LEARNING TO REPAIR: SKILL OPTIMIZATION FOR MULTI-ROUND TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-round text-to-image generation uses feedback to correct compositional errors, yet additional refinement does not always lead to better images. Progress depends on choosing suitable repair actions and ordering them so that later operations preserve earlier corrections. How can repair experience be turned into reusable guidance for these decisions? We address this question through skill optimization, with a focus on designing the space of guidance to be learned. Different failure types respond differently to editing and regeneration, while interactions between repairs make their order consequential. Motivated by these observations, we structure the critic's natural-language skill around failure diagnosis, type-conditional action choice, preservation requirements, and repair priorities, with generator-specific observations stored separately. Starting from human-designed guidance, we use feedback from a small collection of historical trajectories to refine the skill while keeping the underlying models frozen. Experiments on GenEval, GenEval 2, and ConceptMix show improvements over the IRG baseline. Removing either action or ordering guidance reduces the effectiveness of the optimized skill, while transfer from GenEval 2 to ConceptMix shows that the learned guidance retains value beyond its source dataset. These findings support structured skill optimization as a way to improve compositional generation by learning how to repair.

1 INTRODUCTION

Text-to-image generation has advanced rapidly, with leading open and closed models producing high-fidelity images that align closely with general-purpose prompts (Saharia et al., 2022; Betker et al., 2023; Esser et al., 2024; Cai et al., 2025a; Wu et al., 2025). However, satisfying multiple requirements about object counts, attributes, and spatial relations remains challenging. Recent work therefore adopts *multi-round text-to-image generation*: generate an image, verify it against the prompt, and then edit or regenerate it using feedback (Li et al., 2025b; Zhuo et al., 2025; He et al., 2026b; Jaiswal et al., 2026; Ren et al., 2026). This creates opportunities to correct errors, but progress depends on the decisions made between rounds. A VLM must determine both which repair action to use and which failure to address first. An ineffective action wastes a round, while a later repair can undo an earlier success. Figure 1 illustrates this tension: editing fixes a color error, but subsequent regeneration corrects the layout at the cost of losing that color correction.

We study *skill optimization* as a way to improve these repair decisions. The skill is the natural-language guidance supplied to the steering VLM; optimizing it allows experience from previous trajectories to inform subsequent prompts without updating the underlying models. Our starting point is the human-designed critic guidance used in the IRG pipeline (Jaiswal et al., 2026). Rather than treating that guidance as fixed, we refine it using feedback from a small collection of historical repair trajectories. This builds on work that learns reusable agent instructions from task experience (Yang et al., 2026b; Jiang et al., 2026a), and asks how such optimization should be organized for iterative image repair.

A central question is *how to design the skill space*. Treating the entire instruction document as an undifferentiated prompt obscures which repair decisions an update is intended to improve. In our setting, action choice and repair order provide a concrete structure. First, observed action effectiveness varies across failure families: in the diagnostic trajectories of Sec. 3.2, FRESH_START has

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.

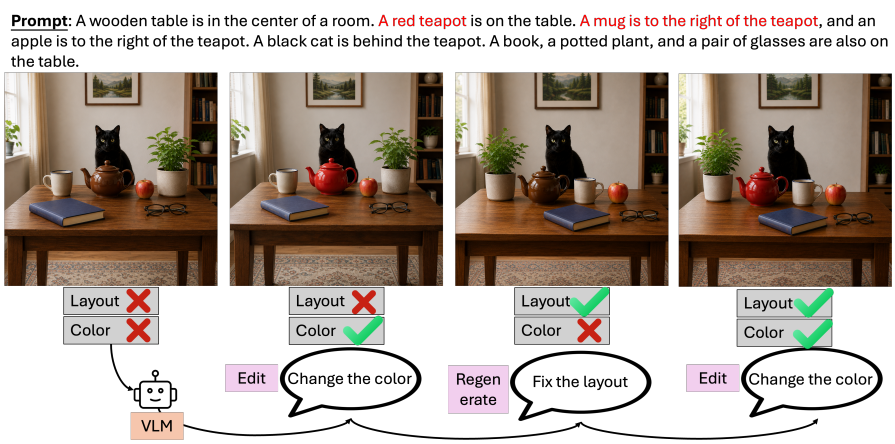


Figure 1: Repair actions can interfere across rounds. An edit corrects the teapot's color (round 1), but regeneration fixes the spatial relation while undoing that correction (round 2), requiring another edit (round 3). The example motivates guidance for both action choice and repair order to preserve progress across rounds.

a higher fix rate than CONTINUE for number errors (46.7% versus 27.4%), while CONTINUE is more effective for style errors (45.5% versus 27.3%). These are observational comparisons, but they motivate guidance that distinguishes repair requirements across families. Second, improving one constraint can disrupt another, as illustrated in Figure 1, motivating explicit repair priorities and instructions to preserve satisfied constraints. The skill space should therefore expose both how failures are repaired and how those repairs are ordered.

We organize the skill into four sections: failure diagnosis, type-conditional action guidance and preservation instructions, repair ordering, and generator-specific observations. This structure turns repair decisions into explicit targets of optimization and makes it possible to examine their contributions by removing individual guidance components. During optimization, trajectories are serialized into per-round feedback, including score changes and flags for regressions and wasted rounds. An optimizer proposes revisions to the skill, which are evaluated on held-out feedback before acceptance. At deployment, the resulting document guides the existing critic within the generation-and-editing loop; the generator, editor, and VLM weights remain frozen. The aim is to distill repair experience into reusable decision rules beyond the initial human-designed guidance.

We evaluate on GenEval (Ghosh et al., 2023a), GenEval 2 (Kamath et al., 2025), and ConceptMix (Wu et al., 2024). Relative to IRG, the overall score improves from 0.81 to 0.89 on GenEval and from 71.5 to 79.6 on GenEval 2. On ConceptMix, gains persist across all seven evaluated complexity levels, ranging from 4.3 to 7.4 percentage points in the reported scores. At $k = 5$, the complete optimized skill achieves 54.7% all-correct accuracy; removing ordering guidance reduces it to 52.2%, and removing action-choice guidance reduces it to 49.8%. These ablations support retaining both components in the skill space. A skill transferred from GenEval 2 reaches 52.7% on ConceptMix at $k = 5$, exceeding the manual baseline of 47.3% and indicating that useful repair guidance can transfer beyond its source dataset.

Our contributions are summarized as follows:

- We study skill optimization for multi-round text-to-image generation, learning reusable critic guidance from historical repair feedback while keeping the underlying models frozen.
- We design a structured skill space around action choice and repair order, together with diagnosis, preservation instructions, and generator-specific observations. Component-removal ablations demonstrate the contribution of both action and ordering guidance to the complete optimized skill.
- We evaluate on three compositional benchmarks, showing improvements over the IRG baseline across constraint types and complexity levels, and demonstrate cross-dataset transfer from GenEval 2 to ConceptMix.

2 Related Work

2.1 Test-time Scaling in Text-to-Image Generation

Text-to-image generation has advanced rapidly with diffusion-based and large multimodal generative models (Saharia et al., 2022; Betker et al., 2023; Esser et al., 2024; Cai et al., 2025a; Wu et al., 2025), shifting the main challenge from producing realistic images to faithfully realizing complex visual intent. Current models still struggle with prompts involving multiple entities, attributes, relations, and compositional constraints (Hu et al., 2023; Ghosh et al., 2023b; Li et al., 2024; Kamath et al., 2025; Wu et al., 2024). To address these failures without retraining, recent work explores test-time scaling through prompt refinement, verifier-guided selection, multimodal critics, and iterative image or prompt revision (Li et al., 2025b; Zhuo et al., 2025; Wu et al.; Jaiswal et al., 2026; Huang et al., 2025; He et al., 2026b), using intermediate feedback to guide subsequent attempts. These approaches provide mechanisms for revision, candidate selection, and use of generation history; GEMS, for example, maintains trajectory-level memory during refinement. Our focus is on learning reusable repair guidance from historical trajectories. We retain the underlying multi-round framework and optimize the critic's instructions for action choice and repair order across subsequent prompts.

2.2 Skills in Multimodal Models

Skills extend language and multimodal agents with reusable procedures and task-specific expertise beyond direct prompt conditioning, treated as modular resources that can be written, selected, routed, or accumulated from experience (Liu et al., 2024; Jiang et al., 2026b; Li et al., 2026; Zheng et al., 2025; 2026; Jiang et al., 2026a). In image generation, agentic systems use such modules to interpret instructions, retrieve external evidence, coordinate specialists, and refine outputs (Chen et al., 2025; Ye et al., 2025; Xiang et al., 2025; Wang et al., 2025; Garg et al., 2026; He et al., 2026b; Han et al., 2026), and retrieval-augmented methods ground generation in external context (Li et al., 2025a; Son et al., 2025; He et al., 2026a; Feng et al., 2026; Chen et al., 2026). These approaches span manually supplied expertise, skill retrieval, and experience-driven learning. SkillWeaver refines reusable skills through practice, while XSkill accumulates experiences and skills from multimodal trajectories. Closest to our setting, text-space optimization methods (Yang et al., 2026b) train natural-language skill documents from scored rollouts, using validation feedback to select updates for frozen agents. We build on this optimization view and investigate how to structure the skill space for multi-round image repair. Our design makes action choice, preservation instructions, and repair order explicit optimization targets, reflecting differences in action effectiveness across failure families and interference between successive repairs.

3 Method

As shown in Figure 2, we treat the VLM's instruction document — its *skill* — as the only trainable state of a frozen multi-round generation pipeline. We first formalize the pipeline and examine how action choice and repair order can affect its progress. These considerations motivate the skill structure and search space, followed by the optimization procedure.

3.1 Multi-round generation pipeline

Pipeline overview. Multi-round generation uses intermediate feedback to guide image revision (Jaiswal et al., 2026; He et al., 2026b; Ren et al., 2026). Existing systems employ image editing, as in IRG (Jaiswal et al., 2026), prompt refinement and regeneration, as in GEMS (He et al., 2026b), or failure-dependent repair choices, as in SCOPE (Ren et al., 2026). We adopt the IRG framework (Jaiswal et al., 2026) and optimize the natural-language guidance supplied to its critic. Given a prompt U, a decomposer constructs atomic binary criteria $\mathcal{C} = \{c_1, \ldots, c_n\}$, such as "are there exactly five books?" A generator $\mathcal{F}_{gen}$ produces the initial image $I_1 = \mathcal{F}_{gen}(U)$. Subsequent rounds use an editor $\mathcal{F}_{edit}$ and a VLM serving as verifier $\mathcal{M}_{ver}$ and critic $\mathcal{M}_{crit}$. The critic is conditioned on a skill document s, the only optimized component; the underlying model weights remain frozen.

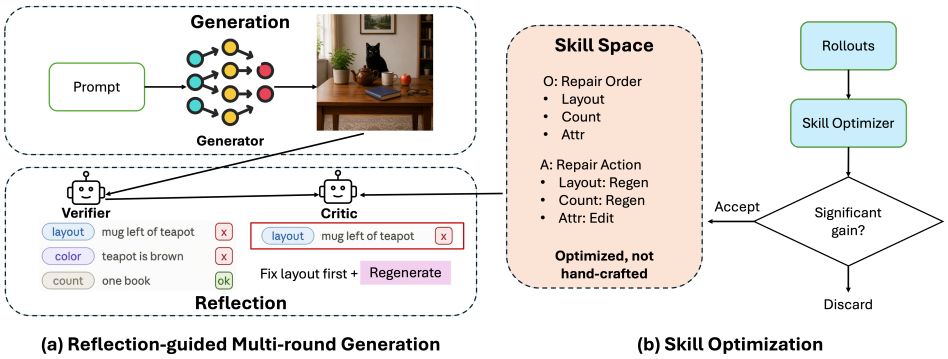


Figure 2: Overview of our framework. **(a)** Reflection-guided multi-round generation: a generator produces an image from the prompt; a VLM verifier identifies *all* failures and their types (here a layout and a color error, with count satisfied), and a VLM critic makes the two decisions the loop hinges on — *which* failure to fix first (layout, boxed) and *how* (regenerate) — before the next round. **(b)** Both decisions are governed by the skill: a natural-language document specifying the repair order over error types and the repair action per type. It is the only trainable component of the otherwise frozen pipeline: rollouts of the current skill are read by an optimizer that proposes edits, and each edit is accepted only if it yields a statistically significant validation gain — the skill is optimized, not hand-crafted.

Verification and decision. At round i, the verifier checks the current image against the criteria:

$$V_i = \mathcal{M}_{ver}(I_i, \mathcal{C}), \qquad v_{i,j} \in \{0, 1\}, \tag{1}$$

where $v_{i,j}$ indicates whether criterion c_j is satisfied. If any criterion remains unmet, the critic uses the target prompt, criteria, current image, verification results, and history $H_i = \{(I_k, V_k, a_k, e_k)\}_{k<i}$ to select an action a_i and an instruction e_i:

$$\begin{aligned}(a_i, e_i) &= \mathcal{M}_{crit}(U, \mathcal{C}, I_i, V_i, H_i;\, s),\\ a_i &\in \{\textsc{edit}, \textsc{regen}, \textsc{back}, \textsc{stop}\}.\end{aligned} \tag{2}$$

The skill guides both which failure to address next and which action to use; these decisions are expressed through the selected action and its repair instruction.

Execution and termination. EDIT applies the instruction to the current image, $I_{i+1} = \mathcal{F}_{edit}(I_i, e_i)$; REGEN generates a new image from the revised prompt, $I_{i+1} = \mathcal{F}_{gen}(e_i)$. These actions correspond to `CONTINUE` and `FRESH_START`, respectively, in our trajectory diagnostics. BACK edits the previous image, $I_{i+1} = \mathcal{F}_{edit}(I_{i-1}, e_i)$, and is unavailable at $i = 1$; STOP terminates the loop. The trajectory τ ends when all criteria pass, the critic stops, or the round budget N_{max} is exhausted. The returned image is the one with the highest fraction of satisfied criteria among the evaluated images.

3.2 Action Choice and Repair Order

The critic in Eq. 2 determines both how to repair a failure and which failure to address next. Figures 1 and 3 motivate explicitly representing these decisions in the skill search space.

Action choice depends on the failure family. Figure 3(a) shows that observed repair success varies with both the action and the failure family. FRESH_START has a higher fix rate than CONTINUE for number errors (46.7% versus 27.4%), whereas CONTINUE has a higher fix rate for style errors (45.5% versus 27.3%). These contrasting patterns motivate guidance conditioned on the failure type, rather than a uniform preference for editing or regeneration. The rates are observational: they compare actions taken at different intermediate states, not alternative actions executed on identical states.

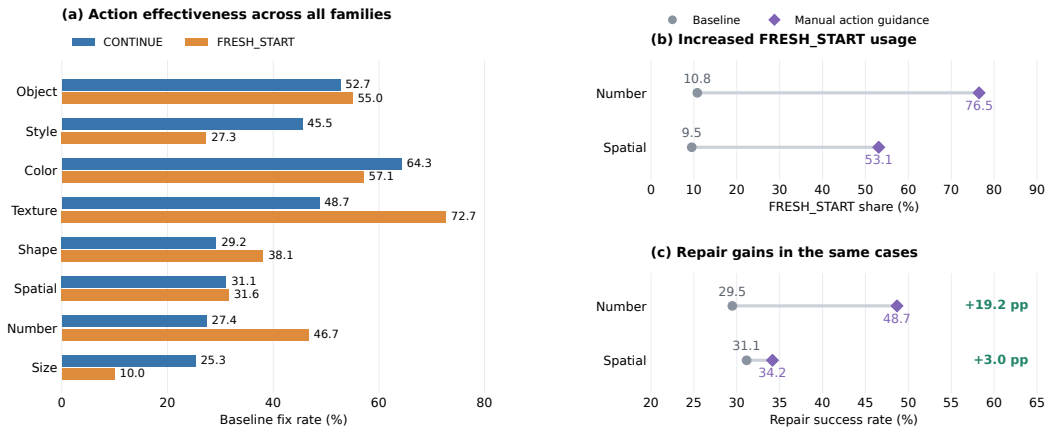


Figure 3: Action choice and repair success. (a) Baseline repair success rates for CONTINUE and FRESH_START across all eight failure families. (b) FRESH_START usage before and after manual action guidance for two selected improving families, number and spatial. (c) Their corresponding repair success rates. The comparison uses the same 501 baseline-failed case IDs; intermediate states are not paired.

Action diversity alone is insufficient. Figure 3(b–c) compares the baseline with a manual action-guidance intervention on the same 501 baseline-failed ConceptMix case IDs. This diagnostic intervention is distinct from learned skill optimization. The hint reduces the overall CONTINUE share from 96.6% to 56.2%, substantially increasing regeneration. However, the change produces uneven outcomes: the repair success rate for number improves by 19.2 percentage points, while object and style decline. A broader action distribution therefore does not ensure better repair. The skill should guide when and for which failures an action is appropriate, rather than simply encourage more regeneration.

Repair order must account for interference. Figure 1 illustrates why action guidance alone is incomplete. Editing first corrects the teapot's color, but subsequent regeneration fixes the layout while undoing that correction, requiring another edit. Thus, progress on one constraint can be lost when a later operation changes the image. Action choice determines how to attempt the current repair; ordering must also account for whether subsequent repairs may disrupt its outcome. The example motivates repair priorities and preservation requirements, without establishing that reversing this particular sequence would necessarily succeed. Together, these observations motivate a skill space that explicitly represents failure-conditioned action choice, preservation requirements, and repair priorities. We next define this space and the procedure for optimizing its guidance.

3.3 Skill Optimization

We define the skill space, specify the reward to optimize, and describe how trajectory feedback guides updates to the skill document.

Skill space. The skill is a natural-language document $s = (s_T, s_M, s_O, s_G)$. The subscripts T, M, O, and G identify four fixed sections:

- s_T (*failure types*): guidance for diagnosing and grouping unmet criteria into failure families, such as count or spatial errors;
- s_M (*repair methods*): failure-conditioned action choices and instruction wording, including which satisfied constraints an edit should preserve;
- s_O (*repair order*): priorities for addressing co-occurring failures while accounting for interference between repairs;
- s_G (*generator notes*): observations specific to the image generator $\mathcal{F}_{gen}$, recorded separately from general guidance.

This structure makes the decisions motivated in Sec. 3.2 explicit and individually editable. The document conditions the critic through Eq. 2; optimizing it does not update the underlying model weights.

Allowed edits. Let $\mathcal{S}$ be the set of documents with this section structure and token length $|s| \leq L_{\max}$, where $|s|$ counts the tokens in s and $L_{\max}$ is the maximum permitted length. An edit is represented by $\delta = (\text{OP}, q, x)$: $\text{OP} \in \{\text{ADD}, \text{DEL}, \text{REP}\}$ specifies addition, deletion, or replacement; $q \in \{T, M, O, G\}$ identifies the target section; and x specifies the affected rule text and, where applicable, the new text. We write $s \oplus \delta$ for the document obtained by applying edit δ to s. Edits remain within their declared sections. The section-specific budget b_q limits the amount of editing allowed in section q during an optimization epoch and decreases over training. Rules should refer to failure types and repair procedures rather than reproduce individual training examples.

Optimization target. For a prompt U drawn from the task distribution $\mathcal{D}$, let $\pi_s(\cdot \mid U)$ denote the distribution of trajectories produced by the pipeline under skill s. A trajectory τ contains the sequence of images, verification results, and repair decisions. Its reward $R(\tau) \in [0, 1]$ is the fraction of atomic criteria satisfied by the returned image. The optimization objective is

$$s^* = \arg\max_{s \in \mathcal{S}} \mathbb{E}_{U \sim \mathcal{D},\, \tau \sim \pi_s(\cdot|U)} \left[R(\tau) \right], \tag{3}$$

where s^* denotes a reward-maximizing skill in $\mathcal{S}$ and $\mathbb{E}$ averages over prompts and generated trajectories. This defines the optimization goal rather than a guarantee that the update procedure finds a global optimum. Criterion-level reward provides optimization feedback; benchmark-specific metrics, including all-correct accuracy, are used for evaluation.

Initialization and rollout. We initialize s_0, the starting skill, from the original IRG critic instructions (Jaiswal et al., 2026). A rollout executes the current pipeline π_s on a training prompt and serializes its trajectory into text. For each round i, the record includes the action a_i, its instruction e_i, verifier outcomes $v_{i,j}$ for each criterion j, and failure flags ϕ_i. Here ϕ_i records regressions, rounds without score improvement, and disagreements between the critic and verifier. For $i \geq 2$, the record also includes changes $(v_{i,j} - v_{i-1,j})_{j=1}^{n}$, where n is the number of criteria for that prompt; the first round supplies the initial verification state. The optimizer receives these textual records rather than the images themselves.

Reflection and update. An optimizer LLM reads a batch of trajectory records and uses the failure flags ϕ_i to identify guidance to revise. It proposes edits δ within the section budgets b_q. Following (Yang et al., 2026a), local edits are accumulated in a fast update loop and consolidated through an epoch-level meta-update. Candidate updates are evaluated on a held-out validation set before acceptance.

Validation gate. We fix generation seeds across validation comparisons to control sampling variation. Let $\bar{R}_{\text{val}}(s)$ denote the mean trajectory reward on the held-out validation set under skill s; the subscript val denotes validation. Let $\hat{\sigma}$ be the bootstrap estimate of the standard error of that mean, and let $\kappa > 0$ be a threshold multiplier fixed using a pilot study. The acceptance rule, adapting the validation-based updates of (Yang et al., 2026a), is

$$\bar{R}_{\text{val}}(s \oplus \delta) - \bar{R}_{\text{val}}(s) > \kappa \hat{\sigma}. \tag{4}$$

Thus, the candidate must exceed the current skill's validation reward by the specified margin. Only reward enters this acceptance rule; the flags ϕ_i guide revision proposals but are not additional optimization objectives.

4 Experiments

4.1 Experiment Settings

Benchmarks. We evaluate on three compositional benchmarks with complementary judges. *ConceptMix* (Wu et al., 2024) builds each prompt from a base object plus k sampled visual concepts (k=1..7, 300 prompts per level) and grades with GPT-4o answering one binary question

Table 1: Detailed Performance Breakdown on GenEval. We report the fine-grained scores across six dimensions: Single Object, Two Object, Counting, Colors, Position, and Attribute Binding, with the overall performance shown in the last column.

Base Model	Method	Single Obj	Two Obj	Counting	Colors	Position	Attr Binding	Overall
Base Models								
Nano Banana	–	–	–	–	–	–	–	0.75
GPT-Image 1	–	0.99	0.92	0.85	0.92	0.75	0.61	0.84
Seedream 4 (Seedream et al., 2025)	–	–	–	–	–	–	–	0.84
Bagel (Deng et al., 2025)	–	–	–	–	–	–	–	0.82
Z-Image (Cai et al., 2025b)	–	1.00	0.94	0.78	0.93	0.62	0.77	0.84
Inference-Time Scaling								
Qwen-Image	Original	0.95	0.87	0.58	0.77	0.36	0.47	0.66
	Rewrite	0.98	0.85	0.60	0.70	0.56	0.41	0.68
	Promptist (Hao et al., 2023)	0.91	0.83	0.54	0.74	0.42	0.54	0.66
	Search (Ma et al., 2025)	1.00	0.94	0.61	0.81	0.63	0.47	0.74
	Maestro (Wan et al., 2025)	1.00	0.93	0.70	0.73	0.62	0.58	0.76
	CRAFT (Kovalev et al., 2025)	1.00	0.90	0.74	0.82	0.70	0.58	0.79
	GEMS (He et al., 2026b)	0.98	0.94	0.73	0.84	0.72	0.54	0.79
	IRG (Jaiswal et al., 2026)	0.98	0.94	0.78	0.83	0.74	0.61	0.81
	Ours	0.98	0.94	0.91	0.91	0.85	0.77	0.89

per concept; we report *all-correct* accuracy, which credits an image only if every question passes. *GenEval* (Ghosh et al., 2023a) contains 553 short prompts over six categories (single/two object, counting, colors, position, attribute binding), judged by an object-detection pipeline with CLIP-based color classification; the overall score averages per-category accuracy over four images per prompt. *GenEval 2* (Kamath et al., 2025) extends coverage to 800 templated prompts over 40 objects, 18 attributes, and 9 relations including transitive verbs, graded by a per-atom VQA judge (Soft-TIFA); we report atom-level accuracy per dimension and strict prompt-level accuracy as the overall score.

Foundation models and generation settings. We use Qwen-Image (Wu et al., 2025) as the generator and Qwen-Image-Edit (Wu et al., 2025) as the editor throughout the multi-round loop. Images are generated at 1024×1024 resolution with 50 denoising steps, a classifier-free guidance scale of 4.0, and the default negative prompt; editing uses 50 steps at guidance 4.0 with the source image as input. The multi-round loop runs for at most 4 rounds on ConceptMix and at most 5 rounds on GenEval and GenEval 2, terminating early upon full constraint satisfaction or a STOP action; parallel-sampling baselines are compute-matched to the average number of generation and editing calls the loop actually consumes on each benchmark (3.2 / 3.4 / 3.4), with the loop's verifier selecting the highest-scoring candidate. We fix the random seed to 42 for all baseline and validation generations. For the steering VLM, we evaluate Gemini-2.5-Pro, Gemini-2.5-Flash (Comanici et al., 2025), and GPT-5.4-mini (OpenAI, 2026) (temperature 0). Unless otherwise noted, skills are trained and evaluated with the same VLM.

Skill optimization settings. We use 25 training examples to obtain trajectory feedback for skill updates and 25 validation examples to evaluate candidate updates. Final performance is evaluated on the full benchmark test sets reported above. The optimization examples serve to learn and select the skill; they are not the test sets used to report benchmark results.

4.2 Main Results

Skill optimization strengthens compositional constraint satisfaction. On GenEval (Table 1), our method achieves the highest overall score among the evaluated methods, improving from IRG's 0.81 to 0.89. Gains concentrate in counting, position, and attribute binding ($+13$, $+11$, and $+16$ percentage points), while single-object and two-object scores remain unchanged. This pattern supports optimizing repair guidance to improve how objects satisfy compositional requirements with a fixed generator and editor. Attribute binding remains the weakest dimension at 0.77, indicating that correctly assigning properties to objects remains challenging.

Gains extend to broader constraints but vary across families. On GenEval 2 (Table 2), our method again achieves the highest overall score, reaching 79.6 versus IRG's 71.5. Consistent with GenEval, object and attribute scores change little, while position, count, and verb scores improve

Table 2: Detailed Performance Breakdown on GenEval2. We report the fine-grained scores across five dimensions: Object, Attribute, Count, Position, and Verb, with the overall performance shown in the last column.

Base Model	Method	Object	Attribute	Count	Position	Verb	Overall
Base Models							
Nano Banana	–	99.0	91.4	70.1	70.2	86.7	44.6
Bagel (Deng et al., 2025)	–	92.9	75.9	55.6	50.6	57.8	23.1
Inference-Time Scaling							
Qwen-Image	Original	97.9	65.2	66.9	58.1	22.1	29.0
	Rewrite	97.3	81.1	64.0	73.8	30.3	39.8
	Promptist (Hao et al., 2023)	97.3	64.4	63.0	59.4	28.6	27.4
	Search (Ma et al., 2025)	98.3	71.1	73.9	66.1	28.2	42.1
	Maestro (Wan et al., 2025)	97.3	84.9	68.0	80.9	31.0	48.0
	CRAFT (Kovalev et al., 2025)	97.9	91.3	79.0	89.9	53.7	66.3
	GEMS (He et al., 2026b)	96.9	91.4	81.3	91.4	45.5	70.4
	IRG (Jaiswal et al., 2026)	98.4	95.3	86.4	69.8	44.8	71.5
	Ours	98.6	95.1	89.7	81.8	48.3	79.6

Table 3: Comparison of all-correct accuracy (%) across different numbers of constraints (k).

Method	$\mathbf{k=1}$	$\mathbf{k=2}$	$\mathbf{k=3}$	$\mathbf{k=4}$	$\mathbf{k=5}$	$\mathbf{k=6}$	$\mathbf{k=7}$
GEMS (He et al., 2026b)	84.3	67.7	53.7	45.5	38.7	24.4	23.7
IRG (Jaiswal et al., 2026)	88.7	77.7	64.0	53.3	47.3	32.0	28.0
Ours	**93.7**	**83.0**	**71.0**	**58.0**	**54.7**	**36.3**	**32.3**

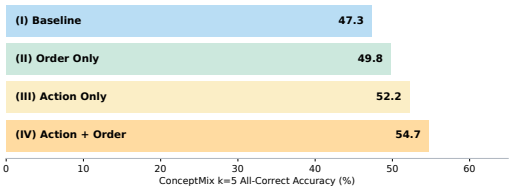
by 12.0, 3.3, and 3.5 points. The benefit therefore extends to a broader set of compositional requirements. However, GEMS retains a higher position score and CRAFT a higher verb score: the aggregate advantage does not imply uniform superiority. These family-dependent outcomes motivate explicit repair guidance, whose action and ordering components are examined below.

Improvements persist across compositional complexity levels. On ConceptMix (Table 3), adding SkillOpt to the IRG framework improves all-correct accuracy by 4.3–7.4 percentage points across all seven complexity levels. Accuracy rises from 47.3% to 54.7% at $k = 5$ and from 28.0% to 32.3% at $k = 7$, compared with GEMS's 23.7% at $k = 7$. The consistent gains support learning repair guidance beyond fixed human-designed rules within the same generation-and-refinement pipeline. Nevertheless, accuracy declines as constraints accumulate, and roughly two-thirds of the $k = 7$ prompts remain unsatisfied.

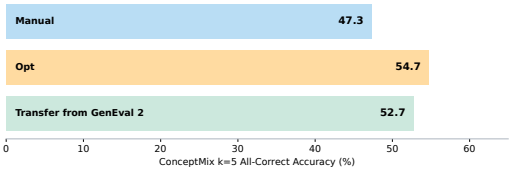
4.3 Ablation Studies

Both guidance components contribute to the optimized skill. Figure 4a compares the original, unoptimized baseline with the complete optimized skill and two component-removal variants on ConceptMix at $k = 5$. Starting from the same optimized skill, we remove repair-order guidance to obtain *Action Only*, and remove action-choice guidance to obtain *Order Only*. These variants ablate the guidance text while retaining the available repair actions; they are not separately optimized skills. The complete skill achieves 54.7% all-correct accuracy, compared with 47.3% for the unoptimized baseline. Removing order guidance reduces accuracy to 52.2%, a drop of 2.5 percentage points; removing action guidance reduces it to 49.8%, a drop of 4.9 points. Thus, both components contribute to the performance of the learned skill, with a larger observed degradation when action-choice guidance is removed. The remaining gap between each partial variant and the complete skill supports retaining explicit guidance for both which repair action to use and which failure to address first.

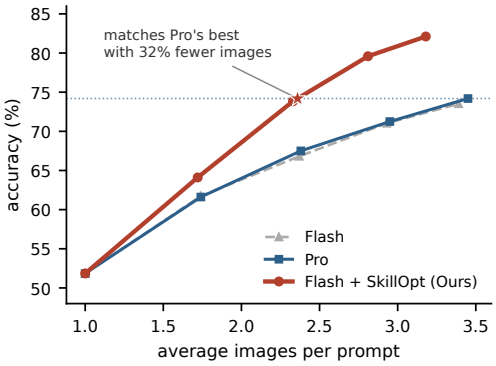
Skill optimization improves the accuracy–budget trade-off. Figure 4c contrasts upgrading the critic with optimizing its instructions. Flash and Pro remain close under the original instructions, while Flash with the skill separates from both as additional generations become available. At approximately 2.2 images per prompt, it reaches about 74% accuracy versus 68% for Pro. The marked



(a) Removing guidance from the optimized skill.



(b) Skill optimization and cross-dataset transfer.



(c) VLM capability versus skill optimization.

Figure 4: Ablation studies. (a) On ConceptMix ($k = 5$), the unoptimized baseline scores 47.3% and the complete optimized skill scores 54.7%. Removing order guidance from that skill yields Action Only (52.2%); removing action-choice guidance yields Order Only (49.8%). (b) On ConceptMix ($k = 5$), manual guidance, optimized guidance, and guidance transferred from GenEval 2 achieve 47.3%, 54.7%, and 52.7%, respectively. (c) Flash and Pro have similar accuracy–budget curves under the original instructions, while Flash with the optimized skill achieves higher accuracy beyond the initial generation. The marked point matches Pro's best plotted accuracy using 32% fewer images.

operating point matches Pro's best plotted accuracy using 32% fewer images; the optimized Flash curve subsequently reaches approximately 82%. Thus, in this comparison, changing the guidance provides a larger benefit than the Flash-to-Pro upgrade alone. Image counts measure generation budget rather than total inference cost, and the absence of an optimized-Pro curve leaves open whether model capability and skill optimization yield further combined gains.

Optimized guidance transfers across datasets. Figure 4b compares manual guidance, optimized guidance, and guidance transferred from GenEval 2 on ConceptMix at $k = 5$. Optimization raises all-correct accuracy from 47.3% to 54.7%, a gain of 7.4 percentage points. The transferred guidance reaches 52.7%, exceeding manual guidance by 5.4 points and falling only 2.0 points below Opt. Thus, the transfer setting retains approximately 73% of the improvement observed with Opt. These results suggest that the learned guidance captures repair strategies useful beyond its source dataset, while the remaining gap indicates a benefit from adapting guidance to the target setting.

5 Conclusion

We study how to design and optimize reusable skills for multi-round text-to-image generation. Our central design is to make action choice and repair order explicit within the critic's natural-language guidance, alongside failure diagnosis and preservation requirements. This gives skill optimization a structured space in which historical repair feedback can improve guidance beyond human-designed rules while the underlying models remain frozen. Experiments on GenEval, GenEval 2, and ConceptMix show improved compositional performance over IRG, with gains persisting across ConceptMix complexity levels. Removing either action or ordering guidance reduces the effectiveness of the optimized skill, supporting the contribution of both decision components. Transfer from GenEval 2 to ConceptMix further shows that the learned guidance retains value beyond its source dataset. Together, these findings support treating repair guidance as a learnable component of iterative generation. Gains remain uneven across failure families, and jointly satisfying many constraints is still challenging; improving robustness to verifier errors and testing transfer across generative backends are useful directions for future work.

AI USE STATEMENT

We used AI tools to assist with writing and translation, plotting code, and feedback on experimental design and result interpretation. AI use within the research pipeline is described in the main text. The authors are responsible for the final content.

ETHICS STATEMENT

This work studies compositional image generation using existing generative models and benchmarks. Improved instruction following may support beneficial creative applications, but may also facilitate the production of misleading or harmful synthetic imagery. The proposed skill optimization does not itself establish content safety or remove biases inherited from the underlying models. Automated visual evaluators may also produce erroneous or biased judgments, which should be considered when interpreting the results.

REPRODUCIBILITY STATEMENT

Section 3 describes the generation pipeline, structured skill space, optimization objective, and update procedure. The experiments section reports the evaluated benchmarks, model configurations, generation settings, and ablation results. Reproduction should account for the stochasticity of image generation and potential changes in externally hosted model APIs.

REFERENCES

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

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

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

Chieh-Yun Chen, Min Shi, Gong Zhang, and Humphrey Shi. T2i-copilot: A training-free multi-agent text-to-image system for enhanced prompt interpretation and interactive generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19396–19405, 2025.

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

Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

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

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

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

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, 2023b.

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

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

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

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

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

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

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

Guanyu Jiang, Zhaochen Su, Xiaoye Qu, et al. Xskill: Continual learning from experience and skills in multimodal agents. *arXiv preprint arXiv:2603.12056*, 2026a.

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

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

V Kovalev, A Kuvshinov, A Buzovkin, D Pokidov, and D Timonin. Craft: Continuous reasoning and agentic feedback tuning for multimodal text-to-image generation. *arXiv preprint arXiv:2512.20362*, 2025.

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

Chuanhao Li, Jianwen Sun, Yukang Feng, Mingliang Zhai, Yifan Chang, and Kaipeng Zhang. Ia-t2i: Internet-augmented text-to-image generation. *arXiv preprint arXiv:2505.15779*, 2025a.

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

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

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

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

OpenAI. GPT-5.4 mini Model Documentation. `https://developers.openai.com/api/docs/models/gpt-5.4-mini`, 2026.

Tianfei Ren, Zhipeng Yan, Yiming Zhao, Zhen Fang, Yu Zeng, Guohui Zhang, Hang Xu, Xiaoxiao Ma, Shiting Huang, Ke Xu, et al. Scope: Structured decomposition and conditional skill orchestration for complex image generation. *arXiv preprint arXiv:2605.08043*, 2026.

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

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

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

Xingchen Wan, Han Zhou, Ruoxi Sun, Hootan Nakhost, Ke Jiang, Rajarishi Sinha, and Sercan Ö Arık. Maestro: Self-improving text-to-image generation via agent orchestration. *arXiv preprint arXiv:2509.10704*, 2025.

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

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

Shiyu Wu, Mingzhen Sun, Weining Wang, Yequan Wang, and Jing Liu. Visualprompter: Semantic-aware prompt optimization with visual feedback for text-to-image synthesis. In *The Fourteenth International Conference on Learning Representations*.

Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. Conceptmix: A compositional image generation benchmark with controllable difficulty. *Advances in Neural Information Processing Systems*, 37:86004–86047, 2024.

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

Yifan Yang, Ziyang Gong, Weiquan Huang, Qihao Yang, Ziwei Zhou, Zisu Huang, Yan Li, Xuemei Gao, Qi Dai, Bei Liu, et al. Skillopt: Executive strategy for self-evolving agent skills. *arXiv preprint arXiv:2605.23904*, 2026a.

Yifan Yang, Ziyang Gong, Weiquan Huang, Qihao Yang, Ziwei Zhou, Zisu Huang, Yan Li, Xuemei Gao, Qi Dai, Bei Liu, et al. Skillopt: Executive strategy for self-evolving agent skills. *arXiv preprint arXiv:2605.23904*, 2026b.

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

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

YanZhao Zheng, ZhenTao Zhang, Chao Ma, YuanQiang Yu, JiHuai Zhu, Yong Wu, Tianze Xu, Baohua Dong, Hangcheng Zhu, Ruohui Huang, et al. Skillrouter: Skill routing for llm agents at scale. *arXiv e-prints*, pp. arXiv–2603, 2026.

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