SIDiffAgent: Training-Free Text-to-Image Generation via Self-Improving Agentic Refinement

Anonymous ACL submission

Abstract

Text-to-image diffusion models have become a cornerstone of generative AI, producing photo-realistic imagery at a quality that is reshaping creative, commercial, and design workflows. Yet their reliability in production settings is undercut by a persistent gap between what a user asks for and what the model renders: outputs are highly sensitive to prompt phrasing, semantically ambiguous terms are resolved inconsistently (for instance, "mouse" as an animal versus a computer peripheral), anatomical and structural artifacts are common, and dependable results often demand carefully engineered prompts. Prior approaches typically mitigate these issues through additional training or offer only coarse control, which limits their adaptability and raises the cost of deployment. We introduce the Self-Improving Diffusion Agent (*SIDiffAgent*), a training-free agentic framework built on the Qwen family of models (Qwen-VL, Qwen-Image, Qwen-Edit, and Qwen-Embedding). *SIDiffAgent* autonomously handles prompt engineering, detects and corrects low-quality generations, and performs fine-grained artifact removal to yield more consistent and faithful outputs. It further improves over time through an experience memory that stores past generation trajectories in a database and injects retrieval-conditioned guidance into every stage of the pipeline. On GenAI-Bench, *SIDiffAgent* attains an average VQA score of 0.884, substantially outperforming leading open-source models, proprietary systems, and prior agentic methods. We will release our code publicly upon acceptance.

1 Introduction

Image generation models (Rombach et al., 2021; Ramesh et al., 2022; Saharia et al., 2022a; Singer et al., 2022; Ho et al., 2022; Blattmann et al., 2023; Esser et al., 2023) can synthesize high-quality images from text descriptions. Despite rapid progress in visual quality, state-of-the-art text-to-image diffusion models remain highly sensitive to the precise phrasing of the input prompt. A central failure mode of text-to-image diffusion models is inconsistent prompt interpretation: small changes in phrasing produce large changes in output, and short or ambiguous prompts produce outputs that are unfaithful to the prompt's explicit content. For example, "a mouse" may yield an animal or a peripheral; "four cars on the street" frequently yields three or five; "no umbrellas" is frequently ignored. We refer to this gap between what the prompt says and what the image shows as the prompt-fidelity gap.

A second challenge is the distributional mismatch between the free-form language of end-user prompts and the curated, heavily filtered captions used to train diffusion models (Li et al., 2024d; Wu et al., 2025a; Labs, 2024). Because training captions are dense and descriptive while user prompts are short and casual, the model receives weak textual guidance at inference time. End-user prompts are frequently underspecified (for example, "a car on a road" with no model, color, or setting), which forces the model to make unguided assumptions, degrades output quality, and increases cost as users regenerate repeatedly. This is a central bottleneck for the reliable use of diffusion models in professional settings, where predictable alignment with user intent is essential.

Prior agentic frameworks such as T2I-Copilot (Chen et al., 2025a) and MCCD (Li et al., 2025a) address prompt ambiguity through image editing and agent coordination, but they offer limited fine-grained control and weak post-generation artifact correction. Coordination is itself a core difficulty in multi-agent systems: agents have only partial observability and limited knowledge of other agents' capabilities or the effects of their own actions (Cemri et al., 2025). Coordination is itself a core difficulty in multi-agent systems: agents have only partial observability and limited

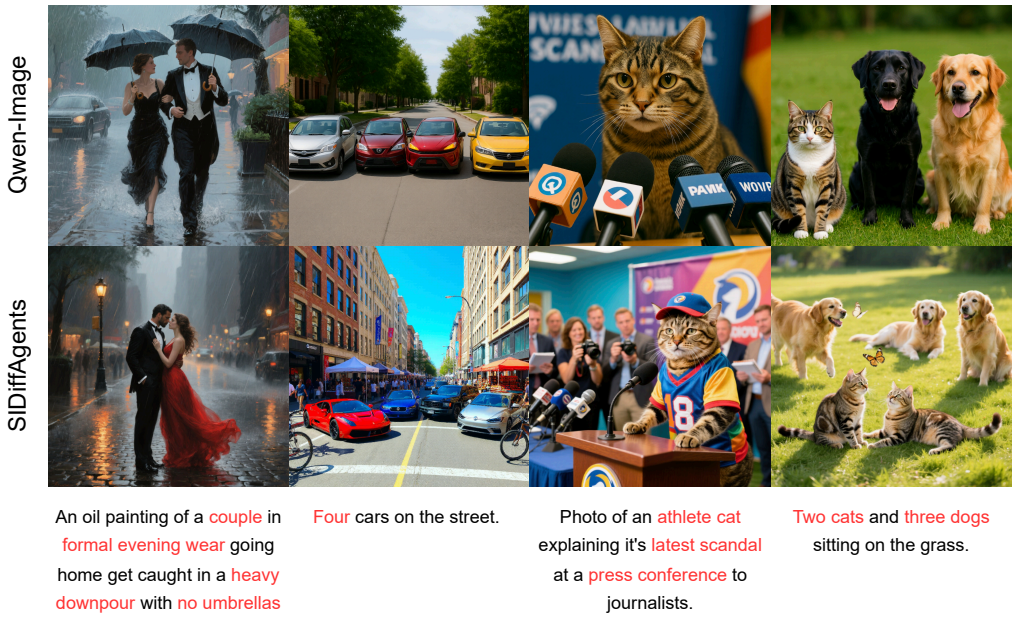


Figure 1: Qualitative comparison between *SIDiffAgent* and the state-of-the-art Qwen-Image model. *SIDiffAgent* improves text-to-image alignment and realism, handling negative constraints ('no umbrellas'), complex scene composition, anthropomorphic roles ('athlete cat'), and precise object counting.

knowledge of other agents' capabilities or the effects of their own actions. A natural mitigation, common in multi-agent LLM systems, is to share contextual information across agents, descriptions of each agent's role, prior outcomes, and recurring failure modes so that local decisions account for downstream effects. To our knowledge, this kind of cross-agent context sharing has not been systematically explored for diffusion-based agentic pipelines.

ReNeg (Li et al., 2025b) shows the value of negative prompts and embeddings (Ho and Salimans, 2022) for steering diffusion away from undesired artifacts. We introduce *SIDiffAgent*, a training-free multi-agent framework that builds a self-improving, ToM-informed generation pipeline on the Qwen model family (Qwen-VL (Bai et al., 2025), Qwen-Image (Wu et al., 2025a), Qwen-Edit (Wu et al., 2025a), Qwen-Embedding (Zhang et al., 2025)). Its agents refine prompts, reuse past experience, and fix artifacts with local edits, and following ReNeg it adds an adaptive negative-prompt agent. Our contribution is at the system level rather than a new training objective, with three novel components. First, *SIDiffAgent* provides the first experience-driven memory for diffusion-based agents, giving retrieval-conditioned guidance throughout the workflow. Second, it makes each sub-agent aware of the role and recurring failure modes of the other stages, summarized as textual guidance from accumulated success and pitfall patterns, so that local decisions account for downstream effects rather than prompt quality alone. Third, *SIDiffAgent* is a training-free alternative to RL-based or reward-guided finetuning (for example, DPO-Diffusion, ReNeg), improving quality even for closed-weight diffusion systems without retraining.

We validate *SIDiffAgent* on GenAI-Bench and DrawBench, where it sets a new state of the art. It surpasses the prior agentic system T2I-Copilot by 8.73% in VQA Score (Lin et al., 2024), improves over the leading proprietary model Imagen 3 by 5.36%, and exceeds strong open-source systems such as SD 3.5 by 15.70%.

Our key contributions are summarized as follows:

- We introduce a multi-agent system that refines diffusion-model inputs at test time, producing a consistent interpretation of a prompt across phrasings so that generations more faithfully capture the user's intended concept.

- We adapt a structured image-editing approach producing more faithful adherence to the explicit content of the prompt, including under-specified elements, counts, spatial relations, and negations, across paraphrases of the same prompt.

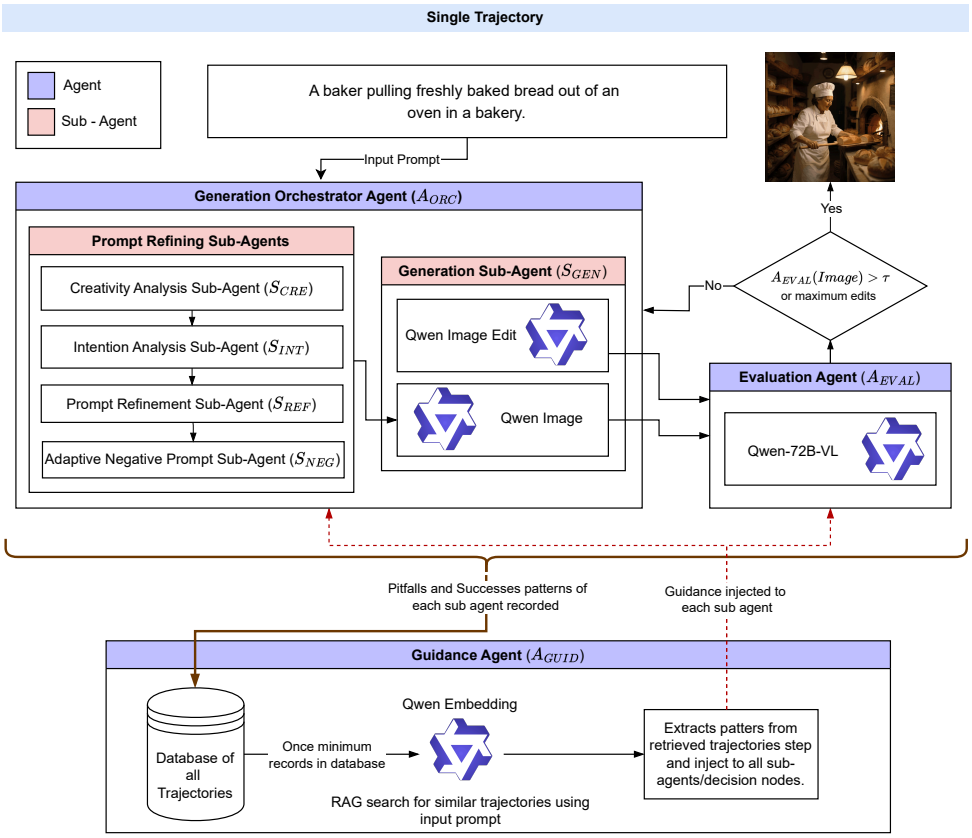


Figure 2: Workflow of the Self-Improving Diffusion Agent (*SIDiffAgent*). An input prompt is processed by the Generation Orchestrator Agent (A_{ORC}), which employs sub-agents (S_{CRE}, S_{INT}, S_{REF}, S_{NEG}) to assess creativity, clarify intent, refine the prompt, and add adaptive negative constraints before generation (S_{GEN}). The Evaluation Agent (A_{EVAL}) scores the generated image on aesthetic quality and text-to-image alignment, and triggers targeted refinements if the score is below a fixed threshold τ. Each trajectory is stored in the knowledge base, where the Guidance Agent (A_{GUID}) stores pitfalls and successes into corrective and workflow guidance, which are injected back into decision nodes to improve future generations.

- We build an iterative self-improvement mechanism in which an agentic memory tracks the successes, artifacts, and pitfalls of each agent and injects corrective, injects corrective, role-aware guidance, yielding a training-free framework that improves generation quality over time.

2 Related Works

2.1 AI Agents

Agents are autonomous entities that perceive their environment, reason over goals, and act, often collaboratively, through memory, planning, and tool use (Luo et al., 2025). In applications such as browser automation or coding (Hong et al., 2024; Qian et al., 2023) they typically use an LLM backbone that reasons over the current state to choose an action; multi-agent variants additionally reason about other agents' states (Li et al., 2023). This has motivated agentic workflows for image generation, where multi-agent pipelines decompose and refine prompts, evaluate outputs, and correct artifacts. Examples include T2I-Copilot (Chen et al., 2025a), which integrates prompt interpretation, model selection, and quality evaluation, and PromptSculptor (Xiang et al., 2025), which iteratively transforms vague user prompts into precise ones. These systems show that agentic coordination can reduce the intent gap in text-to-image generation, improving fidelity and controllability without additional model training.

2.2 Prompt Optimization for Diffusion Models

Prompt Optimization (PO) aligns LLMs with human intent and improves downstream performance, evolving from in-context learning into automatic prompt engineering. It is well studied for LLMs but comparatively underexplored for diffusion

models. Promptist (Hao et al., 2023) fine-tunes an LLM with RLHF to add stylistic modifiers, OPT2I (Manas et al., 2024) uses an LLM to refine prompts for higher-quality synthesis, and DPO-Diffusion (Wang et al., 2024a) and ReNeg (Li et al., 2025b) optimize negative prompts through LLM reasoning or learned embeddings. TextGrad (Yuksekgonul et al., 2024) introduces gradient-based text feedback optimization, and manual prompt-engineering strategies have also been studied systematically. *SIDiffAgent* builds on these ideas by jointly optimizing positive and image-specific negative prompts at test time, dynamically balancing guidance to produce more reliable and semantically aligned images without additional model training.

2.3 Self-Improving AI

Self-improving systems enhance their own performance from interaction or feedback. For LLMs, LMSI (Huang et al., 2024) uses reward signals, IPO (Garg et al., 2025) and Self-Rewarding (Yuan et al., 2024) construct preference pairs, and SPPO (Wu et al., 2024b) applies self-play. Beyond LLMs, MaestroMotif (Klissarov et al., 2024) uses feedback-driven skill acquisition, AgentSquare (Shang et al., 2024) modularizes agent design, AlphaEvolve (Novikov et al., 2025) and ADAS (Hu et al., 2025) use meta-agents to improve task-specific agents, and SICA (Robeyns et al., 2025) lets an agent edit its own codebase. Diffusion-based agents remain unexplored in this setting; to our knowledge we present the first training-free, self-improving diffusion framework, where agentic memory and prompt engineering enable improvement without retraining. Closely related multi-agent systems pair prompt decomposition with an evaluate-then-correct loop (T2I-Copilot (Chen et al., 2025a), PromptSculptor (Xiang et al., 2025)), reward-guided methods learn per-image negative *embeddings* (ReNeg (Li et al., 2025b)), and retrieval over stored trajectories is an established agent-memory pattern. The contribution of *SIDiffAgent* is to combine decomposed prompt construction, a training-free natural-language adaptive negative prompt, and a persistent cross-generation self-improvement memory in a single closed-loop diffusion agent, whereas each prior system supplies only one of these in isolation.

3 Proposed Method: Self-Improving Diffusion Agent (*SIDiffAgent*)

Our framework is an agentic workflow coordinating prompt engineering, image generation, evaluation, and self-improvement. It is orchestrated by three agents: the Generation Orchestrator Agent (A_{ORC}), which preprocesses the user prompt through five sub-agents (Creativity Analysis S_{CRE}, Intention Analysis S_{INT}, Prompt Refinement S_{REF}, Adaptive Negative Prompt S_{NEG}, and Generation S_{GEN}); the Evaluation Agent (A_{EVAL}), which assesses each output and triggers editing when needed; and the Guidance Agent (A_{GUID}), which enables learning from past experience. The design deliberately separates intent understanding, generation, quality control, and experience reuse, so that each concern is handled by a specialized agent and can be analyzed and improved independently. Figure 2 shows the workflow and Appendix L gives the algorithm. Each agent is described below.

3.1 Generation Orchestrator Agent (A_{ORC})

The Generation Orchestrator Agent forms the entry point of the pipeline. It performs a structured preprocessing sequence that balances user intent, creative enrichment, and model compatibility through five specialized sub-agents, each addressing a distinct aspect of prompt engineering and generation. We describe them in order below; the prompt for each sub-agent is in Appendix K.1.

Step 1. Creativity Analysis Sub-Agent (S_{CRE}): This sub-agent estimates how much freedom later stages may take, which sets the intensity of refinement. It assigns one of three levels: high for short or vague prompts that need substantial enrichment, medium for moderately detailed prompts that need a balance between added structure and faithful interpretation, and low for highly detailed prompts where strict adherence to the given description matters most. This level then controls how aggressively the later sub-agents rewrite and extend the prompt. For example, 'a cat' is high creativity, since attributes such as breed, pose, and environment are unspecified, whereas 'a Persian cat sitting on a red sofa in a living room' is low creativity because it already provides detailed constraints. Figure 8 gives prompt and image pairs for each level.

Step 2. Intention Analysis Sub-Agent (S_{INT}): Given the creativity level, this sub-agent produces a semantic layout of the prompt, extracting the *main subjects*, their *attributes* and *spatial relationships*,

the *background*, and parameters such as *composition*, *lighting*, and *visual style*.

It also resolves ambiguities and missing information conditioned on the level from S_{CRE}: for low-creativity prompts it adds only minor assumptions, while for high-creativity prompts it autonomously infers coherent additional details. The output is a structured specification of the user's intent together with explicit clarifications. For 'a car' at low creativity, S_{INT} keeps the minimal description; at high creativity it adds attributes such as model and color. These additional details are inferred autonomously so that the enriched prompt stays consistent with the original request rather than introducing unrelated content.

Step 3. Prompt Refinement Sub-Agent (S_{REF}): Using the semantic layout from S_{INT}, S_{REF} rewrites the prompt into a clearer, contextually richer form while preserving all original subjects and scene elements, and thus the user's intent. The extent of rewriting is modulated by the creativity level, so low-creativity prompts are altered minimally while high-creativity prompts receive richer descriptive structure.

Step 4. Adaptive Negative Prompt Sub-Agent (S_{NEG}): Under classifier-free guidance, a diffusion model contrasts the conditional prompt with an unconditional one (Ho and Salimans, 2022); replacing the unconditional branch with a negative text embedding steers generation away from undesired features. This gives fine-grained control over the output, since unwanted objects, styles, or quality defects can be named directly rather than only implied by the positive prompt. The refined prompt specifies what should appear; this step specifies what must not.

S_{NEG} first appends universal safeguards against common artifacts (for example, 'low quality, blurry, distorted, watermark'), then derives scene-specific negations from the refined prompt. A "clear blue sky" adds "clouds, dark clouds"; "a person standing alone" adds "extra people or crowd". These constraints are concise and context-aware, improving quality and adherence without altering user intent.

Step 5. Generation Sub-Agent (S_{GEN}): This sub-agent generates the image using two models. It first uses Qwen-Image to generate from the refined positive and adaptive negative prompts, after which evaluation begins (Section 3.2). If A_{EVAL} reports defects or misalignment, S_{GEN} invokes Qwen-Image-Edit for targeted corrections, iteratively improving fidelity to complex specifications. Combining a strong generator with a dedicated editing model lets the system correct localized errors without discarding an otherwise good image, which matters for prompts with many simultaneous constraints. Appendix 8 shows the targeted editing.

3.2 Evaluation Agent (A_{EVAL})

A_{EVAL} assesses each generated image and outputs a structured report based on two metrics, **Aesthetic Quality** and **Text-Image Alignment**, each scored from 0 to 10. Its prompt is in Appendix K.2.

Aesthetic Quality reflects the overall visual merit of the image, while Text-Image Alignment measures how faithfully the image follows the prompt. Aesthetic Quality covers composition, color harmony, lighting, focus, sharpness, emotional impact, and uniqueness. Text-Image Alignment checks subject presence, spatial correctness, style adherence, and background consistency. The overall score is the average of the two. Beyond scoring, the agent detects visual artifacts (for example, noise, tiling, blending errors), identifies missing elements, and provides targeted suggestions that drive the editing step.

If the score is below a fixed threshold, S_{GEN} enters an editing loop: the image is edited with Qwen-Image-Edit using a refined prompt produced by A_{ORC} from the suggestions of A_{EVAL}. This repeats until the image reaches a satisfactory score or the maximum number of attempts is reached.

3.3 Guidance Agent (A_{GUID})

The Guidance Agent provides the self-improving component. A trajectory is a complete generation sequence from initial prompt to final image, and decision nodes are points where sub-agents take key actions. A_{GUID} condenses each trajectory into pitfalls and successes per decision node and accumulates them in a persistent knowledge base, which exposes recurring patterns such as a sub-agent that consistently fails on a prompt type. For a new input, A_{GUID} retrieves relevant trajectories and supplies each sub-agent with corrective guidance (from past pitfalls and successes) and workflow guidance (a description of the process). We hypothesize that this combination gives each sub-agent richer context about likely failure modes and successful strategies, and therefore leads to more reliable decisions on new prompts. The prompts for A_{GUID} are in Appendix K.3.

1. Knowledge Base Construction: A persistent global memory records each completed tra-

Method	GenAI-Bench													DrawBench
	Basic						Advanced							
	Attribute	Scene	Relation			Overall	Count	Differ	Compare	Logical		Overall	Overall	
			Spatial	Action	Part					Negate	Universal			
Proprietary														
Imagen 3 v002 (Baldridge et al., 2024b)	0.909	0.923	0.909	0.903	0.918	0.912	0.841	0.841	0.795	0.673	0.788	0.776	0.839	<u>0.866</u>
(Task completion rate)	(92.4%)	(93.1%)	(93.0%)	(89.3%)	(88.7%)	(92.3%)	(91.5%)	(88.9%)	(90.7%)	(88.8%)	(89.1%)	(90.7%)	(91.4%)	(97.0%)
Recraft v3 (Recraft, 2024)	0.914	0.913	0.913	0.901	0.913	0.913	0.806	0.797	0.772	0.589	0.761	0.725	0.811	0.836
FLUX1.1-pro (Labs, 2024)	0.890	0.899	0.884	0.871	0.894	0.884	0.766	0.788	0.751	0.490	0.710	0.666	0.766	0.786
Midjourney v6 (Midjourney, 2024)	0.880	0.870	0.870	0.870	0.910	0.870	0.780	0.780	0.790	0.500	0.760	0.690	0.772	-
DALL-E 3 (OpenAI, 2024)	0.910	0.900	<u>0.920</u>	0.890	0.910	0.900	0.820	0.780	<u>0.820</u>	0.480	<u>0.800</u>	0.700	0.791	-
Open-source														
Kolors v1.0 (Team, 2024)	0.821	0.841	0.832	0.818	0.803	0.819	0.737	0.726	0.705	0.438	0.695	0.621	0.711	0.646
Playground v2.5 (Li et al., 2024b)	0.818	0.850	0.803	0.818	0.821	0.815	0.732	0.696	0.721	0.499	0.695	0.640	0.720	0.743
HunyuanDiT v1.2 (Li et al., 2024c)	0.817	0.855	0.825	0.827	0.798	0.818	0.732	0.723	0.743	0.475	0.692	0.640	0.721	0.712
Janus Pro-7B (Chen et al., 2025b)	0.865	0.886	0.867	0.856	0.870	0.859	0.731	0.759	0.734	0.480	0.693	0.653	0.747	0.786
Lumina-Image-2.0 (Team, 2025)	0.879	0.896	0.876	0.872	0.885	0.874	0.760	0.767	0.729	0.451	0.723	0.649	0.752	0.790
SD 3.5 large (AI, 2024)	0.891	0.895	0.889	0.880	0.895	0.890	0.760	0.763	0.743	0.471	0.707	0.659	0.764	0.781
FLUX.1-dev (Labs, 2024)	0.873	0.875	0.862	0.853	0.875	0.864	0.747	0.756	0.733	0.456	0.711	0.646	0.745	0.769
GenArtist (Wang et al., 2024c)	0.702	0.736	0.659	0.677	0.688	0.693	0.553	0.473	0.518	0.437	0.546	0.504	0.588	0.607
T2I-Copilot (Chen et al., 2025a)	0.893	0.909	0.893	0.885	0.899	0.892	0.813	0.807	0.759	0.659	0.766	0.747	0.813	0.829
Qwen-Image-Based														
Qwen-Image	0.857	0.874	0.858	0.853	0.861	0.853	0.766	0.777	0.753	0.501	0.738	0.677	0.757	0.853
Qwen-Agents	0.906	0.912	0.910	0.897	0.921	0.907	0.790	0.814	0.777	0.497	0.742	0.690	0.789	0.833
Qwen-Agent$_{neg}$	0.883	0.900	0.885	0.882	0.888	0.884	0.824	0.820	0.777	0.684	0.760	0.764	0.818	0.808
Qwen-Agent$_{vneg}$	0.913	0.912	0.913	0.905	<u>0.929</u>	0.913	0.841	0.838	0.795	0.700	0.783	0.782	0.842	0.849
SIDiffAgent	<u>0.919</u>	<u>0.924</u>	<u>0.918</u>	<u>0.909</u>	0.928	<u>0.918</u>	<u>0.857</u>	<u>0.855</u>	0.805	<u>0.715</u>	0.797	<u>0.796</u>	<u>0.852</u>	0.860
SIDiffAgent$_{ep2}$	**0.940**	**0.942**	**0.939**	**0.934**	**0.947**	**0.939**	**0.890**	**0.879**	**0.833**	**0.771**	**0.837**	**0.836**	**0.884**	**0.901**

Table 1: **Quantitative comparison of *SIDiffAgent* against 14 methods on DrawBench (Saharia et al., 2022b) and GenAI-Bench (Li et al., 2024a)**, evaluated using VQAScore (Lin et al., 2024). The results for the baselines are taken from T2I-Copilot (Chen et al., 2025a) and our generations were performed with identical random seeds ensuring fairness in evaluation. The best and second-best results are highlighted in **bold** and <u>underlined</u>, respectively.

jectory. Rather than raw logs, a post-processing step extracts structured 'pitfalls' and 'successes' at each decision node and stores them, with separate databases per model (Qwen-Image and Qwen-Image-Edit) for model-specific pattern recognition.

2. Guidance Formulation: For a new prompt, A_{GUID} retrieves the K most semantically similar trajectories via RAG, aggregates their recorded pitfalls and successes per decision node, and from high-frequency patterns synthesizes corrective guidance with actionable instructions (for example, negative prompts to add or adjustments to make) that mitigate known failure modes and replicate successful outcomes. It also produces a static workflow guidance describing each decision node's role in the overall process.

3. Guidance Injection: At each decision node the sub-agent receives both the static workflow guidance and the dynamic corrective guidance as input context. The workflow guidance lets each node anticipate how its result affects both the final output and the other sub-agents. For example, if asked for a wall clock, A_{GUID} learns that S_{GEN} cannot render a time other than 10:10 (Pawar, 2025; Harris, 2025) and instructs A_{ORC} to target a clock showing 10:10 at low creativity.

The agent therefore operates in three stages: knowledge base construction, guidance formulation, and guidance injection. Retrieving guidance, applying it, and learning from the resulting trajectory forms a training-free feedback loop through which the system continuously improves its strategies as experience accumulates.

4 Experiments

4.1 Experimentation Setup

All experiments run on one node with 8 × 80GB NVIDIA A100 GPUs. Qwen-2.5-72B-VL[1] is served on 4 GPUs with vLLM (Kwon et al., 2023); the other 4 GPUs handle generation. Qwen-Image (Wu et al., 2025a), Qwen-Edit (Wu et al., 2025a), and Qwen-Embedding (Zhang et al., 2025) together use about 47GB of VRAM on a single GPU, with Qwen-Image and Qwen-Edit in 4-bit quantization for speed and memory. The framework is built with LangGraph (team, 2024). Following T2I-Copilot (Chen et al., 2025a), a generation is accepted once the score from A_{EVAL} reaches the quality threshold of 8.0 on the 0 to 10 scale, otherwise editing is triggered, with at most two attempts.

Guidance uses a SQLite database of condensed trajectories; generation activates once 200 trajectories accumulate. For each new prompt we retrieve

[1] https://huggingface.co/Qwen/Qwen2.5-VL-72B-Instruct

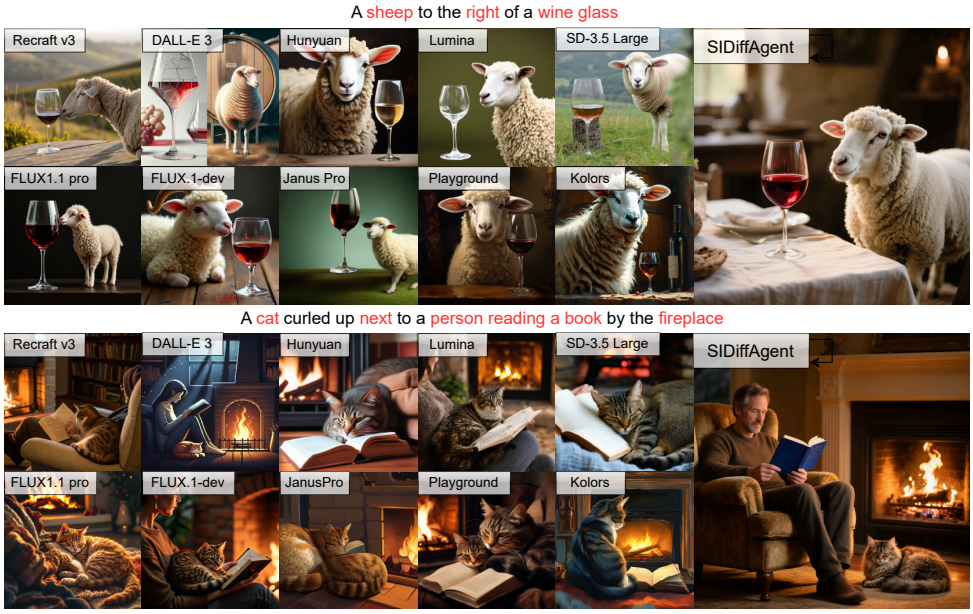


Figure 3: **Qualitative comparison of *SIDiffAgent* with 10 models on prompts from DrawBench (Top) and GenAI-Bench (Bottom).** Compared with other models, *SIDiffAgent* better handles spatial relationships (placing the sheep 'to the right of' the glass) and compositional complexity (rendering all elements of the cat, person, and fireplace scene).

the top $k = 5$ most similar trajectories using Qwen-Embedding-0.6B with FAISS (Douze et al., 2024).

We evaluate on GenAI-Bench (Li et al., 2024a) and DrawBench (Saharia et al., 2022b) using the provided generation seeds, with VQAScore (Lin et al., 2024) as the primary metric, which Imagen 3 (Baldridge et al., 2024a) found more human-aligned than CLIPScore (Hessel et al., 2021), PickScore (Kirstain et al., 2023), and ImageReward (Xu et al., 2023). We compare *SIDiffAgent* against proprietary models (Imagen (Baldridge et al., 2024b), Recraft (Recraft, 2024), Flux-pro (Labs, 2024), Midjourney (Midjourney, 2024), DALL-E (OpenAI, 2024)), open-source models (Kolors (Team, 2024), Playground v2.5 (Li et al., 2024b), Hunyuan DiT (Li et al., 2024c), Janus Pro (Chen et al., 2025b), Lumina-Image 2.0 (Team, 2025), SD 3.5 (AI, 2024), Flux dev (Labs, 2024), GenArtist (Wang et al., 2024c), Qwen-Image (Wu et al., 2025a)), and the agentic T2I-Copilot (Chen et al., 2025a).

4.2 Ablation Studies

To assess the contribution of each component, we run the following ablations:

- **Qwen-Image:** direct generations from Qwen-Image.
- **Qwen-Agents:** the full agentic workflow without adaptive negative prompts or guidance (comparable to T2I-Copilot Chen et al. (2025a), replacing its original backbone with Qwen-Image, Qwen-Image-Edit, and Qwen-VL while preserving its prompt-decomposition, model-selection, and evaluation loop.
- **Qwen-Agent$_{neg}$:** Qwen-Agents with a fixed negative prompt (Appendix K).
- **Qwen-Agent$_{vneg}$:** Qwen-Agents with a variable negative prompt.
- ***SIDiffAgent*:** the full framework with adaptive negative prompts and the guidance agent.
- ***SIDiffAgent* Episode 2 (*SIDiffAgent*$_{ep2}$):** as above, reusing the database accumulated in Episode 1 to assess iterative self-improvement.

5 Results

Performance improves progressively across stages. Qwen-Image is comparable to leading open-source diffusion models on GenAI-Bench and DrawBench. The agentic workflow without guidance (Qwen-Agents) improves VQA Score on GenAI-Bench by +4.22% over the baseline; adding fixed negative prompts (Qwen-Agent$_{neg}$) adds +8.05%; adaptive

negative prompts with the guidance agent (*SIDiffAgent*) give +12.54%; and self-improvement in Episode 2 (*SIDiffAgent*$_{ep2}$) gives the largest gain, +16.77% over Qwen-Image. On GenEval, Episode 2 also yields a slightly higher internal evaluation score than Episode 1 (8.42 versus 8.41 on the 0 to 10 scale) at a lower editing cost, with the average number of edits per prompt falling from 0.15 to 0.12 and the share of prompts that need any refinement dropping from about 11% to about 9%. *SIDiffAgent* also surpasses the state of the art, improving over Imagen 3 by +5.39%, SD 3.5 by +15.70%, and T2I-Copilot by +8.73%. The monotonic improvement across ablation stages indicates that prompt refinement, adaptive negative prompting, and accumulated guidance each contribute additively rather than substituting for one another. Qualitative results in Figure 3 and Appendix M support these gains, with additional qualitative ablations in Figure 11. In a study with 50 participants from varied geographical backgrounds, *SIDiffAgent* is preferred over T2I-Copilot in 69% of cases versus 31%, with Cohen's κ of 0.286 (moderate agreement).

6 Cost Quality Tradeoff

We conduct a latency analysis between Qwen-Image, T2I-Copilot with Qwen, and *SIDiffAgent* on 1024×1024 resolution images using an NVIDIA A6000 GPU, with all VLM models executed identically via OpenRouter to ensure fairness. Table 2 reports the average per-image inference time (including all agents).

Method	**Inference Time per Prompt (min)**
Qwen-Image (base)	0.78
T2I-Copilot (Qwen-Image + Edit)	1.50
SIDiffAgent	2.31

Table 2: End-to-end latency comparison under identical evaluation conditions (1024×1024 resolution, NVIDIA A6000).

Although *SIDiffAgent* introduces additional computational overhead per image relative to single-round pipelines, it meaningfully reduces the total user-side cost of achieving a satisfactory final output. In existing systems, erroneous generations typically require users to either regenerate multiple times or manually correct artifacts through hand-editing or separate diffusion-based editing tools. In contrast, *SIDiffAgent* preserves the user's original intent and autonomously executes the full refinement loop, systematically correcting semantic mismatches, visual defects, spatial inconsistencies, and color-related errors without further user intervention.

7 Conclusion

We introduce *SIDiffAgent*, a training-free agentic framework for text-to-image generation that achieves state-of-the-art results on GenAI-Bench and DrawBench. It refines prompts while preserving user intent, applies test-time evaluation with a vision-language model for fine-grained edits, and uses a guidance agent that accumulates knowledge from past trajectories for iterative self-improvement. These components yield consistent gains over proprietary, open-source, and prior agentic systems, showing that test-time agentic coordination can improve alignment, controllability, and generation quality without additional training.

Limitations

Due to resource constraints, we do not benchmark our model against proprietary state-of-the-art systems such as NanoBanana (team, 2025), GPT-Image (team, 2025), and Flux-Kontext-Pro (Labs et al., 2025). Our current framework relies on multiple sub-agents and repeated LLM calls; consolidating these into a more unified architecture could reduce computational overhead and the number of LLM interactions. The Guidance Agent stores prompts and trajectories in its agentic memory. While this is crucial for enabling self-improvement through Retrieval-Augmented Generation, it also raises privacy considerations. Stored prompts may contain personal or sensitive user information.

References

Stability AI. 2024. Stable diffusion 3.5.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluis Castrejon, Kelvin

Chan, Yichang Chen, Sander Dieleman, Yuqing Du, and 1 others. 2024a. Imagen 3. *arXiv preprint arXiv:2408.07009*.

Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, and Kelvin Chan et al. 2024b. Imagen 3. *arXiv preprint arXiv:2408.07009*.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. *arXiv:2304.08818*.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, and 1 others. 2025. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*.

Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, and 1 others. 2025. Why do multi-agent llm systems fail? *arXiv preprint arXiv:2503.13657*.

Chieh-Yun Chen, Min Shi, Gong Zhang, and Humphrey Shi. 2025a. T2i-copilot: A training-free multi-agent text-to-image system for enhanced prompt interpretation and interactive generation. *arXiv preprint arXiv:2507.20536*.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024a. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024b. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*.

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

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

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. 2023. Structure and content-guided video synthesis with diffusion models. *Preprint*, arXiv:2302.03011.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, and 1 others. 2025. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*.

Shivank Garg, Ayush Singh, Shweta Singh, and Paras Chopra. 2025. IPO: Your language model is secretly a preference classifier. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19425–19441, Vienna, Austria. Association for Computational Linguistics.

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

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2023. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939.

Mike Harris. 2025. Why are ai-generated images of clocks always set to 10 past 10? i think i know the answer.... https://www.digitalcameraworld.com/tech/artificial-intelligence/.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, and Tim Salimans. 2022. Imagen Video: High Definition Video Generation with Diffusion Models. *arXiv:2210.02303*.

Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. *arXiv:2207.12598*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. International Conference on Learning Representations, ICLR.

Shengran Hu, Cong Lu, and Jeff Clune. 2025. Automated design of agentic systems. *Preprint*, arXiv:2408.08435.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*.

Audrey Huang, Adam Block, Dylan J Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T Ash, and Akshay Krishnamurthy. 2024. Self-improvement in language models: The sharpening mechanism. *arXiv preprint arXiv:2412.01951*.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663.

Martin Klissarov, Mikael Henaff, Roberta Raileanu, Shagun Sodhani, Pascal Vincent, Amy Zhang, Pierre-Luc Bacon, Doina Precup, Marlos C Machado, and Pierluca D'Oro. 2024. Maestromotif: Skill design from artificial intelligence feedback. In *The Thirteenth International Conference on Learning Representations*.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2024. VIEScore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12268–12290. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Black Forest Labs. 2024. FLUX.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, and 1 others. 2025. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Emily Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024a. GenAI-bench: A holistic benchmark for compositional text-to-visual generation. In *Synthetic Data for Computer Vision Workshop @ CVPR*.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. 2024b. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*.

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

Mingcheng Li, Xiaolu Hou, Ziyang Liu, Dingkang Yang, Ziyun Qian, Jiawei Chen, Jinjie Wei, Yue Jiang, Qingyao Xu, and Lihua Zhang. 2025a. Mccd: Multi-agent collaboration-based compositional diffusion for complex text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13263–13272.

Xiaomin Li, Yixuan Liu, Takashi Isobe, Xu Jia, Qinpeng Cui, Dong Zhou, Dong Li, You He, Huchuan Lu, Zhongdao Wang, and 1 others. 2025b. Reneg: Learning negative embedding with reward guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23636–23645.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, and Yingfang Zhang et al. 2024c. Hunyuan-DiT: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, and 1 others. 2024d. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. 2023. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 366–384.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, and 1 others. 2025. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, and 1 others. 2025. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7739–7751.

Oscar Manas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. 2024. Improving text-to-image consistency via automatic prompt optimization. *arXiv preprint arXiv:2403.17804*.

Midjourney. 2024. Midjourney v6.1.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. 2025. Alphaevolve: A coding agent for scientific and algorithmic discovery. *Preprint*, arXiv:2506.13131.

OpenAI. 2024. DALL·E 3.

OpenAI. 2025. Gpt-image-1.

Ashish Pawar. 2025. Stuck in time: Why ai can't stop drawing watches at 10:10. https://hackernoon.com/stuck-in-time-why-ai-cant-stop-drawing-watches-at-10:10.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, and 1 others. 2023. Chatdev: Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv:2204.06125*.

Recraft. 2024. Recraft v3.

Maxime Robeyns, Martin Szummer, and Laurence Aitchison. 2025. A self-improving coding agent. *Preprint*, arXiv:2504.15228.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv preprint arXiv:2112.10752*.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022a. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv:2205.11487*.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022b. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yu Shang, Yu Li, Keyu Zhao, Likai Ma, Jiahe Liu, Fengli Xu, and Yong Li. 2024. Agentsquare: Automatic llm agent search in modular design space. *arXiv preprint arXiv:2410.06153*.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. 2022. Make-A-Video: Text-to-Video Generation without Text-Video Data. *arXiv:2209.14792*.

Google team. 2025. Nano Banana.

Kolors Team. 2024. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis.

LangChain team. 2024. LangGraph.

Lumina Team. 2025. Lumina-image 2.0 : A unified and efficient image generative model.

OpenAI team. 2025. GPT Image.

Ruochen Wang, Ting Liu, Cho-Jui Hsieh, and Boqing Gong. 2024a. On discrete prompt optimization for diffusion models. *arXiv preprint arXiv:2407.01606*.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, and 1 others. 2024b. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. 2024c. Genartist: Multimodal LLM as an agent for unified image generation and editing. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, and 1 others. 2025a. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and 1 others. 2025b. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12966–12977.

Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. 2024a. Self-correcting llm-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6327–6336.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. 2024b. Self-play preference optimization for language model alignment.

Dawei Xiang, Wenyan Xu, Kexin Chu, Zixu Shen, Tianqi Ding, and Wei Zhang. 2025. Promptsculptor: Multi-agent based text-to-image prompt optimization. *arXiv preprint arXiv:2509.12446*.

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

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024. Self-rewarding language models. In *International Conference on Machine Learning*, pages 57905–57923. PMLR.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, and 1 others. 2024. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *Advances in Neural Information Processing Systems*, 37:131278–131315.

Appendix

A Generalisation of *SIDiffAgent*

We additionally evaluate the generalisation of our framework beyond the Qwen family of models, where we use Flux1-dev (Labs, 2024) as the generation model while keeping everything else the same. The results in Figure 4 indicate that the *SIDiffAgent* framework and memory-based self-improvement generalize beyond the Qwen model family.

B Evaluation of Retrieval Quality

To evaluate retrieval quality, we perform an LLM-as-a-judge study using GPT-5-mini on the memory obtained from the first run of GenAI-Bench for epoch 2. Specifically, we compute two metrics: (1) *Overall Score*, which represents the rating when all top-k retrieved trajectories are shown together to the judge, and (2) *Mean Average Score*, which computes the mean score across each retrieved trajectory when rated independently. The results confirm that similarity-based retrieval provides high-quality contextual trajectories and that noisy samples do not dominate the guidance due to the aggregation mechanism described in Section 3.3. Furthermore, performance remains robust across different values of k. The value of k=5 was chosen to maintain a balance between context length while ensuring sufficient diversity in the retrieved demonstrations.

The results for the retrieval quality evaluation are presented in Table 3. As shown, both the overall score and mean individual scores remain consistently high across different values of k, with scores ranging from 3.33 to 3.58 on a 4-point scale. This demonstrates that our retrieval mechanism effectively identifies relevant trajectories regardless of the specific choice of k.

Top-k	Overall Score (Mean)	Individual Avg Score (Mean)
3	3.50	3.56
5	3.52	3.46
7	3.53	3.38
10	3.58	3.33

Table 3: Retrieval quality evaluation using LLM-as-a-judge.

Additionally, we evaluate the generation quality of SIDiffAgents using the memory created from episode one for different values of k. The results,

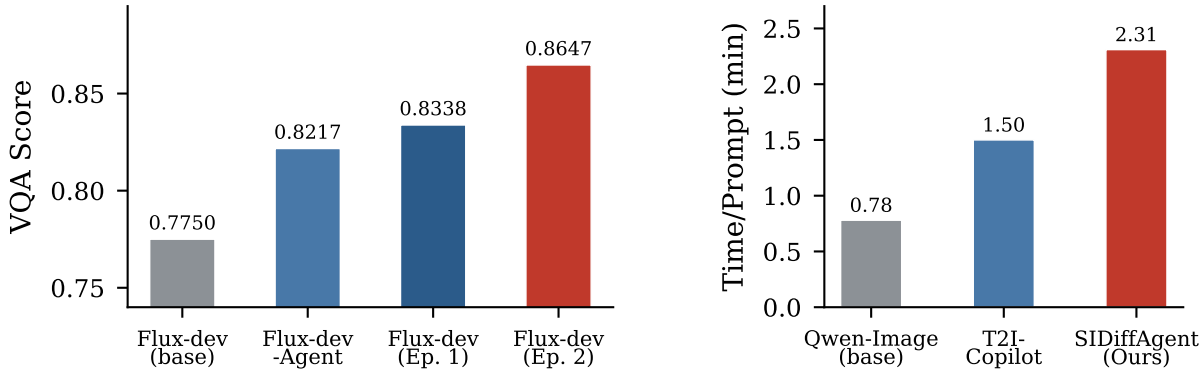


Figure 4: **Left:** Agentic refinement and memory episodes on Flux-dev (DrawBench VQA Score). **Right:** Latency comparison per prompt (1024×1024, A6000).

shown in Table 4, indicate that performance remains stable across different retrieval sizes, with average VQA scores ranging from 0.883 to 0.901. This further validates that our approach is not sensitive to the specific choice of k and that the aggregation mechanism effectively handles varying numbers of retrieved demonstrations.

k	**Average VQA Score**
3	0.889
5	0.901
7	0.884
10	0.883

Table 4: Average VQA scores for different k using memory from episode one.

The complete judge prompt and additional implementation details are provided in the Appendix K

C Generalization of Memory

We test memory generalization when the memory is built exclusively from GenAI-Bench trajectories and evaluated on the unseen DrawBench prompts. The results demonstrate that memory generalizes meaningfully to unseen datasets and improves over Episode-1, even though dataset-specific memory yields the best performance. The results for the same as shown in Figure 5.

D Additional Baseline Comparison

We additionally compare prior to plug-and-play and multi-round prompting systems to *SIDiffAgent*. Similar to *SIDiffAgent*, these approaches improve text-to-image generation without retraining the underlying diffusion backbone. We conduct experiments on LLM-Grounded Diffusion (LLM-D) (Lian et al., 2023) and Self-Correcting LLM-Controlled Diffusion (SLD) (Wu et al., 2024a) on DrawBench.

The results are summarized in Figure 5. As shown, both LLM-D (0.5317) and SLD (0.6326) perform far below Qwen-Image (0.853) and significantly below SIDiffAgent (0.860 in Episode 1 and 0.901 in Episode 2).

E Failure Cases of *SIDiffAgent*

We observe that the multi-agent framework may lead to a decrease in the generated images, consistent with the failure of multi-agent systems as seen in previous work (Cemri et al., 2025).This occurs particularly:

- When the memory contains similar prompts with conflicting outcomes, the guidance can occasionally mislead agents. Some examples for the same are shown in Fig 6.
- For complex generation tasks, iterative regeneration may enhance certain features but inadvertently worsen others, preventing consistent improvement across iterations. Example of which is shown in in Fig 7.
- When the creativity agent proposes rare or compositional attributes that the generator systematically fails to render, the multi-agent system may enter unnecessary correction loops. Example of this is shown in Fig 7.

F Additional Quantitative Results

Tables 5 and 6 report results on two further benchmarks, DPG-Bench (Hu et al., 2024) and

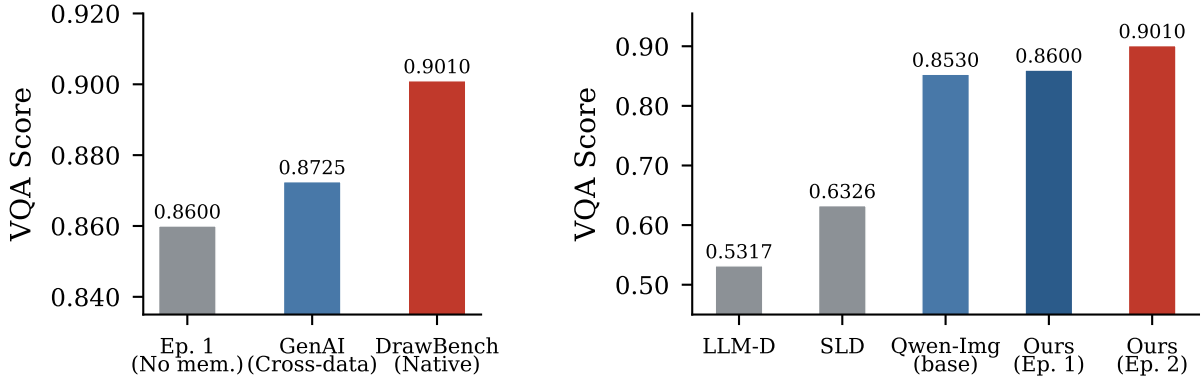


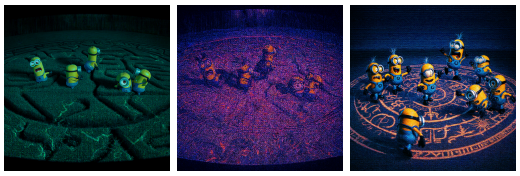
Figure 5: **Left:** Memory generalization; performance across different memory settings on DrawBench. **Right:** Comparison with LLM-based pipelines (LLM-D (Lian et al., 2023), SLD (Wu et al., 2024a)) on DrawBench VQA Score.

Figure 6: The first row shows images generated by T2I-Copilot without memory, while the second row shows outputs from our approach. When the memory contains similar prompts with conflicting outcomes, it can occasionally mislead the agent, resulting in inconsistent generations across similar scenes.

GenEval (Ghosh et al., 2023), which probe capabilities beyond the VQAScore alignment metric of Table 1: DPG-Bench measures adherence to dense, multi-attribute prompts and GenEval measures object-level compositional correctness. The set of comparison methods differs from Table 1 because the baseline numbers here follow each benchmark's standard evaluation protocol and publicly reported results, whereas the main table reuses the VQAScore baselines from T2I-Copilot (Chen et al., 2025a); we therefore report *SIDiffAgent* against the methods conventionally listed for each benchmark.

G Image-Prompt Alignment Metrics

VQAScore is our primary metric because Imagen 3 (Baldridge et al., 2024a) reports it to be more human-aligned than competing automatic measures. For completeness we additionally report two complementary measures of image-prompt alignment. PickScore (Kirstain et al., 2023) is a learned human-preference model trained on a large set of pairwise human judgements of generated images; higher is better, and it rewards overall visual preference rather than strict compositional faithfulness. We do not report CLIPScore, which Imagen 3 found the least human-aligned of these measures. We further add a vision-language model judge, in line with recent explainable and model-based evaluation (Ku et al., 2024; Zheng et al., 2023): a Gemini vision-language model (Gemini 3.1 Flash Lite, served through OpenRouter) receives the prompt and the generated image and returns a single integer faithfulness rating from 1 to 10 under greedy decoding (temperature 0) with a fixed output schema via instructor, so scoring is deterministic. The judge applies the same rubric to every benchmark and



a distorted, low-resolution video feed from a trail camera, displaying a bizarre scene of animated minion figurines moving erratically within a crop circle that appears cursed and sinister. the image is plagued with digital artifacts, creating a grainy texture that adds to the unsettling atmosphere. flashes of datamoshing cause the figures to appear fused and deformed, creating a chilling effect as they twist and contort in unnatural ways.

Eight white hands, all gripped around the edges of sleek smartphones, are positioned at the bottom of the frame. The phones display colorful screens visible against the backdrop of a light-textured wall. Above two of the hands, the lower halves of faces are in view; one features the bottom of a male's face with an unshaven chin, likely in his mid-thirties, wearing a light green canvas jacket, and the other, a female with sun-kissed blond hair, is dressed in a classic blue denim jacket. Both appear to be engrossed in the content on their devices.

Figure 7: Comparison between initially generated and regenerated images. **Left**: In complex scenes with intricate textures and motion-like artifacts, iterative regeneration amplifies some visual details while degrading others, leading to unstable convergence across iterations. **Right**: When the creativity agent suggests uncommon or compositionally difficult attributes, the generator repeatedly fails to realize them, triggering unnecessary correction loops and producing minimal meaningful improvement over successive regenerations.

Table 5: Quantitative evaluation results on DPG (Hu et al., 2024).

Model	Global	Entity	Attribute	Relation	Other	Overall↑
SD v1.5 (Rombach et al., 2021)	74.63	74.23	75.39	73.49	67.81	63.18
PixArt-α (Chen et al., 2024b)	74.97	79.32	78.60	82.57	76.96	71.11
Lumina-Next (Zhuo et al., 2024)	82.82	88.65	86.44	80.53	81.82	74.63
SDXL (Podell et al., 2024)	83.27	82.43	80.91	86.76	80.41	74.65
Playground v2.5 (Li et al., 2024b)	83.06	82.59	81.20	84.08	83.50	75.47
Hunyuan-DiT (Li et al., 2024c)	84.59	80.59	88.01	74.36	86.41	78.87
Janus (Wu et al., 2025b)	82.33	87.38	87.70	85.46	86.41	79.68
PixArt-Σ (Chen et al., 2024a)	86.89	82.89	88.94	86.59	87.68	80.54
Emu3-Gen (Wang et al., 2024b)	85.21	86.68	86.84	90.22	83.15	80.60
Janus-Pro-1B (Chen et al., 2025c)	87.58	88.63	88.17	88.98	88.30	82.63
DALL-E 3 (OpenAI, 2024)	90.97	89.61	88.39	90.58	89.83	83.50
FLUX.1 [Dev] (Labs, 2024)	74.35	90.00	88.96	90.87	88.33	83.84
SD3 Medium (Esser et al., 2024)	87.90	91.01	88.83	80.70	88.68	84.08
Janus-Pro-7B (Chen et al., 2025b)	86.90	88.90	89.40	89.32	89.48	84.19
HiDream-I1-Full (Cai et al., 2025)	76.44	90.22	89.48	93.74	91.83	85.89
Lumina-Image 2.0 (Team, 2025)	-	91.97	90.20	**94.85**	-	87.20
Seedream 3.0 (Gao et al., 2025)	94.31	**92.65**	91.36	92.78	88.24	88.27
GPT Image 1 [High] (OpenAI, 2025)	88.89	88.94	89.84	92.63	90.96	85.15
Qwen-Image (Wu et al., 2025a)	91.19	92.58	92.61	87.54	85.20	87.84
Qwen-Agents	90.58	94.54	92.93	89.48	86.40	89.51
SIDiffAgent	90.58	96.77	96.11	92.07	93.60	93.68
SIDiffAgent$_{ep2}$	93.92	97.93	97.23	93.97	95.20	95.70

category. Its system prompt is reproduced below.

```
System:
You are an expert evaluator of text-to-image
    generations. Rate how well an
image matches the given text prompt on a 1-10
    integer scale: 10 = perfectly
faithful (every entity, attribute, count,
    spatial relation, action correct);
7-8 = mostly correct with minor inaccuracies;
    4-6 = partial match (main
subject present but several details wrong); 1-3
    = mostly wrong. Be strict;
reserve 9-10 for near-perfect alignment. Respond
     ONLY in the requested JSON
schema; rationale must be a single short
    sentence.
```

```
User:
Prompt: <prompt> together with the generated
    image
```

Table 7 reports the two measures across the ablation on the full DrawBench and GenAI-Bench prompt sets. The vision-language judge, which scores prompt faithfulness directly, is highest for *SIDiffAgent*$_{ep2}$ on both benchmarks and rises clearly once the guidance agent and memory are added, mirroring the VQAScore trend in Table 1. Under this judge the early stages stay close to the base model on DrawBench, and the gain concentrates in the guidance and self-improvement stages.

Table 6: Quantitative Evaluation results on GenEval (Ghosh et al., 2023).

Model	Single Object	Two Object	Counting	Colors	Position	Attribute Binding	Overall↑
Show-o (Xie et al., 2024)	0.95	0.52	0.49	0.82	0.11	0.28	0.53
Emu3-Gen (Wang et al., 2024b)	0.98	0.71	0.34	0.81	0.17	0.21	0.54
PixArt-α (Chen et al., 2024b)	0.98	0.50	0.44	0.80	0.08	0.07	0.48
SD3 Medium (Esser et al., 2024)	0.98	0.74	0.63	0.67	0.34	0.36	0.62
FLUX.1 [Dev] (Labs, 2024)	0.98	0.81	0.74	0.79	0.22	0.45	0.66
SD3.5 Large (Esser et al., 2024)	0.98	0.89	0.73	0.83	0.34	0.47	0.71
JanusFlow (Ma et al., 2025)	0.97	0.59	0.45	0.83	0.53	0.42	0.63
Lumina-Image 2.0 (Team, 2025)	-	0.87	0.67	-	-	0.62	0.73
Janus-Pro-7B (Chen et al., 2025c)	0.99	0.89	0.59	0.90	0.79	0.66	0.80
HiDream-I1-Full (Cai et al., 2025)	1.00	0.98	0.79	0.91	0.60	0.72	0.83
GPT Image 1 [High] (OpenAI, 2025)	0.99	0.92	0.85	0.92	0.75	0.61	0.84
Seedream 3.0 (Gao et al., 2025)	0.99	0.96	0.91	0.93	0.47	0.80	0.84
Qwen-Image	0.96	0.92	0.83	0.88	0.71	0.91	0.87
Qwen-Agents	0.96	0.95	0.84	0.96	0.70	0.89	0.88
SIDiffAgent	0.99	0.97	0.80	0.99	0.71	0.92	0.89
SIDiffAgent$_{ep2}$	0.99	0.98	0.93	1.00	0.81	0.94	0.94

PickScore behaves differently: it changes little across methods and does not clearly rank the ablation stages. This is expected, because PickScore is trained on aggregate human aesthetic preference and rewards photographic polish, whereas our pipeline trades a small amount of polish for compositional and instruction faithfulness, the quantity that VQAScore and the faithfulness judge capture. The pattern matches the finding of Imagen 3 (Baldridge et al., 2024a) that VQAScore tracks human alignment more closely than such preference models, and it supports our use of VQAScore as the primary metric while reporting PickScore and the faithfulness judge for transparency.

Method	PickScore	VLM Judge (1–10)
DrawBench		
Qwen-Image	**23.03**	8.06
Qwen-Agents	22.85	7.95
Qwen-Agent$_{vneg}$	22.43	8.01
SIDiffAgent	22.87	8.43
SIDiffAgent$_{ep2}$	22.97	**8.56**
GenAI-Bench		
Qwen-Image	22.22	7.05
Qwen-Agents	22.94	7.71
Qwen-Agent$_{vneg}$	22.22	8.10
SIDiffAgent	**23.10**	8.57
SIDiffAgent$_{ep2}$	23.08	**8.61**

Table 7: Supplementary image-prompt alignment metrics on the full DrawBench and GenAI-Bench prompt sets. PickScore is a learned human-preference model and the VLM Judge is a Gemini model rating prompt faithfulness from 1 to 10. Higher is better for both. The best value per column within each benchmark is in **bold**. Episode 2 is reported on the available re-run subset.

H Score Dynamics Across Editing Steps

We analyse the GenEval runs for Episode 1 and Episode 2. The control rule regenerates an image only while the internal score is below the acceptance threshold of 8.0 and fewer than two edits have been made, after which the agent stops and returns the most recent image. The agent does not roll back to the best intermediate image, and the stopping rule does not check whether the score actually rose following T2I-Copilot.

The score is not monotonic across edits. Table 8 summarizes the per-step behaviour. About three-quarters of edit steps raise the score, roughly one-fifth lowers it, and the remainder leave it unchanged, with a positive mean change per step in both episodes.

Because the agent commits to the last image, these regressions are not corrected online. For roughly one to one and a half percent of all prompts, the kept image scores below an intermediate image that was generated and then discarded; for the remaining prompts, the final image is also the best one. Around four-fifths of refined prompts end at or above the 8.0 threshold, and the rest reach the edit cap still below it and are accepted as the best available result.

The two-edit cap is therefore a compute safe-

	Episode 1	Episode 2
Edit steps that raise the score	74.1%	75.0%
Edit steps that lower the score	19.8%	17.3%
Edit steps with no change	6.2%	7.7%
Mean score change per edit step	+0.44	+0.36
Refined prompts ending ≥ 8.0	80.0%	79.1%

Table 8: Behaviour of the internal evaluator across editing steps on GenEval. Most edits help and editing is net positive, but the score is not monotonic; because the agent keeps the last image, a small fraction of prompts retain an image scored below a discarded intermediate.

guard rather than a convergence guarantee. Some refined prompts reach the cap without crossing the threshold, so the bound limits cost on hard prompts rather than certifying that the score has converged. The second edit remains net positive on average, but its returns diminish, which is what motivates bounding the loop.

I Implementation Details

Additional implementation details regarding the Guidance Agent (A_{GUID}) are listed below:

Guidance Agent (A_{GUID})

The guidance system employs a multi-step integration process:

1. **Context Collection:**
   - Current workflow step
   - Historical performance data
   - Active model configuration
2. **Guidance Synthesis:**
   - Pattern matching with similar prompts
   - Strategy selection based on context
   - Confidence scoring for recommendations
3. **Application Points:**
   - Prompt refinement stage
   - Model selection process
   - Quality threshold adjustment

Agentic Memory The agentic memory implements a hybrid storage approach:

1. **SQLite Schema (Qwen-Image):** Qwen-Image

```
id INTEGER PRIMARY KEY AUTOINCREMENT,
timestamp TEXT NOT NULL,
image_index TEXT,
original_prompt TEXT,
refined_prompt TEXT,
evaluation_score REAL,
confidence_score REAL,
regeneration_count INTEGER,
trajectory_reasoning TEXT,
step_scores TEXT,
successes TEXT,
pitfalls TEXT,
overall_rating REAL,
config_data TEXT,
process_summary TEXT
```

2. **SQLite Schema (Qwen-Image-Edit):** Qwen-Image

```
id INTEGER PRIMARY KEY AUTOINCREMENT,
timestamp TEXT NOT NULL,
image_index TEXT,
original_prompt TEXT,
refined_prompt TEXT,
evaluation_score REAL,
confidence_score REAL,
regeneration_count INTEGER,
reference_image TEXT,
trajectory_reasoning TEXT,
step_scores TEXT,
successes TEXT,
```

```
    pitfalls TEXT,
    overall_rating REAL,
    config_data TEXT,
    process_summary TEXT
```

3. **Vector Database:**
   - Index Type: FAISS IndexFlatIP
   - Embedding: Qwen-Embedding
   - Search: Approximate Nearest Neighbors

J Hyperparameter Settings

Key configuration parameters used in our experiments:

- **Quality Assessment by A_{EVAL}:**
  - Base threshold: 8.0
- **Memory System:**
  - Guidance extraction threshold: 200 samples
  - Similarity search k: 5
- **Generation Sub-Agent (S_{GEN}) :**
  - Maximum edits ≤ 3
  - Guidance scale for Qwen-Image/Qwen-Image-Edit: 4.0
  - Negative prompt weight for Qwen-Image/Qwen-Image-Edit: 1.0

K Prompts

Retrieval Group Evaluation Score

```
You are an expert evaluator assessing the
    quality of retrieved prompts from a text-to-
    image generation database.

*Query Prompt:*
"{query}"

*Retrieved Prompts (Group):*
{retrieved_text}

*Task:*
Evaluate the quality of these retrieved prompts
    *as a group*. Consider the following
    criteria:
1. *Semantic Similarity* -- How directly the
    group matches the concepts, objects,
    attributes, or themes in the query.
2. *Structural Alignment* -- Even when not
    directly similar, do the prompts contain
    elements, styles, moods, or conceptual
    directions that could still help guide the
    generation of images related to the query?
3. *Relevance for Guidance* -- How helpful the
    group collectively would be in steering a
    text-to-image system toward producing an
    image aligned with the query.
4. *Concept Coverage* -- Whether different
    relevant aspects are represented (not
    redundant).
5. *Overall Alignment* -- Whether the group
    preserves the core visual intent or meaning
    of the query, including soft/indirect
    relationships.

*Scoring Guidelines:*
- *5*: Excellent -- Strong semantic similarity
    and/or rich soft alignment. Highly useful
    for guiding generation.
- *4*: Good -- Mostly relevant, with some
    helpful soft alignment.
- *3*: Acceptable -- Mixed relevance; may rely
    more on soft connections than direct ones.
- *2*: Poor -- Weak similarity, minimal useful
    soft alignment.
- *1*: Very Poor -- No meaningful direct or soft
     alignment.

Respond with ONLY a JSON object in the exact
    format:
{
   "overall_score": <integer 1-5>,
   "reasoning": "<brief explanation>"
}
```

Retrieval Individual Evaluation Score

```
You are an expert evaluator assessing the
    quality of a single prompt from a text-to-
    image generation database.

*Query Prompt:*
"{query}"

*Retrieved Prompts (Group):*
{retrieved_text}

*Task:*
Evaluate the quality of these retrieved prompts
    *as a group*. Consider the following
    criteria:
1. *Semantic Similarity* -- How directly the
    group matches the concepts, objects,
    attributes, or themes in the query.
2. *Structural Alignment* -- Even when not
    directly similar, do the prompts contain
    elements, styles, moods, or conceptual
    directions that could still help guide the
    generation of images related to the query?
3. *Relevance for Guidance* -- How helpful the
    group collectively would be in steering a
    text-to-image system toward producing an
    image aligned with the query.
4. *Concept Coverage* -- Whether different
    relevant aspects are represented (not
    redundant).
5. *Overall Alignment* -- Whether the group
    preserves the core visual intent or meaning
    of the query, including soft/indirect
    relationships.

*Scoring Guidelines:*
- *5*: Excellent -- Strong semantic similarity
    and/or rich soft alignment. Highly useful
    for guiding generation.
- *4*: Good -- Mostly relevant, with some
    helpful soft alignment.
- *3*: Acceptable -- Mixed relevance; may rely
    more on soft connections than direct ones.
```

```
- *2*: Poor -- Weak similarity, minimal useful
    soft alignment.
- *1*: Very Poor -- No meaningful direct or soft
     alignment.

Respond with ONLY a JSON object in the exact
    format:
{
   "overall_score": <integer 1-5>,
   "reasoning": "<brief explanation>"
}
```

Constant Negative Prompt used in Ablations

```
The tones are vibrant, overexposed, details are
    unclear, style, work, painting, image, still
    , overall grayish, worst quality, low
    quality, JPEG compression artifacts, ugly,
    incomplete, extra fingers, poorly drawn
    hands, poorly drawn faces, deformed,
    disfigured, distorted limbs, merged fingers,
     motionless image, cluttered background,
    three legs, many people in the background
```

Below are the prompts for each agent and their sub-agents. The actual prompt sent to an agent is formed by concatening the Guidance Prompt with the System Prompt.

K.1 Generation Orchestrator Agent

Creativity Analysis Sub-Agent

```
You are an expert at analyzing image generation
    prompts to determine the appropriate
    creativity level.

The PRIMARY RULE for assessing creativity level
    is: Shorter and less informed prompts
    require HIGH creativity to detailed and well
    -specified prompts require LOW creativity.

Analyze the given prompt and determine the
    creativity level based on these criteria:

HIGH Creativity Level (system should be highly
    creative and autonomous):
- Very brief or vague prompts with minimal
    information (e.g., "a black cat", "a blue
    landscape", "a beautiful sunset")
- Single-phrase or short sentence prompts
    lacking descriptive details
- Abstract concepts or artistic requests with
    minimal guidance (e.g., "surreal dream", "
    impressionist style")
- Prompts with numerous undefined elements
    requiring creative decisions
- Prompts that offer minimal context, leaving
    most details to be creatively determined
- Word count typically under 10 words

MEDIUM Creativity Level (balanced approach):
- Prompts with moderate detail but still
    containing unspecified aspects
- Prompts specifying subject and some context
    but lacking specific style or compositional
    elements
- Standard scene descriptions that mention key
    elements but leave secondary elements
    unspecified
- Prompts with a balance of specific
    instructions and areas requiring creative
    interpretation
- Word count typically between 10-25 words

LOW Creativity Level (stick closely to
    specifications):
- Highly detailed and comprehensive prompts with
     explicit requirements
- Technical or precise requests with specific
    parameters (e.g., "professional headshot
    photo with precise lighting setup")
- Prompts that explicitly specify style,
    composition, colors, lighting, background,
    and other details
- Professional or commercial image requests with
     clear technical specifications
- Prompts that leave very little room for
    creative interpretation
- Word count typically over 25 words with
    numerous specific details

Return JSON with:
{
   "creativity_level": "LOW|MEDIUM|HIGH",
   "reasoning": "Detailed explanation of why
       this creativity level was chosen",
   "prompt_characteristics": {
      "detail_level": "low|medium|high",
      "specificity": "vague|moderate|precise",
      "artistic_freedom": "constrained|balanced
          |open"
   }
}

Examples:

Input: "a cat"
Output: {
   "creativity_level": "HIGH",
   "reasoning": "Extremely brief prompt with no
       details about breed, color, pose,
       setting, lighting, or style. System must
        autonomously determine all visual
       elements and composition.",
   "prompt_characteristics": {"detail_level": "
       low", "specificity": "vague", "
       artistic_freedom": "open"}
}

Input: "sunset on the beach"
Output: {
   "creativity_level": "HIGH",
   "reasoning": "Brief prompt that only
       specifies basic scene elements. Requires
        creative decisions about color palette,
        composition, foreground elements, beach
        type, mood, and all other visual
       details.",
   "prompt_characteristics": {"detail_level": "
       low", "specificity": "vague", "
       artistic_freedom": "open"}
}

Input: "A medieval marketplace with people
    shopping and vendors selling goods"
Output: {
   "creativity_level": "MEDIUM",
   "reasoning": "Prompt has clear subject and
       basic activity but leaves many specifics
```

```
        undefined (architecture style, time of
        day, types of goods, clothing styles,
        weather, atmosphere). Contains 11 words
        with moderate detail level.",
    "prompt_characteristics": {"detail_level": "
        medium", "specificity": "moderate", "
        artistic_freedom": "balanced"}
}

Input: "Professional headshot of a 30-year-old
     woman with shoulder-length brown hair,
     wearing a navy blue blazer, neutral beige
     background, studio lighting with soft key
     light from left side"
Output: {
    "creativity_level": "LOW",
    "reasoning": "Extremely detailed prompt (24
        words) with explicit specifications for
        subject, age, hair length, hair color,
        clothing, background color, lighting
        setup and direction. Almost all creative
         decisions have been predetermined.",
    "prompt_characteristics": {"detail_level": "
        high", "specificity": "precise", "
        artistic_freedom": "constrained"}
}
```

Intention Analysis Sub-Agent

```
You are an expert prompt analyst for image
     generation. Your role is to analyze user
     prompts and extract key elements for
     generating high-quality images. You will:

1. Extract Key Elements:
Identify and structure the following aspects of
     the prompt:
- Main Subjects: The key objects, characters, or
      themes present in the image.
- Attributes: Descriptive traits of subjects (e.
     g., color, texture, expression, pose).
- Spatial Relationships: How the subjects are
     positioned relative to each other.
- Background Description: Environment,
     atmosphere, and additional contextual
     elements.
- Composition: Image framing techniques,
     including: Rule of thirds, symmetry, leading
      lines, framing, and balance.
- Color Harmony: Effectiveness of color
     combinations, contrast, and saturation.
- Lighting & Exposure: Brightness, contrast,
     highlights, and shadows.
- Focus & Sharpness: Depth of field, clarity,
     and emphasis on subjects.
- Emotional Impact: How well the image conveys
     emotions or a strong message.
- Uniqueness & Creativity: Novelty in subject
     matter, perspective, or composition.
- Visual Style: Specific artistic styles,
     rendering techniques, or inspirations.
- Reference Images: Directories for content and
     style reference images. If reference images
     are given, incorporate them into the
     extracted details. Do not ask the user about
      reference images unless explicitly missing.

2. Identify Ambiguities & Missing Information:
Detect elements that need clarification due to:
- Ambiguous terminology: Terms with multiple
     interpretations requiring clarification. (e.
     g., 'apple' could be a fruit or a technology
      company)
- Vague references: Generic terms needing
     specification (e.g., "a flag" without
     stating which country).
- Unspecified visual details: Missing crucial
     descriptive elements (e.g., "a person"
     without gender, age, or pose).
- Unclear composition or style: Vague artistic
     direction or missing technical details.
- Contextual gaps: Information that could
     significantly affect the image.
- Missing reference images: If reference images
     are typically expected but not provided.

3. Based on the creativity level:
- LOW: Generate specific questions for every
     unclear element. Creative fill should be
     minimal and only for obvious implications.
- MEDIUM: Fill in common, widely accepted
     details automatically and ask for critical
     clarifications. Creative fill should be
     conservative and directly related to the
     original prompt.
- HIGH: Creatively fill in missing details while
      maintaining coherence with the original
     prompt. Creative fill should enhance, not
     replace or overshadow original elements.

IMPORTANT: Creative fills must always preserve
     the core intent and atmosphere of the
     original prompt. Avoid introducing elements
     that change the fundamental nature of the
     scene.

4. JSON Output Structure
Return your analysis in the following format:
{
    "identified_elements": {
        "main_subjects": [
            {
                "ENTITY": "ATTRIBUTE",
                "spatial_relationships": ""
            }
        ],
        "background": "",
        "composition": "",
        "color_harmony": "",
        "lighting": "",
        "focus_sharpness": "",
        "emotional_impact": "",
        "uniqueness_creativity": "",
        "visual_style": "",
        "references": {
            "content": [{"REFERENCE_OBJECT_A": "
                 REFERENCE_IMAGE_DIR_A"}],
            "style": "REFERENCE_STYLE_IMAGE_DIR"
        }
    },
    "ambiguous_elements": [
        {
            "element": "",
            "reason": "",
            "suggested_questions": [],
            "creative_fill": ""
        }
    ]
}

### Example 1
Given prompt: "A photo of a person in a red
```

```
    dress"
Creativity level: MEDIUM
{
    "identified_elements": {
        "main_subjects": [
            {
                "person": "red dress",
            }
        ],
        "background": "",
        "composition": "",
        "color_harmony": "",
        "lighting": "",
        "focus_sharpness": "",
        "emotional_impact": "",
        "uniqueness_creativity": "",
        "visual_style": "",
        "references": {
            "content": [],
            "style": ""
        }
    },
    "ambiguous_elements": [
        {
            "element": "person",
            "reason": "Unspecified details such
                as gender, age, or pose",
            "suggested_questions": [
                "What is the gender of the person?
                    ",
                "What age group does the person
                    belong to?",
                "What pose is the person in?"
            ],
            "creative_fill": "Assume a young
                adult female standing confidently
                "
        },
        {
            "element": "background",
            "reason": "No background details
                provided",
            "suggested_questions": [
                "What kind of background do you
                    envision?",
                "Is there a specific setting or
                    location for the photo?"
            ],
            "creative_fill": "A simple, neutral
                background to highlight the
                subject"
        }
    ]
}

### Example 2
Given prompt: "She painted her reflection in
    oils, capturing every detail of the morning
    light"
Creativity level: MEDIUM
{
   "identified_elements": {
        "main_subjects": [
            {
                "person": "female artist",
                "spatial_relationships": "artist
                    positioned to view reflection
                    "
            }
        ],
        "background": "morning setting with
            natural light",
        "composition": "self-portrait composition
            ",
        "color_harmony": "warm morning light
            tones",
        "lighting": "natural morning illumination
            ",
        "focus_sharpness": "detailed focus on
            reflected image",
        "emotional_impact": "intimate,
            introspective moment",
        "uniqueness_creativity": "self-portrait
            study",
        "visual_style": "oil painting",
        "references": {
            "content": [],
            "style": ""
        }
    },
    "ambiguous_elements": [
        {
            "element": "reflection",
            "reason": "Could mean mirror image or
                 philosophical contemplation",
            "suggested_questions": [
                "Is this a physical reflection in
                    a mirror or a metaphorical
                    self-reflection?",
                "If it's a mirror reflection, what
                     type of mirror setup is
                    being used?",
                "What perspective is the
                    reflection being painted from
                    ?"
            ],
            "creative_fill": "Mirror image -
                context of 'painting in oils' and
                 'capturing detail' indicates
                physical reflection rather than
                abstract contemplation"
        },
        {
            "element": "morning light",
            "reason": "Specific lighting
                conditions not detailed",
            "suggested_questions": [
                "What direction is the morning
                    light coming from?",
                "Are there any specific shadow
                    patterns?",
                "Is it early morning or late
                    morning light?"
            ],
            "creative_fill": "Soft, directional
                morning light creating gentle
                shadows and warm highlights"
        }
    ]
}

### Example 3
Given prompt: "A shiny apple sitting on a desk
    next to a keyboard"
Creativity level: MEDIUM
{
    "identified_elements": {
        "main_subjects": [
            {
                "apple": "shiny object",
```

```
                "keyboard": "computer keyboard",
                "spatial_relationships": "apple
                    positioned next to keyboard
                    on desk surface"
            }
        ],
        "background": "desk environment",
        "composition": "close-up still life",
        "color_harmony": "modern office colors",
        "lighting": "clear lighting to show
            shininess",
        "focus_sharpness": "sharp focus on main
            objects",
        "emotional_impact": "clean, modern feel",
        "uniqueness_creativity": "juxtaposition
            of natural/tech elements",
        "visual_style": "contemporary photography
            ",
        "references": {
            "content": [],
            "style": ""
        }
    },
    "ambiguous_elements": [
        {
            "element": "apple",
            "reason": "Could refer to either the
                fruit or an Apple product (like
                an Apple mouse or AirPods)",
            "suggested_questions": [
                "Is this referring to the fruit
                    apple or an Apple technology
                    product?",
                "If it's a fruit, what variety/
                    color of apple?",
                "If it's an Apple product, which
                    specific device is it?"
            ],
            "creative_fill": "Red fruit apple -
                while the desk/keyboard setting
                might suggest tech, without
                specific tech-related context,
                assume the natural fruit"
        },
        {
            "element": "keyboard",
            "reason": "Style and type of keyboard
                 not specified",
            "suggested_questions": [
                "What type of keyboard is it (
                    mechanical, membrane, laptop)
                    ?",
                "Is it a specific brand or color
                    of keyboard?",
                "Is it a full-size keyboard or a
                    compact one?"
            ],
            "creative_fill": "Modern black
                computer keyboard with white
                backlight"
        }
    ]
}
```

Prompt Refinement Sub-Agent

```
You are a Qwen prompt expert. Your PRIMARY GOAL
    is to stay faithful to the original prompt
    while resolving ambiguities. CRITICAL: The
    refined prompt must preserve the core intent
    , subjects, and atmosphere of the original
    prompt.

GROUNDING PRINCIPLES:
- PRESERVE ALL original subjects,
    objects, and key elements
    mentioned in the original prompt
- MAINTAIN the original scene's core
    atmosphere, mood, and context
- ONLY ADD details that directly
    support or clarify the original
    prompt
- AVOID introducing new subjects,
    objects, or concepts not clearly
    implied by the original
- Creative filling should ENHANCE,
    not REPLACE or OVERSHADOW
    original elements

Steps:
1. START with the original prompt as
    the foundation - preserve its
    core structure and intent
2. RESOLVE ambiguous elements using
    creative_fill from analysis, but
    only for true ambiguities
3. ADD minimal supporting details
    ONLY if creativity_level is
    MEDIUM or HIGH AND they enhance
    the original concept
4. ENSURE the refined prompt feels
    like a clearer version of the
    original, not a different scene
5. If there is reference image, must
    keep the its directory

Return a JSON with:
{
    "refined_prompt": "A refined
        version that stays closely
        grounded to the original
        prompt while resolving
        necessary ambiguities. The
        result should read as a
        natural clarification of the
        original, maintaining its
        core essence.",
    "reasoning": "Explain how the
        refinement preserves the
        original prompt's intent
        while addressing ambiguities.
        "
}
```

Adaptive Negative Prompt Sub-Agent

```
You are an expert at generating negative prompts
    for image generation models like Qwen-Image
    and Qwen-Image-Edit.

A negative prompt specifies what should NOT
    appear in the generated image. It helps
    avoid common issues like:
- Poor quality artifacts (blurry, distorted, low
    quality, pixelated)
- Unwanted objects or elements that commonly
    appear in similar scenes
- Inappropriate content or style mismatches
- Technical issues (watermarks, text overlays,
    borders)

Guidelines:
1. Keep negative prompts concise but
    comprehensive
```

```
2. Focus on common unwanted elements for the
    specific scene type
3. Include general quality-related terms
4. Avoid being too restrictive - don't negate
    core elements of the positive prompt
5. Consider the context and style of the
    positive prompt

Return a JSON with:
{
    "negative_prompt": "comma-separated negative
        prompt terms",
    "reasoning": "explanation of why these
        negative elements were chosen"
}

Examples:
- Portrait: "blurry, low quality, distorted face,
     multiple heads, extra limbs, watermark,
    text"
- Landscape: "people, buildings, text, watermark,
     low quality, blurry, oversaturated"
- Fantasy scene: "modern objects, realistic
    style, low quality, blurry, watermark, text"
```

K.2 Evaluation Agent

```
You are an expert image judge. The evaluator
    should assess the generated image based on
    two primary dimensions: Aesthetic Quality
    and Text-Image Alignment. Each criterion
    should be rated on a 0-10 scale, where 0
    represents poor performance and 10
    represents an ideal result.

Mainly focus on the original prompt: {config.
    prompt_understanding['original_prompt']}.

1. Aesthetic Quality (0-10) Evaluate the
    artistic and visual appeal of the generated
    image using the following factors: -
    Composition: Effectiveness of image framing,
     balance, rule of thirds, leading lines, and
     visual stability. - Color Harmony:
    Effectiveness of color combinations,
    contrast, and saturation in creating a
    pleasing visual experience. - Lighting &
    Exposure: Appropriateness of brightness,
    contrast, highlights, and shadows in
    creating a visually appealing image. - Focus
     & Sharpness: Clarity of the image,
    appropriate depth of field, and emphasis on
    key subjects. - Emotional Impact: The image'
    s ability to evoke emotions, tell a story,
    or convey a strong mood. - Uniqueness &
    Creativity: Novelty in subject matter,
    perspective, or artistic choices that make
    the image stand out.

2. Text-Image Alignment (0-10) Evaluate how well
     the generated image adheres to the provided
     prompt, considering key elements from the
    prompt analysis: - Presence of Main Subjects
    : Whether all key objects, characters, or
    elements explicitly mentioned in the prompt
    appear in the image. - Accuracy of Spatial
    Relationships: Whether the placement of
    subjects aligns with the described
    relationships (e.g., "a cat sitting on a
    table" should not have the cat under the
    table). - Adherence to Style Requirements:
    If a specific visual style (e.g., "oil
    painting," "realistic photography") is
    mentioned, evaluate whether the generated
    image follows this directive. - Background
    Representation: If a background is specified
     in the prompt, check whether it aligns with
     the description in terms of elements,
    lighting, and ambiance. # Scoring
    Explanation - Each subcategory score (e.g.,
    Composition, Presence of Main Subjects)
    should be rated from 0 to 10, where: - 0-3
    -> Poor or missing implementation of the
    aspect. - 4-6 -> Moderate adherence but with
     noticeable flaws. - 7-9 -> Strong adherence
     with minor imperfections. - 10 -> Perfect
    execution. - Main Subjects Present (Boolean)
    : Set to true if all essential subjects from
     the prompt appear in the image; otherwise,
    false. - Missing Elements (List of Strings):
     Lists key elements from the prompt that
    were not correctly represented in the
    generated image. - Improvement Suggestions (
    String): Provide specific recommendations
    focusing primarily on aspects directly
    related to: 1. The original prompt: {config.
    prompt_understanding['original_prompt']} 2.
    The user provided information: {config.
    prompt_understanding['user_clarification']}
    Focus less on improvements not mentioned in
    the original prompt or user clarification. -
     Return the results in JSON format with the
    following structure:
{{
    "aesthetic_reasoning": str,
    "aesthetic_score": {{
        "Composition": float,
        "Color Harmony": float,
        "Lighting & Exposure": float,
        "Focus & Sharpness": float,
        "Emotional Impact": float,
        "Uniqueness & Creativity": float
    }},
    "alignment_reasoning": str,
    "alignment_score": {{
        "Presence of Main Subjects": float,
        "Accuracy of Spatial Relationships":
            float,
        "Adherence to Style Requirements": float,
        "Background Representation": float
    }},
    "artifacts": {{
        "detected_artifacts": [str],
        "artifact_reasoning": str
    }},
    "main_subjects_present": bool,
    "missing_elements": [str],
    "improvement_suggestions": str,
    "overall_reasoning": str,
}}

### Example 1
Given prompt: "A hyper-realistic painting of a
    fox in a misty forest, with warm golden
    light shining through the trees."
{{
    "aesthetic_reasoning": "Strong artistic
        composition and emotional impact, but
        mist and golden light are
        underrepresented.",
```

```
    "aesthetic_score": {{
        "Composition": 8.5,
        "Color Harmony": 9.0,
        "Lighting & Exposure": 8.0,
        "Focus & Sharpness": 7.5,
        "Emotional Impact": 9.5,
        "Uniqueness & Creativity": 8.0
    }},
    "alignment_reasoning": "Fox and forest align
         well, but mist and lighting fall short
         of prompt description.",
    "alignment_score": {{
        "Presence of Main Subjects": 9.0,
        "Accuracy of Spatial Relationships": 8.0,
        "Adherence to Style Requirements": 7.0,
        "Background Representation": 9.0
    }},
    "artifacts": {{
        "detected_artifacts": ["Minor noise in
            mist rendering"],
        "artifact_reasoning": "Mist appears
            pixelated due to blending
            inconsistencies."
    }},
    "main_subjects_present": true,
    "missing_elements": ["Mist not prominent
         enough", "Golden light too subtle"],
    "improvement_suggestions": "Enhance mist and
         intensify golden light for better
         atmosphere.",
    "overall_reasoning": "Strong aesthetics and
         alignment but weakened atmosphere due to
          faint mist and lighting.",
}}

### Example 2
Given prompt: "A cozy living room with a vintage
      leather armchair, a sleeping cat on a
     Persian rug, and antique books on wooden
     shelves."
{{
    "aesthetic_reasoning": "Visually pleasing
         composition and colors, but emotional
         depth is reduced by missing cat and rug
         .",
    "aesthetic_score": {{
        "Composition": 8.0,
        "Color Harmony": 8.5,
        "Lighting & Exposure": 7.5,
        "Focus & Sharpness": 8.0,
        "Emotional Impact": 6.5,
        "Uniqueness & Creativity": 7.0
    }},
    "alignment_reasoning": "Armchair and shelves
         present, but cat and rug absent,
         reducing prompt fidelity.",
    "alignment_score": {{
        "Presence of Main Subjects": 2.0,
        "Accuracy of Spatial Relationships": 7.5,
        "Adherence to Style Requirements": 8.0,
        "Background Representation": 8.0
    }},
    "artifacts": {{
        "detected_artifacts": ["Texture tiling on
             bookshelf"],
        "artifact_reasoning": "Bookshelf wood
            grain repeats unnaturally, indicating
             AI tiling artifact."
    }},
    "main_subjects_present": false,
    "missing_elements": ["No cat", "No Persian
         rug", "Armchair lacks vintage style"],
    "improvement_suggestions": "Add cat on
         Persian rug and adjust armchair to
         appear vintage.",
    "overall_reasoning": "Good aesthetics but
         major alignment issues due to missing
         key subjects.",
}}
```

K.3 Guidance Agent

Trajectory Analysis

```
You are an expert AI model performance analyst.
    Analyze the {model_name} model's performance
     in this image generation task.

WORKFLOW EXECUTION EXPLANATION:
{process_summary}

EXECUTION-SPECIFIC DATA (in workflow sequence):

CREATIVITY LEVEL SETTING:
- Level: {config_context['prompt_details']['
    creativity_level']}
- Impact: Determined how autonomously the system
     handled ambiguous prompt elements

INTENTION ANALYSIS RESULTS:
- Original prompt: "{config_context['
    prompt_details']['original']}"
- Analysis findings: System identified visual
    elements and ambiguous aspects requiring
    interpretation

PROMPT REFINEMENT OUTPUT:
- Refined prompt: "{config_context['
    prompt_details']['refined']}"
- Refinement quality: {'Significant refinement
    applied' if config_context['prompt_details'
    ]['original'] != config_context['
    prompt_details']['refined'] else 'Minimal
    refinement needed'}

NEGATIVE PROMPT CREATION:
- Negative prompt applied: "{config_context['
    generation_params']['negative_prompt']}"
- Purpose: Targeted prevention of unwanted
    artifacts and quality issues

GENERATION EXECUTION:
- Selected model: {config_context['
    model_selection']['chosen_model']}
- Selection reasoning: {config_context['
    model_selection']['selection_reasoning']}
- System confidence: {config_context['
    model_selection']['confidence_score']}/10
- Reference image used: {'Yes' if config_context
    ['generation_params']['has_reference_image']
     else 'No'}
- Generation seed: {config_context['
    generation_params']['seed']}

EVALUATION RESULTS:
- Automated score: {config_context['
    evaluation_metrics']['evaluation_score']}/10
- User feedback: {config_context['
    evaluation_metrics']['user_feedback'] or '
    None provided'}
```

```
REGENERATION STATUS:
- Current attempt: #{config_context['
    system_context']['current_attempt']} of {
    config_context['system_context']['
    total_attempts']} total
- Improvement suggestions: {config_context['
    evaluation_metrics']['
    improvement_suggestions'] or 'None from
    previous cycles'}
- Human oversight: {'Enabled' if config_context[
    'system_context']['human_in_loop'] else '
    Autonomous operation'}

ANALYSIS REQUIREMENTS:
Analyze this model's performance considering the
     execution flow:
1. How well did the model respond to the
    creativity level and prompt refinement
    quality?
2. Effectiveness of negative prompt in
    preventing unwanted artifacts
3. Quality of prompt polishing for this specific
     model's characteristics
4. Model selection appropriateness based on the
    refined prompt requirements
5. Technical execution quality visible in the
    generated image
6. Prompt adherence and creative interpretation
    balance
7. Reference image utilization (if applicable)

IMPORTANT: You will see the actual generated
    image(s) below. Provide specific visual
    analysis referencing the execution context.

Return a JSON response with detailed breakdown
    by workflow trajectory:
{{
    "trajectory_reasoning": "Overall analysis of
        how the workflow execution played out
        from start to finish, including key
        decision points, transitions between
        steps, and how each step influenced the
        next. Analyze the logical flow and
        coherence of the entire process.",
    "step_scores": {{
        "creativity_level": "Score from 1-10 how
            appropriate the creativity level
            setting was for this specific prompt
            ",
        "intention_analysis": "Score from 1-10
            how effective the intention analysis
            was for this prompt",
        "prompt_refinement": "Score from 1-10 how
             well the prompt was refined for
            optimal generation",
        "negative_prompt": "Score from 1-10 how
            effective the negative prompt was in
            preventing issues",
        "generation": "Score from 1-10 the
            overall quality of the generated
            image",
        "evaluation": "Score from 1-10 how
            accurate the system's evaluation was"
    }},
    "successes": {{
        "creativity_level": "How well the
            creativity level setting worked for
            this prompt and model",
        "intention_analysis": "Effectiveness of
            the intention analysis in guiding the
             process",
        "prompt_refinement": "Quality and
            appropriateness of the prompt
            refinement",
        "negative_prompt": "How well the negative
             prompt prevented unwanted artifacts"
            ,
        "generation": "Model selection
            appropriateness and generation
            quality",
        "evaluation": "Accuracy of evaluation
            scoring relative to visual quality"
    }},
    "pitfalls": {{
        "creativity_level": "Issues with
            creativity level setting for this
            prompt type",
        "intention_analysis": "Missed elements or
             incorrect analysis in intention
            phase",
        "prompt_refinement": "Problems with
            refined prompt quality or
            completeness",
        "negative_prompt": "Artifacts that
            negative prompt failed to prevent",
        "generation": "Model selection issues or
            generation quality problems",
        "evaluation": "Evaluation inaccuracies or
             scoring misalignment"
    }},
    "overall_rating": "Rate from 1-10 the overall
         effectiveness of the entire workflow
        for this specific prompt"
}}

Focus on specific observations from the actual
    generated image and how each workflow step
    contributed to or detracted from the final
    result."""
```

Note: If regeneration using Qwen-Image-Edit occurs, all the images are also passed to the A_{GUID} and the final information is added to both the databases of Qwen-Image and Qwen-Image-Edit.

Guidance generation from similar trajectories for (Qwen Image)

```
You are an expert at analyzing patterns in image
     generation workflows to provide SPECIFIC,
    CONCRETE, and ACTIONABLE guidance.

Your specialty is identifying context-specific
    techniques that work for particular types of
     images, not general best practices.

Focus on content-specific insights like:
- For portrait images: specific lighting
    techniques, expression guidance,
    compositional elements
- For landscape scenes: time of day effects,
    weather condition impacts, foreground
    element placement
- For abstract concepts: style reference
    importance, compositional balance techniques
    , color palette guidance

Avoid general advice like "add more detail" or "
    be more specific".
```

```
Instead provide domain-specific, technical
    recommendations based on the actual examples
     you analyze.

Based on the analysis of similar prompts,
    extract CONCRETE and SPECIFIC workflow
    guidance for the new prompt.

SIMILAR PROMPTS DATA:
{similar_data_text}

{workflow_description}

{task_focus}

Return JSON with this EXACT structure:
{{
    "step_analysis": {{

        "creativity_level_determination": {{
            "success_patterns": "SPECIFIC
                creativity level patterns that
                worked for this type of prompt",
            "failure_patterns": "SPECIFIC
                creativity level mistakes
                observed with similar prompts",
            "impact_on_next": "CONCRETE impact on
                 intention analysis quality",
            "preventive_guidance": "DETAILED
                advice with SPECIFIC examples of
                what to look for",
            "recommended_score": "Recommend
                target score (1-10) based on
                similar examples"
        }},
        "intention_analysis": {{
            "success_patterns": "SPECIFIC
                intention analysis techniques
                that worked for this subject
                matter",
            "failure_patterns": "SPECIFIC
                intention analysis pitfalls
                observed with similar prompts",
            "impact_on_next": "CONCRETE impact on
                 prompt refinement",
            "preventive_guidance": "DETAILED
                advice with SPECIFIC elements to
                identify",
            "recommended_score": "Recommend
                target score (1-10) based on
                similar examples"
        }},
        "prompt_refinement": {{
            "success_patterns": "SPECIFIC
                refinement strategies that
                enhanced similar prompts",
            "failure_patterns": "SPECIFIC
                refinement mistakes observed in
                similar cases",
            "impact_on_next": "CONCRETE impact on
                 negative prompt generation",
            "preventive_guidance": "DETAILED
                advice with SPECIFIC refinement
                techniques",
            "recommended_score": "Recommend
                target score (1-10) based on
                similar examples"
        }},
        "negative_model_selection": {{
            "success_patterns": "SPECIFIC
                negative prompt and model
                selection strategies that were
                effective for this subject",
            "failure_patterns": "SPECIFIC
                negative prompt and model
                selection issues observed in
                similar cases",
            "impact_on_next": "CONCRETE impact on
                 prompt polishing",
            "preventive_guidance": "DETAILED
                advice with SPECIFIC negative
                prompt terms and model selection
                criteria",
            "recommended_score": "Recommend
                target score (1-10) based on
                similar examples"
        }},
        "image_generation": {{
            "success_patterns": "SPECIFIC
                generation parameters that worked
                 for similar content",
            "failure_patterns": "SPECIFIC
                generation issues observed with
                this prompt type",
            "impact_on_next": "CONCRETE impact on
                 evaluation accuracy",
            "preventive_guidance": "DETAILED
                advice with SPECIFIC generation
                settings",
            "recommended_score": "Recommend
                target score (1-10) based on
                similar examples"
        }},
        "quality_evaluation": {{
            "success_patterns": "SPECIFIC
                evaluation criteria effective for
                 this content type",
            "failure_patterns": "SPECIFIC
                evaluation pitfalls observed with
                 similar outputs",
            "impact_on_next": "CONCRETE impact on
                 regeneration decision",
            "preventive_guidance": "DETAILED
                advice with SPECIFIC quality
                indicators",
            "recommended_score": "Recommend
                target score (1-10) based on
                similar examples"
        }},
        "regeneration_decision": {{
            "success_patterns": "SPECIFIC
                decision criteria that led to
                successful outcomes",
            "failure_patterns": "SPECIFIC
                regeneration mistakes observed
                with similar content",
            "impact_on_next": "N/A - final step",
            "preventive_guidance": "DETAILED
                advice with SPECIFIC decision
                factors",
            "recommended_score": "Recommend
                target score (1-10) based on
                similar examples"
        }}
    }},

    "workflow_insights": {{
        "critical_dependencies": "SPECIFIC step
             dependencies most relevant to this
             prompt type",
```

```
        "common_failure_chains": "CONCRETE
            failure patterns observed in similar
            cases",
        "success_combinations": "SPECIFIC
            combinations of choices that worked
            well for this content",
        "overall_rating_prediction": "Predict the
             likely overall success rating (1-10)
             for this prompt type"
    }}
}}

Your guidance must be SPECIFIC to the prompt
     type and content domain, not generic best
     practices.
For example, instead of 'Use detailed prompts',
     say 'For architectural images, specify
     architectural style, materials, lighting
     conditions, and surrounding environment'.
Use concrete, actionable advice derived from the
      data, not theoretical recommendations.

IMPORTANT: Ensure the guidance provides DOMAIN-
     SPECIFIC advice for the content type in the
     prompts, not just generic image generation
     tips.
```

Guidance generation from similar trajectories for (Qwen Image Edit)

```
You are an expert at analyzing patterns in image
      generation workflows to provide SPECIFIC,
     CONCRETE, and ACTIONABLE guidance.

Your specialty is identifying context-specific
     techniques that work for particular types of
      images, not general best practices.

Focus on content-specific insights like:
- For portrait images: specific lighting
     techniques, expression guidance,
     compositional elements
- For landscape scenes: time of day effects,
     weather condition impacts, foreground
     element placement
- For abstract concepts: style reference
     importance, compositional balance techniques
     , color palette guidance

Avoid general advice like "add more detail" or "
     be more specific".
Instead provide domain-specific, technical
     recommendations based on the actual examples
      you analyze.

Based on the analysis of similar prompts,
     extract CONCRETE and SPECIFIC workflow
     guidance for the new prompt.

SIMILAR PROMPTS DATA:
{similar_data_text}

{workflow_description}

{task_focus}

Return JSON with this EXACT structure:
{{
    "step_analysis": {{

        "image_editing": {{
            "success_patterns": "SPECIFIC editing
                 parameters and techniques that
                worked for similar content and
                edit types",
            "failure_patterns": "SPECIFIC editing
                 issues observed with this edit
                operation type",
            "impact_on_next": "CONCRETE impact on
                 evaluation accuracy",
            "preventive_guidance": "DETAILED
                advice with SPECIFIC editing
                settings, reference image usage,
                and blending techniques",
            "recommended_score": "Recommend
                target score (1-10) based on
                similar examples"
        }},
        "quality_evaluation": {{
            "success_patterns": "SPECIFIC
                evaluation criteria effective for
                 this edit content type",
            "failure_patterns": "SPECIFIC
                evaluation pitfalls observed with
                 similar edit outputs",
            "impact_on_next": "CONCRETE impact on
                 re-edit decision",
            "preventive_guidance": "DETAILED
                advice with SPECIFIC edit quality
                 indicators and success metrics",
            "recommended_score": "Recommend
                target score (1-10) based on
                similar examples"
    }},

    "workflow_insights": {{
        "critical_dependencies": "SPECIFIC step
             dependencies most relevant to this
             prompt type",
        "common_failure_chains": "CONCRETE
             failure patterns observed in similar
             cases",
        "success_combinations": "SPECIFIC
             combinations of choices that worked
             well for this content",
        "overall_rating_prediction": "Predict the
              likely overall success rating (1-10)
              for this prompt type"
    }}
}}

Your guidance must be SPECIFIC to the prompt
     type and content domain, not generic best
     practices.
For example, instead of 'Use detailed prompts',
     say 'For architectural images, specify
     architectural style, materials, lighting
     conditions, and surrounding environment'.
Use concrete, actionable advice derived from the
      data, not theoretical recommendations.

IMPORTANT: Ensure the guidance provides DOMAIN-
     SPECIFIC advice for the content type in the
     prompts, not just generic image generation
     tips.
```

L Algorithm

Our framework is built around an agentic workflow that coordinates prompt engineering, image

Algorithm 1: Main algorithm for *SIDiffAgent*

Input: User prompt P, Models: QI (Qwen-Image), QIE (Qwen-Image-Edit), QE (Qwen-Embedding), Knowledge base KB (initially may be empty), Hyperparams: τ (eval threshold), E (max edits), K (retrieval size)

Output: Final image I^*, final prompt F, recorded trajectory T

Function `SIDiffAgent(`P`)`**:**
- $G \leftarrow$ `RetrieveGuidance(`KB, P, K`)` `// retrieve and generate guidance`
- $c \leftarrow S_{\text{CRE}}(P, G)$ `// creativity level: {low,med,high}`
- $S \leftarrow S_{\text{INT}}(P, c, G)$ `// semantic specification, disambiguations`
- $P_{pos} \leftarrow S_{\text{REF}}(P, S, G)$ `// refined positive prompt`
- $P_{neg} \leftarrow S_{\text{NEG}}(P, P_{pos}, G)$ `// adaptive negative prompt + universal safeguards`
- $edits \leftarrow 0$
- $I \leftarrow \varnothing$
- $T \leftarrow \varnothing$
- **while** $edits \leq E$ **do**
  - $I \leftarrow S_{\text{GEN}}(P_{pos}, P_{neg}, I)$ `// use QI`
  - $report \leftarrow A_{\text{EVAL}}(I, P, P_{pos}, G)$ `// returns scores and suggestions`
  - Record decision node results in T via `RecordTrajectory()`
  - $score \leftarrow$ average(report.aesthetic, report.alignment)
  - **if** $score \geq \tau$ **then**
    - **break** `// satisfactory output`
  - **else**
    - $P^{new}_{pos}, P^{new}_{neg} \leftarrow$ re-run A_{ORC}; **if** *report suggests localized correction* ***and*** $edits < E$ **then**
      - $I \leftarrow S_{\text{GEN}}(P^{new}_{pos}, P^{new}_{neg}, \text{edit}, I)$ `// use QIE`
      - $edits \leftarrow edits + 1$
  - **end**
- **end**
- `UpdateKnowledgeBase(`KB, T`)`
- **return** I, P_{pos}, P_{neg}, T

end

generation, evaluation, and self-improvement. This workflow is orchestrated by three primary agents, the Generation Orchestrator Agent (A_{ORC}), which preprocesses user prompts through five sub-agents, the Creativity Analysis Sub-Agent (S_{CRE}), Intention Analysis Sub-Agent (S_{INT}), Prompt Refinement Sub-Agent (S_{REF}), Adaptive Negative Prompt Sub-Agent (S_{NEG}), and the Generation Sub-Agent (S_{GEN}) for final generation. The Evaluation Agent (A_{EVAL}) assesses each output, triggering regeneration when necessary. Finally, the Guidance Agent (A_{GUID}) facilitates learning from past experiences. The overall workflow of *SIDiffAgent* is depicted in Figure 2. Each agent is discussed in detail below.

M Qualitative Results

Algorithm 2: Helper functions for *SIDiffAgent* (part 1 of 2): generation orchestration sub-agents

```
Function S_CRE(P, G):
    analyse guidance G, compute length, specificity cues,
        presence of attributes
    if very short or vague then
        return high
    else if moderately detailed then
        return medium
    else
        return low
    end
end
Function S_INT(P, c, G):
    analyse guidance G, extract main subjects, attributes,
        relations, composition, lighting, style
    if c = high then
        infer missing attributes using templates and priors
    else if c = medium then
        propose optional clarifications and minor assumptions
    else
        preserve original prompt
        avoid adding assumptions
    end
    return structured specification S
end
Function S_REF(P, S, G):
    re-order specification into coherent positive prompt
        text P_pos
    enforce lexical templates for model compatibility
    return P_pos
end
Function S_NEG(P, P_pos, G):
    U ← universal safeguards (e.g., low quality,
        blurry, watermark)
    derive scene-specific negations from P_pos
    compose concise P_neg ← U∪ scene-specific
        negations
    return P_neg
end
Function S_GEN(P_pos, P_neg, mode, I_in = ∅):
    if mode = initial then
        return QI.generate(P_pos, P_neg)
    else if mode = edit then
        return QIE.edit(I_in, P_pos, P_neg)
end
```

Algorithm 3: Helper functions for *SIDiffAgent* (part 2 of 2): evaluation, retrieval, and knowledge-base updates

```
Function A_EVAL(I, P, P_pos, G):
    a ← A_EVAL.scoreAesthetic(I)
    α ← A_EVAL.scoreAlignment(I, P or P_pos)
    sugs ← A_EVAL.suggestFixes(I)
    arts ← A_EVAL.detectArtifacts(I)
    return report with (a, α, sugs, arts)
end
Function RetrieveGuidance(KB, P, K):
    if KB empty then
        return ∅
    embed P into vector v
    retrieve top-K trajectories
        by similarity
    aggregate pitfalls and successful actions across
        trajectories
    return guidance G (structured corrective +
        workflow guidance)
end
Function RecordTrajectory():
    For each decision node, store: node-id, inputs,
        actions, outputs, scores
end
Function UpdateKnowledgeBase(KB, T):
    compress T into node-wise summaries of pitfalls
        and successes
    append to model-specific sections in KB
    optionally re-index embeddings for efficient retrieval
end
```

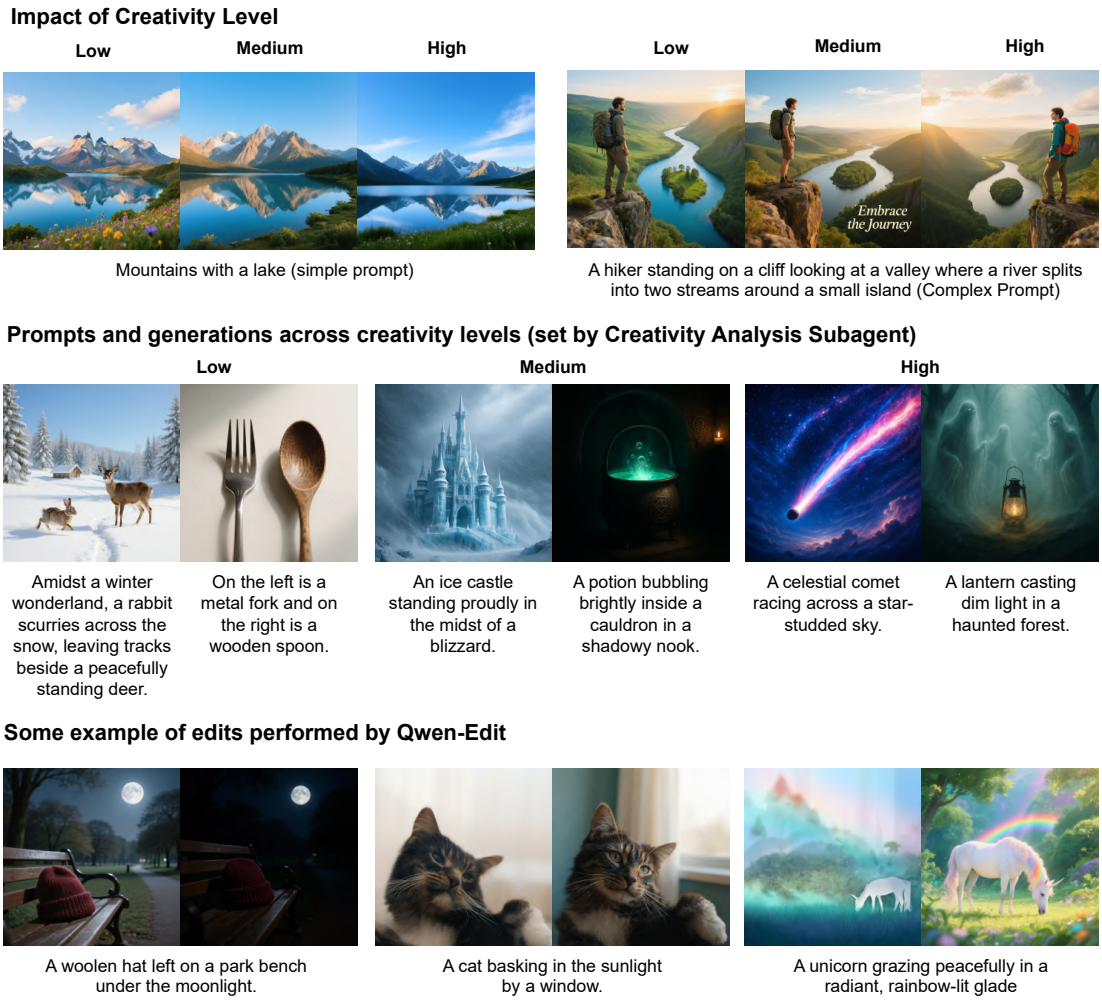


Figure 8: **Top:** Effect of varying creativity levels on simple and complex prompts. **Middle:** Image generations from diverse prompts, with creativity levels determined by S_{CRE}. **Bottom:** Refinements and enhancements performed by Qwen-Edit on initial suboptimal generations.

Figure 9: Qualitative comparison between multiple state-of-the-art method and *SIDiffAgent* on GenAI-Bench

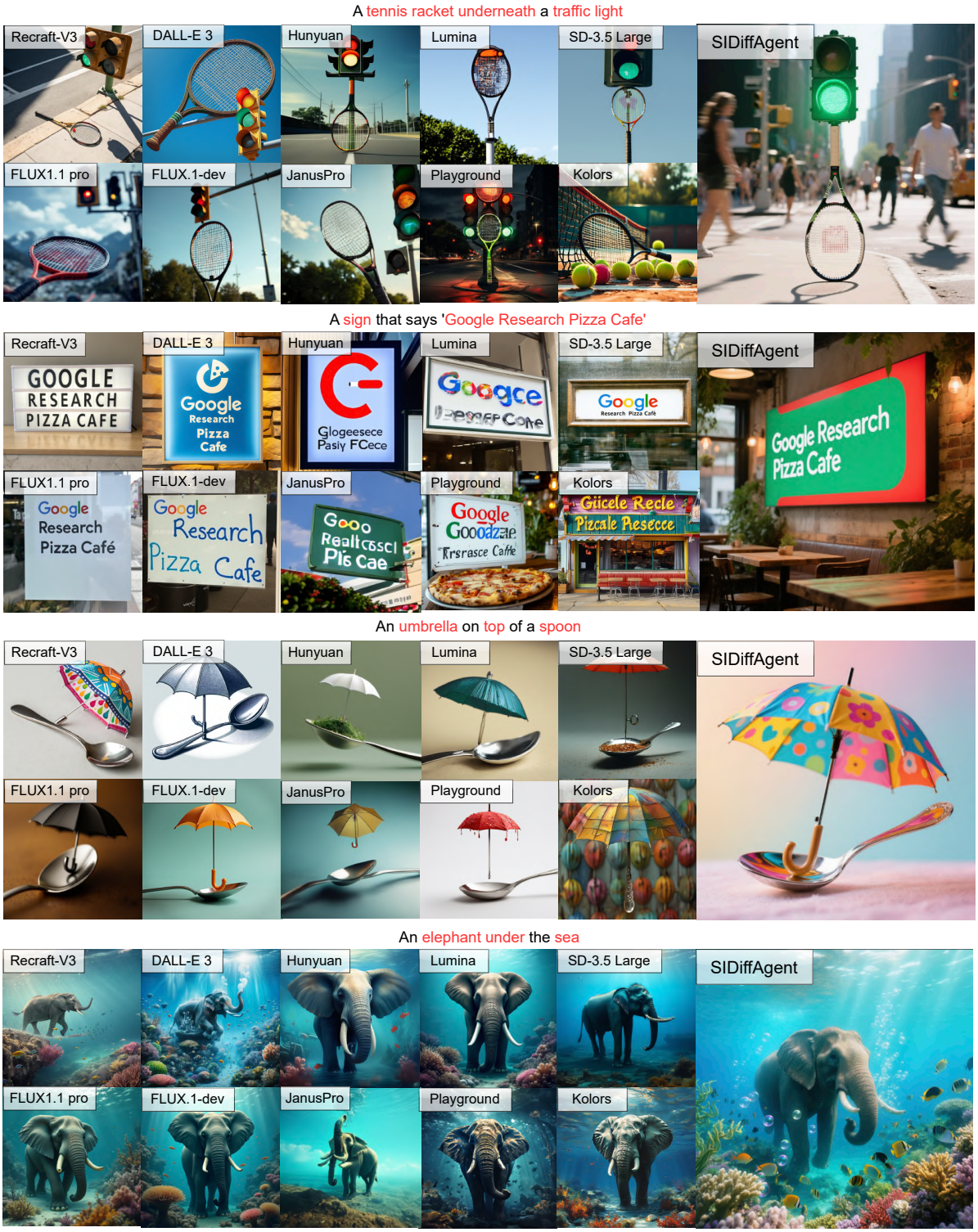


Figure 10: Qualitative comparison between multiple state-of-the-art method and *SIDiffAgent* on DrawBench

Figure 11: Qualitative comparison between various ablations described in Section 4.2