SCOPE: Structured Decomposition and Conditional Skill Orchestration for Complex Image Generation

Anonymous ACL submission

Abstract

While text-to-image models have made strong progress in visual fidelity, faithfully realizing complex visual intents remains challenging because many requirements often need to be tracked across grounding, generation, and verification. We refer to these requirements as **semantic commitments** and formalize their lifecycle discontinuity as the *Conceptual Rift*, where commitments are locally resolved or checked but may fail to remain identifiable as the same operational units throughout the generation lifecycle. To address this, we propose **SCOPE**, a specification-guided skill orchestration framework that maintains semantic commitments in an evolving structured specification and conditionally invokes retrieval, reasoning, and repair skills around unresolved or violated commitments. To evaluate commitment-level intent realization, 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. SCOPE substantially outperforms all evaluated baselines on Gen-Arena, achieving 0.60 EGIP, and further achieves strong results on WISE-V (0.907) and MindBench (0.61), demonstrating the effectiveness of persistent commitment tracking for complex image generation.

1 Introduction

Text-to-image generation is shifting from producing plausible images toward realizing user-specified visual intent. Earlier progress was often measured by whether generated images looked convincing and aligned with prompts at a coarse level (Saharia et al., 2022; Betker et al., 2023; Esser et al., 2024). As these systems enter practical and creative workflows, however, prompts increasingly describe scenes whose defining details must hold for the result to be faithful (Hu et al., 2023; Huang et al., 2023; Ghosh et al., 2023; Li et al., 2024). This changes the standard for generation: an image should not only appear natural, but also correctly instantiate the scene the user intended to depict.

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

Faithfully realizing complex visual intents is challenging because the requirements imposed by a prompt do not all become actionable at the same time. Some requirements are explicit from the beginning, such as *the car is red*. Others become determinate *only after* the system grounds external context, such as identifying *the winner of the 2024 Wimbledon men's final*, or reasons about implicit scene semantics, such as inferring the correct answer to a visual puzzle (Li et al., 2025a; Son et al., 2025; He et al., 2026a; Feng et al., 2026; Chen et al., 2026). We refer to the requirements imposed by a complex visual intent as **semantic commitments**: *conditions whose fulfillment deter-*

mines whether the generated image realizes the user's intent. The challenge is therefore not only to discover these commitments, but also to preserve them as identifiable units that can be realized, verified, and revised throughout the generation lifecycle.

Recent work has increasingly moved beyond one-shot prompt-to-image generation by introducing multi-step interventions such as retrieval, planning, and iterative refinement (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). These interventions improve different aspects of complex generation, from resolving missing information to correcting visible failures. However, multi-step generation alone is insufficient to ensure lifecycle continuity: a commitment may be resolved before generation, checked after generation, and revised in a later step, yet these operations may not remain tied to the same identifiable unit. As a result, even when a system retrieves relevant information or identifies a real failure, the generation target may still be diluted, the error may be misattributed, or the repair action may be poorly targeted. We refer to this lifecycle discontinuity as the **Conceptual Rift**. This raises a central question: ***how can semantic commitments remain representable, verifiable, and actionable as unified operational units throughout the generation lifecycle?***

To address this challenge, we propose **SCOPE**, a specification-guided skill orchestration framework for complex image generation. Figure 1 shows representative SCOPE outputs across diverse generation scenarios. SCOPE makes semantic commitments explicit in an evolving structured specification, which serves as a shared interface across the generation lifecycle. Guided by this specification, SCOPE conditionally invokes retrieval, reasoning, and repair skills to ground missing external information, infer implicit requirements, and revise violated commitments. It also verifies generated images against the commitments represented in the current specification. By writing skill outputs and verification results back to the specification, SCOPE keeps complex visual intents actionable across stages, allowing generation to proceed through a unified lifecycle rather than a sequence of disconnected local operations.

Due to complex generation is organized around semantic commitments, evaluation should also reveal which commitments are fulfilled or violated. Existing evaluations often rely on holistic alignment scores, while checklist-style protocols may still treat conditions as largely independent, obscuring whether a failure is a root error or a downstream consequence of missing prerequisite content. We introduce **Gen-Arena**, an entity- and constraint-level benchmark that represents each prompt with a structured evaluation specification. By linking constraints to their prerequisite entities, Gen-Arena supports entity-first evaluation and enables diagnosis of missing entities, violated constraints, and downstream failures caused by unmet prerequisites.

In summary, our contributions are as follows:

- We identify the **Conceptual Rift** in complex image generation, where semantic commitments behind a complex visual intent lose continuity across the generation lifecycle, and propose **SCOPE**, which addresses this rift by maintaining these commitments in an evolving specification and orchestrating retrieval, reasoning, and repair skills around it.

- We introduce **Gen-Arena**, a human-annotated benchmark for commitment-level intent realization, with entity- and constraint-level evaluation specifications and **Entity-Gated Intent Pass Rate (EGIP)** as a strict entity-first pass criterion.

- Experiments show that SCOPE substantially outperforms all evaluated baselines on Gen-Arena, achieving 0.60 EGIP, and further achieves strong results on WISE-V (0.907) and MindBench (0.61), demonstrating the value of persistent commitment tracking for complex image generation.

2 Related Work

2.1 Agentic Image Generation

Recent work has begun to use multimodal agents to mediate 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., 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

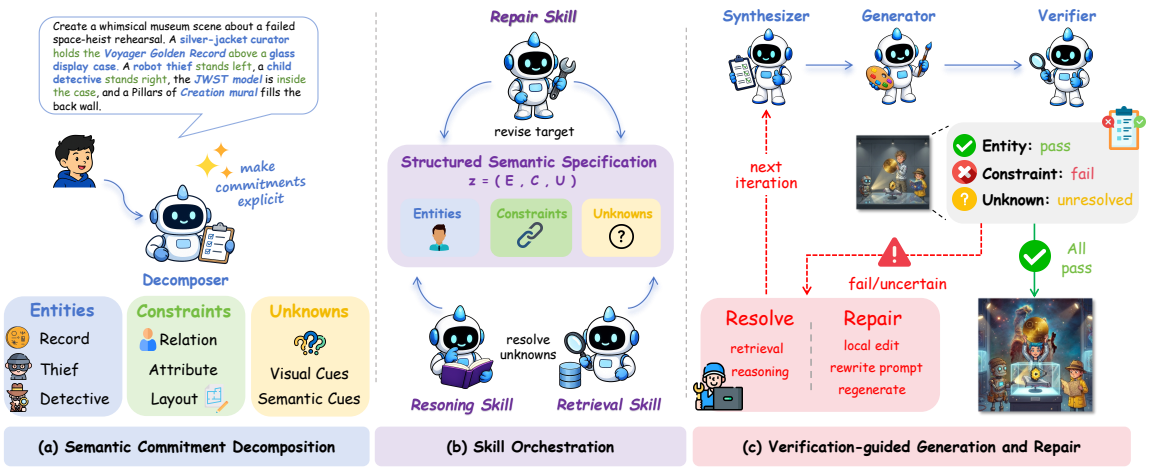

Figure 2: Overview of SCOPE. a)The user prompt is decomposed into an evolving structured semantic specification containing entities, constraints, and unknowns.b) Retrieval and reasoning resolve open unknowns and provide grounding evidence for generation. c) Verification identifies unresolved or violated commitments, after which SCOPE invokes targeted repair to update the prompt or image while preserving the structured specification. Details of the full orchestration procedure are provided in Algorithm 1.

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 agentic mediation for complex generation. However, their intermediate representations are usually tailored to particular interventions, rather than to maintaining the same semantic commitments across the full generation lifecycle. As a result, resolved information, verification outcomes, and repair decisions may not remain reliably tied to the same underlying commitments.

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. However, most existing work treats skills primarily as reusable agent resources. Less explored is how skill use should be grounded in the evolving semantic commitments of a specific task, so that each invocation addresses a concrete unresolved or violated item and its result remains usable in later stages.

3 SCOPE

3.1 Overall Pipeline

SCOPE is designed to keep semantic commitments operational across the generation lifecycle. It represents the current visual intent as an evolving structured specification and uses this specification as the shared interface for generation, verification, and targeted skill invocation. Given a user prompt and an optional reference image, SCOPE iteratively operates over a fixed core pipeline of **Decomposer** → **Synthesizer** → **Generator** → **Verifier**. Specifically, the **Decomposer** transforms the user request into the specification, the **Synthesizer** consolidates resolved information from the current specification into a coherent generation prompt, the **Generator** performs image generation or editing, and the **Verifier** evaluates entities and constraints item by item. In addition to this fixed core loop, SCOPE conditionally orchestrates retrieval, reasoning, and repair skills to address unresolved semantics and localized generation failures as they arise. Figure 2 shows the overall SCOPE architecture, where the structured specification provides the shared interface for generation, verification, and targeted skill invocation.

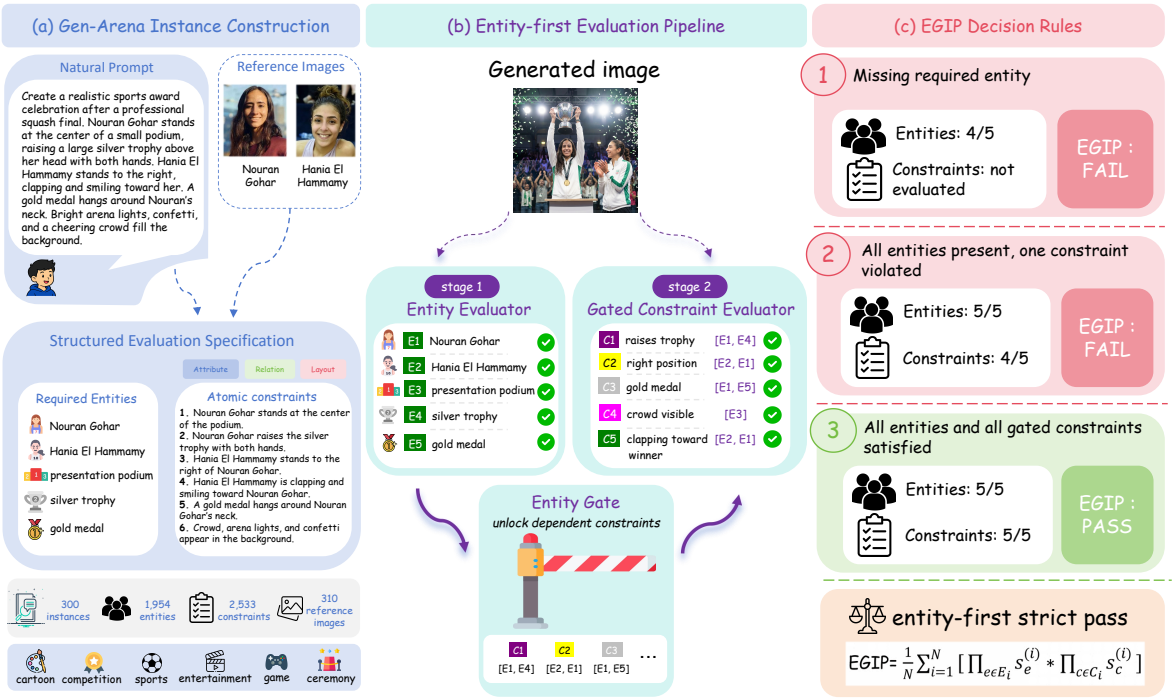

Figure 3: Overview of Gen-Arena construction and EGIP evaluation.a) Natural user prompts are converted into required visible entities.b) Atomic constraints are written over these entities and linked to their prerequisite entities.c) EGIP applies an entity-first rule: constraints are credited only when their prerequisite entities are correctly realized.

3.2 Structured Semantic Specification

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. Importantly, unknowns in U are not treated as independent questions. Each unknown is attached to a prompt-, entity-, or constraint-level owner, indicating which commitment it is meant to resolve. This ownership link allows retrieval or reasoning results to update the corresponding part of the specification, and later allows verification failures to be mapped back to the same entities or constraints. When a verified failure is no longer associated with unresolved unknowns, the mapped entity or constraint becomes the target of repair. Thus, the specification provides the operational interface for carrying resolved information, verification outcomes, and repair targets across iterations.

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 is anchored to a concrete unknown or violated commitment, and its output updates the same specification: retrieval and reasoning close unresolved unknowns, while repair records how violated items are revised. In this way, SCOPE adapts the generation lifecycle according to what remains unresolved or violated in the current specification.

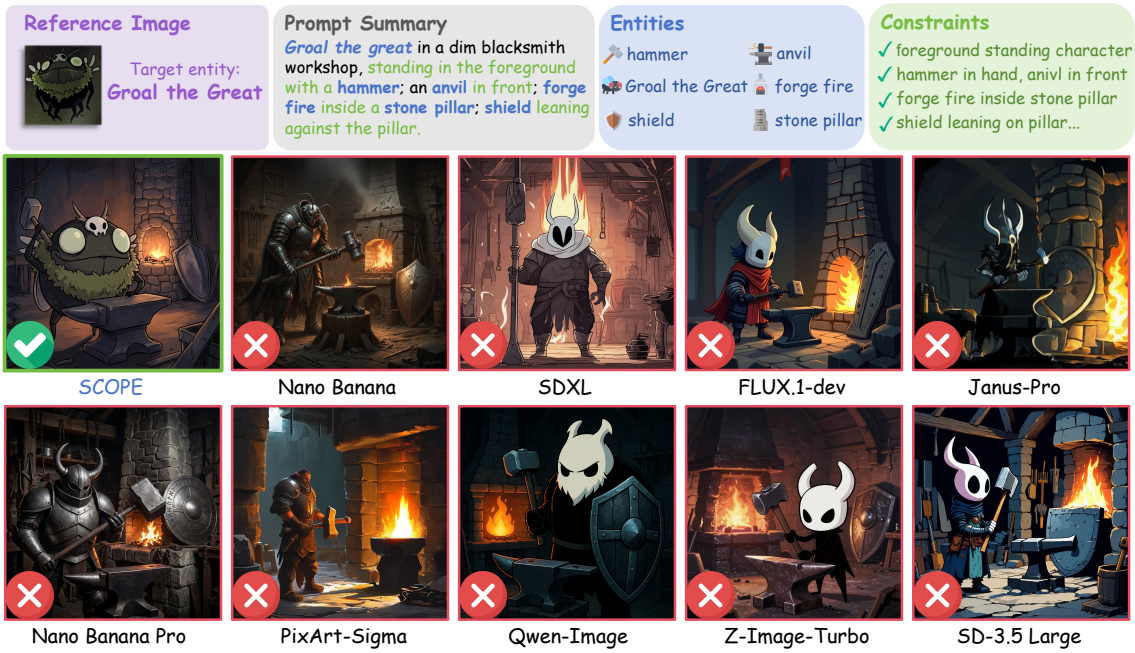

Figure 4: Qualitative comparisons between direct prompting and SCOPE-guided generation using the same backend. Direct prompting often misses entities, weakens identity fidelity, or violates relation and layout constraints, while SCOPE localizes and repairs these failures through explicit commitments.

3.4 Verification-Guided Resolution and Repair

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

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

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. If no output satisfies all entities and constraints within T attempts, SCOPE selects the best candidate according to the verifier, prioritizing entity correctness before downstream constraint satisfaction.

4 Gen-Arena

Gen-Arena evaluates whether complex image generation fulfills structured visual intents rather than only matching prompts at a coarse level. Each instance pairs a natural-language prompt with an evaluation specification that identifies the required entities and the constraints they must satisfy, enabling commitment-level evaluation of generated images.

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 con-

Method	Cartoon	Game	Sports	Ent.	Comp.	Ceremony	Overall
Closed-Source Models							
Nano Banana (Google DeepMind, 2025b)	0.02	0.00	0.16	0.14	0.10	0.00	0.07
Nano Banana Pro (Google DeepMind, 2025a)	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
Z-Image-Turbo (Cai et al., 2025)	0.00	0.00	0.09	0.02	0.00	0.00	0.01
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	0.03	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
SCOPE	**0.52**	**0.46**	**0.72**	**0.62**	**0.52**	**0.74**	**0.60**
Δ vs. NBP	**+0.48**	**+0.46**	**+0.38**	**+0.46**	**+0.34**	**+0.20**	**+0.39**

Table 1: Main results on Gen-Arena. We report Entity-Gated Intent Pass Rate (EGIP) across categories and overall. Ent. and Comp. denote entertainment and competition. **Bold** indicates the best result, underline indicates the strongest non-agentic baseline, and Δ is computed against Nano Banana Pro.

straints over these targets, including attributes, relations, and layouts. Each constraint is linked to the entities it depends on, allowing evaluation to distinguish missing-entity failures from unsatisfied constraints over correctly realized entities. 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.

For evaluation, Gen-Arena uses an entity-first strict pass rule. The evaluator first checks whether all required entities are correctly realized. If any required entity is missing or incorrectly depicted, the instance is marked as failed. Only when all required entities are satisfied does the evaluator check the associated constraints. Let $s_e^{(i)} \in \{0,1\}$ indicate whether entity $e \in E_i$ is correctly realized, and let $s_c^{(i)} \in \{0,1\}$ indicate whether constraint $c \in C_i$ is satisfied. We define **Entity-Gated Intent Pass Rate (EGIP)** as

$$\text{EGIP} = \frac{1}{N}\sum_{i=1}^{N}\left(\prod_{e\in E_i} s_e^{(i)} \prod_{c\in C_i} s_c^{(i)}\right). \quad (2)$$

Thus, EGIP measures strict instance-level intent fulfillment: an example passes only when all required entities and constraints are satisfied. Additional construction details, category statistics, and evaluation prompts are provided in Appendix B.

5 Experiments

We evaluate SCOPE from two perspectives: whether it improves commitment-level realization on Gen-Arena, and whether the same framework transfers to existing complex generation benchmarks.

5.1 Experimental Setup

Baselines and Benchmarks. We evaluate SCOPE on Gen-Arena and two external benchmarks: WISE-V (Niu et al., 2025) and MindBench from Mind-Brush (He et al., 2026a). Gen-Arena uses EGIP to measure strict commitment-level intent fulfillment across six categories, while the external benchmarks test transfer to existing evaluations involving world knowledge and reasoning-intensive visual generation. On Gen-Arena, we compare against closed-source models Nano Banana (Google DeepMind, 2025b) and Nano Banana Pro (Google DeepMind, 2025a), as well as open-source models SDXL (Podell et al., 2023), SD-3.5-large (Esser et al., 2024), FLUX.1-dev (Black Forest Labs, 2024), Qwen-Image (Wu et al., 2025a), Z-Image-Turbo (Cai et al., 2025), PixArt-Sigma (Chen et al., 2024), and Janus-Pro-7B (Chen et al., 2025b).

Implementation Details. We use GPT-5.4 as the multimodal language model backend and Nano Banana Pro as the image generation and editing backend. Retrieval is implemented with the Google Search API. Each method is evaluated under a fixed inference budget of at most three image-generation or image-editing calls per instance. For Gen-Arena, Gemini 3-Pro is used as the official evaluator, which judges entities and constraints item by item. When an entity is associated with reference images, the evaluator is instructed to compare the generated image against the provided references rather than relying only on

Method	WISE-V							MindBench		
	Culture	Time	Space	Biology	Physics	Chemistry	Overall	Knowledge	Reasoning	Overall
Closed-Source Models										
Nano Banana Pro (Google DeepMind, 2025a)	**0.898**	0.817	0.933	0.817	0.867	0.875	0.876	0.40	0.44	0.41
GPT-Image-1.5 (OpenAI, 2025)	0.890	0.692	0.883	0.800	0.758	0.775	0.825	0.22	0.18	0.21
Open-Source Models										
Bagel w/ CoT (Deng et al., 2025)	0.780	0.633	0.567	0.375	0.550	0.508	0.628	-	-	-
Qwen-Image (Wu et al., 2025a)	0.628	0.525	0.558	0.342	0.483	0.250	0.510	0.02	0.02	0.02
Z-Image (Cai et al., 2025)	0.548	0.467	0.508	0.325	0.475	0.175	0.453	0.02	0.00	0.02
FLUX.1-dev (Black Forest Labs, 2024)	0.523	0.400	0.533	0.175	0.375	0.242	0.416	0.01	0.03	0.02
FLUX.1-schnell (Black Forest Labs, 2024)	0.465	0.325	0.467	0.208	0.383	0.100	0.364	-	-	-
SD-3.5-large (Esser et al., 2024)	0.490	0.408	0.442	0.300	0.375	0.208	0.404	0.02	0.03	0.01
SD-3.5-medium (Esser et al., 2024)	0.483	0.375	0.375	0.183	0.392	0.200	0.376	0.01	0.00	0.01
SDXL (Podell et al., 2023)	0.493	0.367	0.242	0.267	0.333	0.183	0.364	0.02	0.01	0.01
Bagel (Deng et al., 2025)	0.413	0.350	0.308	0.200	0.442	0.258	0.352	0.00	0.03	0.02
Janus-Pro-7B (Chen et al., 2025b)	0.370	0.350	0.283	0.283	0.400	0.233	0.334	-	-	-
Mind-Brush (He et al., 2026a)	-	-	-	-	-	-	-	0.38	0.24	0.31
SCOPE	0.883	**1.000**	**0.944**	**0.861**	**0.889**	**0.917**	**0.907**	**0.59**	**0.63**	**0.61**
Δ vs. NBP	**-0.015**	**+0.183**	**+0.011**	**+0.044**	**+0.022**	**+0.042**	**+0.031**	**+0.19**	**+0.19**	**+0.20**

Table 2: External benchmark results on WISE-V (Niu et al., 2025) and MindBench from Mind-Brush (He et al., 2026a). **Bold** indicates the best result, underline indicates the second-best result, and Δ is computed against Nano Banana Pro (NBP).

the text prompt.

For the baselines, Direct (single) generates one image from the original prompt. Direct (best-of-3) generates three independent images from the same prompt and reports the best candidate selected by the official evaluator. Self-refine w/o spec uses the same three-attempt budget as SCOPE, but replaces structured specification tracking with free-form critique and prompt rewriting. SCOPE uses the same maximum budget, allowing each attempt to be either a new generation or a targeted edit according to the verifier feedback.

To audit the reliability of automatic evaluation, we additionally conduct human validation on 120 sampled Gen-Arena instances; agreement statistics and human EGIP results are reported in Appendix D.

5.2 Main Results

We report main results in two parts. Table 1 evaluates methods on Gen-Arena with commitment-level metrics, while Table 2 summarizes results on external benchmarks.

Table 1 presents the quantitative comparison on Gen-Arena. SCOPE achieves a significant improvement in overall EGIP compared to direct generation baselines, surpassing Nano Banana Pro by 39 percentage points. The improvement is consistent across all six categories, with especially strong results on Sports and Ceremony. These categories frequently require identity grounding, event-specific relations, and precise scene composition, suggesting that SCOPE benefits from keeping retrieved evidence, inferred requirements, and verification feedback tied to the same structured specification. In contrast, direct generation baselines often fail under the strict pass criterion even when individual visual elements appear plausible, showing that stronger generation alone is insufficient for complete structured intent realization. Figure 4 compares generated images from baseline models and SCOPE on a representative Gen-Arena prompt.

As shown in Table 2, on WISE-V, SCOPE achieves the best overall WiScore of 0.907 and ranks first in five of the six reported categories, improving over Nano Banana Pro by 3.5% overall. On MindBench, SCOPE achieves 0.63 in Reasoning and 0.61 overall, improving over Nano Banana Pro by 48.8% in overall accuracy. These results further corroborate that SCOPE effectively maintains explicit commitments across the generation lifecycle, thereby realizing more faithful complex image generation. Additional diagnostic results are provided in Appendix C.

5.3 Ablation Study

Table 3 studies the contribution of the main SCOPE skills on Gen-Arena. Direct (single) obtains 0.21 EGIP, while Direct (best-of-3) rises to 0.40 by reporting the best official-evaluator score among three independent direct generations. Self-

Method	EGIP	Gated Cons.
Direct (single)	0.21	0.59
Direct (best-of-3)	0.40	0.71
Self-refine w/o spec	0.39	0.73
SCOPE w/o R&R	0.22	0.60
SCOPE w/o repair	0.42	0.72
SCOPE	**0.60**	**0.83**

Table 3: Ablation results on Gen-Arena. Direct uses the Nano Banana Pro backend with the original prompt; best-of-3 selects the best result among three independent generations using the official evaluator. Self-refine w/o spec uses the same three-generation budget but replaces structured commitments with free-form critique and prompt rewriting. R&R denotes retrieval and reasoning, and Gated Cons. denotes gated constraint pass rate.

refine w/o spec reaches only 0.39 EGIP despite using the same three-generation budget, suggesting that free-form critique and prompt rewriting do not reliably convert partial improvements into complete instance-level success. Within SCOPE, removing retrieval and reasoning skills drops EGIP to 0.22, close to Direct (single), indicating that structured decomposition alone is insufficient unless exposed commitments are semantically resolved. Removing only repair skills reaches 0.42 EGIP, showing that retrieval and reasoning provide substantial initial-generation gains by resolving underspecified commitments. Full SCOPE reaches 0.60 EGIP, improving over w/o repair by 18 percentage points. These results suggest that SCOPE benefits from coupling a persistent structured specification with complementary skills: retrieval and reasoning resolve underspecified commitments before generation, while verification-guided repair targets commitments that remain violated later in the generation lifecycle.

5.4 Analysis

We further analyze SCOPE by asking two questions about the generation lifecycle.

Q1: Is structured decomposition alone sufficient? The ablation results show that the structured specification should not be understood as a stronger prompt by itself. Its main function is to preserve addressable semantic units across stages. Without retrieval and reasoning, unresolved unknowns remain exposed in the specification; without repair, violations detected after generation cannot be targeted. Thus, the specification is useful because it provides a stable interface through which different skills can operate on the same commitments.

Q2: Why does SCOPE improve EGIP? Table 6 helps explain where SCOPE gains come from. Direct generation baselines pass many individual entity checks, but EGIP requires every required entity in an instance to be correct before its constraints can be credited. Thus, even an item-level Entity Pass of 0.82 for Nano Banana Pro still leaves many instances with at least one missing or incorrectly grounded entity, which then blocks dependent attributes, relations, or layouts. Qwen-Image and Z-Image-Turbo show a similar pattern: their Entity Pass exceeds 0.83, yet their EGIP remains near zero because entity errors and downstream constraint violations accumulate at the instance level. SCOPE improves EGIP by keeping entities and their associated constraints tied to the same structured specification across the generation lifecycle. This reduces cascading entity-to-constraint failures, raising Entity Pass to 0.92, Gated Constraint Pass to 0.83, and overall EGIP to 0.60.

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. Experiments show that SCOPE substantially outperforms all evaluated baselines on Gen-Arena and achieves strong results on external benchmarks, suggesting that keeping commitments identifiable across the generation lifecycle is a practical path toward more faithful complex image generation.

Limitations

While SCOPE improves commitment-level intent realization, it also has several limitations. First, SCOPE requires multiple calls to the MLLM, image generator, verifier, and optionally retrieval services across up to three iterations per instance, leading to higher inference-time cost and latency

than one-shot generation. This overhead could be reduced through adaptive early stopping or more selective skill invocation. Second, SCOPE depends on item-level verification to route subsequent repair actions. Verification errors can therefore propagate through the lifecycle: a false negative may trigger unnecessary repair of a correctly realized commitment, while a false positive may leave a genuine failure unresolved. Improving verifier calibration is therefore important for making verification-guided repair more reliable.

References

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3.pdf. OpenAI technical report.

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

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, and 1 others. 2025. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*.

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

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

Shuang Chen, Quanxin Shou, Hangting Chen, Yucheng Zhou, Kaituo Feng, Wenbo Hu, Yi-Fan Zhang, Yunlong Lin, Wenxuan Huang, Mingyang Song, and 1 others. 2026. Unify-Agent: A unified multimodal agent for world-grounded image synthesis. *arXiv preprint arXiv:2603.29620*.

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

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, and 1 others. 2025. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12606–12633.

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

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

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

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

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

Ruiyan Han, Zhen Fang, XinYu Sun, Yuchen Ma, Ziheng Wang, Yu Zeng, Zehui Chen, Lin Chen, Wenxuan Huang, Wei-Jie Xu, and 1 others. 2026. UniCorn: Towards self-improving unified multimodal models through self-generated supervision. *arXiv preprint arXiv:2601.03193*.

Jun He, Junyan Ye, Zilong Huang, Dongzhi Jiang, Chenjue Zhang, Leqi Zhu, Renrui Zhang, Xiang Zhang, and Weijia Li. 2026a. Mind-Brush: Integrating agentic cognitive search and reasoning into image generation. *arXiv preprint arXiv:2602.01756*.

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

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

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

Wenxuan Huang, Shuang Chen, Zheyong Xie, Shaosheng Cao, Shixiang Tang, Yufan Shen, Qingyu Yin, Wenbo Hu, Xiaoman Wang, Yuntian Tang, and 1 others. 2025. Interleaving reasoning for better text-to-image generation. *arXiv preprint arXiv:2509.06945*.

Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, and 1 others. 2026. Vision-DeepResearch: Incentivizing deepresearch capability in multimodal large language models. *arXiv preprint arXiv:2601.22060*.

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

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

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

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

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

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Arsh Koneru, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. 2025b. 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*, pages 15657–15668.

Xiangyi Li, Wenbo Chen, Yimin Liu, Shenghan Zheng, Xiaokun Chen, Yifeng He, Yubo Li, Bingran You, Haotian Shen, Jiankai Sun, and 1 others. 2026. SkillsBench: Benchmarking how well agent skills work across diverse tasks. *arXiv preprint arXiv:2602.12670*.

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

Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, and 1 others. 2025. WISE: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*.

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

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, and 1 others. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

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

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

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

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

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

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

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

Yu Zeng, Wenxuan Huang, Zhen Fang, Shuang Chen, Yufan Shen, Yishuo Cai, Xiaoman Wang, Zhenfei Yin, Lin Chen, Zehui Chen, and 1 others. 2026. Vision-DeepResearch Benchmark: Rethinking visual and textual search for multimodal large language models. *arXiv preprint arXiv:2602.02185*.

Boyuan Zheng, Michael Y. Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, and 1 others. 2025. SkillWeaver: Web agents can self-improve by discovering and honing skills. *arXiv preprint arXiv:2504.07079*.

YanZhao Zheng, ZhenTao Zhang, Chao Ma, Yuan-Qiang Yu, JiHuai Zhu, Yong Wu, Tianze Xu, Baohua Dong, Hangcheng Zhu, Ruohui Huang, and 1 others. 2026. SkillRouter: Skill routing for llm agents at scale. *arXiv preprint arXiv:2603.22455*.

Le Zhuo, Liangbing Zhao, Sayak Paul, Yue Liao, Renrui Zhang, Yi Xin, Peng Gao, Mohamed Elhoseiny, and Hongsheng Li. 2025. 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*, pages 15329–15339.

A SCOPE Generation and Repair Lifecycle

Algorithm 1 SCOPE Generation and Repair Lifecycle

Require: User prompt p, maximum attempts 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**
    $z_t \leftarrow \begin{cases} \textsc{AddUnknown}(z_t, i), & \text{if } i \text{ exposes an unknown} \\ \textsc{RepairSpec}(z_t, i), & \text{otherwise} \end{cases}$
  end for
  $(a_{t+1}, \rho_{t+1}) \leftarrow \textsc{PlanRepair}(F_t, z_t, y_t)$
end for
return y^*

B Gen-Arena Details

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 a type, attribute, relation, or layout, and is linked to the entity ids on which it depends. This dependency annotation supports the entity-first evaluation rule used by EGIP.

B.2 Benchmark Statistics

Table 4 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.

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

B.3 Constraint-Type Distribution

Table 5 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.

Category	Attr.	Rel.	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

Table 5: Constraint-type distribution in Gen-Arena. Attr., Rel., and Layout denote attribute, relation, and layout constraints, respectively.

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 produces item-level binary judgments for required entities and constraints. 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 Additional Gen-Arena Results

Building on the aggregate diagnostic results in Table 6, Table 7 breaks down diagnostic results by category for Nano Banana Pro and SCOPE. The gains are consistent across all six categories, suggesting that the improvement is not driven by a single retrieval-heavy category but by more reliable gated constraint fulfillment across diverse visual intents.

Method	EGIP	Entity Pass	Gated Constraint Pass
Nano Banana (Google DeepMind, 2025b)	0.07	0.78	0.49
Nano Banana Pro (Google DeepMind, 2025a)	0.21	0.82	0.59
Qwen-Image (Wu et al., 2025a)	0.02	0.83	0.49
Z-Image-Turbo (Cai et al., 2025)	0.01	0.84	0.48
PixArt-Sigma (Chen et al., 2024)	0.00	0.67	0.31
FLUX.1-dev (Black Forest Labs, 2024)	0.01	0.78	0.42
SD-3.5-large (Esser et al., 2024)	0.00	0.76	0.39
SDXL (Podell et al., 2023)	0.00	0.56	0.23
Janus-Pro-7B (Chen et al., 2025b)	0.00	0.78	0.40
SCOPE	0.60	0.92	0.83

Table 6: Diagnostic Gen-Arena results. Entity Pass measures item-level entity satisfaction, while Gated Constraint Pass measures constraint satisfaction after applying entity prerequisites.

Category	NBP		SCOPE	
	EGIP	Gated	EGIP	Gated
Cartoon	0.04	0.64	0.52	0.85
Game	0.00	0.27	0.46	0.76
Sports	0.34	0.76	0.72	0.91
Entert.	0.16	0.47	0.62	0.79
Comp.	0.18	0.66	0.52	0.80
Ceremony	0.54	0.81	0.74	0.89

Table 7: Category-wise diagnostic results on Gen-Arena. NBP denotes Nano Banana Pro, and Gated denotes gated constraint pass rate.

D Human Validation of Gen-Arena Evaluation

We sample 120 Gen-Arena instances, 20 from each category, to validate the reliability of the automatic evaluator. Human annotators judge required entities and constraints using the same structured evaluation specification as EGIP. Table 8 reports both inter-annotator agreement and human–Gemini agreement, while Table 9 shows the resulting system ranking on the same subset.

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

Table 8: Human validation of the Gen-Arena automatic evaluator on 120 sampled instances. Agreement is computed over item-level entity and constraint judgments, while EGIP agreement indicates whether the final instance-level pass decision is consistent.

Method	Gemini EGIP	Human EGIP
Nano Banana Pro	0.24	0.23
SCOPE	**0.60**	**0.58**

Table 9: System ranking under automatic and human EGIP on the human-validation subset.

E Skill Instruction Templates

We provide the instruction templates used by SCOPE to invoke external skills during the generation lifecycle. Figure 6 summarizes the condensed templates for the main orchestration, retrieval, reasoning, and repair skills. The templates define when each skill is triggered, what operation it performs, and how its output is written back to the structured specification.

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.
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.
Constraints are interpreted with respect to their prerequisite entities. If a prerequisite entity is missing or incorrectly realized, its dependent constraint is not credited.

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`.

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.

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, verification trace, and evaluation results.

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 when verification reports failed or uncertain items and no new unknowns remain.
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.

F Representative Case Visualizations

We provide six representative SCOPE traces across Gen-Arena categories and failure types. Each visualization shows the prompt, structured commitments, initial output, verification feedback, repair action, and final repaired output. These examples illustrate how SCOPE maps a visual failure back to the corresponding entity or constraint before invoking a targeted repair.

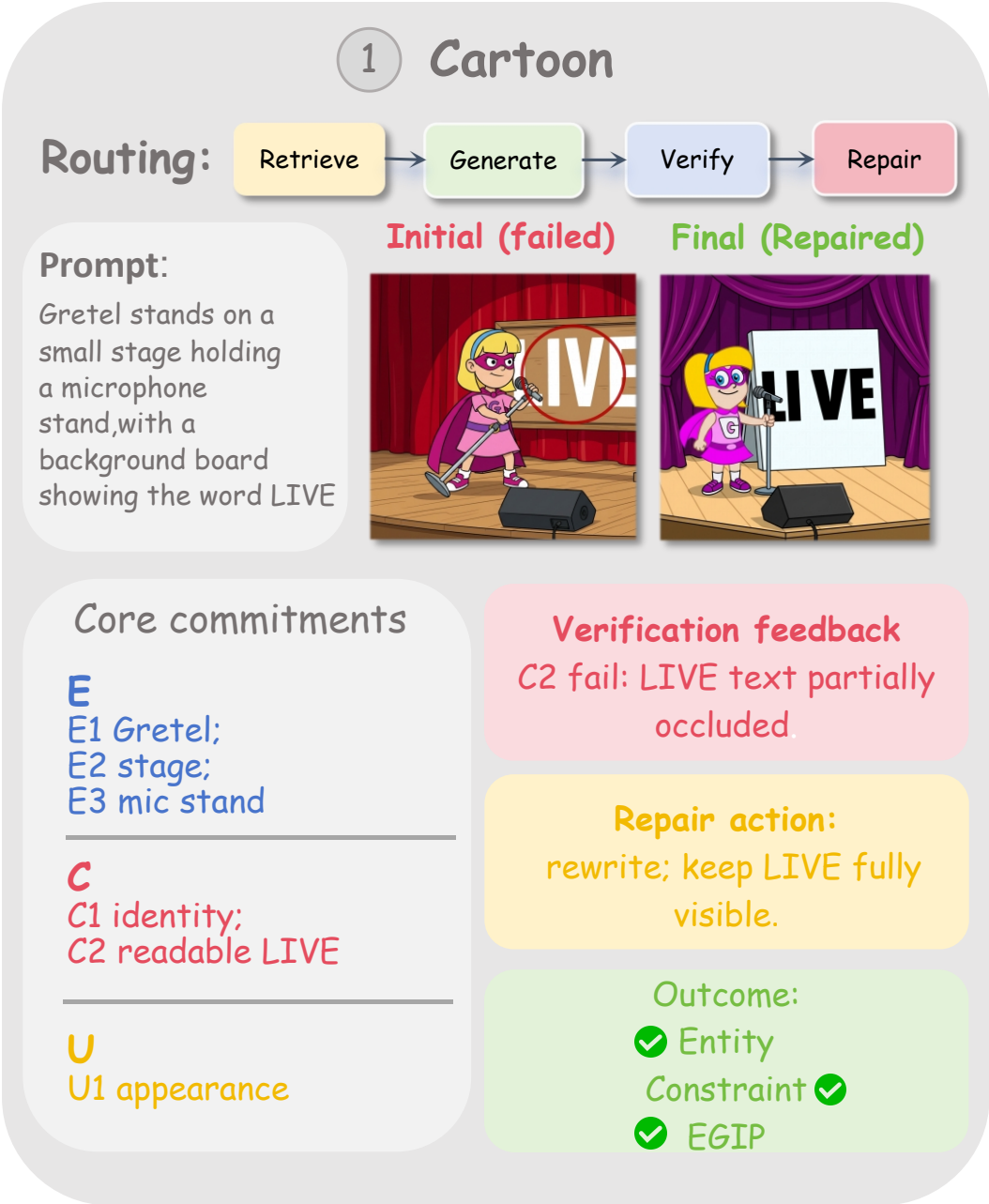

Figure 7: **Visualization result for a cartoon-character case.** The initial image captures the stage setting but partially occludes the required word *LIVE*. The repair keeps the board fully visible while preserving the target entity, stage, and microphone stand.

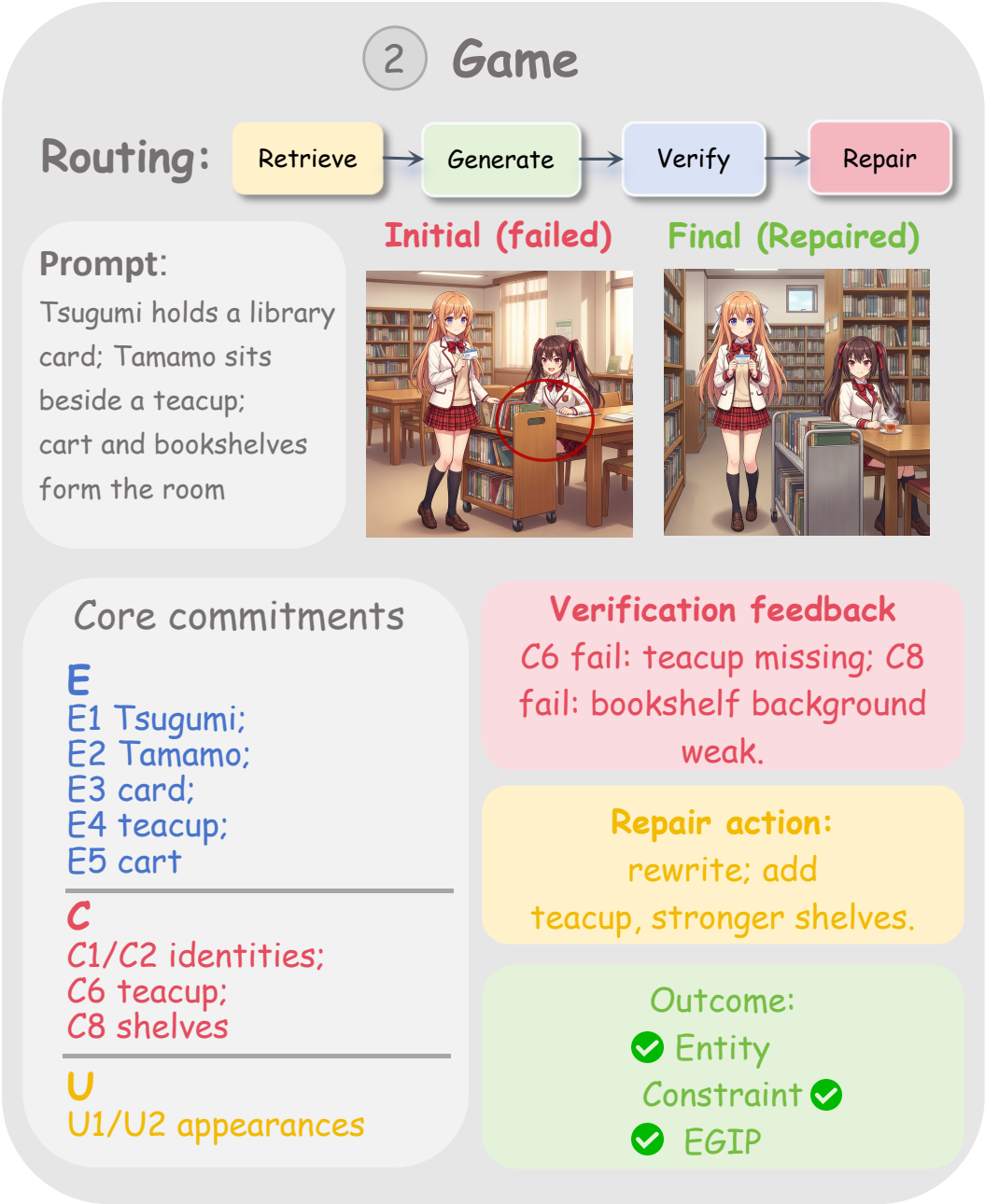

Figure 8: **Visualization result for a game-character case.** The initial output weakens the target identity or omits a signature visual attribute. SCOPE uses retrieved evidence to restate the identity commitment and repairs the missing entity-attribute constraint.

Figure 9: **Visualization result for a sports-relation case.** The initial image fails to clearly realize the overhead racket and shuttlecock constraints during the jump. SCOPE localizes the failure to the action-related constraints and rewrites the prompt to enforce the overhead pose and shuttle placement.

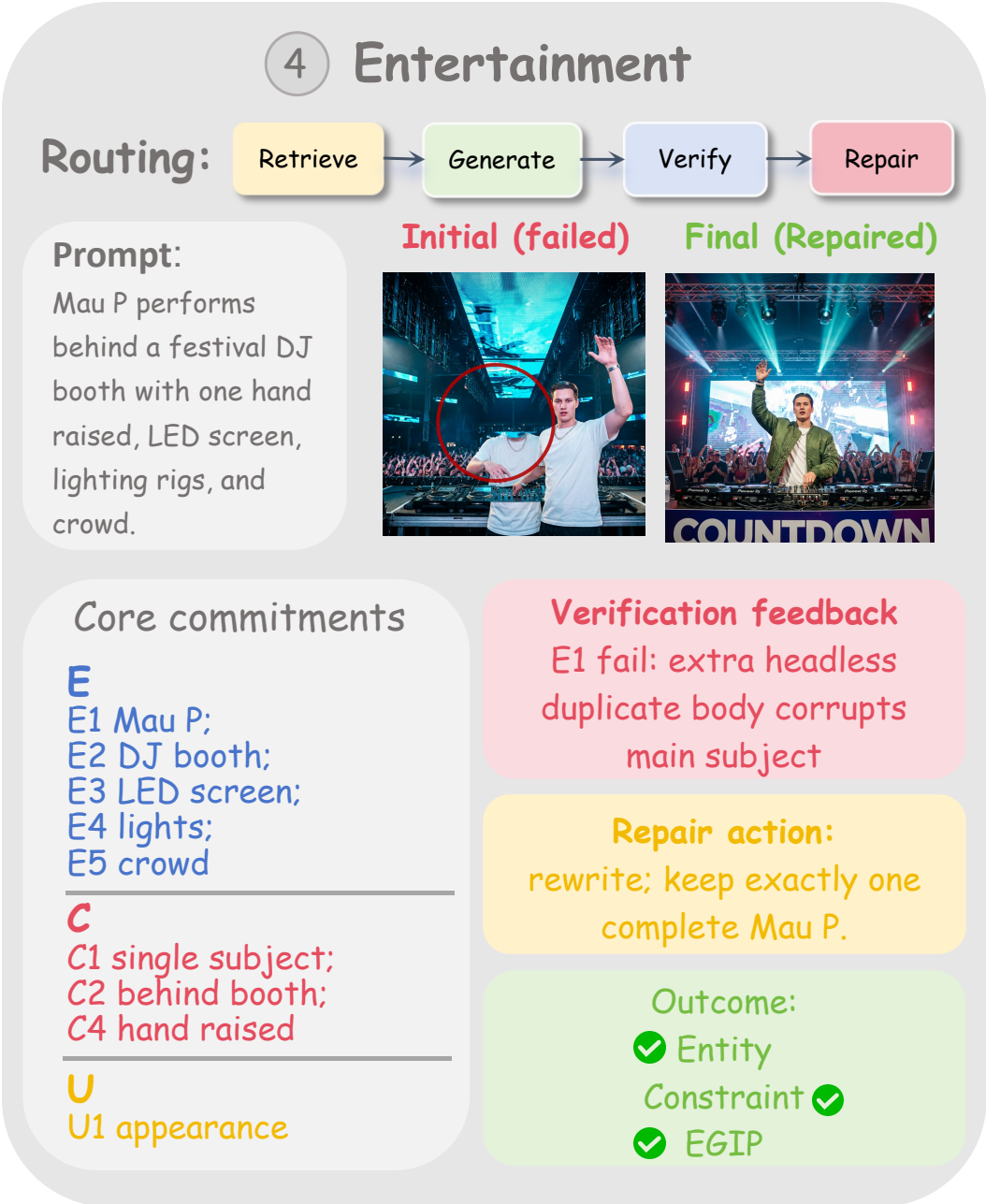

Figure 10: **Visualization result for an entertainment-action case.** The initial image introduces an extra corrupted duplicate body, violating the single-subject and identity commitments. SCOPE maps the failure to the entity-level commitment and repairs the scene by enforcing exactly one complete performer.

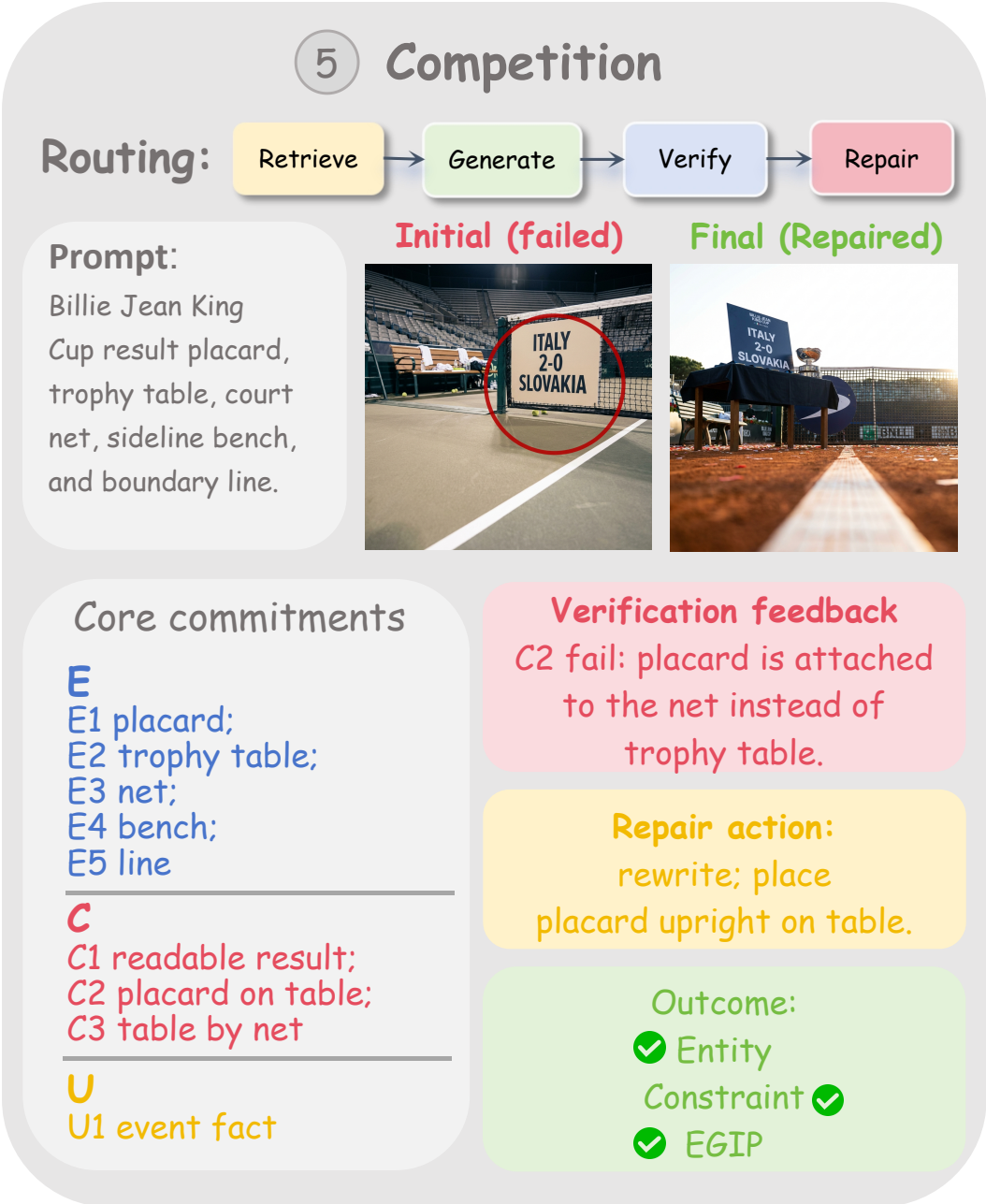

Figure 11: **Visualization result for a competition/event case.** The initial image places the result placard on the net rather than upright on the trophy table. SCOPE localizes the failure to the placard-table layout constraint and repairs the scene while preserving the event setting.

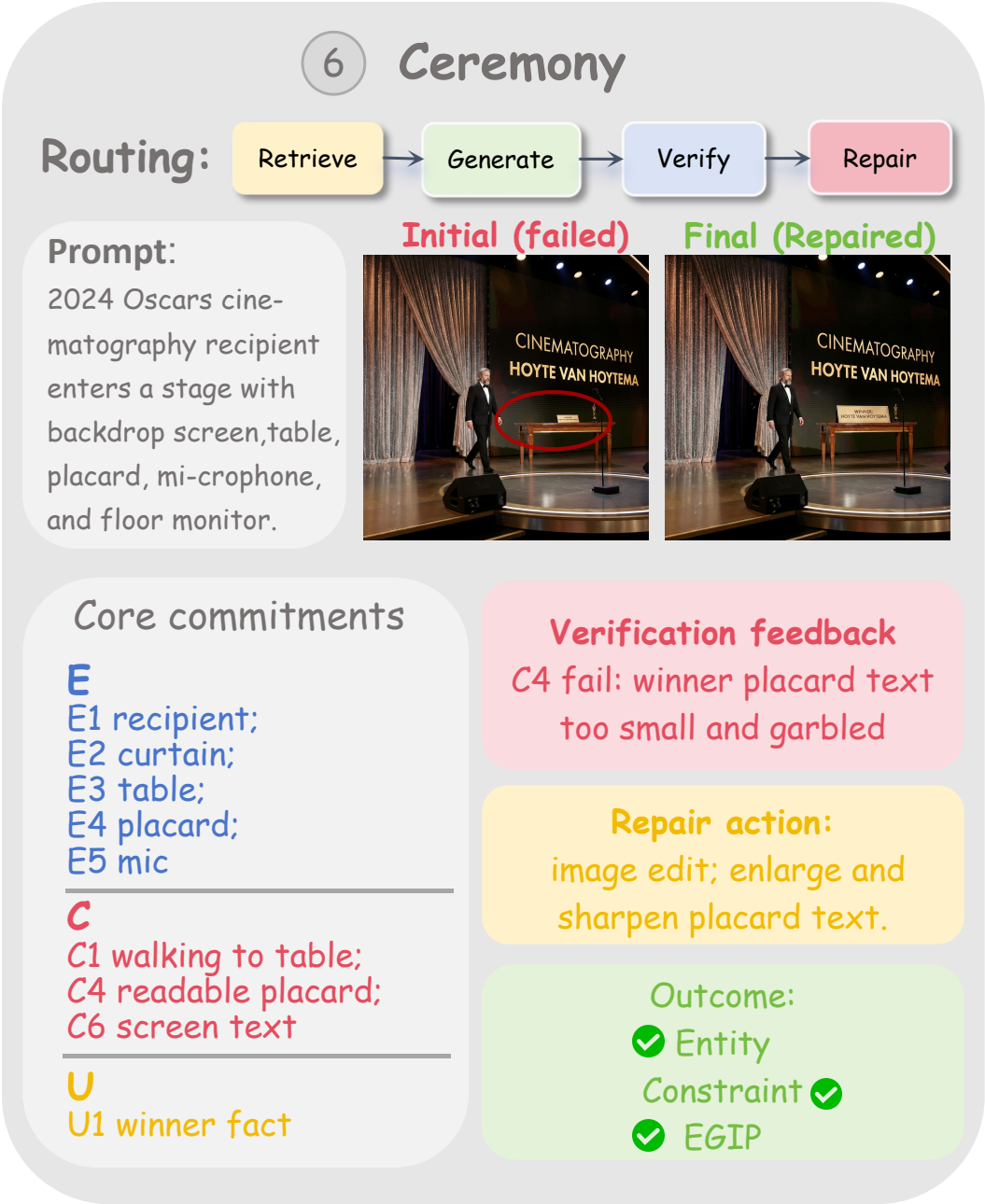

Figure 12: **Visualization result for an award-ceremony case.** The initial image contains the required ceremony setting but renders the winner placard too small and garbled. SCOPE uses an image-edit repair to enlarge and sharpen the placard text while preserving the stage layout.