristled brush head) and a weathered wooden bench with metal armrests in a serene outdoor park. The bench rests on a grassy lawn dotted with fallen leaves and small shrubs. The toothbrush lies centrally on the bench's seat not embedded in the armrest, its bristles slightly angled. Soft natural light casts gentle shadows, highlighting the contrast between the everyday toothbrush and the rustic, weathered bench, creating a calm, everyday-scene aesthetic.
Reviewer diagnosis. Both requested subjects (a toothbrush and a park bench) are present and clearly visible in the image. There are no major technical or content flaws: the image has sharp, clear details, natural accurate exposure, harmonious natural outdoor coloring, no significant artifacts, noise or distortion. All requirements of the request are fully met.

L Human Evaluation

We conduct a pairwise human evaluation as an empirical validation of the automatic reviewer used throughout training and evaluation. The goal is not to replace benchmark evaluation, but to test whether reviewer-induced preferences are reasonably aligned with human judgments on generated images.

Sample construction. We sample some image pairs from two pools: T2I-ReasonBench and GenEval results. For each image, the reviewer produces a scalar score

$$\rho = 0.3 \cdot \text{Visual Quality} + 0.7 \cdot \text{Instruction Comprehension}.$$

Pairwise reviewer preferences are derived from score differences: a pair is treated as a reviewer tie when $|\rho_A - \rho_B| < 0.3$, and otherwise the image with the higher ρ is preferred. To cover both easy and ambiguous cases, pairs are sampled across four score-gap bins: tie (< 0.3), small (0.3–0.8), medium (0.8–1.6), and large (≥ 1.6).

Annotation protocol. Annotators are shown the original prompt and two anonymized candidate images in randomized left-right order. Source run information, turn identifiers, and reviewer scores are hidden. When available, prompts are displayed in both English and Chinese. Each annotator selects whether the left image is better, the right image is better, or the two images are tied. The interface assigns each annotator a deterministic random subset of 40 pairs.

Metric and result. Agreement is computed by comparing the human pairwise preference with the reviewer-induced pairwise preference. In the current study, we collect 320 annotations from eight annotators. The MLLM reviewer's relative ranking agrees with human preference on approximately 70.3% of decisive non-tie comparisons. Given the subjectivity of multimodal evaluation and the ambiguity of near-tie image pairs, this result suggests that the reviewer provides a usable trajectory-optimization signal, while still leaving room for richer human studies in future evaluations.

M Limitations and Future Directions

Like all agentic workflows, Generation Navigator inherits two structural costs compared with single-shot generation: additional inference latency from iterative action making, and dependence on an external reviewer as the environment signal. We mitigate both within the current framework. For latency, the no-CoT navigator variant reduces per-turn action-making overhead by 2.4× while still achieving an 18.3-point accuracy improvement compared to the baseline (Section 4.6). The turn budget is also user-configurable; even with a small number of turns, the navigator substantially outperforms single-shot baselines (Appendix I), providing a practical operating point that balances quality and cost. Meanwhile, compared with the recent agent-based method IRG [Huang et al., 2026], the no-CoT Navigator is 4.7× faster on WISE while improving the overall score from 0.77 to 0.88, suggesting that our closed-loop design offers a more favorable quality–latency trade-off than prior agentic workflows (Appendix J). For reviewer dependence, the learned state-conditioned action policy is not tied to any single scorer: replacing the inference-time reviewer across five architecturally distinct models yields only a 3.03-point spread (Section 4.5), preserving a large gain over one-shot generation and confirming that the navigator acquires general visual-state action logic rather than reviewer-specific preferences. This decoupling also suggests a natural upgrade path: as foundational multimodal models continue to advance, stronger reviewers can be plugged in at inference time to provide richer feedback without retraining the navigator.

LLM Usage

We acknowledge the use of a large language model (LLM) to assist in the preparation of this manuscript. The LLM's role was strictly limited to improving grammar and refining language. It did not contribute to any of the core research components, such as the initial ideas, experimental design, data analysis, or interpretation of the results.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the central claims: complex T2I generation is formulated as a state-conditioned action-making problem, Generation Navigator outputs structured actions among STOP, REFINE, and REGENERATE, and PRE-GRPO optimizes trajectory-level quality, retention, and efficiency. These claims are supported by experiments on T2I-ReasonBench, WISE, and GenEval, together with ablations on reward components, action behavior, sampling budget, and transfer across components.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Appendix M.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present formal theoretical results, theorems, or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper discloses the main components needed to reproduce the experimental results, including the navigator, generator, reviewer, training stages, trajectory construction procedure, reward definition, benchmark suite, maximum turn budget, reward weights, prompts, contamination checks, and evaluation protocols. Additional implementation details and prompt templates are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The current submission does not provide open access to the full codebase. The code will be made open-source upon completion of the review process.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the main experimental setup, including the Qwen3-VL-8B-Instruct navigator, Doubao-Seed1.5 reviewer, FLUX.2-Klein-9B generator, 103K trajectories, one SFT epoch, 300 PRE-GRPO optimization steps, maximum turn budget of three, reward weights, evaluated benchmarks, baseline categories, and appendix-level hyperparameter studies.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports inference latency and quality–latency trade-offs for the agentic generation pipeline, including full-CoT and no-CoT navigator variants and the effect of different maximum turn budgets.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [N/A]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper credits the existing models, datasets, and benchmarks used in the experiments, including Qwen3-VL, FLUX.2-Klein, Doubao-Seed, T2I-ReasonBench, WISE, GenEval, and other baseline systems, with citations and model or benchmark references where applicable.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We illustrate the usage of the LLM in the appendix, which is mainly employed for polishing and revision.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.