SCOPE: STRUCTURED DECOMPOSITION AND CONDITIONAL SKILL ORCHESTRATION FOR COMPLEX IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The emergence of agentic image generation has expanded visual synthesis into a multi-stage process of knowledge grounding, reasoning, and iterative refinement. Yet, faithfully realizing complex user intent depends on preserving its underlying requirements across these stages. We refer to these requirements as **semantic commitments** and identify the *Conceptual Rift* as a break in their continuity: commitments may be resolved or checked locally without remaining explicitly tracked across grounding, generation, verification, and repair. We address this with **SCOPE**, a specification-guided skill orchestration framework that maintains semantic commitments in an evolving structured specification and conditionally invokes retrieval, reasoning, and repair skills to address unresolved or violated commitments. To evaluate whether generated images jointly satisfy these commitments, we introduce **Gen-Arena**, a human-annotated benchmark with entity- and constraint-level specifications, together with **Entity-Gated Intent Pass Rate (EGIP)**, a strict entity-first pass criterion. Experiments on Gen-Arena, WISE-V, and MindBench show that SCOPE consistently outperforms the evaluated agentic baselines in overall performance across two generation backends, while reducing workflow orchestration cost and latency.



Figure 1: Representative SCOPE results on complex image-generation prompts spanning knowledge-intensive events, reference-heavy characters, relation-sensitive scenes, and multi-entity compositions.

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.

1 INTRODUCTION

Recent advances in generative modeling are expanding the role of image synthesis in creative workflows, where users increasingly rely on these systems to translate complex ideas into visual content (Saharia et al., 2022; Betker et al., 2023; Esser et al., 2024). This growing role places greater demands on semantic fidelity: successful generation must preserve the identities, factual details, and relationships that give a requested scene its intended meaning. Despite improvements in visual quality, these requirements remain difficult to satisfy jointly, allowing an image to appear convincing while misrepresenting the user's intent (Hu et al., 2023; Huang et al., 2023; Ghosh et al., 2023; Li et al., 2024).

These requirements differ in how they become actionable. Some are explicit in the prompt, such as an object's color, whereas others require external grounding to identify a referenced entity or reasoning to infer an implicit condition (Li et al., 2025a; Son et al., 2025; He et al., 2026a; Feng et al., 2026; Chen et al., 2026). For example, a request depicting the winner of the 2024 Wimbledon men's final requires resolving the person's identity before specifying their appearance in the scene. We refer to these requirements as **semantic commitments**: conditions whose fulfillment determines whether the generated image realizes the user's intent.

Recent agentic approaches incorporate retrieval, planning, and iterative refinement to resolve missing information and improve generated images (Li et al., 2025a; Son et al., 2025; Feng et al., 2026; Chen et al., 2025a; Li et al., 2025b; Wu et al., 2025b; Ye et al., 2025). However, these capabilities do not by themselves ensure that the same commitments remain explicitly tracked across stages. A resolved identity may be lost during prompt revision, while verification feedback may become detached from the requirement it concerns, leaving subsequent repair without a precise target. We call this loss of cross-stage continuity the **Conceptual Rift**. This raises a central question: ***How can semantic commitments be consistently tracked and acted upon across grounding, generation, verification, and repair?***

To address this question, we propose **SCOPE**, a specification-guided skill orchestration framework for complex image generation. SCOPE maintains semantic commitments in an evolving structured specification that links each requirement to its unresolved information, supporting evidence, and verification status. This shared representation guides the conditional invocation of retrieval and reasoning skills to resolve missing information, and repair skills to correct outputs that violate the specified requirements. Skill outputs and verification feedback are written back to the corresponding commitments, allowing subsequent stages to act on updated information while preserving the original generation goals. Figure 1 presents representative SCOPE outputs.

To evaluate whether complex visual intents are faithfully realized, we introduce **Gen-Arena**, a human-annotated benchmark covering specific characters, people, and events that require joint satisfaction of identity, factual, and compositional requirements. Each prompt is paired with an entity- and constraint-level evaluation specification, with reference images provided when needed to verify identity and appearance. These annotations are withheld from the generation workflow and used exclusively for evaluation. We adopt **Entity-Gated Intent Pass Rate (EGIP)**, a strict entity-first criterion that counts an instance as successful only when all required entities and constraints are satisfied.

Our contributions are threefold. **(1) Framework:** We propose SCOPE, which addresses the Conceptual Rift through an evolving specification that preserves semantic commitments across grounding, generation, verification, and repair. **(2) Benchmark:** We introduce Gen-Arena, a human-annotated benchmark with entity- and constraint-level specifications and a strict entity-first evaluation criterion, EGIP. **(3) Evaluation:** Experiments on Gen-Arena, WISE-V, and MindBench demonstrate consistent performance and efficiency gains over the agentic baselines across two generation backends.

2 RELATED WORK

2.1 AGENTIC IMAGE GENERATION

Recent work uses multimodal agents to improve complex image generation beyond direct prompt conditioning (Chen et al., 2025a; Ye et al., 2025; Xiang et al., 2025; Wang et al., 2025; Garg et al.,

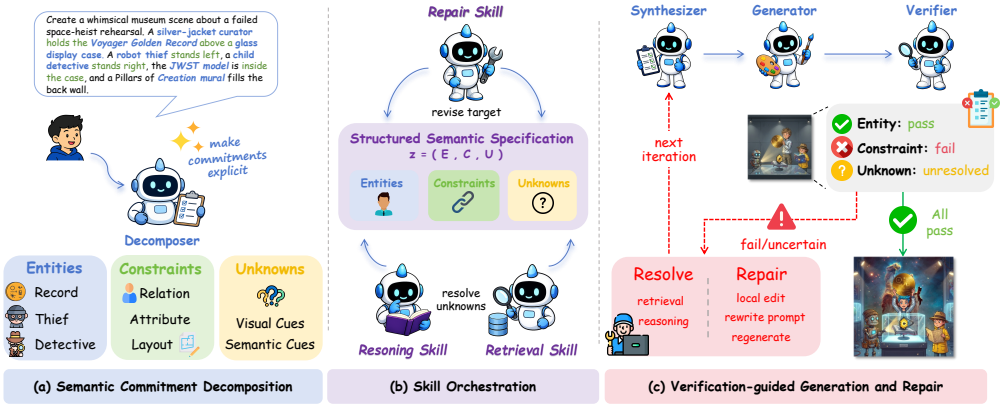

Figure 2: Overview of SCOPE. (a) The user prompt is decomposed into an evolving structured semantic specification containing entities, constraints, and unknowns. (b) Retrieval and reasoning resolve open unknowns and provide grounding evidence for generation. (c) Verification identifies unresolved or violated commitments; SCOPE routes remaining information gaps to retrieval or reasoning and visual failures to targeted repair. The repair update ("revise target") records corrective instructions for the failed item while preserving the original requirement. Algorithm 1 summarizes the procedure.

2026; He et al., 2026b; Han et al., 2026). Existing methods improve this process from different angles: some strengthen the interpretation of user requests (Chen et al., 2025a; Ye et al., 2025; Xiang et al., 2025), some ground generation or multimodal reasoning with retrieved visual or factual evidence (Li et al., 2025a; Son et al., 2025; He et al., 2026a; Feng et al., 2026; Chen et al., 2026; Huang et al., 2026; Zeng et al., 2026), and others refine outputs through reflection or feedback (Li et al., 2025b; Zhuo et al., 2025; Wu et al., 2025b; Jaiswal et al., 2026; Venkatesh et al., 2025; Huang et al., 2025). These approaches demonstrate the value of structured interpretation, grounding, and iterative feedback. SCOPE builds on these capabilities by organizing generation around a shared specification of the user's requirements.

2.2 Skills in Language and Multimodal Agents

Prior work broadly views skills as reusable procedural knowledge that extends language agents beyond one-off tool use (Liu et al., 2024; Jiang et al., 2026b; Li et al., 2026). Such skills may be written by humans, distilled from demonstrations or trajectories, or selected from large repositories according to the current task (Zheng et al., 2025; 2026). Recent multimodal systems further show that skill abstractions can support complex visual reasoning and generation workflows: XSkill (Jiang et al., 2026a) accumulates task-level skills from visual-tool trajectories, while GEMS (He et al., 2026b) introduces memory and domain skills for agent-native multimodal generation. SCOPE adopts this view of skills and studies their conditional orchestration for resolving missing information and repairing generation failures.

3 SCOPE Framework

3.1 Framework Overview

SCOPE maintains semantic commitments in an evolving structured specification shared by generation, verification, and skill execution. Given a user prompt and an optional reference image, the **Decomposer** initializes the specification once. SCOPE then iterates through **Synthesizer** $\rightarrow$ **Generator** $\rightarrow$ **Verifier**: the Synthesizer consolidates resolved information into a generation prompt, the Generator generates or edits an image, and the Verifier checks entities and constraints item by item. Retrieval, reasoning, and repair skills are conditionally invoked to address information gaps and visual failures. Figure 2 illustrates the overall architecture.

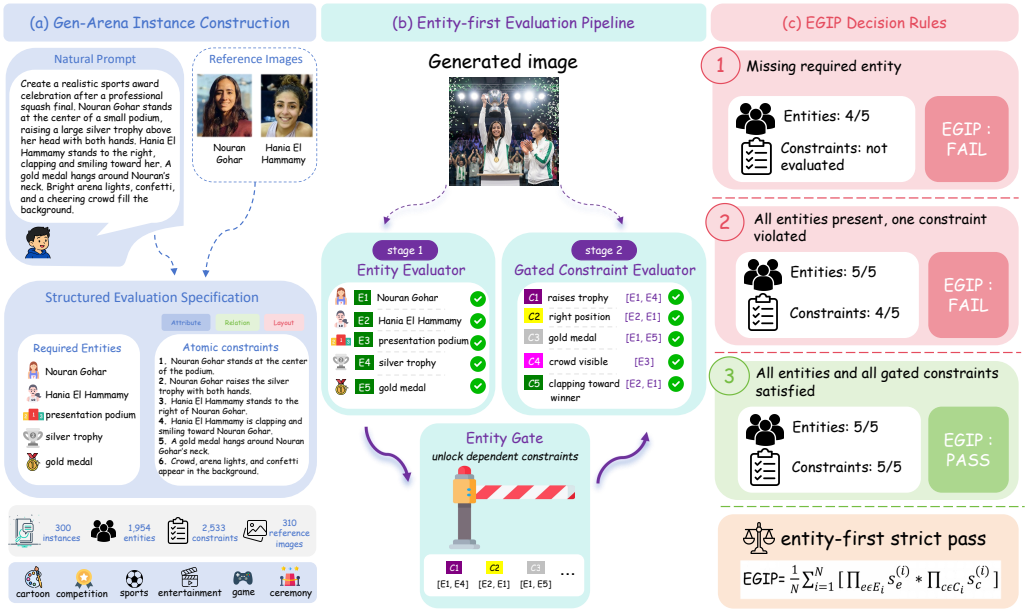

Figure 3: Overview of Gen-Arena construction and EGIP evaluation. (a) A natural-language request is annotated with required entities, atomic constraints, and reference images. (b) All entities are checked before any constraints; a failed entity causes all constraint checks to be skipped. (c) An instance passes only when all required entities and constraints are satisfied.

3.2 Structured Semantic Specification

Representation. To keep semantic commitments identifiable across the generation lifecycle, SCOPE represents each request as an evolving structured semantic specification $z = (E, C, U)$. Here, E denotes the target entities that should be instantiated in the image, C denotes verifiable commitments over these entities, and U denotes unresolved information that prevents reliable realization of a commitment. We group constraints in C into three types. Attribute constraints specify entity-level requirements such as identity, appearance, quantity, and visible text. Relation constraints specify interactions or semantic relationships between entities. Layout constraints specify the placement of entities within the scene and their composition with the surrounding environment.

Persistent state and updates. The specification is a persistent state shared across iterations. Entities, constraints, and unknowns remain addressable through stable item IDs, and each unknown is linked to a prompt-, entity-, or constraint-level owner. This link identifies which part of the request the missing information concerns. Retrieval evidence and reasoning results are written back to the corresponding owner, and the associated unknown is marked as resolved when the information gap is closed. The Synthesizer then incorporates the resolved information into the generation prompt.

Verification feedback. Verification feedback is attached to the same entity or constraint IDs, preserving the connection between an intended requirement and its realization in the current image. If feedback exposes a remaining information gap, SCOPE associates the unknown with the affected owner for further resolution. If the requirement is already resolved but remains visually unsatisfied, the corresponding item becomes a repair target. Thus, resolving an unknown and satisfying a visual requirement are distinct: the former supplies information for generation, while the latter is assessed on the generated image. This shared state keeps evidence, verification feedback, and repair targets linked to the same commitments throughout the workflow.

3.3 Conditional Skill Orchestration

SCOPE orchestrates retrieval, reasoning, and repair skills through the current specification rather than applying them as fixed pipeline steps. Retrieval is invoked when a commitment depends on missing

external evidence, such as factual information or reference identity cues. Reasoning is invoked when an implicit or underspecified commitment must be resolved before reliable synthesis. Repair is invoked when verification maps a visual failure back to an entity or constraint that no longer requires additional grounding or reasoning. Each skill updates the specification at the corresponding unknown or violated item: retrieval and reasoning resolve information gaps, while repair records corrective actions without changing the original entity or constraint requirements.

3.4 Verification and Repair

Item-level verification. Given the current output image y_t at iteration t and the specification $z = (E, C, U)$, SCOPE verifies the image against the explicit item set $I = E \cup C$, rather than relying on a holistic image-level judgment. The Verifier returns itemized review results $R_t = \{r_i^{(t)}\}_{i \in I}$, where each $r_i^{(t)}$ records a verdict $v_i^{(t)} \in \{\texttt{pass}, \texttt{fail}, \texttt{uncertain}\}$ and a textual reason. We define the set of items requiring further action as

$$F_t = \{\, i \in I \mid v_i^{(t)} \in \{\texttt{fail}, \texttt{uncertain}\} \,\}. \tag{1}$$

Failure routing and repair. For each item in F_t, SCOPE maps the verification result back to the current specification. If the item is associated with an unresolved or newly exposed unknown, the issue is treated as a remaining semantic gap, and retrieval or reasoning is invoked to continue resolution. Otherwise, the issue is treated as a visual realization failure: the commitment is already specified, but the generated image does not satisfy it. In this case, SCOPE invokes repair on the violated entity or constraint. The repair skill selects among prompt rewriting, image editing, and regeneration according to the scope of the failure. Prompt rewriting is used when the synthesized prompt does not faithfully express the current specification, image editing is used for localized defects, and regeneration is used when the failure is broad or too entangled for reliable local correction. Thus, verification serves as the routing mechanism between continued semantic resolution and targeted visual repair, keeping post-generation actions grounded in the same specification used before generation. Algorithm 1 summarizes how SCOPE operates across the generation lifecycle.

Stopping and output selection. SCOPE stops early when all entity and constraint reviews pass. If no output satisfies all entities and constraints within the limit of T image-generation/editing calls, SCOPE selects the best candidate according to its runtime verifier, prioritizing entity correctness before constraint satisfaction.

4 Gen-Arena

Benchmark scope. Gen-Arena evaluates whether image generation can jointly satisfy identity, factual, and compositional requirements in natural-language requests involving specific characters, people, and events. Correctly identifying the requested entities is necessary but insufficient: their appearance, attributes, relations, and layouts must also satisfy the request. The benchmark complements evaluations of compositional generation and world knowledge by examining these requirements together in concrete scenarios. Each instance pairs a prompt with a human-annotated evaluation specification of the required entities and constraints, with reference images when needed to establish identity or appearance.

Construction and annotation. Gen-Arena is manually constructed through a human annotation pipeline covering six categories: cartoon, game, sports, entertainment, competition, and ceremony. Annotators first write natural user prompts and collect reference images when identity or appearance cannot be specified reliably by text alone. They then identify visible target entities in each prompt and annotate atomic constraints over these targets, including attributes, relations, and layouts. Each constraint is linked to the entities it concerns. The human annotations and benchmark reference images are reserved for evaluation and are not exposed to the generation workflow. The resulting benchmark contains 300 instances, 1,954 entities, 2,533 constraints, and 310 reference images. Figure 3 summarizes the Gen-Arena construction pipeline and the entity-gated evaluation protocol.

Evaluation protocol. Gen-Arena uses an entity-first strict pass rule. Let $s_e^{(i)} \in \{0, 1\}$ and $s_c^{(i)} \in \{0, 1\}$ denote the binary correctness judgments for entity $e \in E_i$ and constraint $c \in C_i$, respectively.

Table 1: Main results on Gen-Arena. We report Entity-Gated Intent Pass Rate (EGIP) across categories and overall. For Mind-Brush, GEMS, and SCOPE, Overall reports the mean ± sample standard deviation over three independent full runs; category scores are from the original run. Ent. and Comp. denote entertainment and competition. **Bold** and underlining indicate the highest and second-highest distinct scores (means for Overall) within each generation backend, respectively.

Method	Cartoon	Game	Sports	Ent.	Comp.	Ceremony	Overall
Closed-Source Models							
Nano Banana (Google DeepMind, 2025a)	0.02	0.00	0.16	0.14	0.10	0.00	0.07
Nano Banana Pro (Google DeepMind, 2025b)	0.04	0.00	0.34	0.16	0.18	0.54	0.21
Open-Source Models							
Qwen-Image (Wu et al., 2025a)	0.00	0.00	0.10	0.02	0.00	0.00	0.02
PixArt-Sigma (Chen et al., 2024)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FLUX.1-dev (Black Forest Labs, 2024)	0.00	0.00	0.04	0.02	0.00	0.00	0.01
SD-3.5-large (Esser et al., 2024)	0.00	0.00	0.00	0.02	0.00	0.00	0.00
SDXL (Podell et al., 2023)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Janus-Pro-7B (Chen et al., 2025b)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Inference-Time Scaling							
Generator: Qwen-Image-2512 (Wu et al., 2025a)							
Original	0.02	0.00	0.14	0.04	0.02	0.02	0.04
Mind-Brush (He et al., 2026a)	0.22	0.14	0.40	0.32	0.28	0.44	0.30 ± 0.02
GEMS (He et al., 2026b)	<u>0.30</u>	<u>0.22</u>	<u>0.46</u>	<u>0.40</u>	<u>0.32</u>	<u>0.46</u>	<u>0.36</u> ± 0.01
SCOPE	**0.48**	**0.40**	**0.68**	**0.58**	**0.48**	**0.74**	**0.56** ± 0.01
Generator: Z-Image-Turbo (Cai et al., 2025)							
Original	0.00	0.00	0.08	0.02	0.00	0.00	0.02
Mind-Brush (He et al., 2026a)	0.18	0.12	0.36	0.28	0.22	0.40	0.25 ± 0.02
GEMS (He et al., 2026b)	<u>0.26</u>	<u>0.20</u>	<u>0.44</u>	<u>0.34</u>	<u>0.28</u>	<u>0.46</u>	<u>0.32</u> ± 0.02
SCOPE	**0.42**	**0.36**	**0.64**	**0.52**	**0.44**	**0.68**	**0.51** ± 0.01

Across N instances, **Entity-Gated Intent Pass Rate (EGIP)** is defined as

$$\mathrm{EGIP} = \frac{1}{N}\sum_{i=1}^{N}\left[\prod_{e\in E_i} s_e^{(i)} \cdot \prod_{c\in C_i} s_c^{(i)}\right]. \tag{2}$$

The product is evaluated with an entity-first short-circuit rule: if any required entity fails, the instance receives zero and all constraint checks are skipped. Only when every entity passes are all constraints evaluated. Thus, an instance passes only when every required entity and constraint is satisfied; skipped constraints receive no individual judgment. Additional construction details, category statistics, and evaluation prompts are provided in Appendix B.

5 Experiments

We evaluate SCOPE on three benchmarks with two generation backends, comparing performance, component contributions, and workflow efficiency.

5.1 Experimental Setup

Baselines. We compare SCOPE with closed- and open-source image generators and two agentic methods: Mind-Brush (He et al., 2026a) and GEMS (He et al., 2026b). Original denotes direct generation from the input prompt.

Benchmarks and Metrics. We use EGIP on Gen-Arena and the official protocols on WISE-V (Niu et al., 2025) and MindBench (He et al., 2026a). Gen-Arena Overall reports the mean and sample standard deviation over three independent full runs on the same 300 prompts per backend, with configurations and call budgets fixed. We measure workflow overhead in terms of monetary cost

Table 2: External benchmark results on WISE-V (Niu et al., 2025) and MindBench (He et al., 2026a). All scores are on a 0–1 scale. – denotes unavailable results. **Bold** and underlining indicate the highest and second-highest distinct scores in each column, respectively. Rankings are computed separately within the Qwen-Image-2512 and Z-Image-Turbo groups; ties receive the same formatting. Standalone generator scores in the upper block are unmarked.

Method	WISE-V							MindBench		
	Culture	Time	Space	Biology	Physics	Chemistry	Overall	Knowledge	Reasoning	Overall
Closed-Source Models										
Nano Banana Pro (Google DeepMind, 2025b)	0.90	0.82	0.93	0.82	0.87	0.88	0.88	0.40	0.44	0.42
GPT-Image-1.5 (OpenAI, 2025)	0.89	0.69	0.88	0.80	0.76	0.78	0.83	0.22	0.18	0.20
Open-Source Models										
Bagel w/ CoT (Deng et al., 2025)	0.78	0.63	0.57	0.38	0.55	0.51	0.63	–	–	–
Z-Image (Cai et al., 2025)	0.55	0.47	0.51	0.33	0.48	0.18	0.45	0.02	0.00	0.01
FLUX.1-dev (Black Forest Labs, 2024)	0.52	0.40	0.53	0.18	0.38	0.24	0.42	0.01	0.03	0.02
FLUX.1-schnell (Black Forest Labs, 2024)	0.47	0.33	0.47	0.21	0.38	0.10	0.36	–	–	–
SD-3.5-large (Esser et al., 2024)	0.49	0.41	0.44	0.30	0.38	0.21	0.40	0.02	0.03	0.03
SD-3.5-medium (Esser et al., 2024)	0.48	0.38	0.38	0.18	0.39	0.20	0.38	0.01	0.00	0.01
SDXL (Podell et al., 2023)	0.49	0.37	0.24	0.27	0.33	0.18	0.36	0.02	0.01	0.01
Bagel (Deng et al., 2025)	0.41	0.35	0.31	0.20	0.44	0.26	0.35	0.00	0.03	0.02
Janus-Pro-7B (Chen et al., 2025b)	0.37	0.35	0.28	0.28	0.40	0.23	0.33	–	–	–
Inference-Time Scaling										
Generator: Qwen-Image-2512 (Wu et al., 2025a)										
Original	0.65	0.53	0.60	0.47	0.57	0.36	0.56	0.06	0.04	0.05
Mind-Brush (He et al., 2026a)	0.78	0.64	0.73	0.62	0.72	0.59	0.71	0.46	0.30	0.38
GEMS (He et al., 2026b)	0.87	0.77	0.85	0.74	0.82	0.75	0.82	0.40	0.30	0.35
SCOPE	**0.88**	**0.96**	**0.95**	**0.86**	**0.88**	**0.87**	**0.90**	**0.50**	**0.54**	**0.52**
Generator: Z-Image-Turbo (Cai et al., 2025)										
Original	0.66	0.58	0.65	0.49	0.61	0.43	0.60	0.04	0.02	0.03
Mind-Brush (He et al., 2026a)	0.67	0.57	0.65	0.51	0.62	0.46	0.61	0.40	0.26	0.33
GEMS (He et al., 2026b)	**0.87**	0.78	0.86	0.76	**0.84**	0.75	0.83	0.36	0.26	0.31
SCOPE	0.86	**0.92**	**0.91**	**0.80**	**0.84**	**0.77**	**0.85**	**0.45**	**0.49**	**0.47**

(USD) and execution time (hours) per complete benchmark run, including controller inference and retrieval. These measurements exclude the cost and inference time of the image generation backends.

Generative Backends. We run each agentic method with Qwen-Image-2512 (Wu et al., 2025a) and Z-Image-Turbo (Cai et al., 2025), reporting the two configurations separately. Qwen-Image-Edit-2511 provides visual conditioning and image editing when required and is shared across the agentic methods.

Method Configurations. All three agentic methods share Kimi-K3 and a limit of three generation/editing calls per instance. SCOPE uses the Google Search API for retrieval. Gemini 3-Pro evaluates Gen-Arena using the entity-first protocol and available benchmark references (Section 4). Blinded human evaluation covers all three methods on the same 120 prompts with Qwen-Image-2512 (Appendix C). Implementation details and skill instructions appear in Appendix D.

5.2 Main Results

Results on Gen-Arena. Table 1 shows that SCOPE outperforms the agentic baselines across all six categories with both generation backends. This overall performance advantage persists across three independent runs, with run-to-run variation small relative to the performance gaps between methods. Blinded human evaluation on a shared subset preserves the ranking obtained by the automatic evaluator (Appendix C).

Results on external benchmarks. Table 2 shows that SCOPE also achieves the strongest overall performance among the agentic methods on WISE-V and MindBench with both generation backends. The gains on MindBench span both Knowledge and Reasoning, while performance on individual WISE-V categories varies. These results demonstrate SCOPE's effectiveness on both knowledge-intensive and reasoning-oriented generation tasks.

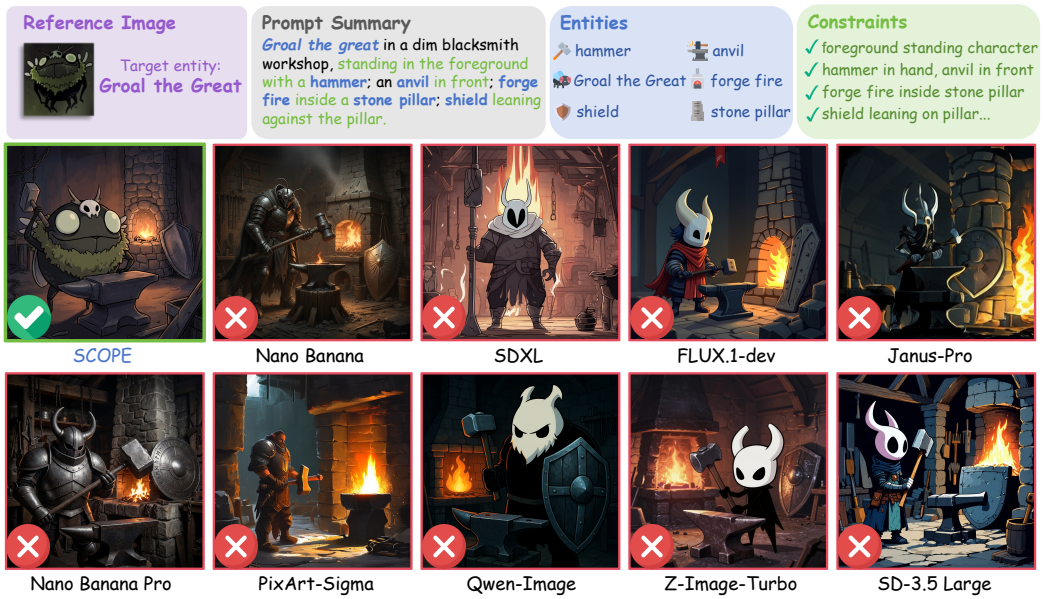

Figure 4: Qualitative comparison on a Gen-Arena example requiring character identity and spatial constraints. SCOPE uses Qwen-Image-2512. The displayed reference image is used only for evaluation and is not provided to the generation workflow; green checks and red crosses indicate whether the complete request is satisfied.

Qualitative comparison. Figure 4 compares generated images for a reference-dependent character in a blacksmith workshop. SCOPE preserves the target identity together with the requested objects and spatial relations, while the other examples exhibit errors in identity or constraint satisfaction.

5.3 Ablation Study

We conduct ablations on Gen-Arena under both generation backends to examine what contributes to SCOPE's performance (Table 3). The comparisons distinguish the role of persistent, item-linked specification tracking from those of retrieval and reasoning for resolving missing information and repair for correcting visual failures. We also compare against direct generation with single-sample and best-of-three selection to assess whether additional sampling alone can explain the gains.

Table 3: Ablation on Gen-Arena (EGIP; higher is better), with results from a single run. Direct uses the original prompt; best-of-3 generates three candidates, scores each with the official evaluator, and selects the highest-scoring candidate as an oracle reference. The w/o specification variant uses free-form records and feedback while retaining retrieval, reasoning, repair, and the same budget limit. **Bold** indicates the best result per backend.

Method	Qwen-Image-2512	Z-Image-Turbo
Direct (single)	0.04	0.02
Direct (best-of-3)	0.09	0.03
SCOPE w/o specification	0.40	0.35
SCOPE w/o retrieval & reasoning	0.12	0.08
SCOPE w/o repair	0.42	0.38
SCOPE	**0.56**	**0.51**

Contribution of the specification. We replace the structured specification and its item-linked updates with free-form records of evidence and feedback. The controller, generation and editing

Table 4: Cost and latency on Gen-Arena, WISE-V, and MindBench. Values report workflow orchestration cost (USD) and execution time (hours) for one complete run of each benchmark. Lower is better; **bold** and underlining mark the lowest and second-lowest values within each generation backend, respectively.

Generator	Method	Gen-Arena		WISE-V		MindBench	
		Cost ($)	Latency (h)	Cost ($)	Latency (h)	Cost ($)	Latency (h)
Qwen-Image-2512	Mind-Brush	3.80	4.00	4.40	4.34	6.00	6.00
	GEMS	9.00	6.50	9.17	7.94	10.00	11.20
	SCOPE	**1.90**	**3.20**	**1.97**	**3.65**	**2.80**	**5.00**
Z-Image-Turbo	Mind-Brush	3.60	3.90	4.20	4.26	5.70	5.90
	GEMS	9.20	8.20	9.63	7.83	12.60	10.80
	SCOPE	**1.80**	**3.00**	**1.93**	**3.44**	**2.70**	**4.70**

backends, tool access, and image-call budget remain fixed; verification and repair retain the same stages and available operations but use free-form records instead of item-linked specification entries. SCOPE outperforms this variant with both generators, supporting the value of persistent commitment tracking.

Interaction with retrieval, reasoning, and repair. Removing retrieval and reasoning while retaining the specification causes the largest degradation, indicating that organizing unresolved requirements cannot substitute for obtaining or inferring the information needed to realize them. Repair also improves complete-intent success relative to its removal, showing the value of revising candidates after verification. These results highlight complementary roles: the skills resolve missing information and revise visual failures, while the specification preserves which commitments their outputs and feedback concern.

Comparison with direct sampling. Direct (single) generates one image from the original prompt. Direct (best-of-3) generates three candidates, scores each with the official evaluator, and selects the highest-scoring candidate. SCOPE outperforms direct best-of-three selection under both backends, suggesting that additional direct sampling alone does not account for the gains of the complete workflow.

5.4 Cost and Latency

Table 4 compares the workflow orchestration cost and latency of the agentic methods on Gen-Arena, WISE-V, and MindBench. SCOPE achieves the lowest cost and shortest execution time across all three benchmarks with both generation backends. Together with the stronger overall performance in Tables 1 and 2, these results show that SCOPE combines performance gains with lower expenditure and shorter execution time than the compared workflows.

6 Conclusion

We introduced SCOPE, a specification-guided skill orchestration framework for complex image generation. SCOPE addresses the Conceptual Rift by representing the semantic commitments imposed by complex visual intents in an evolving structured specification. It then orchestrates retrieval, reasoning, and repair skills around commitments that remain unresolved or violated. We further introduced Gen-Arena, a human-annotated benchmark for commitment-level evaluation, together with EGIP as a strict entity-first pass metric. On Gen-Arena, WISE-V, and MindBench, SCOPE achieves consistent gains over the agentic baselines with Qwen-Image-2512 and Z-Image-Turbo, together with lower workflow orchestration cost and latency.

Ethics Statement

We follow the ethical principles of ICLR 2027 and prioritize responsible research practices throughout this study. Our methodology and findings are reported transparently and honestly, with attention to potential harms and biases in the data and generation process. We take measures to mitigate these risks, respect individual privacy, and ensure that data use complies with applicable regulations and ethical standards.

Reproducibility Statement

We present SCOPE, a specification-guided framework for complex image generation. To support reproducibility and further research, we provide the source code in the supplementary materials. Section 5.1 describes the model configurations, generation backends, and evaluation settings. Appendices A–D provide the orchestration algorithm, Gen-Arena construction and statistics, human validation protocol, and implementation details, including the skill instruction templates.

Large Language Models Usage Disclosure

Large language models were used to assist with manuscript preparation and revision, including grammar correction, clearer phrasing, organization of the text, and LaTeX formatting. The research ideas and experimental design were developed by the authors. The authors made the final decisions on the manuscript and take responsibility for its scientific content, experimental results, and conclusions.

References

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. `https://cdn.openai.com/papers/dall-e-3.pdf`, 2023. OpenAI technical report.

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

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Chieh-Yun Chen, Min Shi, Gong Zhang, and Humphrey Shi. T2I-Copilot: A training-free multi-agent text-to-image system for enhanced prompt interpretation and interactive generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19396–19405, 2025a.

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

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. *arXiv preprint arXiv:2603.29620*, 2026.

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

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

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*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633, 2024.

Kaituo Feng, Manyuan Zhang, Shuang Chen, Yunlong Lin, Kaixuan Fan, Yilei Jiang, Hongyu Li, Dian Zheng, Chenyang Wang, and Xiangyu Yue. Gen-Searcher: Reinforcing agentic search for image generation. *arXiv preprint arXiv:2603.28767*, 2026.

Shivank Garg, Ayush Singh, and Gaurav Kumar Nayak. SiDiffAgent: Self-improving diffusion agent. *arXiv preprint arXiv:2602.02051*, 2026.

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

Google DeepMind. Nano Banana image generation. `https://ai.google.dev/gemini-api/docs/image-generation`, 2025a.

Google DeepMind. Gemini 3 Pro Image: Nano Banana Pro. `https://deepmind.google/models/gemini-image/pro/`, 2025b.

Ruiyan Han, Zhen Fang, XinYu Sun, Yuchen Ma, Ziheng Wang, Yu Zeng, Zehui Chen, Lin Chen, Wenxuan Huang, Wei-Jie Xu, et al. UniCorn: Towards self-improving unified multimodal models through self-generated supervision. *arXiv preprint arXiv:2601.03193*, 2026.

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. *arXiv preprint arXiv:2602.01756*, 2026a.

Zefeng He, Siyuan Huang, Xiaoye Qu, Yafu Li, Tong Zhu, Yu Cheng, and Yang Yang. GEMS: Agent-native multimodal generation with memory and skills. *arXiv preprint arXiv:2603.28088*, 2026b.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Wenxuan Huang, Shuang Chen, Zheyong Xie, Shaosheng Cao, Shixiang Tang, Yufan Shen, Qingyu Yin, Wenbo Hu, Xiaoman Wang, Yuntian Tang, et al. Interleaving reasoning for better text-to-image generation. *arXiv preprint arXiv:2509.06945*, 2025.

Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, et al. Vision-DeepResearch: Incentivizing deepresearch capability in multimodal large language models. *arXiv preprint arXiv:2601.22060*, 2026.

Shantanu Jaiswal, Mihir Prabhudesai, Nikash Bhardwaj, Zheyang Qin, Amir Zadeh, Chuan Li, Katerina Fragkiadaki, and Deepak Pathak. Iterative refinement improves compositional image generation. *arXiv preprint arXiv:2601.15286*, 2026.

Guanyu Jiang, Zhaochen Su, Xiaoye Qu, and Yi R. Fung. XSkill: Continual learning from experience and skills in multimodal agents. *arXiv preprint arXiv:2603.12056*, 2026a.

Yanna Jiang, Delong Li, Haiyu Deng, Baihe Ma, Xu Wang, Qin Wang, and Guangsheng Yu. SoK: Agentic skills–beyond tool use in llm agents. *arXiv preprint arXiv:2602.20867*, 2026b.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5290–5301, 2024.

Chuanhao Li, Jianwen Sun, Yukang Feng, Mingliang Zhai, Yifan Chang, and Kaipeng Zhang. IA-T2I: Internet-augmented text-to-image generation. *arXiv preprint arXiv:2505.15779*, 2025a.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Arsh Koneru, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Reflect-DiT: Inference-time scaling for text-to-image diffusion transformers via in-context reflection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15657–15668, 2025b.

Xiangyi Li, Wenbo Chen, Yimin Liu, Shenghan Zheng, Xiaokun Chen, Yifeng He, Yubo Li, Bingran You, Haotian Shen, Jiankai Sun, et al. SkillsBench: Benchmarking how well agent skills work across diverse tasks. *arXiv preprint arXiv:2602.12670*, 2026.

Anthony Zhe Liu, Jongwook Choi, Sungryull Sohn, Yao Fu, Jaekyeom Kim, Dong-Ki Kim, Xinhe Wang, Jaewon Yoo, and Honglak Lee. SkillAct: Using skill abstractions improves llm agents. In *ICML 2024 Workshop on LLMs and Cognition*, 2024.

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. WISE: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

OpenAI. The new ChatGPT images is here. `https://openai.com/index/new-chatgpt-images-is-here/`, 2025.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Moo Hyun Son, Jintaek Oh, Sun Bin Mun, Jaechul Roh, and Sehyun Choi. World-to-image: Grounding text-to-image generation with agent-driven world knowledge. *arXiv preprint arXiv:2510.04201*, 2025.

Kavana Venkatesh, Connor Dunlop, and Pinar Yanardag. CREA: A collaborative multi-agent framework for creative image editing and generation. *arXiv preprint arXiv:2504.05306*, 2025.

Kaishen Wang, Ruibo Chen, Tong Zheng, and Heng Huang. ImAgent: A unified multimodal agent framework for test-time scalable image generation. *arXiv preprint arXiv:2511.11483*, 2025.

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

Shiyu Wu, Mingzhen Sun, Weining Wang, Yequan Wang, and Jing Liu. VisualPrompter: Semantic-aware prompt optimization with visual feedback for text-to-image synthesis. *arXiv preprint arXiv:2506.23138*, 2025b.

Dawei Xiang, Wenyan Xu, Kexin Chu, Tianqi Ding, Zixu Shen, Yiming Zeng, Jianchang Su, and Wei Zhang. PromptSculptor: Multi-agent based text-to-image prompt optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 774–786, 2025.

Wen Ye, Zhaocheng Liu, Yuwei Gui, Tingyu Yuan, Yunyue Su, Bowen Fang, Chaoyang Zhao, Qiang Liu, and Liang Wang. GenPilot: A multi-agent system for test-time prompt optimization in image generation. *arXiv preprint arXiv:2510.07217*, 2025.

Yu Zeng, Wenxuan Huang, Zhen Fang, Shuang Chen, Yufan Shen, Yishuo Cai, Xiaoman Wang, Zhenfei Yin, Lin Chen, Zehui Chen, et al. Vision-DeepResearch Benchmark: Rethinking visual and textual search for multimodal large language models. *arXiv preprint arXiv:2602.02185*, 2026.

Boyuan Zheng, Michael Y. Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, et al. SkillWeaver: Web agents can self-improve by discovering and honing skills. *arXiv preprint arXiv:2504.07079*, 2025.

YanZhao Zheng, ZhenTao Zhang, Chao Ma, YuanQiang Yu, JiHuai Zhu, Yong Wu, Tianze Xu, Baohua Dong, Hangcheng Zhu, Ruohui Huang, et al. SkillRouter: Skill routing for llm agents at scale. *arXiv preprint arXiv:2603.22455*, 2026.

Le Zhuo, Liangbing Zhao, Sayak Paul, Yue Liao, Renrui Zhang, Yi Xin, Peng Gao, Mohamed Elhoseiny, and Hongsheng Li. From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15329–15339, 2025.

A SCOPE ALGORITHM

Algorithm 1 SCOPE Generation and Repair Lifecycle

Require: User prompt p, maximum image-generation/editing calls T

$z_0 = (E_0, C_0, U_0) \leftarrow \textsc{Decompose}(p)$; $y^*, y_0, F_0, \rho_1 \leftarrow \varnothing$; $a_1 \leftarrow \textsc{Generate}$
for $t = 1$ to T **do**
  $z_t \leftarrow z_{t-1}$
  for all unresolved $u \in U_t$ **do**
    $e_u \leftarrow \begin{cases} \textsc{Retrieve}(u), & \text{if external evidence is needed} \\ \textsc{Reason}(u, z_t), & \text{otherwise} \end{cases}$
    $z_t \leftarrow \textsc{UpdateSpec}(z_t, u, e_u)$
  end for
  $s_t \leftarrow \textsc{SynthesizePrompt}(z_t, \rho_t)$
  $y_t \leftarrow \begin{cases} \textsc{Edit}(y_{t-1}, s_t, \rho_t), & a_t = \textsc{Edit} \wedge y_{t-1} \neq \varnothing \\ \textsc{Generate}(s_t), & \text{otherwise} \end{cases}$
  $R_t \leftarrow \textsc{Verify}(y_t, z_t)$; $y^* \leftarrow \textsc{UpdateBest}(y^*, y_t, R_t)$
  if $\textsc{AllPass}(R_t)$ **then**
    return y_t
  end if
  $F_t \leftarrow \textsc{MapFailures}(R_t, z_t)$
  for all $i \in F_t$ **do**
    $z_t \leftarrow \begin{cases} \textsc{AddUnknown}(z_t, i), & \text{if } i \text{ exposes an unknown} \\ \textsc{RecordRepair}(z_t, i), & \text{otherwise} \end{cases}$
  end for
  $(a_{t+1}, \rho_{t+1}) \leftarrow \textsc{PlanRepair}(F_t, z_t, y_t)$
end for
return y^*

B GEN-ARENA CONSTRUCTION AND EVALUATION

B.1 CONSTRUCTION PIPELINE

Gen-Arena is constructed through a human annotation pipeline. Annotators first write natural user-style prompts for each category, avoiding template-like wording while keeping every requested element visually verifiable. When a prompt requires a specific identity, character design, outfit, or appearance that cannot be reliably specified by text alone, annotators attach one or more reference images. They then identify all visible target entities required by the prompt and write atomic constraints over these entities. Each constraint is assigned one of three types—attribute, relation, or layout—and linked to the IDs of the entities it concerns. These links specify the targets of each constraint.

B.2 BENCHMARK STATISTICS

Table 5 reports the category-wise statistics of Gen-Arena. Each category contains 50 instances. Reference images are included when the intended identity or appearance cannot be fully specified by text alone.

Table 5: Gen-Arena statistics. Ent. and Cons. denote the numbers of annotated entities and constraints. Refs. counts reference images linked to instances.

Category	Cases	Ent.	Cons.	Refs.
Cartoon	50	357	490	56
Game	50	334	489	104
Sports	50	340	398	50
Entertainment	50	348	451	50
Competition	50	256	306	0
Ceremony	50	319	399	50
Total	300	1954	2533	310

Table 6: Distribution of annotated constraints across the six Gen-Arena categories. Counts refer to individual attribute, relation, and layout constraints rather than benchmark instances; each instance may contain multiple constraints. Relation constraints form the largest group, reflecting the benchmark's emphasis on interactions among entities.

Category	Attribute	Relation	Layout
Cartoon	122	245	123
Game	84	262	143
Sports	63	258	77
Entertainment	84	280	87
Competition	96	142	68
Ceremony	61	237	101
Total	510	1424	599

B.3 Constraint-Type Distribution

Table 6 reports the distribution of constraint types. Relation constraints account for the largest portion because many Gen-Arena instances require multi-entity interactions or event-specific relationships to be correctly realized.

B.4 Annotation and Quality Control

Gen-Arena is manually designed and verified. During annotation, prompts are required to remain natural user requests, entities must correspond to visible targets, and constraints must be verifiable from a single generated image whenever possible. We further check four types of quality issues. First, reference images must correspond to the intended target and must be usable for visual identity or appearance comparison. Second, every annotated entity must be a concrete visible target rather than an abstract requirement. Third, constraints must be atomic so that a binary pass/fail judgment is meaningful. Fourth, each constraint dependency must point to the entities required for judging that constraint; for example, a relation constraint involving two people depends on both people.

B.5 Official Evaluation Protocol

The official evaluator receives the generated image, the natural-language prompt, optional reference images, and the structured evaluation specification. It first produces binary judgments for all required entities. If any entity fails, the instance fails and all constraint checks are skipped. Otherwise, it judges every constraint and the instance passes only if all constraints pass. EGIP is then computed from these judgments using the entity-gated rule defined in Section 4. The condensed evaluator prompt template is shown in Figure 5.

C Human Validation and Cross-Method Evaluation

C.1 Annotation Protocol

Samples and annotators. We randomly sample 120 Gen-Arena instances, with 20 from each of the six categories, and evaluate one final output from each of SCOPE, GEMS, and Mind-Brush per instance, yielding 360 images in total. All three methods use Kimi-K3 and Qwen-Image-2512. Three annotators independently assess each image using shared annotation instructions. Method identities are hidden, and image presentation order is randomized. They receive the input prompt, generated image, human-annotated entity and constraint checklist, and reference images when needed to establish identity or appearance. Annotators do not see other annotators' responses or the automatic evaluator's judgments.

Entity-first annotation. Each annotator first judges every required entity as pass or fail. If any entity fails, the annotator assigns an instance-level EGIP decision of 0 and skips all constraints.

Condensed Gen-Arena Evaluation Prompt Template

Input.
Generated image, natural-language prompt, optional reference images, entity list, and constraint list with prerequisite entity IDs.

Evaluation Goal.
Judge whether the image fulfills the specified visual intent. Focus on entity correctness, reference consistency, and constraint satisfaction rather than image aesthetics or overall plausibility.

Entity Judging.
Evaluate every required entity first. An entity passes only if it is visible and correctly realized. For entities with `reference_ids`, compare the generated image against the corresponding reference images rather than relying only on the text prompt.

Constraint Judging.
Only after all required entities pass, evaluate each attribute, relation, or layout constraint from visual evidence in the generated image. Use fail when the required evidence is absent, ambiguous, assigned to the wrong entity, or replaced by a generic substitute.

Gating Rule.
If any required entity fails, mark the entire instance as failed and skip all constraint checks. Evaluate constraints only when every required entity passes. The instance passes only if every constraint also passes.

Output Format.
Return valid JSON with two arrays, `entities` and `constraints`. Each item contains `id`, `verdict` in {`pass`, `fail`}, a brief `reason`, and visible `evidence`. If any entity fails, return an empty constraint array because constraint checks are skipped; this instance has EGIP 0.

Figure 5: Condensed official evaluation prompt template for Gen-Arena. The evaluator produces item-level entity and constraint judgments, and EGIP is computed from these judgments using the entity-gated rule.

Otherwise, the annotator evaluates each constraint and assigns an EGIP decision of 1 only if all constraints pass. Skipped constraints are recorded as unjudged rather than failed.

Majority-vote reference. We form human reference labels by majority vote over the independent annotations. Entity reference labels are obtained from the three entity judgments. For each constraint, a reference label is available only when at least two annotators explicitly assign the same label; skipped judgments do not count as votes. Constraints without such a majority are excluded from human–evaluator constraint agreement. The human instance-level reference is the majority vote over the three annotators' final EGIP decisions, rather than an EGIP decision recomputed from the majority-vote entity and constraint labels.

C.2 Human–Evaluator Agreement

Agreement computation. Table 7 reports raw agreement rates on the 120 SCOPE outputs. Inter-annotator agreement is computed before majority-vote aggregation: we calculate the fraction of matching judgments for each annotator pair and average over the three pairs. Human–Gemini agreement compares the majority-vote reference labels with the corresponding Gemini judgments. Entity agreement uses all required entity items, and EGIP agreement uses all 120 images for the corresponding method. Constraint agreement is restricted to items for which both sides have valid judgments: both annotators for a human pair, or an available human majority label and an explicit Gemini judgment for human–Gemini comparison. The denominator is the number of eligible constraint items in each comparison; jointly skipped constraints are not counted as agreements.

C.3 Blind Cross-Method Human Evaluation

We apply the same annotation and majority-vote protocol to all three methods on the paired set of 120 prompts. Gemini EGIP and Human EGIP are the fractions of images passing the automatic evaluation and the majority-vote human evaluation, respectively. Human–Gemini agreement is the fraction of images for which the two instance-level decisions match, counting both jointly passing and jointly failing instances. Table 8 reports the scores and agreement for each method; SCOPE's

Table 7: Agreement on 120 SCOPE outputs with Kimi-K3 and Qwen-Image-2512. Inter-annotator agreement is computed before aggregation; human–Gemini agreement uses majority-vote human labels. Constraint agreement includes only eligible, explicitly judged items.

Evaluation	Entity	Constraint	EGIP
Inter-annotator agreement	0.86	0.81	0.84
Human–Gemini agreement	0.91	0.87	0.89

Table 8: Blind human evaluation on 120 shared Gen-Arena prompts using Kimi-K3 and Qwen-Image-2512. Human EGIP uses majority vote; agreement compares human and Gemini instance-level decisions.

Method	Gemini EGIP	Human EGIP	Human–Gemini Agreement
Mind-Brush	0.32	0.30	0.90
GEMS	0.37	0.35	0.90
SCOPE	0.57	0.56	0.89

agreement is the same value reported in Table 7. Human evaluation preserves the method ranking obtained by Gemini, with SCOPE outperforming both agentic baselines. The comparable agreement rates across methods extend evaluator validation beyond SCOPE outputs.

D Implementation Details

We describe the runtime interface, benchmark input isolation, and skill execution rules used by SCOPE.

D.1 Runtime Configuration and Execution

Runtime interface and benchmark inputs. The controller produces structured JSON inputs for decomposition, synthesis, verification, and skill execution. Stage tools validate these inputs and persist the resulting state and artifacts, which subsequent stages read. For Gen-Arena, the generation workflow receives only the natural-language prompt. Human-annotated entities, constraints, and benchmark reference images are stored separately for final evaluation and are not supplied to decomposition, runtime verification, retrieval, reasoning, or repair. SCOPE constructs its runtime specification from the prompt; any visual references used during generation are obtained through its retrieval skill.

Skill execution. Retrieval plans bind each query to an open external-reference unknown and separate factual text queries from visual identity or appearance queries. The returned resolution contains the unknown identifier, a generation-facing note, and supporting evidence. Reasoning similarly returns a closed result for an open semantic unknown, such as an exact value, selected entity, or visible text string. Requirements that are already explicit do not trigger additional reasoning. Resolved unknowns are marked as consumed when synthesis incorporates their results into the generation prompt; this status does not indicate visual satisfaction.

Runtime verification and repair. Runtime review records contain the target identifier, verdict, reason, and visible evidence. Failed or uncertain items also identify the relevant repair family. An uncertain verdict is not accepted as a pass. If verification reveals missing knowledge or an unresolved inference, the workflow records an owner-linked unknown and returns to retrieval or reasoning before synthesis. Otherwise, repair targets the failed or uncertain items. Prompt rewriting addresses unclear generation instructions, image editing targets localized defects, and regeneration handles structural or repeated repair failures. Repairs preserve the original identity, count, relations, and exact-text

requirements. These diagnostic reviews guide the workflow; final benchmark scoring follows the separate EGIP protocol.

Stopping and output selection. Generation and image editing share the three-call budget. Prompt rewriting itself does not create another image, but the subsequent generation consumes a call. The workflow stops early when all runtime reviews pass. At the call limit, SCOPE selects a candidate using its runtime verifier, prioritizing entity correctness before constraint satisfaction. The run records the structured state, stage sequence, generation prompts, candidate images, verification results, and repair history.

D.2 Skill Instruction Templates

Figure 6 summarizes the condensed templates for the main orchestration, retrieval, reasoning, and repair skills.

Condensed SCOPE Skill Instruction Templates

Main Orchestration Skill.
Trigger. Run once for each complex image-generation prompt.
Operation. Decompose the prompt into entities, constraints, and unknowns; invoke retrieval or reasoning when needed; then synthesize, generate, verify, repair, and finalize.
Output. Final image, prompt, and runtime verification records.

Retrieval Skill.
Trigger. Run when open `external_reference` unknowns exist.
Operation. Create short text or image queries for external grounding; use text search for factual lookup and image search for visual identity or appearance grounding.
Output. Retrieval resolutions with unknown IDs, generation-facing notes, and supporting evidence.

Reasoning Skill.
Trigger. Run when open `semantic_reasoning` unknowns exist.
Operation. Resolve each unknown into a closed semantic result, such as a final value, selected option, inferred relation, or exact visible text.
Output. Reasoning resolutions for the next synthesized prompt.

Repair Skill.
Trigger. Run for failed or uncertain items that have no unresolved information gaps requiring retrieval or reasoning.
Operation. Select among prompt rewriting, image editing, regeneration, or no action. Prompt rewriting handles prompt-specification mismatch, image editing handles localized defects, and regeneration handles broad or entangled failures.
Output. Selected review IDs, repair action, optional updated prompt, and focused repair patch.

Figure 6: Condensed skill instruction templates used by SCOPE.