RS-GEN: A MULTI-STAGE AGENTIC FRAMEWORK FOR REASONING AND SEARCH-AUGMENTED IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

While image generation and editing have progressed significantly, existing models still struggle with ambiguous intent, complex reasoning, and out-of-distribution (OOD) tasks due to static knowledge gaps and reasoning limitations. To overcome these bottlenecks, we propose **RS-Gen**, a plug-and-play, training-free, multi-stage agentic framework for image generation and editing. By introducing a closed-loop "question-posing and problem-solving" mechanism, RS-Gen autonomously plans retrieval actions to bridge information gaps and execute deep logical reasoning. Experiments show that RS-Gen yields absolute gains of 0.31 (on a 0–1 scale) and 19.7 over Qwen-Image and Qwen-Image-Edit-2511 on WISE_Verified and RISEBench, respectively, achieving new SOTA results among open-source models and considerably pushing the capability boundaries of foundational backbones.

1 INTRODUCTION



Figure 1: Qualitative generation examples produced by RS-Gen.

Despite remarkable advances in visual generation, existing models remain hindered from true "intelligent generation" by three bottlenecks: (1) **Weak implicit intent parsing**: difficulty resolving ambiguous object references, especially in multi-turn interactions. (2) **Lack of visual logical reasoning**: inability to reliably handle visual puzzles and state transitions while maintaining causal and physical consistency. (3) **Static knowledge boundaries**: outdated training data causes factual errors and hallucinations on long-tail entities and out-of-distribution (OOD) concepts.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address these limitations, researchers have explored unified understanding and generation architectures, such as Transfusion Zhou et al. (2024), JanusFlow Ma et al. (2024), and Bagel Deng et al. (2025). While improving intention understanding and logical reasoning, their monolithic designs remain limited in multi-step reasoning and out-of-distribution (OOD) knowledge processing. Recent advances in AI agents have further inspired image agent systems that combine search and reasoning, including commercial products such as Seedream Chen et al. (2025), Nano Banana Pro Google DeepMind (2025), and FLUX-2 Max Black Forest Labs (2026), as well as open-source efforts such as Mind-Brush He et al. (2026) and Unify-Agent Chen et al. (2026a). By leveraging multimodal reasoning and external search, these systems improve contextual knowledge and generation quality. However, existing open-source solutions remain limited in complex problem solving and often require extensive training and costly model adaptation, restricting their generality across open-source image models. This leaves a substantial gap between open-source research and commercial systems in advancing toward "intelligent generation."

Inspired by OpenClaw OpenClaw Team (2026), we propose RS-Gen, a training-free, plug-and-play multi-stage agentic framework that augments image generation and editing with reasoning and search. RS-Gen formulates these tasks as a "question-posing and problem-solving" closed loop, enabling existing open-source image models to better handle implicit intentions, logical reasoning, and real-time information. As shown in Figure 1, RS-Gen achieves accurate visual generation on challenging tasks requiring up-to-date knowledge and complex graphical reasoning.

Specifically, the core advantages of RS-Gen include:

- Multi-turn implicit intention parsing: Leveraging agentic memory and multimodal reasoning, RS-Gen accurately resolves co-references and grounds target objects across multi-turn interactions, reconstructing ambiguous and implicit intentions into explicit prompts for generation and editing.
- Knowledge gap and logical flaw detection: By evaluating instruction complexity in advance, the framework identifies potential information deficits and logical dilemmas, leveraging a proactive questioning mechanism to guide subsequent targeted retrieval and deep reasoning.
- Retrieval-augmented visual generation: Integrating external search tools, RS-Gen breaches the static knowledge boundaries of foundation models, providing real-time factual and logical grounding for complex visual generation and manipulation.
- Autonomous planning and self-correction: Prior to generation, it follows the ReAct Yao et al. (2023) ("Observe-Think-Act") paradigm to autonomously plan steps, invoke tools, and dynamically adjust strategies, resolving knowledge and logical deficits beforehand; during generation, it introduces a "Generate-Review-Refine" self-correction closed loop, considerably enhancing output reliability and robustness.

2 Related Work

2.1 Image Generation and Editing

Recent years have witnessed milestone advancements in visual generation. From Latent Diffusion Models (LDMs) represented by Stable Diffusion Rombach et al. (2021), which perform iterative denoising in low-dimensional latent spaces, to subsequent cutting-edge models like SD3 Esser et al. (2024) and FLUX.1 Black Forest Labs (2024) that incorporate Flow Matching and large-scale Transformer architectures, visual fidelity and realism have been pushed to unprecedented heights. Despite these dramatic qualitative leaps, existing models remain trapped in a "semantic gap." Highly dependent on static text encoders such as CLIP Radford et al. (2021) and T5 Raffel et al. (2020), their instruction parsing and following capabilities degrade significantly when handling complex prompts involving ambiguous intentions or co-references due to semantic representation bottlenecks. Fundamentally, their generation process remains restricted to a superficial "text-to-pixel" statistical mapping, lacking deep reasoning capabilities for the higher-order semantics and complex logical flaws underlying intricate instructions.

As user demands evolve toward fine-grained visual modifications, image editing has emerged as a highly challenging frontier. Unlike generation from scratch, editing imposes stricter multidimen-

sional constraints: models must accurately execute instructions while strictly preserving the semantic coherence and visual texture consistency of unedited regions. To this end, InstructPix2Pix Brooks et al. (2023) pioneered the data-driven "instruction-guided editing" paradigm. Subsequent representative works—such as MagicBrush Zhang et al. (2023), HQ-Edit Hui et al. (2025), and Emu Edit Sheynin et al. (2024)—have fine-tuned diffusion models on large-scale "source-instruction-target" triplet datasets, significantly enhancing editing precision and generalization across fine-grained tasks like local manipulation, style transfer, and attribute modification.

To further alleviate the bottlenecks in understanding complex instructions, recent studies have incorporated Vision-Language Models (VLMs) into editing architectures. Pioneering works—such as OmniGen2 Wu et al. (2025c), Qwen-Image-Edit Wu et al. (2025a), and Step1X-Edit Liu et al. (2025)—innovatively replace traditional static encoders (e.g., CLIP or T5) with VLMs, utilizing their deep semantic representations as conditioning signals for generation. By leveraging the robust multimodal understanding and logical reasoning of VLMs, these architectures considerably enhance semantic parsing, context awareness, and fine-grained control, substantially improving comprehension when handling long-form text and intricate instructions.

2.2 Unified Multimodal Understanding and Generation Models

To fundamentally bridge the gap between understanding and generation, Unified Multimodal Understanding and Generation Models have emerged. Departing from traditional decoupled paradigms (i.e., independent text encoders combined with image generators), these architectures map visual understanding and generation into a shared representational space for joint modeling. Representative works—such as Chameleon Chameleon Team (2024), Show-o Xie et al. (2025), Janus Wu et al. (2025b), and Bagel Deng et al. (2025)—have demonstrated the immense potential of native unification; leveraging inherent robust multimodal perception, these models achieve a profound grasp of physical common sense and logical laws, thereby significantly mitigating realistic "visual hallucinations" in synthesized outputs.

Nevertheless, despite the enhanced semantic comprehension afforded by VLMs and unified architectures, deploying these models on real-world, complex tasks remains impeded by four core bottlenecks:

- Contextual co-reference ambiguity in multi-turn interactions: Real-world scenarios often involve continuous, highly conversational turns. Lacking memory and disambiguation mechanisms, existing models struggle to ground and track target objects across multi-turn dialogues.
- Static knowledge cut-offs and visual hallucinations: Constrained by static training data and knowledge cut-offs, models cannot perceive real-time information or long-tail entities, rendering them highly susceptible to visual hallucinations when processing out-of-distribution (OOD) concepts.
- A lack of explicit logical reasoning: Relying on superficial "text-to-pixel" statistical mappings, models depend heavily on explicit prompts; when confronting tasks involving implicit physical laws or visual logical puzzles, they lack the capacity for deep reasoning and pattern mining.
- Prohibitive training and iteration costs: Unified models rely heavily on massive interleaved text-image datasets and large-scale computational resources, resulting in prohibitive training and fine-tuning costs.

2.3 Image Agent Systems

To overcome the performance bottlenecks of monolithic architectures in complex multimodal understanding and generation, the artificial intelligence community is witnessing a paradigm shift from passive instruction-following models toward autonomous AI agents. Modern agents have transcended basic unimodal text interactions and question-answering, evolving into compound architectures that integrate high-order capabilities such as multimodal perception, long- and short-term memory, autonomous multi-step planning, and external tool invocation.

Within agentic architectures, the ReAct paradigm ("Reason-Act-Observe") and cutting-edge applications like OpenClaw have demonstrated the immense potential of dynamic feedback and tool synergy in executing automated workflows. Inspired by these advances, researchers have explored Image Agents to reframe visual generation and editing as agent-driven closed-loop tasks. Representative works—such as Mind-Brush He et al. (2026), Unify-Agent Chen et al. (2026a), and Gen-Searcher Feng et al. (2026)—leverage Multimodal Large Language Models (MLLMs) as central decision controllers to invoke external search tools, effectively bridging intrinsic knowledge gaps and mitigating factual hallucinations. These efforts signify a pivotal evolution in visual generation: moving away from superficial "text-to-pixel" mappings toward autonomous agentic paradigms equipped with complex task planning and closed-loop execution.

Nevertheless, existing Image Agents remain at an exploratory stage, facing two significant challenges when confronting intricate tasks:

- **High coupling and training costs:** Most frameworks are model-specific and require costly fine-tuning (e.g., Unify-Agent uses 143k customized multimodal trajectories), limiting plug-and-play generalization.
- **Rigid workflows and poor fault tolerance:** Linear pipelines lack self-reflection and error correction, allowing intermediate errors to propagate and degrade factual and logical consistency.

3 RS-GEN

We propose RS-Gen, a multi-stage agentic framework that augments image generation and editing with reasoning and search (Figure 2). Moving beyond traditional end-to-end black-box mapping, RS-Gen is collaboratively driven by four specialized agents: prior to generation, it leverages explicit reasoning and retrieval feedback loops to resolve ambiguity, bridge knowledge gaps, and complete reasoning, converting vague inputs into precise instructions and reference images; during generation, it incorporates a "generate-verify-modify" self-correcting closed loop to further enhance output accuracy and reliability.

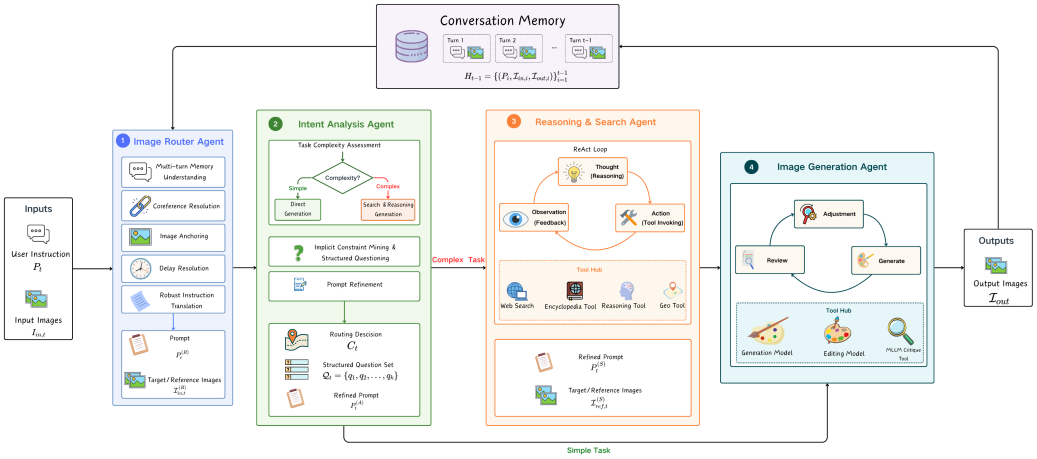


Figure 2: The overall architecture of the RS-Gen framework.

3.1 IMAGE ROUTER AGENT

Serving as the system entry point, this agent leverages memory mechanisms to resolve contextual coreferences and ground target regions within images, thereby transforming ambiguous, unstructured user inputs into clear, structured operational instructions. Formally, given the t-th turn's raw user instruction P_t, input image set $\mathcal{I}_{in,t}$, and multimodal dialogue history $H_{t-1} = \{(P_i, \mathcal{I}_{in,i}, \mathcal{I}_{out,i})\}_{i=1}^{t-1}$,

the agent outputs the disambiguated instruction $P_t^{(R)}$ along with the grounded target/reference image set $\mathcal{I}_{in,t}^{(R)}$:

$$(P_t^{(R)}, \mathcal{I}_{in,t}^{(R)}) = F^{(R)}(P_t, \mathcal{I}_{in,t}, H_{t-1}). \tag{1}$$

To prevent hallucinations caused by unsupported prior inferences, we introduce a "Deferred Parsing" mechanism. This module is strictly restricted to image routing and co-reference resolution, explicitly prohibiting heuristic completion. When instructions involve higher-order semantics—such as external knowledge or logical deduction—they are preserved as unresolved variables within $P_t^{(R)}$ and deferred to the downstream "Reasoning & Search Agent," thereby guaranteeing the factual grounding and logical consistency of the final output.

Furthermore, to mitigate the inherent weakness of diffusion models in following negative instructions, we incorporate a negative-to-affirmative semantic inversion strategy when formulating $P_t^{(R)}$. By prompting the agent to translate negative constraints (e.g., "do not keep stickers") into equivalent positive visual attributes (e.g., "a clean, solid-colored surface"), we effectively mitigate feature contamination at the source.

3.2 Intent Analysis Agent

Serving as a task router and complexity evaluator, the agent assesses the task difficulty of the input $(P_t^{(R)}, \mathcal{I}_{in,t}^{(R)})$ and performs dynamic routing: straightforward tasks (C_t = direct_generation) are directly dispatched for generation, whereas complex tasks (C_t = search_reason_generation)—those necessitating external knowledge or logical reasoning—prompt the agent to identify knowledge gaps and extract target questions for resolution. Formally, via the mapping function $F^{(A)}$, it outputs a tuple comprising a routing identifier C_t, a structured question set $\mathcal{Q}_t$, and a reconstructed instruction $P_t^{(A)}$:

$$(C_t, \mathcal{Q}_t, P_t^{(A)}) = F^{(A)}(P_t^{(R)}, \mathcal{I}_{in,t}^{(R)}). \tag{2}$$

When the task routing is determined as C_t = search_reason_generation ($P_t^{(A)} = P_t^{(R)}$), the agent uncovers implicit constraints behind the instruction and converts them into an explicit, comprehensive question set $\mathcal{Q}_t = \{q_1, q_2, \ldots, q_k\}$ for downstream agents via multi-dimensional questioning strategies. (1) Defensive Questioning. Except for generic basic entities, the strategy strictly prohibits the model from relying on internal prior knowledge for heuristic completion. Instead, it enforces explicit queries targeting their exact objective visual features, thereby erecting a "cognitive firewall" against hallucinations. (2) Process Decoupling and Final-State Grounding. For instructions involving complex dynamic reasoning (e.g., physical laws or visual puzzles), the strategy compels the agent to bypass intermediate reasoning details and anchor queries directly on final-state visual attributes—such as terminal physical appearances and spatial layouts—that are directly renderable by diffusion models.

When the task routing is determined as C_t = direct_generation, the agent bypasses the questioning phase (i.e., $\mathcal{Q}_t = \emptyset$) and performs adaptive refinement on $P_t^{(R)}$ based on the task type and input modality, generating a render- and edit-friendly instruction $P_t^{(A)}$ for underlying models: (1) High-Density Semantic Expansion (for text-to-image tasks). It automatically supplements fine-grained visual details—such as composition, lighting, and textures—to mitigate the information sparsity of the initial text. (2) Final-State Visual Grounding (for image-to-image tasks). It filters out procedural action descriptions and retains only absolute objective features of the expected final state, thereby minimizing the risk of feature confusion in diffusion models. (3) Structured Action Decomposition (for image editing tasks). It reconstructs ambiguous instructions into a structured paradigm of "Action + Target Object + Resulting State" to guide underlying models in executing precise modifications.

3.3 Reasoning & Search Agent

Activated when the routing decision is C_t = search_reason_generation, this module is responsible for invoking external tools to perform logical reasoning and knowledge completion. Given the input instruction $P_t^{(A)}$, image set $\mathcal{I}_{in,t}^{(R)}$, structured question set $\mathcal{Q}_t$, and tool library $\mathcal{T}$, it leverages the ReAct paradigm to execute multi-round retrieval and reasoning. Ultimately, it outputs a logically consistent and information-complete operational instruction $P_t^{(S)}$ alongside a high-quality visual reference set $\mathcal{I}_{ref,t}^{(S)}$:

$$\left(P_t^{(S)}, \mathcal{I}_{ref,t}^{(S)}\right) = F^{(S)}\left(P_t^{(A)}, \mathcal{I}_{in,t}^{(R)}, \mathcal{Q}_t, \mathcal{T}\right). \tag{3}$$

To address the inherent uncertainties of tool invocation (e.g., retrieval failures or noisy results), this module introduces a cross-modal cascaded retrieval and adaptive fallback mechanism: (1) Text-to-Image Cascaded Anchoring. For ambiguous entities, directly invoking the image search tool τ_{img} is strictly prohibited. Instead, the agent must first use web search τ_{web} to determine the exact objective identifier (e.g., a specific model name) and subsequently use it as the query for image retrieval, eliminating semantic ambiguity at the source. (2) Fallback Tolerance and Visual Substitution. If image retrieval fails or yields subpar quality, the agent sequentially triggers fallback strategies based on priority: switching alternative tools, relaxing semantic constraint boundaries, or seeking visually similar alternative entities. This ensures reliable factual and visual grounding even under extreme information deficiency.

3.4 Image Generation Agent

Moving beyond the traditional static single-step generation pipeline, this agent encapsulates underlying generation models, editing models, and Vision-Language Models (VLMs) into an expert toolset $\mathcal{T}_{gen}$ to construct an iterative self-correction closed loop based on the ReAct paradigm. Given the dynamically routed upstream instruction $P_t^{(G)} \in \{P_t^{(A)}, P_t^{(S)}\}$ and the reference image set $\mathcal{I}_{ref}^{(G)} \in \{\mathcal{I}_{in,t}^{(R)}, \mathcal{I}_{ref,t}^{(S)}\}$, the agent dynamically invokes tools to synthesize the target image set $\mathcal{I}_{out,t}$ that highly aligns with user intent:

$$\mathcal{I}_{out,t} = F^{(G)}(P_t^{(G)}, \mathcal{I}_{ref}^{(G)}, \mathcal{T}_{gen}). \tag{4}$$

Breaking away from the traditional open-loop paradigm, this module executes an iterative "Synthesis-Verification-Correction" closed loop. In the k-th iteration: (1) Dynamic Synthesis. The agent parses the current strategy p_k (where initial state $p_0 = P_t^{(G)}$) and schedules $\mathcal{T}_{gen}$ to generate a candidate image I_k. (2) Multimodal Verification. A VLM is invoked to examine the fine-grained semantic alignment between I_k and $P_t^{(G)}$, as well as image quality, outputting a structured diagnostic report v_k. (3) Strategy Correction. If v_k detects semantic deviations or visual artifacts, the agent derives a correction strategy Δp_k and updates parameters: $p_{k+1} = \text{Update}(p_k, \Delta p_k, P_t^{(G)})$. This loop iterates until the candidate image fully aligns with the user intent or the maximum iteration limit is reached, terminating to output the final result $\mathcal{I}_{out}$. This correction mechanism effectively mitigates the poor instruction-following and visual quality issues typical of traditional open-loop models when handling complex instructions, significantly enhancing the reliability and robustness of the system output.

4 Experiments

4.1 Benchmarks and Evaluation Protocols

To evaluate RS-Gen on complex instruction understanding, knowledge retrieval, and logical reasoning, we conduct experiments on two challenging benchmarks: (1) Image Generation (WISE Niu et al. (2025)): evaluates world-knowledge integration and complex semantics across 25 sub-fields using WiScore, which measures Consistency, Realism, and Aesthetic Quality. We use the official

Table 1: Performance on WISE_Verified Niu et al. (2025). Best results are bold.

Model	Cult.	Time	Spac.	Bio.	Phys.	Chem.	Avg.
Commercial Models							
Seedream-4.5	0.90	0.75	0.82	0.72	0.69	0.73	0.81
GPT-Image-1	0.89	0.69	0.88	0.80	0.76	0.78	0.83
Nano Banana Pro	0.90	0.82	0.93	0.82	0.87	0.88	**0.88**
Generation Only Models							
SD-1.5	0.45	0.31	0.21	0.21	0.22	0.15	0.31
SD-XL-0.9	0.49	0.37	0.24	0.27	0.33	0.18	0.36
FLUX.1-schnell	0.47	0.33	0.47	0.21	0.38	0.10	0.36
SD-3.5-medium	0.48	0.38	0.38	0.18	0.39	0.20	0.38
SD-3-medium	0.47	0.41	0.40	0.20	0.38	0.26	0.39
FLUX.2-klein-4B	0.44	0.37	0.47	0.32	0.39	0.33	0.40
SD-3.5-large	0.49	0.41	0.44	0.30	0.38	0.21	0.40
FLUX.1-dev	0.52	0.40	0.53	0.18	0.38	0.24	0.42
FLUX.2-klein-9B	0.49	0.39	0.55	0.38	0.48	0.23	0.44
Z-Image	0.55	0.47	0.51	0.33	0.48	0.18	0.45
Qwen-Image-2512	0.60	0.48	0.60	0.35	0.49	0.26	0.50
Qwen-Image	0.63	0.53	0.56	0.34	0.48	0.25	0.51
FLUX.2-dev	0.67	0.57	0.66	0.37	0.53	0.38	**0.57**
Unified Models							
Janus-Pro-1B	0.31	0.23	0.23	0.22	0.31	0.20	0.27
Janus-1.3B	0.32	0.28	0.18	0.23	0.34	0.18	0.27
Janus-Pro-7B	0.37	0.35	0.28	0.28	0.40	0.23	0.33
Bagel	0.41	0.35	0.31	0.20	0.44	0.26	0.35
Uniworld-V1	0.52	0.49	0.55	0.23	0.40	0.17	0.43
HunyuanImage-3.0	0.53	0.39	0.48	0.31	0.45	0.24	0.44
DeepGen 1.0	0.65	0.41	0.72	0.39	0.59	0.45	0.57
BAGEL (w/ CoT)	0.78	0.63	0.57	0.38	0.55	0.51	**0.63**
RS-Gen (Ours)	**0.90**	**0.88**	**0.91**	**0.60**	**0.69**	**0.78**	**0.82**

WISE_Verified Niu et al. (2025) benchmark with Qwen3.5-35B-A3B as the judge. (2) Image Editing (RISEBench Zhao et al. (2025)): evaluates temporal, causal, spatial, and logical reasoning, with GPT-4.1-2025-04-14 as the judge and three metrics: Instruction Reasoning, Appearance Consistency, and Visual Plausibility.

4.2 Experimental Settings

As a plug-and-play framework, RS-Gen uses consistent configurations across all benchmarks, with Doubao-Seed-2.0-Pro-260215 as the central driving model and Qwen-Image Wu et al. (2025a) and Qwen-Image-Edit-2511 Wu et al. (2025a) for generation and editing, respectively. Its expert tool library includes four utilities: Web Search (Baidu and Google, Top-5 results), Geographic Info (LocationIQ), Encyclopedia (Wikipedia), and Deep Reasoning (GPT-5.4).

4.3 Main Results

Table 1 presents the quantitative evaluation on WISE_Verified: (1) Non-commercial SOTA. RS-Gen obtains the top score of 0.82, significantly outperforming the runner-up BAGEL (w/ CoT) (0.63) and achieving leading performance across all six knowledge sub-fields. (2) Commercial competitiveness. Compared to the best-performing proprietary model, Nano Banana Pro (0.88), our system narrows the performance gap to merely 0.06, demonstrating robust competitiveness. (3) Substantial baseline enhancement. By yielding an absolute leap of 0.31 over the baseline Qwen-Image (0.51), RS-Gen thoroughly validates its core value in bridging foundational knowledge deficits and reinforcing complex reasoning.

Table 2 reports results on RISEBench. Three findings stand out: (1) Strong performance among non-commercial methods: RS-Gen achieves 39.1, surpassing Qwen-Image-Edit-2511 (19.4) and

Table 2: Performance on RISEBench Zhao et al. (2025). Best results are bold; "-" indicates unavailable results. Gem-2.0 denotes Gemini-2.0-Flash.

Model	Sub-Dim.			Accuracy				
	I.R.	A.C.	V.P.	Temp.	Caus.	Spat.	Logi.	Avg.
Seedream-4.0	58.9	67.4	91.2	17.6	13.3	11.0	7.1	12.2
GPT-Image-1-mini	54.1	71.5	93.7	25.9	31.1	33.0	9.4	25.3
GPT-Image-1	62.8	80.2	94.9	36.5	34.4	37.0	10.6	30.0
Nano Banana	61.2	86.0	91.3	29.4	48.9	37.0	18.8	33.9
Nano Banana Pro	77.0	85.5	94.4	43.5	63.3	48.0	37.6	48.3
GPT-Image-2	73.8	89.3	94.9	45.9	66.7	50.0	34.1	49.4
GPT-Image-1.5	69.7	92.5	94.9	57.6	62.2	62.0	21.2	**51.4**
FLUX.1-Canny	20.2	13.1	77.5	0.0	0.0	0.0	0.0	0.0
HiDream-Edit	30.3	12.6	74.9	0.0	0.0	0.0	0.0	0.0
EMU2	22.6	38.2	78.3	1.2	1.1	0.0	0.0	0.5
OmniGen	22.0	32.6	55.3	1.2	1.1	0.0	1.2	0.8
Step1X-Edit	25.1	41.5	73.5	0.0	2.2	2.0	3.5	1.9
Ovis-U1	33.9	52.7	72.9	1.1	3.3	4.0	2.4	2.8
FLUX.1-Kontext-Dev	26.0	71.6	85.2	2.3	5.5	13.0	1.2	5.8
BAGEL	36.5	53.5	73.0	2.4	5.6	14.0	1.2	6.1
Qwen-Image-Edit-2509	37.2	66.4	86.9	4.7	11.1	17.0	2.4	9.2
Gem-2.0-pre	49.9	68.4	84.9	11.8	14.4	11.0	2.4	10.0
BAGEL (CoT)	45.9	73.8	80.1	5.9	17.8	21.0	1.2	11.9
Gem-2.0-exp	48.9	68.2	82.7	9.4	16.7	23.0	4.7	13.9
Qwen-Image-Edit-2511	49.9	71.0	91.5	21.2	18.9	31.0	4.7	**19.4**
Mind-Brush	61.5	79.4	86.5	-	-	-	-	24.7
RS-Gen	69.9	91.1	93.6	36.4	50.0	40.0	29.4	**39.1**

Table 3: Ablation study on WISE_Verified. Acronyms: IR (Image Router Agent), IA (Intent Analysis Agent), RS (Reasoning & Search Agent), IG (Image Generation Agent), QI (Qwen-Image), LCI (LongCat-Image), ZI (Z-Image), Seed (Doubao-Seed-2.0-Pro-260215), MiMo (MiMo-v2-Omni).

Model	Cult.	Time	Spac.	Bio.	Phys.	Chem.	Avg.
QI	0.63	0.53	0.56	0.34	0.48	0.25	0.51
IR + QI	0.68	0.60	0.70	0.33	0.44	0.26	0.55
IR+IA+QI	0.68	0.64	0.73	0.43	0.54	0.25	0.59
IR+IA+RS+QI	0.88	0.87	0.88	0.54	0.70	0.70	0.79
IR+IA+RS+IG+QI	0.90	0.88	0.91	0.60	0.69	0.78	0.82
LCI	0.69	0.46	0.63	0.37	0.49	0.35	0.55
LCI+RS-Gen	0.86	0.86	0.84	0.62	0.66	0.72	0.79
ZI	0.55	0.46	0.46	0.36	0.43	0.13	0.44
ZI+RS-Gen	0.88	0.90	0.90	0.71	0.75	0.83	0.84
QI+RS-Gen (MiMo)	0.87	0.83	0.84	0.69	0.73	0.78	0.81
QI+RS-Gen (Seed)	0.90	0.88	0.91	0.60	0.69	0.78	0.82

Mind-Brush He et al. (2026) (24.7) by 19.7 and 14.4 points, respectively. (2) Substantial reasoning gains: RS-Gen improves Instruction Reasoning by 20.0 points over the base model, with gains of up to 31.1 on temporal, causal, and spatial reasoning. (3) Benefits of closed-loop reasoning: Its adaptive fallback and iterative self-correction enable robust error recovery, yielding strong performance across complex editing tasks and narrowing the gap with leading proprietary models.

4.4 Ablation Study

To evaluate RS-Gen's effectiveness and generalization, we conduct three ablations: (1) Component contribution: progressively enabling core modules to measure their individual gains; (2) Image model generalization: replacing generation and editing backbones to assess cross-model adaptabil-

Table 4: Ablation study on RISEBench. Acronyms: IR (Image Router Agent), IA (Intent Analysis Agent), RS (Reasoning & Search Agent), IG (Image Generation Agent), QIE (Qwen-Image-Edit), LCIE (LongCat-Image-Edit), SDr-4.0 (Seedream-4.0), Seed (Doubao-Seed-2.0-Pro-260215), MiMo (MiMo-v2-Omni).

Model	**Accuracy**				
	Temp.	**Caus.**	**Spat.**	**Logi.**	**Avg.**
QIE	21.2	18.9	31.0	4.7	19.4
IR+QIE	17.6	27.7	29.0	12.9	22.2
IR+IA+QIE	21.1	38.8	35.0	10.5	26.9
IR+IA+RS+QIE	37.6	48.8	36.0	17.6	35.2
IR+IA+RS+IG+QIE	36.4	50.0	40.0	29.4	39.1
LCIE	4.7	12.2	12.0	4.7	8.6
LCIE+RS-Gen	37.6	40.0	25.0	14.1	29.1
SDr-4.0	17.6	13.3	11.0	7.1	12.2
SDr-4.0+RS-Gen	41.1	52.2	34.0	27.0	38.6
QIE+RS-Gen (MiMo)	34.1	32.2	36.0	27.0	32.5
QIE+RS-Gen (Seed)	36.4	50.0	40.0	29.4	39.1

ity; and (3) Language model generalization: replacing the driving LLM to evaluate universality and robustness across LLMs.

Tables 3 and 4 report ablations on WISE_Verified and RISEBench. Progressive activation of RS-Gen's core modules consistently improves performance. Adding the RS (Reasoning & Search) Agent yields gains of 0.20 and 8.3 on generation and editing, respectively, highlighting the value of retrieval and reasoning. Adding the IG (Image Generation) Agent further achieves the best overall scores (0.82 and 39.1), with its self-correcting loop improving RISEBench logical reasoning from 17.6 to 29.4. Across LongCat-Image, Z-Image, LongCat-Image-Edit, and Seedream-4.0, RS-Gen improves performance by 43.6%–238.4%, demonstrating strong cross-model generalization. Replacing the driving LLM with MiMo yields 0.81 on WISE_Verified and 32.5 on RISEBench, a 67.5% gain over Qwen-Image-Edit-2511, confirming robustness across LLMs.

5 Limitations & Future Work

Although experiments validate the generalizability and robustness of RS-Gen, its substantial gains rely largely on meticulous prompt engineering (e.g., deferred parsing, defensive inquiry), leaving the framework partially dependent on the instruction-following capabilities of underlying models. To address these limitations, our future work focuses on: (1) Adaptive prompt optimization. We plan to incorporate an automatic prompt optimization mechanism that dynamically adjusts strategies based on underlying model characteristics to enhance universal adaptability. (2) Experience accumulation and skill expansion. The system will automatically distill behavioral trajectories into reusable "skills" stored in an external database, enabling autonomous self-evolution and lifelong learning capabilities.

6 Conclusion

We propose RS-Gen, a training-free image agent framework augmented by reasoning and search. This system reframes image generation and editing into a multi-agent collaborative self-correcting closed loop, and effectively overcomes the bottlenecks of foundational models in handling ambiguous intent, complex logic, and knowledge truncation through explicit reasoning, factual retrieval, and dynamic error correction. We envision that agent-based collaborative systems will gradually surpass static end-to-end single models, becoming the core paradigm for future visual generation and editing.

AI USE STATEMENT

We used generative AI tools solely to assist with proofreading and improving the clarity, grammar, and readability of the manuscript. All AI-assisted text was reviewed and revised by the authors. The research ideas, methodology, experiments, results, analysis, and conclusions were developed and conducted by the authors. The authors take full responsibility for the final content of this paper.

REFERENCES

Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024. GitHub repository.

Black Forest Labs. FLUX.2 Max. https://bfl.ai/models/flux-2-max, 2026. Product page, accessed 2026-05-06.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, 2023. doi: 10.1109/CVPR52729.2023.01764.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024. URL https://arxiv.org/abs/2405.09818. Code available at https://github.com/facebookresearch/chameleon.

Shuang Chen, Quanxin Shou, Hangting Chen, Yucheng Zhou, Kaituo Feng, Wenbo Hu, Yi-Fan Zhang, Yunlong Lin, Wenxuan Huang, Mingyang Song, et al. Unify-Agent: A unified multimodal agent for world-grounded image synthesis, 2026a. URL https://arxiv.org/abs/2603.29620.

Sixiang Chen, Zhaohu Xing, Tian Ye, Xinyu Geng, Yunlong Lin, Jianyu Lai, Xuanhua He, Fuxiang Zhai, Jialin Gao, and Lei Zhu. GenEvolve: Self-evolving image generation agents via tool-orchestrated visual experience distillation, 2026b. URL https://arxiv.org/abs/2605.21605.

Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, and Seedream Team. Seedream 4.0: Toward next-generation multimodal image generation, 2025. URL https://arxiv.org/abs/2509.20427.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Bagel: Emerging properties in unified multimodal pretraining, 2025. URL https://arxiv.org/abs/2505.14683.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow Transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 12001–12020, 2024.

Kaituo Feng, Manyuan Zhang, Shuang Chen, Yunlong Lin, Kaixuan Red Fan, Yilei Jiang, Hongyu Li, Dian Zheng, Chenyang Wang, and Xiangyu Yue. Gen-Searcher: Reinforcing agentic search for image generation, 2026. URL https://arxiv.org/abs/2603.28767.

Google DeepMind. Gemini Image Pro: High-quality image generation. https://deepmind.google/models/gemini-image/pro/, 2025. Product page, accessed 2026-01-26.

Jun He, Junyan Ye, Zilong Huang, Dongzhi Jiang, Chenjue Zhang, Leqi Zhu, Renrui Zhang, Xiang Zhang, and Weijia Li. Mind-Brush: Integrating agentic cognitive search and reasoning into image generation, 2026. URL https://arxiv.org/abs/2602.01756.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Cihang Xie, and Yuyin Zhou. HQ-Edit: A high-quality dataset for instruction-based image editing. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=mZptYYttFj.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1X-Edit: A practical framework for general image editing, 2025. URL https://arxiv.org/abs/2504.17761.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Gold Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. JanusFlow: Harmonizing Autoregression and Rectified Flow for unified multimodal understanding and generation, 2024. URL https://arxiv.org/abs/2411.07975.

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

OpenClaw Team. OpenClaw: Personal ai assistant. https://github.com/openclaw/openclaw, 2026. URL https://openclaw.ai/. GitHub repository, accessed 2026-05-06.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text Transformer. *Journal of Machine Learning Research (JMLR)*, 21(140):1–67, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. URL https://arxiv.org/abs/2112.10752.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu Edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8871–8879, 2024. doi: 10.1109/CVPR52733.2024.00847.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977, 2025b. doi: 10.1109/CVPR52734.2025.01210.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Exploration to advanced multimodal generation, 2025c. URL https://arxiv.org/abs/2506.18871.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single Transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=o6Ynz6OIQ6.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. MagicBrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 40100–40112, 2023.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, Hua Yang, Xue Yang, and Haodong Duan. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/fe79898dcf078ec54b6feeea10ebb751-Abstract-Datasets_and_Benchmarks_Track.html.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model, 2024. URL https://arxiv.org/abs/2408.11039.

A Appendix

A.1 Implementation Details

RS-Gen is a training-free agentic framework for image generation and editing with minimal local computational overhead. The local computing platform is primarily responsible for multi-agent collaborative workflow orchestration, reasoning orchestration, and automated evaluation script execution. The detailed experimental configuration is as follows.

- Hardware: All workflow scheduling, task planning, and evaluation pipelines run on a computing platform equipped with a 13th Gen Intel Core i7-13700 CPU (16 cores and 24 threads) and 32.0 GiB of system memory.
- Software: The underlying operating system is Ubuntu 22.04.5 LTS, and the runtime environment is Python 3.12.
- Model and service invocation: Due to the plug-and-play, lightweight architecture of our framework, the central driving LLMs for all agents, online search tools, and evaluation models are remotely accessed via standardized API calls. This design eliminates the need for local GPU resources for large-model inference, demonstrating the framework's broad applicability and cost-effective deployment in resource-constrained environments.

Unless otherwise specified, the proposed RS-Gen framework maintains identical default configurations across all benchmark evaluations. Specifically, all specialized sub-agents in RS-Gen are powered by Doubao-Seed-2.0-Pro-260215 as their backbone LLM (accessed via the Volcengine Ark platform API).

Tool Library Configurations: To support distinct core functionalities, the external tool libraries for specialized agents are configured as follows:

- Reasoning & Search Agent: Integrates four essential API tools: (1) Web search. Combines Baidu Search via the Baidu Qianfan platform and Google Search via the Serper platform for comprehensive online retrieval; (2) Geolocation. Utilizes LocationIQ for reverse geocoding services; (3) Encyclopedia knowledge. Accesses standardized Wikipedia queries via the MediaWiki Action API; (4) Logical reasoning. Incorporates GPT-5.4 as an external reasoning tool for solving complex reasoning tasks.
- Image Generation Agent: Configured with three core visual tools: (1) Image generation. Uses Qwen-Image for fundamental text-to-image synthesis; (2) Image editing. Uses Qwen-Image-Edit-2511 for image manipulation and inpainting; (3) Image verification. Leverages

the multimodal visual understanding capabilities of GPT-5.4 for automated quality inspection and closed-loop verification.

Hyperparameters and Iteration Limits: To optimize system runtime efficiency and effectively prevent infinite reasoning loops, the maximum number of ReAct reasoning iterations for both the Reasoning & Search Agent and the Image Generation Agent is strictly capped at 10. To ensure fair comparisons with baseline models, we perform no manual hyperparameter tuning on Doubao-Seed-2.0-Pro-260215 or any other LLMs evaluated in our ablation studies; all parameters (e.g., Temperature, Top-P) adhere to the default settings of their respective cloud service APIs.

Evaluation Protocol: All experimental results reported in this paper are derived from direct single-pass inference without repeated sampling or manual selection of favorable outputs. We strictly adhere to the official evaluation scripts and standardized protocols published by each respective benchmark during the evaluation phase. The corresponding data preprocessing workflows, VLM evaluators, and their hyperparameter settings completely maintain the official script settings (with the temperature of the judge models explicitly set to 0).

A.2 Additional Experimental Results

Since many existing methods have been evaluated on the WISE_legacy benchmark (the legacy version of WISE Niu et al. (2025)), we conduct additional testing on this benchmark for a more comprehensive comparison. Consistent with previous experiments, RS-Gen adopts Qwen-Image Wu et al. (2025a) as its foundational generator and strictly adheres to the official evaluation protocol using gpt-4o-2024-05-13 as the evaluator. As presented in Table 5, compared to the baseline Qwen-Image (0.62), RS-Gen achieves a substantial improvement and secures the highest overall score (Overall: 0.83) across all evaluated methods, obtaining a higher score than both the proprietary commercial model GPT-4o (0.80) and cutting-edge visual generation agent frameworks (e.g., GenEvolve Chen et al. (2026b) and Mind-Brush He et al. (2026)) under the same evaluation protocol. Notably, on the knowledge-intensive Cultural and Biology subsets, RS-Gen achieves top performances with scores of 0.86 and 0.85, respectively. These significant gains are primarily attributed to our proposed "question-posing and problem-solving" workflow and dynamic error-correction mechanism. By autonomously identifying knowledge gaps and retrieving external factual information, RS-Gen effectively overcomes the knowledge cutoff of foundational models, significantly suppressing factual hallucinations when handling out-of-distribution (OOD) concepts or long-tail entities.

Due to space limitations in the main text, the experiments section only reported partial core ablation results of RS-Gen on the RISEBench benchmark. Here, Table 6 presents the full set of ablation results and detailed evaluation metrics on this benchmark.

A.3 Prompt Templates

To ensure complete experimental reproducibility and intuitively illustrate the instruction guidance and logical orchestration of our framework across multi-stage collaborative tasks, this section details the specific prompt templates utilized by the core modules of RS-Gen (including the Image Router Agent, Intent Analysis Agent , Reasoning & Search Agent, and Image Generation Agent) during actual execution. Note that, for readability, only minor formatting adjustments have been made solely for presentation purposes, while the original prompts remain unchanged.

A.3.1 Image Router Agent

```
# Role & Objective
You are an intelligent routing and instruction optimization
    assistant for a multi-turn image generation and editing system
    . Your task is to analyze the user's multi-turn dialogue
    history (containing text and image IDs), determine the **core
    intent** of the latest interaction, identify the associated **
    image IDs**, and perform **disambiguation and optimization**
    on the raw instruction.

# Tasks & Rules
```

Table 5: Performance comparison on WISE_legacy. Models are grouped into Dedicated T2I, Unified MLLM, and Visual Generation Agent. Best results within each group are bold.

Model	Cultural	Time	Space	Biology	Physics	Chemistry	Overall
Dedicated T2I							
SD-v1-5	0.34	0.35	0.32	0.28	0.29	0.21	0.32
SD-2-1	0.30	0.38	0.35	0.33	0.34	0.21	0.32
FLUX.1-schnell	0.39	0.44	0.50	0.31	0.44	0.26	0.40
SD-3-medium	0.42	0.44	0.48	0.39	0.47	0.29	0.42
SD-XL-base-0.9	0.43	0.48	0.47	0.44	0.45	0.27	0.43
SD-3.5-Medium	0.43	0.50	0.52	0.41	0.53	0.33	0.45
SD-3.5-Large	0.44	0.50	0.58	0.44	0.52	0.31	0.46
PixArt-Alpha	0.45	0.50	0.48	0.49	0.56	0.34	0.47
playground-v2.5	0.49	0.58	0.55	0.43	0.48	0.33	0.49
FLUX.1-dev	0.48	0.58	0.62	0.42	0.51	0.35	0.50
Unified MLLM							
Orthus-7B-base	0.07	0.10	0.12	0.15	0.15	0.10	0.10
JanusFlow-1.3B	0.13	0.26	0.28	0.20	0.19	0.11	0.18
Janus-1.3B	0.16	0.26	0.35	0.28	0.30	0.14	0.23
Janus-Pro-1B	0.20	0.28	0.45	0.24	0.32	0.16	0.26
Orthus-7B-instruct	0.23	0.31	0.38	0.28	0.31	0.20	0.27
show-o	0.28	0.36	0.40	0.23	0.33	0.22	0.30
vila-u-7b-256	0.26	0.33	0.37	0.35	0.39	0.23	0.31
show-o-512	0.28	0.40	0.48	0.30	0.46	0.30	0.35
Janus-Pro-7B	0.30	0.37	0.49	0.36	0.42	0.26	0.35
Emu3	0.34	0.45	0.48	0.41	0.45	0.27	0.39
Liquid	0.38	0.42	0.53	0.36	0.47	0.30	0.41
Harmon-1.5B	0.38	0.48	0.52	0.37	0.44	0.29	0.41
OpenUni-B-512	0.37	0.45	0.58	0.39	0.50	0.30	0.43
Manzano-3B	0.42	0.51	0.59	0.45	0.51	0.32	0.46
OpenUni-L-512	0.51	0.49	0.64	0.48	0.63	0.35	0.52
OpenUni-L-1024	0.49	0.53	0.69	0.49	0.56	0.39	0.52
BAGEL	0.44	0.55	0.68	0.44	0.60	0.39	0.52
Manzano-30B	0.58	0.50	0.65	0.50	0.55	0.32	0.54
MetaQuery-XL	0.56	0.55	0.62	0.49	0.63	0.41	0.55
UniWorld-V1	0.53	0.55	0.73	0.45	0.59	0.41	0.55
Hunyuan-Image 3.0	0.58	0.57	0.70	0.56	0.63	0.31	0.57
UniWorld-V2	0.60	0.61	0.70	0.53	0.64	0.32	0.58
NextFlow-RL	0.63	0.63	0.77	0.58	0.67	0.39	0.62
Qwen-Image	0.62	0.63	0.77	0.57	0.75	0.40	0.62
LongCat-Image	0.66	0.61	0.72	0.66	0.72	0.49	0.65
DeepGen1.0	0.72	**0.81**	0.70	0.67	0.82	0.66	0.73
GPT-4o	0.81	0.71	**0.89**	0.83	0.79	0.74	0.80
Visual Generation Agent							
GenAgent	0.78	0.67	0.78	0.72	0.77	0.55	0.72
Gen-Searcher-8B	0.80	0.71	0.82	0.76	0.74	0.75	0.77
Mind-Brush	0.83	0.69	0.84	0.71	**0.85**	0.68	0.78
GenEvolve	0.84	0.74	0.87	0.83	0.81	**0.83**	0.82
RS-Gen	**0.86**	0.74	0.88	**0.85**	0.83	0.74	**0.83**

```
## Task Type Classification
- `generate` (Image Generation): Create a completely new image
    from scratch.
- `edit` (Image Editing): Modify an existing image.
- `chat` (General Dialogue): Pure text-based chit-chat or Q&A that
    does not involve image generation or editing whatsoever.
```

Table 6: Ablation study on RISEBench. IR, IA, RS, and IG denote the Image Router, Intent Analysis, Reasoning & Search, and Image Generation Agents, respectively. QIE, LCIE, and SDr-4.0 denote Qwen-Image-Edit, LongCat-Image-Edit, and Seedream-4.0. Seed and MiMo denote Doubao-Seed-2.0-Pro-260215 and MiMo-v2-Omni.

Model	Sub-Dim.			Accuracy				
	Instr.	App.	Vis.	Temp.	Causal	Spatial	Logical	Overall
QIE	49.9	71.0	91.5	21.2	18.9	31.0	4.7	19.4
IR+QIE	51.5	79.8	90.5	17.6	27.7	29.0	12.9	22.2 (+2.8)
IR+IA+QIE	56.7	81.0	92.0	21.1	38.8	35.0	10.5	26.9 (+4.7)
IR+IA+RS+QIE	66.4	87.1	91.1	37.6	48.8	36.0	17.6	35.2 (+8.3)
IR+IA+RS+IG+QIE	69.9	91.1	93.6	36.4	50.0	40.0	29.4	39.1 (+3.9)
LCIE	37.2	74.6	85.2	4.7	12.2	12.0	4.7	8.6
LCIE + RS-Gen	63.0	86.8	87.7	37.6	40.0	25.0	14.1	29.1 (+20.5)
SDr-4.0	58.9	67.4	91.2	17.6	13.3	11.0	7.1	12.2
SDr-4.0 + RS-Gen	69.0	89.4	93.8	41.1	52.2	34.0	27.0	38.6 (+26.4)
QIE	49.9	71.0	91.5	21.2	18.9	31.0	4.7	19.4
QIE + RS-Gen (MiMo)	68.4	84.7	89.8	34.1	32.2	36.0	27.0	32.5 (+13.1)
QIE + RS-Gen (Seed)	69.9	91.1	93.6	36.4	50.0	40.0	29.4	39.1 (+19.7)

```
## Image ID Extraction
- For `chat` tasks: Strictly fix as `[]`.
- For `generate` tasks: If the user neither uploads nor specifies
    an image reference, strictly fix as `[]`; if the user uploads
    or explicitly designates reference images, extract the
    corresponding image IDs.
- For `edit` tasks: Extract the IDs of the source image to be
    edited and any reference images based on user intent. Priority
     rule: Latest Upload > Explicit Designation > Implicit
    Inheritance (defaulting to inheriting the most recently
    generated image in the context).

## Instruction Optimization
- Prohibition of Image IDs: It is strictly forbidden to include
    any system image IDs in the `prompt`; use natural language
    references instead.
- Coreference Resolution: Resolve pronouns such as "it," "this,"
    or "the previous one" based on dialogue context to explicitly
    specify the subjects and objects.
- Principle of Delayed Resolution: This module is exclusively
    responsible for instruction refinement and coreference
    resolution. It is strictly prohibited from using prior
    knowledge to "pre-answer" or make speculative completions. For
     the following scenarios, the original constraints must be
    preserved and passed transparently to downstream modules:
  1. External Entity / Knowledge Blind Spots: Unknown or ambiguous
       entities, events, concepts, or spatio-temporal coordinates
     (fabricating details is strictly prohibited).
  2. State Dynamics Simulation: Physical state transitions under
     temporal changes, spatial transformations, or environmental
     interactions (describing the final evolved state prematurely
      is strictly prohibited).
  3. Logical Reasoning: Tasks involving mathematical calculations,
       spatial reasoning, or logical deduction (providing the
     final answer directly is strictly prohibited).
```

```
- Negative-to-Affirmative Inversion: Underlying image models
    exhibit poor comprehension of negative constraints. If the
    user instruction contains negative descriptions or constraints
     (e.g., "do not keep the mud on the car paint"), they must be
    translated into affirmative feature descriptions (e.g., "clean
     and solid-colored paint").

# Output Format
{
  "task_type": "xxx",                 // "generate" | "edit" | "chat
     "
  "target_image_id": ["xxx"],         // List of relevant image IDs
  "reasoning": "xxx",                 // Reasoning and analytical
     process
  "prompt": "xxx"                     // Optimized natural language
     instruction
}
```

A.3.2 Intent Analysis Agent

```
# Role & Objective
You are an intelligent routing and instruction optimization
    assistant for a multi-turn image generation and editing system
    . Your task is to analyze the user's multi-turn dialogue
    history (containing text and image IDs), determine the **core
    intent** of the latest interaction, identify the associated **
    image IDs**, and perform **disambiguation and optimization**
    on the raw instruction.

# Tasks & Rules
## Task Type Classification
- `generate` (Image Generation): Create a completely new image
    from scratch.
- `edit` (Image Editing): Modify an existing image.
- `chat` (General Dialogue): Pure text-based chit-chat or Q&A that
     does not involve image generation or editing whatsoever.

## Image ID Extraction
- For `chat` tasks: Strictly fix as `[]`.
- For `generate` tasks: If the user neither uploads nor specifies
    an image reference, strictly fix as `[]`; if the user uploads
    or explicitly designates reference images, extract the
    corresponding image IDs.
- For `edit` tasks: Extract the IDs of the source image to be
    edited and any reference images based on user intent. Priority
     rule: Latest Upload > Explicit Designation > Implicit
    Inheritance (defaulting to inheriting the most recently
    generated image in the context).

## Instruction Optimization
- Prohibition of Image IDs: It is strictly forbidden to include
    any system image IDs in the `prompt`; use natural language
    references instead.
- Coreference Resolution: Resolve pronouns such as "it," "this,"
    or "the previous one" based on dialogue context to explicitly
    specify the subjects and objects.
- Principle of Delayed Resolution: This module is exclusively
    responsible for instruction refinement and coreference
    resolution. It is strictly prohibited from using prior
```

```
    knowledge to "pre-answer" or make speculative completions. For
     the following scenarios, the original constraints must be
    preserved and passed transparently to downstream modules:
   1. External Entity / Knowledge Blind Spots: Unknown or ambiguous
       entities, events, concepts, or spatio-temporal coordinates
      (fabricating details is strictly prohibited).
   2. State Dynamics Simulation: Physical state transitions under
      temporal changes, spatial transformations, or environmental
      interactions (describing the final evolved state prematurely
       is strictly prohibited).
   3. Logical Reasoning: Tasks involving mathematical calculations,
       spatial reasoning, or logical deduction (providing the
      final answer directly is strictly prohibited).
- Negative-to-Affirmative Inversion: Underlying image models
    exhibit poor comprehension of negative constraints. If the
    user instruction contains negative descriptions or constraints
     (e.g., "do not keep the mud on the car paint"), they must be
    translated into affirmative feature descriptions (e.g., "clean
     and solid-colored paint").

# Output Format
{
  "task_type": "xxx",                 // "generate" | "edit" | "chat
     "
  "target_image_id": ["xxx"],         // List of relevant image IDs
  "reasoning": "xxx",                 // Reasoning and analytical
     process
  "prompt": "xxx"                     // Optimized natural language
     instruction
}
```

A.3.3 Reasoning & Search Agent

```
# Role & Objective
You are the Search and Reasoning Hub of an image generation and
    editing system. Your primary responsibility is to coordinate
    multimodal reasoning and external tools to resolve knowledge
    blind spots and logical dilemmas inherent in user instructions
    . You must integrate all retrieved information to output
    precise visual prompts and high-quality reference images that
    can be directly executed by underlying image generation and
    editing models.

# Tool Suite
The available tools include:
- 'Spatial & Geolocation Tool': A reverse geocoding tool. Queries
    and retrieves location names and detailed addresses based on
    latitude and longitude coordinates.
- 'Logical Reasoning Tool': A deep reasoning tool powered by a
    state-of-the-art multimodal large language model, capable of
    handling complex problem-solving, mathematical calculations,
    logical deductions, and graphical reasoning.
- 'Visual Asset Search Tool': Used to retrieve image-modality
    information from the internet.
- 'Knowledge & Information Search Tool': Used to retrieve text-
    modality information from the internet.
- 'Wikipedia Tool Suite': An encyclopedic knowledge retrieval
    suite used to acquire authoritative domain knowledge.
```

```
# Input Data Description
1. Instruction: The user-provided raw text instruction for image
    generation or editing.
2. Image: The user-provided input image(s) (optional).
3. Unresolved Questions: A list of questions that require
    resolution via tool invocation or internal reasoning.

# Actions & Rules
You must adhere to the following rules to execute the 'Observation
    -Thought-Action' cycle (ReAct Loop).

## Optimal Routing Decisions
Dynamically evaluate and determine the optimal solution path for
    each unresolved question:
1. Internal Reasoning (Native Resolution): For common sense and
    simple queries, prioritize relying on native multimodal
    reasoning. If uncertainty persists or the problem remains
    unresolved after reasoning attempts, automatically trigger a
    fallback to invoke external tools.
2. External Invocation (Tool Augmentation): For in-depth image
    understanding, complex puzzle-solving, specific concrete
    appearances, and real-time/latest time-sensitive information,
    you must invoke external domain-specific tools to acquire
    accurate factual anchors and visual assets.

## Capability Boundaries
You possess baseline native multimodal perception and reasoning
    capabilities. To ensure output rigor, adhere to the following
    invocation rules:
1. Autonomous Resolution & Fallback Mechanism: Prioritize native
    capabilities. When native capabilities fail to resolve a query
    , trigger tool invocation for secondary verification.
2. Tool Prioritization: When encountering complex logical
    reasoning, mathematical calculations, or graphical reasoning,
    prioritize invoking the 'Logical Reasoning Tool'. Note that
    tool outputs may occasionally be inaccurate; cross-verify tool
     results with internal reasoning to determine the final answer
    .

## Wikipedia Tool Suite Invocation Rules
1. Entry Retrieval: First, invoke the Wikipedia search tool to
    locate relevant encyclopedic entries.
2. Preliminary Screening: Invoke the summary tool to acquire the
    entry introduction and evaluate whether it satisfies
    information requirements.
3. Precise Extraction: If in-depth content is required, first
    invoke the Table of Contents (TOC) tool to retrieve section
    indices, and subsequently invoke the section tool to extract
    the complete text of the targeted section.

## Search Query Construction Rules
When invoking external retrieval tools, you must strictly execute
    the following query construction rules to ensure Information
    Retrieval (IR) precision:
1. General Query Construction
    - Feature Decoupling & Denoising: Extract core entities (
        people, products, events, artworks, etc.) from the
        question, retaining only essential constraints such as
        space-time, version/model, and state features.
```

```
        Unconditionally strip interrogative words and action
        phrases (e.g., "what does it look like", "what is").
    - Half-Width Space Segmentation: Use half-width spaces to
        separate core entities from constraint terms (search-
        engine-friendly syntax).
    - Single-Target Retrieval: Strictly prohibit merging queries
        for multi-entity retrieval; decompose them into
        independent search tasks and invoke tools separately.
2. Wikipedia-Specific Retrieval Rules
    - De-constraintization: When invoking entry searches, retain
        only pure noun entities (e.g., names of persons, locations
        , proper nouns, titles of works), unconditionally
        discarding all constraints regarding states, space-time,
        or perspectives.
    - Exact Entity Passthrough: Once an entry is identified, the `
        title` parameter passed to subsequent tool invocations
        must be an exact, 100% verbatim copy of the official entry
         title. Truncation, translation, or case alteration is
        strictly prohibited.
    - Index Matching: When retrieving specific sections, you must
        use the numerical section index; passing textual section
        headers is strictly prohibited.
    - Disambiguation: If a search returns similar or homonymic
        entries, perform contextual disambiguation based on the
        user's input to select the most precise entry (e.g.,
        selecting "Apple Inc." over "Apple (fruit)") before
        proceeding.

## Visual Retrieval Strategies
When invoking image search tools for visual references,
    dynamically route to the optimal retrieval path based on the
    certainty and completeness of the target entity's information:
1. Direct Retrieval: When the input already contains explicit,
    concrete, and unambiguous objective entities, prioritize
    directly invoking the image search tool to maximize system
    efficiency.
2. Cascaded Retrieval (Text-to-Image Anchoring): When the target
    involves vague references, generalized constraints, or
    semantic ambiguity (polysemy), you must execute a "text-to-
    image" cascaded retrieval strategy. First, invoke the `
    Knowledge & Information Search Tool` for entity disambiguation
     to lock onto the exact entity noun that best fits the context
    ; subsequently, use this precise keyword to invoke the image
    search tool for visual assets.

## Visual Verification & Re-search Mechanism
Upon acquiring retrieved images, you must evaluate whether they
    serve as qualified visual references and resolve the pending
    questions. If an image fails evaluation, trigger the "Multi-
    Round Retrieval & Adaptive Fallback Mechanism" to re-search
    until standards are met. Typical failure scenarios include:
1. Semantic Misalignment (Text-Image Mismatch): The image content
    completely contradicts the search query, target entity, or
    core intent.
2. Structural Defects: The target subject is severely cropped or
    heavily occluded, or core visual features are unrecognizable
    within the frame.
```

```
3. Visual Degradation: The image exhibits extremely low resolution
    , dense watermark interference, or explicit AI-generated
    artifacts and traces.

### Multi-Round Retrieval & Adaptive Fallback Mechanism
When external retrieval fails or results fail visual verification,
     directly abandoning the task or fabricating information is
    strictly prohibited. Sequentially trigger the following
    adaptive fallback strategies:
1. Tool Switching: Retain the original keyword and switch to
    alternative tools within the same modality to re-try retrieval
    .
2. Query Relaxation & Denoising: Moderately expand the search
    scope by stripping secondary modifiers, or substituting
    synonyms/near-synonyms within the vertical domain. Excessive
    divergence that introduces retrieval noise is strictly
    prohibited.
3. Path Reversion: When direct image searches fail repeatedly,
    proactively revert to the "text-to-image" mechanism, utilizing
     text retrieval to re-deduce and lock onto new, high-precision
     image search keywords.
4. Visually Isomorphic Substitution: Retrieve generalized
    substitutes that share highly isomorphic visual features with
    the target object.
5. Textual Fallback (Text-to-Visual Compensation): When image
    search attempts reach the threshold (maximum of 5 rounds)
    without acquiring valid visual references, trigger cross-modal
     fallback. Transition to deep text retrieval to extract
    objective, high-density textual descriptions of fine visual
    details, directly fusing them into the prompt to compensate
    for the lack of visual reference assets.

## Information Integration & Output
When internal reasoning and external tool invocations successfully
     bridge information gaps and resolve all "unresolved questions
    ," immediately terminate the tool invocation loop and execute
    multimodal information integration:
1. Prompt Refinement: Synthesize all acquired information to
    denoise, reconstruct, and enrich the original instruction,
    constructing a comprehensive prompt for image generation or
    editing.
2. Asset Curation: Select a curated list of high-quality, highly
    relevant visual reference images to serve alongside the prompt
     as the final input for the underlying image generation/
    editing model.

### Visual Asset Extraction
When constructing the final reference image list, strictly adhere
    to the assembly principles of "user input prepending, top-K
    curation, and quality-driven exclusion":
1. Unconditional Prepending of User Input: If original user input
    images exist, unconditionally place their URLs/file paths at
    the very top of the reference image list as the highest-
    priority visual anchors.
2. Top-K Curated Retrieval: For each independent entity/scene,
    curate only 1~2 most relevant images for the list. Execute
    strict hierarchical verification:
   - Structural Completeness & Canonical Perspective: Prioritize
       images showing the complete, uncropped subject with clear
```

```
        visibility and canonical canonical perspective (unless the
        instruction specifies a local detail, unconditionally favor
         full-view images).
    - Artifact & Noise Exclusion: Select high-definition, sharp,
        and highly representative assets; unconditionally reject
        unusable images with severe cropping, heavy occlusion,
        dense watermarks, or explicit AI-generated artifacts (e.g.,
         anatomical/physical logic collapse, typical plastic-like
        AI textures).
    - Dynamic Degradation Tolerance: For historical archives,
        breaking news, or documentary photographs, the screening
        threshold for resolution and clarity may be appropriately
        relaxed.

### Prompt Refinement
1. Text-to-Image: Perform high-density visual semantic expansion.
    Transform the input into a concrete, global description rich
    in visual details, including the subject, action, appearance
    specifics, perspective, environmental background, artistic
    style, lighting, color, and quality-enhancing terms.
2. Image-to-Image (Reference-Guided Generation): Fuse reference
    image features with text instructions, reconstructing them
    into a novel visual description.
    - Feature Extraction & Reconstruction: Accurately extract core
        visual features to be referenced (subject appearance, color
         palette, composition, etc.), converting them into text
        integrated directly into the final scene description.
    - Objective Final-State Description: Strictly prohibit action
        verbs such as "reference," "preserve," or "modify";
        describe the final generated scene directly and objectively
        .
    - Concrete Visualization of Reasoning: For logical,
        mathematical, or graphical reasoning tasks, strictly
        prohibit outputting problem-solving steps or reasoning
        processes in the prompt; directly output the visual
        features of the terminal state so the underlying model can
        render the answer directly.
3. Image Editing: Convert vague intentions into precise "
    imperative editing instructions."
    - Structured Paradigm: Strictly follow the format: '[Editing
        Action] + [Target Object/Region] + [Resulting State]'.
        Utilize spatial orientations or feature modifiers (e.g., "
        the red mug in the top-left corner") to specify targets and
         eliminate ambiguity.
    - Describe Changes Only: Describe only what has changed.
        Strictly prohibit mentioning elements that do not need
        modification, and absolutely ban redundant phrases such as
        "keep," "maintain," or "everything except...".
4. Negative-to-Affirmative Inversion: Underlying image models
    exhibit poor comprehension of negative constraints. If the
    instruction contains negative descriptions or constraints (e.g
    ., "do not keep mud on the car paint"), they must be
    translated into affirmative feature descriptions (e.g., "clean
     and solid-colored paint").

# Output Format
{
    "prompt": "xxx",              // The refined instruction for
        image generation or editing
```

```
    "reference_images": ["xxx"] // File paths/URLs of user-input
        images (if any) and curated reference images (if any)
}
```

A.3.4 IMAGE GENERATION AGENT

```
# Role & Objective
You are the central controller model of an image agent system
    grounded in the ReAct architecture. Your primary
    responsibility is to parse multimodal inputs and dynamically
    orchestrate the underlying tool suite to execute image
    generation or editing tasks. Driven by rigorous quality
    auditing and autonomous self-correction mechanisms, you steer
    the system through closed-loop iterative optimization until
    the visual output perfectly aligns with user intent or
    triggers the predefined maximum iteration threshold.

# Tool Suite
The available tools include:
1. `Image Generation Tool`: Generates images based on textual
    descriptions.
2. `Image Editing Tool`: Edits input images based on textual
    instructions.
3. `Image Audit Tool`: Evaluates generated or edited results
    against input instructions and original input images,
    outputting a detailed audit report.

# Input Data Description
1. Instruction: The user-provided raw text instruction for image
    generation or editing.
2. Image: The user-provided input image(s) (optional).

# Actions & Rules
## Initial Execution
Upon receiving multimodal inputs, first deconstruct user intent
    and dynamically dispatch the appropriate underlying tool:
- Image Generation: For requests requiring from-scratch image
    synthesis, invoke the `Image Generation Tool`.
- Image Editing: For requests requiring modifications to a
    provided input image, invoke the `Image Editing Tool`.

## Quality Audit
Upon acquiring intermediate images generated by tools, you must
    immediately invoke the `Image Audit Tool` to conduct rigorous
    quality evaluation; skipping this step is strictly prohibited.
When invoking the `Image Audit Tool`, its `prompt` and `
    input_images` parameters must strictly pass the user's
    original multimodal input. To prevent semantic drift,
    utilizing intermediate optimized instructions or generated
    outputs as audit reference anchors is strictly prohibited.

## Adaptive Iteration Strategy
Conduct a deep analysis of the audit report feedback and execute
    dynamic routing accordingly:
1. Audit Passed (Convergence): The report explicitly indicates
    that the output aligns with the original instruction without
    significant defects. Immediately terminate the loop and output
     the result.
2. Audit Failed (Adaptive Self-Correction):
```

```
  - Path A: Local Refinement: The current image possesses a solid
      structural baseline with only minor, repairable flaws.
      Formulate specific imperative editing instructions, use the
      current intermediate image as the base input, and invoke the
       'Image Editing Tool' for targeted refinement.
  - Path B: Global Regeneration: If any of the following
      conditions are met, you must synthesize failure experience
      to reconstruct the prompt/instruction and invoke the
      generation/editing tool for global regeneration:
    - The audit report identifies fatal structural defects (e.g.,
        complete composition collapse, severe physical logic
        violations, total disregard of core instructions, or
        severe visual degradation);
    - Multiple rounds of targeted local refinement have been
        executed on the current image without passing the closed-
        loop audit.

## Output Decisions
The system continuously monitors the current iteration count
    against the maximum iteration threshold to execute terminal
    decisions:
- Normal Convergence: Within the maximum iteration threshold, once
     a compliant image passes the audit, immediately extract it
    into the final output list.
- Forced Termination & Top-1 Candidate Fallback: If the maximum
    iteration threshold alarm is triggered and the latest result
    still fails the audit, immediately force-terminate all tool
    invocations. Review all intermediate results and corresponding
     audit reports across historical iterations. Based on the
    principle of "minimum visual defect and minimum semantic
    deviation," evaluate and extract the globally optimal
    candidate into the final output list as a robust fallback.

# Output Format
{
    "images": ["xxx"] // The final output image list
}
```