R3: RUBRIC-GUIDED REGION REWARD FOR PRECISE TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern text-to-image (T2I) models support complex, structured conditions involving multiple entities, attributes, text, and spatial layouts, yet existing image rewards remain insufficiently granular. Global rewards yield a single score for the entire image, and recent rubric-based rewards improve semantic granularity but judge each criterion against the whole image. We introduce **R3** (**R**ubric-guided **R**egion **R**eward), which reformulates image evaluation as an explicit target–evidence matching problem. R3 first structures the generation intent into entity-bound, coarse-to-fine rubrics and unambiguous grounding labels; then independently records what is present in the image through visual observations; finally, it matches visual evidence with the target specification while explicitly accounting for the spatial consistency of the observed targets. This design turns the reward estimation from direct holistic judgment into grounded verification: structured intent in, grounded credit out. We also introduce **GranuBench**, featuring high-quality contemporary generations, complex structured specifications, and explicit spatial targets to systematically evaluate reward granularity. On GranuBench, R3-based Gemini reaches a preference accuracy of 90.8%, outperforming direct-Gemini (80.7%) and existing reward models by a substantial margin. We further construct **R3Traj-100K**, containing 100K high-quality multi-stage R3 supervision trajectories, each recording structured target specifications, localized visual observations, and target–evidence assessments. Using R3Traj-100K, we distill the Gemini-instantiated R3 pipeline into an open 27B VLM with stage-specialized adapters, achieving state-of-the-art performance on multiple reward benchmarks. R3 can also improve image quality at test time by enabling precise region-targeted refinement. Together, these results demonstrate that finer reward granularity benefits not only evaluation but also inference- and training-time optimization.

1 INTRODUCTION

Text-to-image (T2I) generation has rapidly evolved from short-caption synthesis to following detailed and compositional instructions. Modern systems can render multiple entities, bind attributes, synthesize literal text, and reason over relations (Google DeepMind, 2025a;b; 2026; OpenAI, 2025a;b; 2026b; Black Forest Labs, 2025; Krea, 2026; Alibaba, 2025), while recent models further support richer conditioning interfaces for precise spatial and compositional control (Ideogram, 2026; Chen et al., 2026). Yet as generation becomes increasingly structured, image reward remains comparatively coarse. Classical reward models like ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), and HPS (Wu et al., 2023; Ma et al., 2025) compress a prompt–image pair into a single score learned from pairwise or ranking annotations. While effective for ranking and optimization, such scalar rewards reveal little about which requested conditions succeed or fail. Recent work therefore moves toward decomposed reward signals: VisionReward (Xu et al., 2026) evaluates multiple preference dimensions, while rubric-based methods (Feng et al., 2026; Kao et al., 2026; Tian et al., 2026; Liu et al., 2026; Wang et al., 2025; Wei et al., 2025) construct or learn explicit criteria and assess them separately with multimodal judges. These approaches improve the *semantic granularity* of reward through decomposition: instead of asking whether an image is good as a whole, they ask whether individual objects, attributes, relations, text, or quality requirements are satisfied.

However, semantic decomposition solves only half of the granularity problem. In most rubric-based evaluation, each fine-grained criterion is still judged against the *entire image*: the evaluator

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.

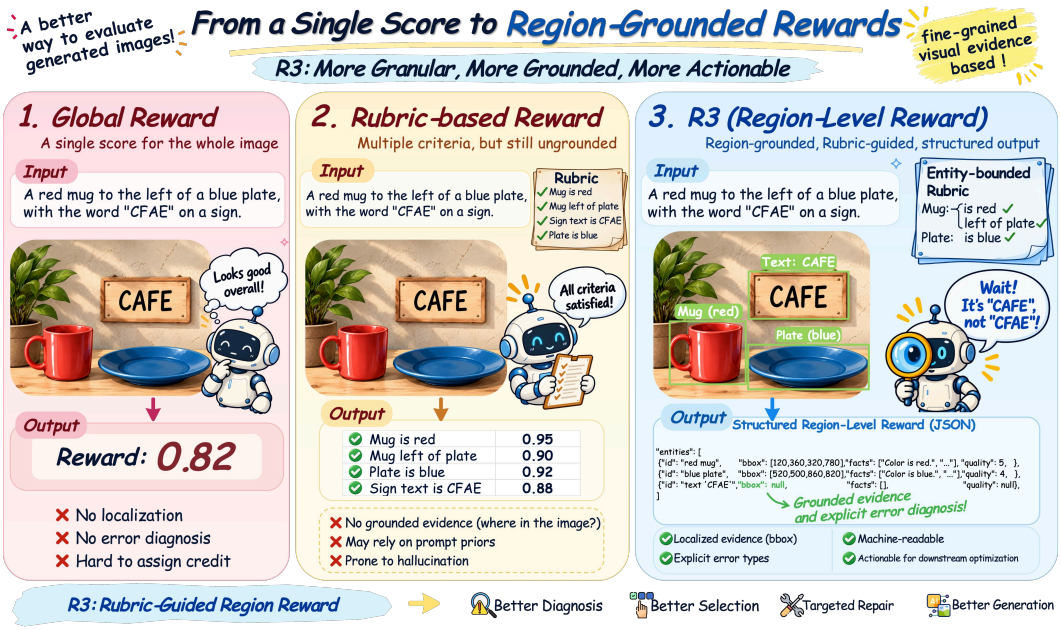


Figure 1: Global reward compresses the entire image into a single score, while rubric-based reward improves semantic granularity by decomposing the evaluation into multiple criteria but still judges each criterion against the whole image. This creates a semantic–visual granularity mismatch: the evaluator knows *what* to judge, but not necessarily *where* the supporting evidence lies. R3 resolves this by binding entity-level rubrics to localized visual evidence, producing structured and grounded rewards for explicit error diagnosis, targeted refinement, and downstream generation optimization.

knows *what* to check, but need not expose *where* the supporting evidence lies. We refer to this as a **semantic–visual granularity mismatch**, where textual supervision becomes fine-grained while visual evidence remains largely image-level and ungrounded in practice. Several prior works have explored more localized visual reward, but only in restricted forms. GenEval (Ghosh et al., 2023) introduces object-centric checks, but such detector-based evaluation remains limited for the diverse and complex conditions supported by modern T2I systems. RichHF (Liang et al., 2024) collects localized human feedback for visual artifacts and text–image misalignment, but its localized error maps are not designed to verify arbitrary entity-bound generation requirements or produce a general-purpose reward. Meanwhile, recent multimodal reward models report that VLM judges may rely on language priors rather than faithfully verifying fine-grained visual evidence (Liu et al., 2026). We therefore argue that **reward granularity should extend from language to vision**: a reward model should not only know *what to judge*, but also expose *what visual evidence its judgment is grounded on*. As illustrated in Fig. 1, R3 explicitly associates structured generation requirements with localized visual evidence, preserving both the semantic identity of each requirement and the region from which its credit is derived. Our goal is: *structured intent in, grounded credit out*.

Grounding a reward, however, requires more than attaching a bounding box to a VLM judgment. When a VLM sees the image together with the target prompt or rubric, it often over-relies on language priors and incorrectly confirms requested conditions despite contradictory visual evidence, leading to hallucinated judgments. We thus formulate region reward as explicit **target–evidence matching**: first specify the target, then independently observe the image, and finally compare the observation against the target. Accordingly, we introduce **R3** (**R**ubric-guided **R**egion **R**eward), a three-stage pipeline summarized in Fig. 2. **(1) Structure the Target.** Given a natural-language or structured prompt, R3 constructs a specification with minimally ambiguous grounding labels, explicit target regions, and entity-bound coarse-to-fine rubrics, preserving each entity's semantic and spatial requirements. **(2) Ground the Evidence.** R3 receives only the image and grounding labels, and grounds what is actually present. For each entity, it records a localized observation including its grounding box, visible attributes, and generation quality, without access to target rubrics or regions. **(3) Assess by Matching Target and Evidence.** R3 freezes the observations and matches them against the target specification. Semantic criteria are assessed solely from observed evidence

in text space, avoiding direct image access and reducing hallucination, while spatial consistency is measured between target and observed boxes. A deterministic aggregation produces entity-level rewards and the final image reward. We ablate all the design choices for all stages in Sec. A.

A region-level reward also calls for a benchmark with comparable granularity. Existing reward benchmarks substantially lags behind modern T2I conditioning. Early benchmarks such as ImageReward (Xu et al., 2023), Pick-a-Pic (Kirstain et al., 2023), and HPD v2 (Wu et al., 2023) are built around short natural-language prompts and image-level preferences, while newer benchmarks such as MMRB2 (Hu et al., 2026) improve image generation quality and prompt complexity with expert-annotated pairs. However, they still indicate mainly *which* image is preferred, without exposing the structured entity-level semantics and spatial targets underlying that preference. Meanwhile, modern T2I systems increasingly support structured generation interfaces, where entities, attributes, text, style, and layouts can be explicitly specified, even through JSON prompts with per-element bounding boxes (Ideogram, 2026; Chen et al., 2026),which creates a growing *specification gap* between modern T2I generation conditions and reward supervision. We therefore introduce **GranuBench**, which pairs high-quality contemporary generations with complex structured specifications, bounding boxes, and fine-grained entity- and rubric-level annotations. On its 1,000 fully annotated preference pairs, **R3-based Gemini** achieves **90.8%** accuracy versus **80.7%** for direct Gemini judging, showing that *reward formulation* remains a major bottleneck even for frontier VLMs.

Gemini-based R3 provides strong instantiation of the formulation, but closed-source inference limits reproducibility and deployment. We therefore construct **R3Traj-100K**, containing 100K complete R3 trajectories that record the outputs of all pipeline stages, and distill the pipeline into an open 27B VLM (Qwen, 2026). To accommodate the heterogeneous capabilities required across stages, we use lightweight stage-specific adapters (Hu et al., 2021) over a shared backbone. Open R3 consistently outperforms conventional reward models on GranuBench and public human-preference benchmarks.

Finally, we study whether finer reward granularity translates better evaluation into better generation at inference time. R3 enables two complementary self-reflection strategies: revising the structured specification based on localized errors and regenerating the image, or converting erroneous regions into box-grounded editing instructions. We evaluate both strategies across contemporary T2I systems and find that region-grounded feedback enables targeted corrections while better preserving already-correct content. Together with its strong performance on GranuBench and public human-preference benchmarks, these results show that finer reward granularity benefits not only image evaluation, but also actionable test-time refinement. Our contributions are summarized as follows:

- We identify the **semantic–visual granularity mismatch** in existing T2I rewards and introduce **R3**, a three-stage target–evidence matching framework that structures generation intent, independently grounds visual observations, and explicitly matches semantic and spatial targets, extending reward granularity from language to localized visual evidence.
- We introduce **GranuBench**, a 1,000-pair modern T2I reward benchmark with structured specifications, explicit target regions, and fine-grained entity-level annotations, together with **R3Traj-100K** for distilling the Gemini-instantiated R3 pipeline into **Open R3**.
- We extend fine-grained reward from evaluation to generation: accurate, region-grounded reward signals enable T2I models to precisely improve generation quality at inference time.

2 Related Work

2.1 From Holistic Preference to Semantic Reward Decomposition

Early T2I reward models such as ImageReward Xu et al. (2023), PickScore Kirstain et al. (2023), and HPS Wu et al. (2023); Ma et al. (2025) learn holistic image–text preference scores from human comparisons, providing effective but largely monolithic supervision for ranking and optimization. With stronger multimodal evaluators, recent work moves toward more structured judgment. UnifiedReward Wang et al. (2025) and UnifiedReward-Think Wang et al. (2026) extend reward modeling with generative scoring and explicit reasoning, while VisionReward Xu et al. (2026) decomposes preference into interpretable dimensions. Rubric-based methods further make evaluation criteria explicit: RubricRL Feng et al. (2026) constructs prompt-specific rubrics, AutoRubric-T2I Kao et al. (2026) and Auto-Rubric as Reward Tian et al. (2026) derive adaptive criteria, and DeltaRubric Liu



Figure 2: R3 separates target specification, visual observation, and final assessment. **(1) Structure the Target:** R3 constructs entity-bound rubrics, grounding queries, and target boxes. **(2) Ground the Evidence:** R3 independently records each entity's observed box, visible attributes, and generation quality from the image without seeing prompt. **(3) Match Target and Evidence:** R3 matches the observations against the target specification in semantic and spatial space to give final rewards.

et al. (2026) frames reward modeling as structured planning and verification. These approaches substantially improve reward granularity in the *semantic space*, but the visual evidence supporting each criterion is still typically not represented as an explicit localized component of the reward. R3 extends this progression by coupling semantic decomposition with region-grounded evidence, associating fine-grained criteria with the visual observations from which their credit is derived.

2.2 From Localized Visual Feedback to Region-Grounded Verification

Several works explore localized supervision in the visual space. RichHF Liang et al. (2024) augments scalar human feedback with localized regions of implausibility and text–image misalignment, while HEIE Yang et al. (2025) combines defect localization with scalar scores and natural-language explanations. Focus-N-Fix Xing et al. (2025) further uses localized feedback to guide targeted T2I fine-tuning. Together, these studies demonstrate the value of localized signals for diagnosing and correcting visual errors. SpatialReward Zhou et al. (2026) introduces a verifiable reward specialized for spatial consistency using grounded visual evidence. R3 instead targets general entity-bound generation requirements and explicitly separates *target specification* from *visual observation*. A target-blind observer first records localized evidence, which is then matched against target-aware rubrics and spatial constraints, enabling structured semantic and spatial credit assignment.

3 R3: Rubric-Guided Region Reward

As illustrated in Fig. 2, R3 follows a three-stage pipeline: **structure the target**, **ground the evidence**, and **match target and evidence**. We first formalize the overall reward formulation in Sec. 3.1, then describe the target construction in Sec. 3.2, the target-blind visual observation in Sec. 3.3, and structured semantic, quality, and spatial matching in Sec. 3.4. Finally, Sec. 3.5 introduces the distillation of R3 into an open VLM model while preserving staged information flow.

3.1 Overview and Problem Formulation

Let $\mathcal{T}$ denote a generation target and I a candidate image. R3 represents $\mathcal{T}$ as a structural specification describing the requested entities, their attributes and relations, and, when available, their intended spatial regions. Instead of directly mapping $(\mathcal{T}, I)$ to a scalar score, R3 decomposes reward estimation into three stages: target construction, visual observation, and target–evidence matching.

From $\mathcal{T}$, R3 then constructs a fixed set of N entity-level target units:

$$\mathcal{S} = \{S_i\}_{i=1}^{N}, \qquad S_i = \left(e_i, q_i, \mathcal{C}_i, b_i^t\right), \tag{1}$$

where e_i is an entity identifier, q_i is a minimally disambiguated grounding query, $\mathcal{C}_i = \{c_{ij}\}_{j=1}^{M_i}$ is the corresponding set of semantic rubrics, and b_i^t is the target bounding box specifying the intended spatial extent of the entity when available. The grounding query is used only to locate the entity, while semantic requirements and spatial requirements are represented by $\mathcal{C}_i$ and the target box b_i^t. The same target contract $\mathcal{S}$ is reused for all candidate images generated from the same request. For each candidate image, R3 then independently constructs a frozen observation for every entity:

$$O_i = f_{\text{obs}}(I, q_i) = \left(z_i, \hat{b}_i, A_i, Q_i\right), \tag{2}$$

where z_i indicates whether the entity is found, $\hat{b}_i$ is the observed bounding box grounded from the candidate image, A_i records visible image evidence, and Q_i measures the visual quality of the entity. The observer does not access the semantic rubrics, target bounding box b_i^t, or full target specification, so visual evidence is collected independently of the criteria used for later assessment.

Finally, R3 matches the frozen observation O_i against the target unit S_i. Semantic rubrics are assessed from the frozen observation, visual integrity is derived from Q_i, and spatial consistency is computed from b_i^t and $\hat{b}_i$ using SAGE metric described in Sec. 3.4. These signals are combined into entity-level rewards r_i and deterministically aggregated into the final image reward $R_{\text{R3}}(\mathcal{T}, I)$.

3.2 Structure the Target

R3 first normalizes the generation request into a **structural prompt (SP)** (Ideogram, 2026; Chen et al., 2026), represented as a JSON-style specification. The SP describes the global scene and style while representing each target entity with its textual description, attributes, relations, literal text when applicable, and intended bounding box. R3 supports both natural-language requests and structured prompts in other formats, which are normalized into the same canonical SP representation.

Given the SP, the target-structuring stage predicts a fixed, candidate-independent target contract. For each target entity e_i, it produces a minimally disambiguated grounding query q_i, an entity-bound rubric set $\mathcal{C}_i = \{c_{ij}\}_{j=1}^{M_i}$, and, when warranted by the input request, a target bounding box b_i^t (Ideogram, 2026; Chen et al., 2026). The grounding query serves only as a visual lookup identifier and contains the minimum information required to uniquely localize the entity, while semantic requirements remain in the rubric set and spatial requirements are represented separately by b_i^t. The semantic rubrics cover requirements such as identity, count, attributes, parts, text, and relations, and are organized into 3 coarse-to-fine levels. **L1** captures defining requirements such as identity, existence, key counts, and exact short text; **L2** covers major attributes, components, materials, colors, and relations; and **L3** captures finer local details such as textures, decorations, and secondary parts.

Neither absolute spatial position nor generic generation defects are encoded in the semantic rubric set $\mathcal{C}_i = \{c_{ij}\}_{j=1}^{M_i}$. Absolute position is represented by the target box b_i^t, while generic generation defects such as malformed anatomy or rendering artifacts are captured later in the visual observation O_i. The resulting target contract is frozen and reused for all candidate images from the same request.

3.3 Ground the Evidence

After structuring the target, R3 independently localizes each target entity and records what is actually present in the candidate image. For each entity e_i with grounding query q_i, R3 receives only the candidate image I and q_i, without access to the full SP, semantic rubrics $\mathcal{C}_i$, nor target box b_i^t. This information isolation prevents requested attributes or locations from directly biasing the visual observation, thereby reducing target-conditioned hallucination. For each entity, R3 produces:

$$O_i = f_{\text{obs}}(I, q_i) = \left(z_i, \hat{b}_i, A_i, Q_i\right), \tag{3}$$

where z_i indicates whether the entity is found, $\hat{b}_i$ is its observed bounding box, A_i records visible image-grounded facts, and Q_i denotes the eneity's visual integrity. The attribute set A_i describes directly observable properties such as identity, count, color, material, shape, parts, and relative configuration. For text entities, R3 explicitly records the visible text rather than assuming that the requested string is present. R3 records this evidence descriptively without judging whether target

rubrics are satisfied. Meanwhile, Q_i provides a 1–5 quality score together with visible generation defects, capturing generic realization failures separately from prompt-specific semantics.

Once generated, O_i is frozen and becomes the sole visual evidence used for subsequent semantic assessment. The next stage therefore strictly judges each rubric solely from $(O_i, \mathcal{C}_i)$ rather than directly from the image, implementing R3's explicit "observe first, judge later" information flow.

3.4 Match Target and Evidence

Given the frozen target contract and visual observations, R3 performs target–evidence matching independently for each entity. For each entity e_i, it evaluates semantic fidelity by matching O_i against the semantic rubrics $\mathcal{C}_i$, visual integrity from Q_i, and spatial concordance between the target box b_i^t and observed box $\hat{b}_i$. These three complementary signals are combined into a region reward r_i, after which all region rewards are deterministically aggregated into the final image reward R_{R3}.

For semantic matching, R3 operates entirely in text space using only the frozen observation O_i and entity-bound rubric set $\mathcal{C}_i$. Each rubric is independently judged with a binary yes/no decision and supporting evidence drawn from O_i. The judgment is strictly evidence-bounded: properties not established by the frozen observation are not inferred from the target specification or external knowledge. These rubric-level decisions are then aggregated into the semantic score of entity e_i. Visual integrity is taken directly from Q_i, capturing generic generation defects independently of semantic satisfaction. Spatial concordance is computed between target b_i^t and observed $\hat{b}_i$ using **SAGE** (Scale-Adaptive Gaussian Entity-Box Concordance) when b_i^t is available; otherwise, the spatial reward is omitted. SAGE measures agreement through a Gaussian similarity over normalized center displacement and relative box scale, assigning higher scores to boxes closer in position and extent. Unlike IoU, SAGE remains continuous even when the boxes do not overlap, a common case in T2I generation. The resulting image reward is a continuous scalar signal usable for image ranking, test-time refinement, and reward-based optimization without pairwise comparison. The exact SAGE formulation, aggregation rule, and ablations are provided in Appendix A.

3.5 Open R3 via Multi-Stage Distillation

While R3 can be instantiated with frontier closed-source models, we further distill the complete pipeline into an open reward model. We construct **R3Traj-100K**, containing 100K complete R3 trajectories that preserve intermediate supervision across structural prompting, rubric construction, visual observation, and target–evidence assessment, rather than only the final scalar reward. Details of trajectory construction are provided in Appendix B. Using R3Traj-100K, we build **Open R3** on a shared Qwen3.8-27B (Qwen, 2026) backbone with four rank-128 LoRA (Hu et al., 2021) adapters, each trained with standard supervised fine-tuning (SFT) for its corresponding stage. At inference time, the backbone and adapters are loaded once and activated according to the R3 pipeline stage.

4 GranuBench: A Fine-Grained Reward Benchmark

As modern T2I systems support increasingly fine-grained and structured text conditions (Chen et al., 2026; Ideogram, 2026; ByteDance Seed, 2026), reward models require supervision at comparable semantic and spatial granularity. However, existing reward benchmarks, from ImageReward (Xu et al., 2023), Pick-a-Pic (Kirstain et al., 2023), and HPDv2 (Wu et al., 2023) to recent GenAI-Bench (Li et al., 2024) and MMRB2 (Hu et al., 2026), still primarily evaluate image-level human preference. They reveal which image is preferred, but provide limited support for evaluating entity-level semantics, spatial targets, and grounded credit assignment. To address this gap, we construct **GranuBench** (**Granularity Benchmark**), a 1,000-case preference benchmark designed around the capabilities and failure modes of modern T2I systems, as illustrated in Fig. 3. Its cases are curated from challenging real-world generation requests and rendered by multiple state-of-the-art T2I models. Beyond pairwise preference labels, each case provides the original natural-language prompt, a machine-readable structural prompt (SP), entity-level grounding queries and coarse-to-fine rubrics, together with explicit spatial targets and region-grounded annotations. To our knowledge, GranuBench is the first T2I reward benchmark to combine structured prompts with region-grounded, entity-level reward annotations. This design yields challenging high-quality image pairs while preserving the fine-grained semantic and spatial supervision required to evaluate modern reward models. The complete construction and annotation protocol is provided in Appendix C.

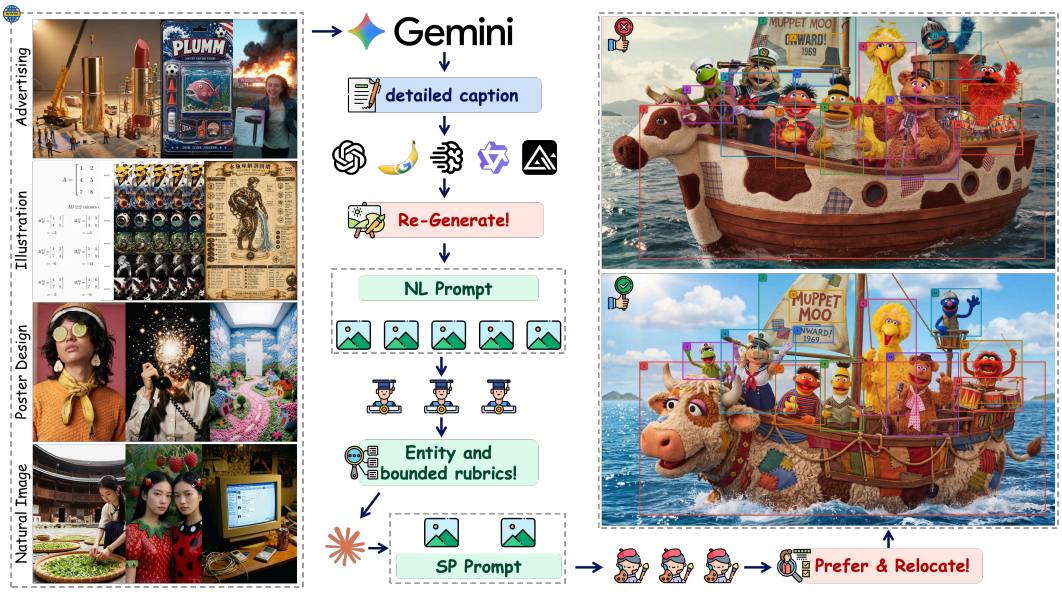


Figure 3: We curate challenging real-world T2I requests from open communities and use Gemini-3.7-Flash (Google, 2026) to produce detailed captions. These captions are used to generate diverse candidates with multiple modern T2I models (OpenAI, 2026b; Google DeepMind, 2026; Ideogram, 2026; Alibaba, 2025; Black Forest Labs, 2025), after which annotators identify the evaluation entities and construct entity-bound rubrics. The resulting annotations are converted into structured prompts using Magic Prompt LLM (Ideogram, 2026). Professional annotators with strong photographic and aesthetic expertise then select preferred candidate, considering semantic fidelity, spatial layout, and visual quality. The resulting benchmark provides pairwise human preferences together with structured, entity-level semantic and spatial annotations for fine-grained reward evaluation.

5 Experiments

We evaluate open R3 from two complementary perspectives. First, Sec. 5.1 evaluates reward-model accuracy on **GranuBench**, the MMRB2 T2I subset (Hu et al., 2026), GenAI-Bench (Li et al., 2024), and HPDv3 (Ma et al., 2025), against representative pointwise rewards and pairwise judges including ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), HPSv2/v3 (Wu et al., 2023; Ma et al., 2025), CLIPScore (Hessel et al., 2021), VQAScore (Lin et al., 2024), UnifiedReward-2.0 (Wang et al., 2025), UnifiedReward-Think (Wang et al., 2026), as well as strong closed-source general-purpose VLMs including GPT-5.6-Luna (OpenAI, 2026a) and Gemini-3.7-Flash (Google, 2026). Second, Sec. 5.2 examines whether R3's fine-grained feedback is actionable at inference time, using it for test-time self-reflection to improve T2I generations without updating the generator.

5.1 Reward Model Evaluation

Tab. 1 reports preference accuracy on GranuBench. Under its entity, attribute, and layout constraints, conventional pointwise reward models remain below the strongest direct pairwise judge. For closed-source pairwise judges, we provide the same frozen canonical SP used by R3 together with the two candidate images, ensuring access to the same structured generation specification. The component-wise results show that neither semantic assessment nor spatial matching alone is sufficient. Removing SAGE substantially weakens R3, while using SAGE alone performs poorly, indicating that geometry by itself cannot explain the benchmark performance. In contrast, the complete R3 formulation consistently achieves the strongest results by combining grounded semantic assessment with explicit spatial verification, which suggests that fine-grained T2I evaluation requires jointly determining *what* is rendered and *where* it is rendered, rather than relying on either signal alone.

R3 also produce no ties with or without SAGE, yielding a more discriminative scalar signal than several conventional rewards. Such behavior is desirable for reward-based optimization, where tied or collapsed scores provide little learning signal when candidates differ only in fine-grained local details. Beyond GranuBench, Open R3 generalizes well to public human-preference benchmarks,

Table 1: **Reward-model evaluation on GranuBench.** We report preference accuracy (%) across the four categories and overall. Pairwise methods jointly take both candidates and directly predict the preferred image, whereas pointwise methods score candidates independently and select the higher-scoring one. For closed-source pairwise judges, we provide exactly the same frozen canonical SP used by R3, together with the two candidate images, so that they have access to the same structured generation specification. We additionally report the non-spatial semantic/quality signal without SAGE and a geometry-only SAGE baseline to isolate the contribution of each component.

Reward Model / Setting	Advertising	Illustration	Natural Image	Poster Design	Overall	Tie
Pairwise Reward / Judge Models						
Gemini-3.7-Flash Direct	83.60	87.90	53.30	83.60	80.70	4
UnifiedReward-Think	71.40	64.40	32.10	69.00	63.30	4
UnifiedReward-2.0	63.40	68.20	43.80	63.40	62.10	9
GPT-5.6-Luna, no reasoning	52.60	64.70	19.70	40.40	48.10	10
Pointwise Reward Models						
PickScore-v1	69.30	64.70	41.60	63.10	62.40	2
HPSv3-7B	64.10	58.80	32.10	60.60	57.20	4
VQAScore	59.20	55.40	48.20	53.30	54.90	11
HPSv2	43.20	49.10	30.70	35.50	41.00	14
Open R3 w/o SAGE	66.20	72.20	55.00	55.80	63.80	0
Open R3 SAGE Only	51.70	59.80	17.40	39.90	47.70	0
Open R3 + SAGE	**90.70**	**93.20**	**83.30**	**90.10**	**90.00**	0
Gemini R3 w/o SAGE	66.60	73.00	55.50	56.50	64.00	0
Gemini R3 SAGE Only	51.90	59.80	17.50	39.90	47.80	0
Gemini R3 + SAGE	**91.60**	**93.40**	**83.90**	**90.60**	**90.80**	0

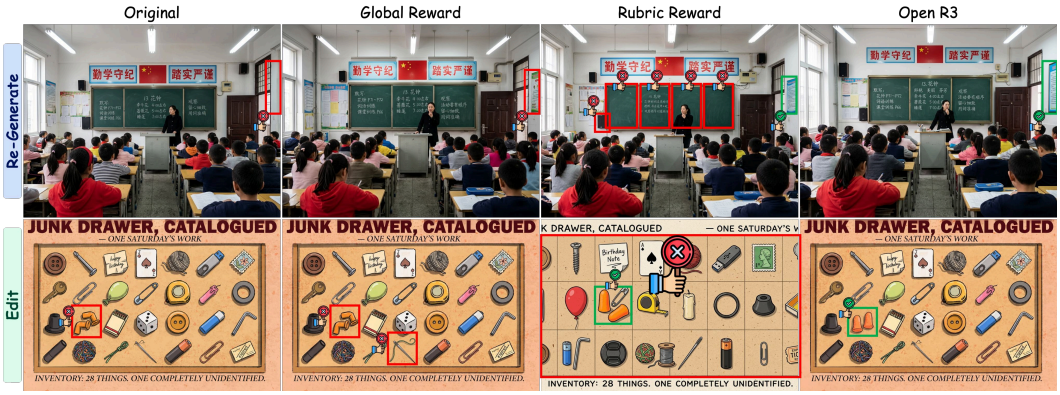


Figure 4: Open R3 identifies the erroneous target and provides localized feedback for regeneration or editing, correcting the intended content while preserving already-correct regions clean.

as reported in Tab. A7 in Appendix. Its advantage is particularly pronounced on recent benchmarks like MMRB2 (Hu et al., 2026) containing compositionally demanding generations. Together, these results support the **semantic–visual granularity mismatch**: as modern T2I systems increasingly expose fine-grained control over entities, attributes, and spatial composition, effective reward modeling likewise benefits from localized semantic assessment coupled with explicit spatial verification.

5.2 Test-Time Self-Reflection

To further examine whether the fine-grained signals provided by Open R3 are actionable, we evaluate it for test-time self-reflection (TTS), as illustrated in Fig. 4. Viewing T2I generation as a long-horizon refinement process, we consider two common paradigms. The first performs *specification revision and regeneration*: reward feedback is used to revise the structural prompt (SP), after which the image is regenerated from scratch. The second performs *reward-guided image editing*: reward feedback is converted into an editing instruction applied directly to the previous generation. We sample 150 images from each of the four GranuBench categories, resulting in 600 test cases, and

Table 2: We compare global, rubric-level, and R3 region-level feedback under two paradigms.

Metric	Global	Rubric	R3 Region
Specification Revision + Regeneration			
Target local-score improvement ↑	50.9%	51.2%	**69.8%**
Target semantic-quality improvement ↑	29.0%	34.0%	**49.0%**
Clean target-rubric repair ↑	23.7%	29.2%	**42.5%**
Correct target rubrics corrupted ↓	28.1%	28.2%	**18.6%**
Non-target entity disappearance ↓	3.81%	3.71%	**3.27%**
Correct non-target rubrics corrupted ↓	22.02%	22.04%	**21.73%**
Net targeted gain, w/o SAGE ↑	−0.16	+3.08	**+12.00**
Human preference for refined output ↑	64.13%	66.71%	**89.75%**
Reward-Guided Image Editing			
Target local-score improvement ↑	55.1%	59.9%	**79.6%**
Target semantic-quality improvement, w/o SAGE ↑	40.3%	46.0%	**64.4%**
Clean target-rubric repair ↑	31.9%	41.0%	**60.2%**
Correct target rubrics corrupted ↓	27.1%	25.5%	**18.5%**
Non-target entity disappearance ↓	7.51%	7.98%	**7.10%**
Correct non-target rubrics corrupted ↓	25.64%	26.57%	**25.41%**
Net targeted gain, w/o SAGE ↑	+3.20	+6.64	**+19.38**
Human preference for refined output ↑	63.21%	66.58%	**94.08%**

evaluate four modern T2I systems supporting structured conditioning: Ideogram4 (Ideogram, 2026), Qwen-Image-SP (Chen et al., 2026), Nano-Banana-2 (Google DeepMind, 2026), and GPT-Image-2 (OpenAI, 2026b). For each case, we compare three levels of reward feedback. **Global reward** produces a holistic score and natural-language feedback for revising the SP or constructing an editing instruction. **Rubric reward** additionally exposes the violated rubrics and explicitly includes them in the revision or editing instruction. **Open R3** provides explicit region-grounded feedback: for regeneration, only the JSON entry corresponding to the erroneous entity is revised; for image editing, the erroneous entity is explicitly identified together with its grounded bounding box.

Results are shown in Tab. 2. Across both paradigms, we evaluate whether the erroneous target is improved and whether refinement damages entities and rubrics that were already correct. To provide an independent evaluation beyond the R3-based metrics, we additionally conduct a blinded human study. For each pre- and post-reflection pair, three annotators are shown the two images in randomized order together with the full SP. Annotators can optionally display the bounding box of any entity in the SP and then independently judge which image better satisfies the overall generation specification; the final preference is determined by majority vote. The human study consistently favors the refined outputs produced with Open R3 feedback, corroborating the automated results.

Open R3 achieves substantially stronger targeted repair while introducing less collateral degradation to clean content. This advantage is especially pronounced because its feedback combines fine-grained semantic credit with explicit visual localization, allowing the refinement step to act on the failed target rather than perturbing the image globally. These results show that fine-grained reward is useful not only for evaluating a generation, but also for providing actionable feedback that enables more reliable self-reflection. Detailed settings and metric definitions are provided in Appendix E.

6 Conclusion

We introduce **R3**, a region-grounded reward formulation for addressing the semantic–visual granularity mismatch in modern text-to-image evaluation. R3 decomposes structured generation requirements into entity-bound criteria, grounds the corresponding visual evidence, and performs explicit target–evidence matching to produce fine-grained region rewards. We further introduce **GranuBench** for evaluating reward granularity and distill the pipeline into **Open R3**, which outperforms conventional holistic and rubric-based rewards on diverse benchmarks. Our test-time self-reflection experiments further show that region-grounded feedback enables more targeted and fine-grained signal, demonstrating that finer reward granularity benefits both evaluation and generation.

Limitation. Our self-reflection study focuses on single-round refinement to isolate reward quality from degradation introduced by repeated image editing. Reliable multi-round refinement, where current editors may progressively alter already-correct content, remains an important direction.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting, language polishing, and structural editing, as well as with literature search and code development.

Generative AI models were also used as components of the data-construction pipeline, as described in Sec. B.

AI-generated suggestions were used only as assistance and were reviewed, revised, and verified by the authors. All technical claims, experimental results, interpretations, and final written content were validated by the authors, who take full responsibility for the work.

REFERENCES

Alibaba. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, November 2025.

ByteDance Seed. Beyond generation, it understands design: Introducing seedream 5.0 pro. https://seed.bytedance.com/en/blog/beyond-generation-it-understands-design-introducing-seedream-5-0-pro, July 2026. Accessed: 2026-09.

Zilong Chen, Chaorui Deng, Kunchang Li, Hongyi Yuan, and Haoqi Fan. Scaling properties of text conditioning in visual generation. *arXiv preprint arXiv:2607.29679*, 2026.

Xuelu Feng, Yunsheng Li, Ziyu Wan, Zixuan Gao, Junsong Yuan, Dongdong Chen, and Chunming Qiao. Rubricrl: Simple generalizable rewards for text-to-image generation. In *European Conference on Computer Vision*, pp. 131–149. Springer, 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. Gemini 3.7 flash. https://ai.google.dev/gemini-api/docs/models/gemini-3.7-flash, 2026. Accessed: 2026-09-16.

Google DeepMind. Nano Banana: Image Editing in Gemini Gets a Major Upgrade. https://blog.google/products-and-platforms/products/gemini/updated-image-editing-model/, August 2025a.

Google DeepMind. Nano Banana Pro: Gemini 3 Pro Image. https://blog.google/innovation-and-ai/products/nano-banana-pro/, November 2025b.

Google DeepMind. Nano Banana 2: Combining Pro Capabilities with Lightning-Fast Speed. https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/, February 2026.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yushi Hu, Reyhane Askari-Hemmat, Melissa Hall, Emily Dinan, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal rewardbench 2: Evaluating omni reward models for interleaved text and image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36904–36915, 2026.

Ideogram. Ideogram 4.0 technical details: Open model at the forefront of design. https://ideogram.ai/blog/ideogram-4.0/, 2026.

Kuei-Chun Kao, Daixuan Huo, Yuanhao Ban, and Cho-Jui Hsieh. Autorubric-t2i: Robust rule-based reward model for text-to-image alignment. *arXiv preprint arXiv:2605.17602*, 2026.

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

Krea. Krea 2 Technical Report. https://www.krea.ai/blog/krea-2-technical-report, June 2026.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *2024 ieee/cvf conference on computer vision and pattern recognition (cvpr)*, pp. 19401–19411. IEEE, 2024.

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

Rui Liu, Dian Yu, Zhenwen Liang, Yucheng Shi, Tong Zheng, Runpeng Dai, Haitao Mi, Pratap Tokekar, et al. Deltarubric: Generative multimodal reward modeling via joint planning and verification. *arXiv preprint arXiv:2605.09269*, 2026.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095. IEEE, 2025.

NVIDIA. Cosmos3-super-text2image. https://huggingface.co/nvidia/Cosmos3-Super-Text2Image, May 2026. Official model card.

OpenAI. Introducing Our Latest Image Generation Model in the API. https://openai.com/index/image-generation-api/, April 2025a.

OpenAI. Introducing GPT-Image-1.5: The New ChatGPT Images. https://openai.com/index/new-chatgpt-images-is-here/, December 2025b.

OpenAI. Gpt-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, 2026a. Accessed: 2026-09-16.

OpenAI. GPT-Image-2. https://developers.openai.com/api/docs/models/gpt-image-2, April 2026b.

Qwen. Qwen3.8-27B. https://huggingface.co/Qwen/Qwen3.8-27B, 2026. Hugging Face model card.

Qwen Team. Qwen-image-2512: Finer details, greater realism. https://qwen.ai/blog?id=qwen-image-2512, December 2025. Official model release blog.

Reve. Introducing reve 2.0. https://blog.reve.com/posts/announcing-reve-2.0/, June 2026. Official model release blog.

Juanxi Tian, Fengyuan Liu, Jiaming Han, Yilei Jiang, Yongliang Wu, Yesheng Liu, Haodong Li, Furong Xu, and Wanhua Li. Auto-rubric as reward: From implicit preferences to explicit multimodal generative criteria. *arXiv preprint arXiv:2605.08354*, 2026.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

Yibin Wang, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, Jiaqi Wang, et al. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *Advances in neural information processing systems*, 38:159130–159157, 2026.

Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, Bairui Li, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Xiaoying Xing, Avinab Saha, Junfeng He, Susan Hao, Paul Vicol, Moonkyung Ryu, Gang Li, Sahil Singla, Sarah Young, Yinxiao Li, et al. Focus-n-fix: Region-aware fine-tuning for text-to-image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18486–18496. IEEE, 2025.

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

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11269–11277, 2026.

Fan Yang, Ru Zhen, Jianing Wang, Yanhao Zhang, Haoxiang Chen, Haonan Lu, Sicheng Zhao, and Guiguang Ding. Heie: Mllm-based hierarchical explainable aigc image implausibility evaluator. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3856–3866. IEEE, 2025.

Sashuai Zhou, Qiang Zhou, Junpeng Ma, Yue Cao, Ruofan Hu, Ziang Zhang, Xiaoda Yang, Zhibin Wang, Jun Song, Cheng Yu, et al. Spatialreward: Verifiable spatial reward modeling for fine-grained spatial consistency in text-to-image generation. *arXiv preprint arXiv:2603.22228*, 2026.

Appendix

In this Appendix, we provide the following materials:

1. Sec. A: design analyses of R3, including faithful target-contract construction, GEF and EBRA, visual-observation granularity, the observe-before-judging information boundary, SAGE, and deterministic reward aggregation;
2. Sec. B: the complete construction pipeline of R3Traj-100K, including target-side quality control, multi-generator image collection, grounded observation, target–evidence matching, and trajectory packaging;
3. Sec. C: the construction and annotation protocol of GranuBench, including seed collection, candidate generation, entity and rubric annotation, human preference labeling, and preference-anchored spatial targets;
4. Sec. D: evaluation of Open R3 on external human-preference reward benchmarks;
5. Sec. E: detailed experimental setup and metric definitions for test-time self-reflection.

A Understanding the Design of R3

The main paper focuses on the final R3 pipeline. Here we provide the design studies that led to its information boundaries, target-construction strategy, observation protocol, and deterministic reward formulation. Unless otherwise specified, these experiments are used for method development and component selection rather than as independent test-set evaluations.

A.1 Constructing a Faithful Target Contract

Constructing the target contract raises two requirements. First, the contract must remain faithful to the generation specification and independent of any particular generated image. Otherwise, the evaluation target may drift from *what was requested* toward *what happened to appear in one reference image*, contaminating the reward definition with image-specific evidence. Second, although grounding queries and rubrics are ultimately useful because they support accurate downstream reward estimation, repeatedly running the complete grounding–observation–matching pipeline is impractical during large-scale data construction. We therefore introduce two lightweight text-only audits that directly measure the properties required from a good target contract: **Groundable Entity Fidelity (GEF)** for grounding queries and **Entity-Bound Rubric Adequacy (EBRA)** for semantic rubrics. Neither metric contributes to the final image reward; both are construction-time quality measures used to compare candidate annotation pipelines. Their definitions and weights are frozen before the pipeline ablations and are not tuned using downstream preference accuracy.

Grounding-query quality with GEF. GEF evaluates whether a candidate query list faithfully converts the SP into a set of usable visual lookup identifiers. For each case, GPT-5.6-Sol (OpenAI, 2026a) with low reasoning receives only the complete SP source catalog and one candidate grounding-query list in a single text-only structured-output call. It does not receive the image, rubrics, human preference labels, downstream reward, or outputs from other construction settings. For every candidate query q_j, the model returns its matched SP source IDs together with binary flags for faithfulness, atomicity, non-redundancy, sufficient disambiguation, and minimal specification, as well as text- and group-scope judgments when applicable. Importantly, the model does not directly predict a GEF score; all components and the final score are computed deterministically from these structured outputs.

Let $\mathcal{U} = \{u_k\}_{k=1}^{K}$ be the SP source atoms and $\mathcal{Q} = \{q_j\}_{j=1}^{J}$ the candidate query list. For query q_j, let π_j denote its predicted matched-source set, and let $f_j, a_j, n_j, d_j, m_j \in \{0, 1\}$ indicate whether it is faithful, atomic, non-redundant, sufficiently disambiguated, and minimally specified, respectively. GEF first computes

$$P = \frac{1}{J}\sum_{j=1}^{J} f_j, \qquad R = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}\left[k \in \bigcup_{j:f_j=1} \pi_j\right], \tag{4}$$

where P is faithful candidate precision and R is faithful source coverage. The remaining general query properties are

$$A = \frac{1}{J}\sum_j a_j, \qquad N = \frac{1}{J}\sum_j n_j, \qquad D = \frac{1}{J}\sum_j d_j, \qquad M = \frac{1}{J}\sum_j m_j, \tag{5}$$

measuring atomicity, non-redundancy, minimal sufficient disambiguation, and query minimality, respectively.

Text and grouped entities require additional scope checks. Let $\mathcal{Q}_T$ and $\mathcal{Q}_G$ denote the subsets for which literal-text and group-scope evaluation is applicable. With corresponding binary judgments t_j and g_j,

$$T = \frac{1}{|\mathcal{Q}_T|}\sum_{j\in\mathcal{Q}_T} t_j, \qquad G = \frac{1}{|\mathcal{Q}_G|}\sum_{j\in\mathcal{Q}_G} g_j. \tag{6}$$

Non-applicable queries do not incur a penalty; when a case contains no applicable text or group query, the corresponding component is neutral. The final audit score is

$$\mathrm{GEF} = 100(0.28P + 0.22R + 0.12A + 0.10N + 0.13D + 0.05M + 0.05T + 0.05G). \tag{7}$$

Thus, different components expose different construction failures: P decreases when a query introduces unsupported content or binds to the wrong source; R detects omitted SP entities; A penalizes merged queries that cannot be localized as one source atom; N detects duplicated coverage; D detects ambiguous queries when multiple same-category instances exist; M penalizes over-specified queries that import unnecessary semantic or image-specific attributes; and T and G preserve the intended literal-text and group scope. GEF therefore measures specification fidelity rather than image-level localization accuracy.

We additionally evaluate operational grounding independently. A fresh Gemini-3.7-Flash (Google, 2026) call receives only the reference image and the exact final query, without the SP, target box, rubrics, or GEF output, and predicts the observed region. The *Grounded* column in Tab. A1 reports the resulting non-null grounding rate. This visual probe is deliberately separated from GEF so that image evidence cannot influence the definition of the target itself.

Grounding-query construction ablation. We compare five construction strategies on a 512-case development set containing 7,392 source entities. All text-generation stages use GPT-5.6-Sol (OpenAI, 2026a) with low reasoning.

Direct generation uses one text-only call to map the complete SP source catalog into an ordered grounding-query list. + *Text-only verification* adds a second case-level call that receives the same source catalog and the generated list and checks omissions, hallucinated entities, merges or splits, duplicate queries, wrong-instance binding, insufficient disambiguation, over-specification, and text/group-scope errors. The verifier may revise query wording but must preserve the source-order one-to-one mapping.

The remaining settings test whether looking at a reference image helps improve the already text-verified queries. + *Full-list visual verification* gives Gemini-3.7-Flash (Google, 2026) the reference image and the complete query list in one call. + *Batched visual verification* applies the same procedure to source-ordered batches of at most eight queries. + *Selective singleton verification* applies one-query-per-call visual verification only to pre-selected difficult cases, while retaining the text-only result elsewhere. These visual verifiers see the image and query text but not the SP, source IDs, target boxes, rubrics, GEF scores, or human preference labels, and may only revise query wording rather than add, delete, or reorder entities.

As shown in Tab. A1, direct text generation is already strong, while one additional text-only verification pass further improves specification fidelity and source binding at low cost. In contrast, increasingly fine-grained visual verification provides limited operational grounding gains while making the query list less faithful to the original specification. Inspection reveals the underlying failure mode: once a verifier sees a particular reference image, it may rewrite the requested target using incidental pixel evidence—for example, replacing the requested location, instance description, or literal text with what happened to be rendered. This changes the question from *what should be present* to *what this image already contains*. We therefore freeze grounding queries before any image is observed and reserve visual grounding exclusively for downstream evidence collection.

Table A1: **Grounding-query construction.** Text-only verification improves specification fidelity and entity binding, while image-aware revision provides little additional grounding benefit and can alter the intended target.

Construction Setting	GEF ↑	Binding ↑	Grounded ↑	Calls/Case ↓
Direct generation	99.45	99.57	98.70	1.00
+ Text-only verification	**100.00**	**100.00**	98.70	2.00
+ Full-list visual verification	98.54	97.40	**99.57**	3.00
+ Batched visual verification	98.77	97.84	**99.57**	4.31
+ Selective singleton verification	99.64	99.13	**99.57**	6.94

Rubric quality with EBRA. Rubric construction faces an analogous problem: a good rubric set must faithfully encode the semantic requirements of the SP without requiring the complete downstream reward pipeline to be executed during annotation. We therefore evaluate candidate rubric dictionaries with EBRA.

All EBRA development calls use GPT-5.6-Sol (OpenAI, 2026a) with low reasoning and remain strictly text-only. Before evaluating any candidate rubric pipeline, two independent catalog-construction passes receive only the SP and the frozen grounding-query list and compile a fixed catalog $\mathcal{O} = \{o_\ell\}_{\ell=1}^{L}$ of atomic semantic obligations. The second pass verifies and reconciles this catalog. Crucially, candidate rubrics are not used to construct the catalog, so each construction strategy is evaluated against the same target semantics.

The candidate rubric dictionary is then audited against this frozen catalog. For every unique rubric c_j, the evaluator determines whether it is entailed by the SP, whether it is bound to the correct entity, which catalog obligations it expresses, whether it is atomic and binary-verifiable, and whether it improperly leaks absolute spatial criteria or generic visual-integrity criteria into the semantic channel. The evaluator again produces only structured item-level judgments and obligation mappings; the six EBRA components and final score are computed locally. Images, grounding results, human preferences, downstream reward scores, and the identities or outputs of competing construction settings are never provided.

Let $\mathcal{C} = \{c_j\}_{j=1}^{M}$ denote the candidate rubrics after collapsing exact normalized duplicates within each entity. For rubric c_j, let $e_j \in \{0,1\}$ denote SP entailment, $b_j \in \{0,1\}$ correct entity binding, $a_j \in \{0,1\}$ atomic binary verifiability, $p_j \in \{0,1\}$ position-channel cleanliness, and $i_j \in \{0,1\}$ generic-integrity-channel cleanliness. Let $\Gamma_j \subseteq \mathcal{O}$ be the semantic obligations to which the evaluator maps the rubric.

SP entailment is the fraction of candidate rubrics directly supported by the specification:

$$E = \frac{1}{M}\sum_{j=1}^{M} e_j. \tag{8}$$

Entity binding is evaluated among SP-supported rubrics:

$$B = \frac{\sum_{j=1}^{M} e_j b_j}{\sum_{j=1}^{M} e_j}. \tag{9}$$

Semantic coverage measures how much of the independently compiled obligation catalog is recovered by at least one entailed, correctly bound candidate rubric:

$$C = \frac{\left|\bigcup_{j:e_j b_j=1} \Gamma_j\right|}{|\mathcal{O}|}. \tag{10}$$

The remaining three dimensions are item-level rates:

$$A = \frac{1}{M}\sum_{j} a_j, \qquad P = \frac{1}{M}\sum_{j} p_j, \qquad I = \frac{1}{M}\sum_{j} i_j. \tag{11}$$

Here, A penalizes compound or otherwise non-binary criteria; P requires semantic rubrics to exclude absolute canvas-position constraints, which are handled separately by SAGE; and I excludes generic defects such as malformed anatomy or rendering artifacts, which belong to the visual-quality channel.

Table A2: **Entity-bound rubric construction.** Decoupling semantic content from hierarchy assignment and constructing rubrics in smaller entity batches provide the best quality–cost trade-off.

Construction Setting	Macro EBRA-LA ↑	Calls ↓
Joint content + hierarchy	91.928	16
+ Unconstrained revision	90.831	32
Decoupled content / hierarchy	94.106	48
Batched construction	**94.871**	69
Singleton construction†	95.728	88†

†Evaluated only on the pre-selected hard cases as a costly upper-bound setting.

The final score is

$$\mathrm{EBRA} = 100(0.275E + 0.225B + 0.250C + 0.150A + 0.050P + 0.050I). \tag{12}$$

The supplied L1/L2/L3 label is intentionally excluded from EBRA. Levels organize semantic granularity and rubric budget, rather than determining whether the underlying semantic criterion is valid. Likewise, exact duplicate text is collapsed before scoring so that copying a rubric cannot increase the score.

Entity-bound rubric construction ablation. We use the same GPT-5.6-Sol low-reasoning model for all rubric-generation variants and compare five strategies that isolate three design choices: whether semantic content and hierarchy should be generated jointly, whether an unconstrained revision step is useful, and how many entities should be processed together.

Joint content + hierarchy uses one case-level call to jointly generate the complete rubric dictionary and its L1/L2/L3 assignments. *+ Unconstrained revision* adds a second case-level verifier that may freely add, delete, or rewrite rubric content. *Decoupled content / hierarchy* instead first generates a level-agnostic semantic inventory, performs case-level reconciliation to repair omissions, unsupported criteria, incorrect binding, compound criteria, and channel leakage, and only afterward invokes an independent allocator that assigns L1/L2/L3 without changing rubric text. *Batched construction* keeps this decoupled pipeline but generates the level-agnostic inventory in source order over at most eight target entities per call, while each batch still receives the complete case context. Finally, *Singleton construction* generates the inventory independently for each entity before applying the same reconciliation and level allocation, serving as a costly fine-grained upper bound.

All variants share the same deterministic schema validator, entity ordering, exact-duplicate handling, and EBRA evaluator; therefore, the comparison isolates the construction strategy rather than differences in evaluation.

The comparison reveals two useful design choices. First, an unconstrained second-pass rewrite is not sufficient by itself: it can remove valid semantic coverage or replace atomic criteria with new compound formulations. By contrast, separating *what should be evaluated* from *how those criteria should be organized* produces substantially better rubric quality, motivating level-agnostic content construction followed by an independent hierarchy allocator. Second, smaller generation batches improve the difficult high-entity-count tail, while singleton construction provides only a costly upper-bound improvement. We therefore adopt batched, level-decoupled rubric construction as the default.

Handling long specifications. Even with batched inventory generation, very long specifications introduce a different bottleneck: a single full-case reconciliation call may compress or omit entity-specific semantics. For cases containing at least 32 entities, we therefore use a local long-context variant. Each source-ordered chunk contains at most eight entities and independently performs rubric generation, verification, and indexed L1/L2/L3 allocation; no final full-case reconciliation is applied. The local SP retains the shared scene context needed to interpret the current entities and their relations.

As shown in Tab. A3, local construction substantially improves EBRA, semantic coverage, and atomic binary verifiability on long specifications. We therefore use the batched global pipeline for ordinary cases and the fully local variant for specifications containing at least 32 entities.

Role of the rubric hierarchy. Finally, we examine whether progressively finer rubric levels should receive special treatment in the reward. We keep the candidate images, visual observations, individ-

Table A3: **Rubric construction for long specifications.** Results are reported on 256 cases containing at least 32 entities.

Setting	EBRA-LA ↑	Coverage ↑	Atomic Binary ↑
Global reconciliation	88.961	0.8783	0.7549
Local long-context construction	**95.051**	**0.9454**	**0.9648**

Table A4: **Effect of rubric depth.** Increasing entity coverage is useful, while progressively adding finer rubrics after full coverage does not monotonically improve reward accuracy.

Rubric Subset	Entity Coverage	R3 + SAGE ↑	R3 w/o SAGE ↑
$\frac{1}{2}$ L1	60.77	80.33	52.00
L1	100	84.33	55.67
L1 + $\frac{1}{2}$L2	100	**86.00**	55.67
L1 + L2	100	85.67	54.33
L1 + L2 + $\frac{1}{2}$L3	100	84.67	**56.00**
L1 + L2 + L3	100	84.33	55.33

ual rubric judgments, quality scores, and SAGE values fixed, and vary only the subset of L1/L2/L3 rubrics included in aggregation.

Tab. A4 shows that completing entity coverage is useful, whereas adding progressively more criteria after full entity coverage does not monotonically improve preference accuracy. We therefore use L1/L2/L3 primarily to organize semantic granularity and rubric budget rather than treating more criteria as intrinsically more important. All retained rubrics consequently receive equal weight within each entity in the default R3 reward.

A.2 Controlling Visual Observation Granularity

R3 must simultaneously localize each entity and record sufficiently complete visual evidence. Processing too many entities in the same multimodal call may preserve entity identities and boxes while degrading attribute coverage and local-quality assessment. We therefore compare visual batches of 4, 6, and 8 entities against singleton observations.

As shown in Tab. A5, grounding remains highly stable across all batch sizes, with nearly identical found-status and box agreement. The main degradation instead appears in local-quality assessment: quality agreement drops substantially as more entities are processed jointly, reducing the overall observation consistency. The four-entity setting achieves the highest composite agreement while retaining near-singleton grounding quality. We therefore use at most four entities per visual call.

A.3 Why Observe Before Judging?

A central information boundary in R3 is that visual evidence is collected without seeing the semantic rubrics, frozen, and only then matched against the target in text space. This separation is intended to prevent the evaluator from using the requested condition itself as evidence that the condition is visually present. We isolate this design on a 512-case diagnostic set by progressively varying what information is available during visual observation and semantic judgment.

The comparison covers three families of evaluation. *Direct image judgment* evaluates each rubric while still viewing the candidate image. We test a flat formulation that provides the full SP and a self-contained rubric, an entity-bound formulation that additionally exposes the corresponding entity specification, and a query-conditioned formulation that removes the full SP but retains the grounding query and bound rubric. We also test *joint observation and judgment*, where the model receives the image, grounding query, and rubric in the same call and simultaneously describes the visual evidence and decides whether the rubric is satisfied.

We then introduce a staged protocol. R3 first obtains a *rubric-blind frozen observation* from the image and grounding query. In *staged + image revisit*, the second-stage judge receives this frozen observation, the bound rubric, and the image again. In *staged + text-only judgment*, the second stage receives only the frozen observation and rubric, so the semantic decision cannot be revised using target-conditioned image inspection. We instantiate this text-only setting with both Gemini and

Table A5: **Visual-observer batching study.** Agreement is measured against singleton observations used as the reference.

Entities / call	Composite ↑	Found Agr. ↑	Box SAGE ↑	Quality Agr. ↑
4	**0.9967**	1.000	0.9990	**0.9688**
6	0.9895	1.000	**0.9994**	0.8958
8	0.9759	1.000	0.9992	0.7604

Table A6: **Information-boundary ablation.** We progressively separate visual observation from target-aware semantic judgment. "Gold-no FP" counts false-positive judgments among 100 manually verified image–rubric pairs for which the requested property is absent.

Evaluation Setting	No-SAGE Acc.	Yes-rate	Gold-no FP
Direct SP + flat rubric	57.5	95.88	96/100
Direct SP + bound rubric	52.5	96.19	91/100
Direct query + bound rubric	52.5	94.33	24/100
Joint observe + judge	55.0	95.99	11/100
Staged + image revisit	52.5	93.70	9/100
Staged + text-only Gemini	**55.0**	48.58	**0/100**
Staged + flat text-only GPT	45.0	47.97	**0/100**
Staged + bound text-only GPT	47.5	**47.38**	**0/100**

GPT-family judges. For GPT, we additionally compare flattened rubrics with entity-bound rubrics to isolate the value of preserving explicit target–evidence binding.

Because the target boxes in this diagnostic set are reference-anchored and the shared spatial channel is highly predictive, we remove SAGE from all comparisons below so that the results expose differences in semantic judgment rather than spatial matching.

As shown in Tab. A6, direct rubric-visible image judgment exhibits a strong tendency to affirm requested properties. The three direct or joint image-visible settings produce rubric yes-rates above 94%, even though the manually verified negative examples contain requested properties that are genuinely absent from the candidate image.

The cleanest controlled comparison is between *staged + image revisit* and *staged + text-only Gemini*. Both use the same rubric-blind frozen observation, the same entity binding, the same rubric, and the same model family; the only difference is whether the second-stage judge can revisit the image. Removing this image access reduces the rubric yes-rate from 93.70% to 48.58% and reduces false positives on the 11 manually verified negative image–rubric pairs from 1 to 0, while preference accuracy changes from 52.5% to 55.0%. This indicates that even after a faithful observation has been collected, allowing the target-aware judge to inspect the image again can reintroduce target-conditioned affirmation.

The GPT text-only variants further test how the frozen evidence should be consumed. Flattening the rubrics removes explicit entity binding, whereas the final R3 setting preserves the correspondence between each frozen observation and its entity-bound rubric. Both retain the desired low yes-rate and produce no false positive on the manually verified negatives. These results support the information flow adopted by R3: collect rubric-blind visual evidence first, freeze it, and perform target–evidence matching afterward without image access.

This experiment is not intended to show that text-only semantic matching alone provides strong pairwise ranking. Observation under-coverage may still introduce false negatives, and the no-SAGE preference accuracies on this small diagnostic set remain modest. The purpose of this design is specifically to reduce target-conditioned false affirmation and make each semantic decision traceable to frozen visual evidence.

A.4 Spatial Matching and Deterministic Aggregation

R3 handles explicit spatial requirements with **SAGE** (**S**cale-**A**daptive **G**aussian **E**ntity-Box Concordance), rather than asking the semantic judge to assess absolute position in language. SAGE compares the target box specified by the SP with the observed box grounded in the candidate image. Let the target and observed boxes have centers (c_x^t, c_y^t) and (c_x^o, c_y^o) and sizes (w_t, h_t) and (w_o, h_o). We define

$$\sigma_x = \max\left(\frac{w_t}{2}, 40\right), \qquad \sigma_y = \max\left(\frac{h_t}{2}, 40\right), \tag{13}$$

$$d^2_{\text{center}} = \left(\frac{c_x^o - c_x^t}{\sigma_x}\right)^2 + \left(\frac{c_y^o - c_y^t}{\sigma_y}\right)^2, \tag{14}$$

and

$$d^2_{\text{logsize}} = \log^2\left(\frac{w_o}{w_t}\right) + \log^2\left(\frac{h_o}{h_t}\right). \tag{15}$$

The spatial concordance is then

$$\text{SAGE}_i = 2^{-0.5 d^2_{\text{center}} - 0.25 d^2_{\text{logsize}}}. \tag{16}$$

SAGE equals one for a perfect match and decreases continuously with center displacement and scale mismatch. Unlike IoU, it can distinguish different degrees of spatial error even when the target and observed boxes do not overlap. The constant 40 introduces a 4% scale floor in the normalized-1000 coordinate system, preventing very small target boxes from becoming excessively sensitive to minor grounding jitter. If a target box exists but the entity is not observed, SAGE is set to zero. If no target box is specified, no spatial penalty is applied, corresponding to $\text{SAGE}_i = 1$ in the implementation. For grouped observations, all visible member boxes are first unioned before comparison with the group target box.

Given the semantic judgments and visual observation, the final reward is aggregated deterministically. For entity e_i, its semantic score is

$$s_i^{\text{sem}} = \frac{1}{|\mathcal{C}_i|} \sum_j \mathbb{1}[y_{ij} = \text{yes}], \tag{17}$$

and its visual-integrity score is

$$s_i^{\text{qual}} = \begin{cases} 0, & z_i = 0, \\ (q_i^{\text{qual}} - 1)/4, & z_i = 1. \end{cases} \tag{18}$$

We first combine semantic fidelity and visual integrity as

$$s_i^{\text{base}} = 0.8 s_i^{\text{sem}} + 0.2 s_i^{\text{qual}}, \tag{19}$$

and then apply spatial concordance through

$$g_i = 0.5 + 0.5\,\text{SAGE}_i, \qquad r_i = s_i^{\text{base}} g_i. \tag{20}$$

The image-level reward is

$$R_{\text{R3}} = 0.85\ \text{Mean}_i(r_i) + 0.15\ \text{Bottom3Mean}_i(r_i). \tag{21}$$

When fewer than three entities are available, the second term uses all entities. L1/L2/L3 rubrics are uniformly averaged within each entity.

The spatial channel contributes substantially to the complete reward. On the 300-case development benchmark, R3 achieves 84.33% preference accuracy with SAGE, compared with 55.33% when the explicit SAGE factor is removed. This comparison shows that explicit target-region information provides an important signal in the current structured-generation setting. However, the benchmark uses reference-anchored target regions, so the result should not be interpreted as evidence that the semantic evaluator alone accounts for the full improvement. We therefore report results without SAGE separately whenever analyzing the semantic component.

B R3TRAJ-100K CONSTRUCTION AND OPEN R3 INFERENCE

To distill Gemini R3 into an open model, we construct **R3Traj-100K**, a collection of 100K complete R3 trajectories spanning diverse prompts and image generators. Rather than retaining only the final scalar reward, each trajectory preserves intermediate supervision across structural prompting, rubric construction, visual observation, and target–evidence reasoning. These trajectories are further decomposed into stage-specific SFT examples for the four LoRA adapters used by Open R3.

Source prompts and structural targets. We begin from an internal high-quality T2I prompt pool curated by our organization. The pool combines high-quality prompts collected from open-source communities with prompts manually authored and refined by employees, covering diverse visual styles, compositions, text-rich designs, multi-entity scenes, and spatially constrained requests. Each selected request is normalized into the canonical R3 structural prompt (SP), which represents entities, attributes, relations, literal text when applicable, and spatial targets when they are specified or meaningfully implied by the request.

For each SP, R3 constructs an ordered set of minimally disambiguated grounding queries. These queries are audited with **GEF**, which measures faithful entity coverage, atomicity, non-redundancy, sufficient disambiguation, minimality, and text/group scope. Queries that fail the audit are reconstructed from the source specification and re-audited without exposing candidate-image evidence. The resulting entity–query mapping is then frozen for subsequent rubric construction and visual observation.

Metric-guided rubric construction. Given the frozen entity list, we construct entity-bound semantic rubrics following the protocol in Sec. A.1. Each rubric-generation call processes at most **eight target entities**. For each batch, R3 first constructs a level-agnostic semantic inventory, followed by verification and reconciliation of omissions, unsupported requirements, incorrect entity binding, compound criteria, and semantic-channel leakage. An independent allocator subsequently assigns L1/L2/L3 labels without modifying the verified rubric text.

The resulting rubric dictionary is independently audited with **EBRA**, which evaluates specification entailment, entity binding, semantic coverage, atomic binary verifiability, and separation from the spatial and visual-integrity channels. GEF and EBRA therefore provide complementary quality control over *what should be located* and *what should be assessed* for each target entity.

Multi-generator candidate generation. After the target contract is frozen, we generate candidate images using nine contemporary T2I systems. R3Traj-100K contains **40K generations from four open models** and **60K generations from five closed-source models**. Specifically, we collect 10K generations each from **Ideogram4** (Ideogram, 2026), **Qwen-Image-2512** (Qwen Team, 2025), **Cosmos-T2I-Super** (NVIDIA, 2026), and **Qwen-Image-SP** (Chen et al., 2026). For the closed-source pool, we collect 12K generations each from **GPT-Image-2** (OpenAI, 2026b), **GPT-Image-1.5** (OpenAI, 2025b), **Nano-Banana-2** (Google DeepMind, 2026), **Seedream 5.0** (ByteDance Seed, 2026), and **Reve 2.0** (Reve, 2026).

Rubric-blind visual observation. Each generated image is processed using the information boundary and observation granularity described in Sec. A.2. For every target entity, the observer receives only the candidate image and its frozen grounding query, without access to the semantic rubrics or target box. Each multimodal call processes at most **four entities and their grounding queries**. The resulting observation

$$O_i = (z_i, \hat{b}_i, A_i, Q_i) \tag{22}$$

records whether the entity is found, its observed bounding box, visible image-grounded facts, and local generation quality. Observations from different batches are merged by stable entity ID before subsequent reasoning.

Target–evidence reasoning. The observations are then frozen and paired with their entity-bound rubrics. For each entity, the reasoning stage receives only $(O_i, \mathcal{C}_i)$ and predicts binary judgments for the corresponding semantic criteria together with supporting evidence. It does not revisit the image or access the target box. Spatial consistency is computed separately and deterministically

using SAGE when a valid target box is available, while semantic satisfaction and visual integrity are obtained from the reasoning output and observation quality.

Stage-specific SFT data. The complete trajectories are decomposed into training examples for the four Open R3 adapters. Because one trajectory may contain multiple entity batches or rubric groups, the number of stage-level SFT examples is larger than the number of original trajectories.

For **SP refinement**, we combine R3Traj-100K with our internal prompt-rewriter data, yielding approximately **250K** training examples. This stage includes both natural-language-to-SP conversion and normalization from other structured prompt formats into the canonical R3 SP. Following structured T2I prompting practices (Ideogram, 2026; Chen et al., 2026), spatial boxes are generated only when appropriate for the request; prompts without a meaningful spatial target retain no target box, in which case SAGE is omitted from the final reward.

For **rubric generation**, the trajectories produce **212,117** SFT examples, with at most eight entities included in each example following Sec. A.1. For **visual observation**, they produce **303,225** multimodal SFT examples, with at most four entity–query pairs processed per example following Sec. A.2. For **target–evidence reasoning**, they produce **324,875** text-only SFT examples from frozen observations and their corresponding entity-bound rubrics. Each adapter is trained independently using standard LoRA SFT on its corresponding stage-specific data.

Open R3 inference. At inference time, the shared Qwen3.8-27B backbone and all four LoRA adapters are loaded once, with the corresponding adapter activated for each R3 stage. Open R3 first normalizes the input request into the canonical SP and constructs the entity/query representation. Conditioned on this representation, rubric construction and visual observation are independent and can be executed in parallel. The rubric adapter produces entity-bound semantic criteria $\mathcal{C}_i$, while the observation adapter processes the candidate image and grounding queries to produce frozen observations O_i. The reasoning adapter then performs target–evidence assessment entirely in text space from $(O_i, \mathcal{C}_i)$. Finally, semantic judgments, visual quality, and SAGE when applicable are combined using the deterministic R3 aggregation rule to obtain region rewards and the final image reward.

Thus, Open R3 reproduces the staged information flow of Gemini R3 through four standard stage-specific SFT tasks, while preserving the explicit separation between target specification, visual observation, semantic reasoning, and deterministic reward aggregation.

C Construction and Annotation of GranuBench

Seed collection and candidate generation. We begin by manually collecting a large pool of real-world T2I requests from open AIGC communities. The requests are organized into four categories: *advertising*, *illustration*, *poster design*, and *natural image*. We retain cases with high visual appeal and substantial compositional difficulty, particularly those involving multiple interacting entities, dense attributes, text rendering, and explicit layout constraints. After manual screening, 1,000 requests are selected as the seed cases of GranuBench.

For each retained request, Gemini-3.7-Flash (Google, 2026) produces a detailed generation caption. We then render five candidate images using GPT-Image-2 (OpenAI, 2026b), Nano-Banana-2 (Google DeepMind, 2026), Ideogram4 (Ideogram, 2026), Qwen-Image (Alibaba, 2025), and FLUX.2 (Black Forest Labs, 2025). Using multiple contemporary generators increases the diversity of both successful realizations and fine-grained failure modes within each case.

Entity and rubric annotation. Three PhD-candidate annotators jointly inspect the original request and all five candidate images to determine which entities should be explicitly evaluated. For every selected entity, annotators provide a minimally disambiguated grounding query and a set of entity-bound semantic rubrics following the same L1/L2/L3 hierarchy defined in Sec. 3.2. The independently produced entity lists, grounding queries, and rubrics are then merged and deduplicated to form a canonical annotation for each case. The original request and consolidated annotations are subsequently converted by Magic Prompt LLM (Ideogram, 2026) into the canonical SP representation used by R3, including the structured semantic and spatial constraints required for fine-grained evaluation.

Human preference annotation. For each case, we randomly sample two images generated by different candidate models to form the final preference pair. A separate group of three full-time imaging professionals independently evaluates each pair. Annotators are shown the original request, both candidate images, and the corresponding entity and rubric annotations. They select the image that better satisfies the intended generation conditions, considering semantic fidelity and spatial layout while maintaining overall visual quality. The final preference label is determined by majority vote, ensuring that the benchmark preference is defined by human judgment rather than by an automatic reward model.

Final benchmark annotations. Each GranuBench case therefore contains the original natural-language request, its canonical SP, entity-level grounding queries, coarse-to-fine semantic rubrics, explicit spatial targets, and a human-labeled preference pair. Together, these annotations support evaluation at both image-preference level and finer entity, semantic, and spatial granularities.

D Evaluation on External Reward Benchmarks

We further evaluate Open R3 on three established human-preference benchmarks: the MMRB2 T2I subset (Hu et al., 2026), GenAI-Bench (Li et al., 2024), and HPDv3 (Ma et al., 2025). These benchmarks provide an independent test of whether the reward formulation developed for fine-grained condition following generalizes beyond GranuBench. For benchmarks specified only by natural-language prompts, Open R3 constructs the structured target using its standard inference pipeline, without introducing benchmark-specific region annotations. We compare against representative pairwise and pointwise reward models using the evaluation protocol and metric defined by each benchmark.

Table A7: Results on MMRB2 (Hu et al., 2026) T2I, GenAI-Bench (Li et al., 2024), and HPDv3 (Ma et al., 2025).

Reward Model	MMRB2 ↑	genai-bench		HPDv3 ↑
		Diff. ↑	τ ↑	
Pairwise Reward Models				
PickScore	57.6	67.2	53.2	65.6
UnifiedReward	61.2	79.0	63.1	–
Pointwise Reward Models				
ImageReward	54.0	65.0	47.8	58.6
HPSv2	54.7	68.4	51.6	65.3
CLIPScore	51.0	50.8	39.6	48.6
VQAScore	58.3†	68.0	45.2	–
HPSv3	59.4	–	–	76.9
R3 w/o SAGE	63.1	69.2	63.7	63.8
R3 + SAGE	**81.1**	**80.2**	**73.2**	**77.1**

As shown in Table A7, Open R3 generalizes consistently across external benchmarks and outperforms conventional reward models on most evaluated metrics. The advantage is particularly pronounced on MMRB2 T2I, which focuses on more contemporary T2I generations and challenging text conditions. Open R3 also performs strongly on GenAI-Bench, while remaining competitive with specialized reward models on HPDv3.

These results complement the GranuBench evaluation in Sec. 5.1. The performance gap becomes most apparent on settings that place stronger demands on compositional and fine-grained condition following, consistent with our hypothesis of a **semantic–visual granularity mismatch**. As T2I models increasingly support detailed entities, attributes, relations, and spatial specifications, holistic image-level scores become a progressively coarser representation of generation quality. R3 instead preserves this structure during evaluation by assigning credit to localized visual evidence associated with individual generation requirements.

E Test-Time Self-Reflection: Experimental Setup and Metrics

Experimental protocol. We evaluate single-round test-time self-reflection on 600 prompts sampled uniformly from the four GranuBench categories, with 150 prompts per category. For each prompt, we obtain one baseline image from each of four structured T2I systems: Ideogram4 (Ideogram, 2026), Qwen-Image-SP (Chen et al., 2026), Nano-Banana-2 (Google DeepMind, 2026), and GPT-Image-2 (OpenAI, 2026b), yielding 2,400 baseline images. Each image is evaluated under two refinement paradigms and three reward policies, resulting in 14,400 refinement trajectories in total. Accordingly, every experimental cell in Tab. 2 contains $N = 2400$ trajectories. Within each paradigm, the baseline image, frozen SP, refinement planner, downstream generator or editor, and final evaluator are held fixed; only the reward feedback exposed to the planner changes.

Reward feedback. We compare three levels of reward granularity. **Global reward** evaluates the complete image holistically in terms of semantic correctness, completeness, spatial composition,

text rendering, style, technical quality, and aesthetic consistency, and returns natural-language feedback without entity-level scores, bound rubrics, grounded observations, or bounding boxes. **Rubric reward** decomposes evaluation into individual rubric judgments, each with a binary decision and supporting evidence, but presents them to the planner as a flattened set without explicit entity binding. **Open R3** instead identifies the target entity explicitly and provides entity-bound semantic failures together with its grounded visual evidence and spatial information. All three policies are restricted to refining a single selected entity in each trajectory.

Specification revision and regeneration. In the first paradigm, reward feedback is used to revise the structural specification before regenerating the image. The planner produces a replacement description for exactly one target entity, and the program replaces only the corresponding entity description in the original SP. A parsed-JSON guard verifies that all remaining entities are unchanged. The revised SP is then passed back to the same T2I model that produced the baseline image. Each model retains its native structured-conditioning adapter, with deterministic conversion from the canonical SP representation to the required prompt and coordinate format. For Open R3, only the JSON entry associated with the identified erroneous entity is revised.

Reward-guided image editing. In the second paradigm, reward feedback is converted into a single-entity natural-language editing instruction applied directly to the baseline image. All reward policies use the same Nano-Banana-2 editor and explicitly instruct it to preserve non-target content. Open R3 additionally provides the grounded bounding box of the erroneous entity. For spatial relocation errors, the instruction contains both the observed box $\hat{b}_i$ and the intended target box b_i^t, allowing the editor to distinguish where the entity currently appears from where it should appear.

Shared entity scoring. All refined images are evaluated using the same frozen R3 scoring rule. For trajectory n and entity e_i, let $y_{n,i,j} \in \{0,1\}$ denote the judgment for rubric $c_{ij} \in \mathcal{C}_i$. The semantic score is

$$s_{n,i}^{\text{sem}} = \frac{1}{|\mathcal{C}_i|} \sum_{j=1}^{|\mathcal{C}_i|} y_{n,i,j}. \tag{23}$$

Using the visual-quality score $Q_{n,i} \in \{1, \ldots, 5\}$ from the observation,

$$s_{n,i}^{\text{qual}} = \begin{cases} 0, & z_{n,i} = 0, \\ (Q_{n,i} - 1)/4, & z_{n,i} = 1. \end{cases} \tag{24}$$

We define the non-spatial semantic-quality score as

$$s_{n,i}^{\text{base}} = 0.8 s_{n,i}^{\text{sem}} + 0.2 s_{n,i}^{\text{qual}}, \tag{25}$$

and the complete region reward as

$$r_{n,i} = s_{n,i}^{\text{base}} \left(0.5 + 0.5\, \text{SAGE}_{n,i}\right). \tag{26}$$

Thus, s^{base} measures semantic satisfaction and local visual quality without rewarding spatial correction, whereas r additionally incorporates explicit spatial agreement through SAGE.

Let superscripts b and a denote measurements before and after refinement, respectively, and let t_n denote the index of the target entity selected in trajectory n. We use $\epsilon = 10^{-12}$ for strict floating-point comparisons.

Target local-score improvement. This metric measures the fraction of trajectories in which the complete region reward of the selected target entity increases:

$$\text{TargetLocalImprove} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\left[r_{n,t_n}^{a} - r_{n,t_n}^{b} > \epsilon\right]. \tag{27}$$

Because r_{n,t_n} includes SAGE, this metric jointly reflects semantic satisfaction, visual quality, and spatial correction.

Target semantic-quality improvement. To isolate content repair from explicit spatial matching, we separately measure whether the target's non-spatial score improves:

$$\text{TargetContentImprove} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left[s_{n,t_n}^{\text{base},a} - s_{n,t_n}^{\text{base},b} > \epsilon\right]. \tag{28}$$

This metric excludes SAGE and therefore does not count a bounding-box correction alone as semantic-quality improvement.

Clean target-rubric repair. A target is considered cleanly repaired only when at least one previously unsatisfied target rubric changes from *no* to *yes*, while none of its previously satisfied rubrics changes from *yes* to *no*. Formally,

$$\text{CleanRepair} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left[\begin{array}{l} \exists j : y_{n,t_n,j}^{b} = 0 \wedge y_{n,t_n,j}^{a} = 1, \\ \forall k : y_{n,t_n,k}^{b} = 1 \Rightarrow y_{n,t_n,k}^{a} = 1 \end{array}\right]. \tag{29}$$

This stricter trajectory-level metric prevents an intervention from being counted as successful when it fixes one requirement by breaking another previously correct requirement of the same target.

Correct target rubrics corrupted. We measure target-side regression as the fraction of initially satisfied target rubrics that become unsatisfied after refinement:

$$\text{TargetRubricCorruption} = \frac{\sum_n \sum_j \mathbb{1}\left[y_{n,t_n,j}^{b} = 1 \wedge y_{n,t_n,j}^{a} = 0\right]}{\sum_n \sum_j \mathbb{1}\left[y_{n,t_n,j}^{b} = 1\right]}. \tag{30}$$

This is a rubric-level micro-average and therefore directly measures how often refinement damages requirements that were already correctly satisfied on the entity being edited.

Non-target entity disappearance. We explicitly track whether refinement removes content outside the selected target. Using the entity-presence indicator $z_{n,i}$, the disappearance rate is

$$\text{NonTargetDisappear} = \frac{\sum_n \sum_{i \neq t_n} \mathbb{1}\left[z_{n,i}^{b} = 1 \wedge z_{n,i}^{a} = 0\right]}{\sum_n \sum_{i \neq t_n} \mathbb{1}\left[z_{n,i}^{b} = 1\right]}. \tag{31}$$

Only non-target entities that are present before refinement enter the denominator, so the metric measures collateral deletion rather than pre-existing missing content.

Correct non-target rubrics corrupted. To capture subtler collateral damage that does not remove an entity entirely, we compute the fraction of initially correct non-target rubrics that become incorrect:

$$\text{NonTargetRubricCorruption} = \frac{\sum_n \sum_{i \neq t_n} \sum_j \mathbb{1}\left[y_{n,i,j}^{b} = 1 \wedge y_{n,i,j}^{a} = 0\right]}{\sum_n \sum_{i \neq t_n} \sum_j \mathbb{1}\left[y_{n,i,j}^{b} = 1\right]}. \tag{32}$$

This rubric-level micro-average measures collateral semantic degradation across content that the refinement was not intended to modify.

Net targeted gain without SAGE. Finally, we report the average change in the target's non-spatial semantic-quality score:

$$\Delta_{\text{target}} = \frac{1}{N}\sum_{n=1}^{N}\left(s_{n,t_n}^{\text{base},a} - s_{n,t_n}^{\text{base},b}\right). \tag{33}$$

Tab. 2 reports $100\Delta_{\text{target}}$. Positive values therefore indicate net improvement in the selected entity's semantic satisfaction and visual quality, independently of any gain obtained purely through SAGE-based spatial correction.

Except for the two rubric-corruption metrics, which are explicitly micro-averaged over rubrics, all reported rates are computed at the trajectory level over the same $N = 2400$ trajectories in each experimental cell. This prevents cases containing more entities or rubrics from receiving greater weight in the primary trajectory-level comparisons.