arXiv:2610.18688v1 [cs.CV] 13 Oct 2026

RSIgen: Reciprocal Self-Improvement for Agentic Visual Generation

Yidi Liu[1], Haozhe Wang[3,*], Jie Xiao[2], Kai Zhu[2], Xueyang Fu[1,*], Zheng-Jun Zha[1]

[1]University of Science and Technology of China
[2]Wan Team, Alibaba Group
[3]The Hong Kong University of Science and Technology
*Corresponding authors

Experience from agentic visual generation can improve the native understanding (agent) and generation roles of a unified multimodal backbone, but task success alone does not determine who should learn. Generator updates can also make action supervision stale. We introduce RSIgen, a task-centric reciprocal self-improvement framework linking learning allocation, supervision refresh, and recursive curriculum construction. Dynamic Responsibility Routing (DRR) uses matched interventions on evidence, inputs, and supported actions to select recipients and construct compatible training targets. After generator learning, RSIgen refreshes agent supervision by comparing tested actions' relative utility with branch inputs fixed; retained roles construct the next curriculum. Across BAGEL, Qwen-Image, and Wan-Image-7B, RSIgen improves AgentGen-Bench Full Overall9 by 3.2–3.3 points over a three-cycle SearchGen adaptation with action costs. Wan controls support allocation gains at matched role-level training exposure, refresh gains at a matched adaptation budget, and benefits from updated curriculum sources while recipients continue learning. Blind evaluation on 200 real-user requests yields pooled tie-adjusted preferences of 70.0% for task completion and 64.9% for visual quality against the same baseline.

Date: October 12, 2026

1 Introduction

Agentic visual generation turns an image request into decisions about evidence acquisition, input construction, generation, and repair (Chen et al., 2026a; Zhang et al., 2026b). We study the *native-role setting*: the agent and generator are the understanding and generation roles of the same unified multimodal backbone. Task success depends on both the agent's orchestration and the generator's ability to realize the resulting inputs.

A useful action and a useful training target answer different questions. Consider an architecture poster that places buildings in the wrong chronological order. If correcting the layout instruction fixes the image while the evidence and generator stay unchanged, the correction can train the agent. A validated search-assisted image of the same buildings, paired with an evidence-free request, may instead teach stable historical knowledge to the generator. The final image alone does not identify the learning recipient or supply a compatible target.

Learning by either role reshapes its partner's next learning problem. If Direct succeeds on reserved uses of newly learned knowledge while Search remains reliable, an earlier search-required label may become stale. Shared architecture does not automatically recalibrate the agent's policy to the generator's updated capabilities. This *capability–policy calibration* aims at appropriate evidence use: volatile facts and instance-specific references retain external dependencies. Better agent inputs and evidence, in turn, change the generator's next learning opportunities (Figure 1).

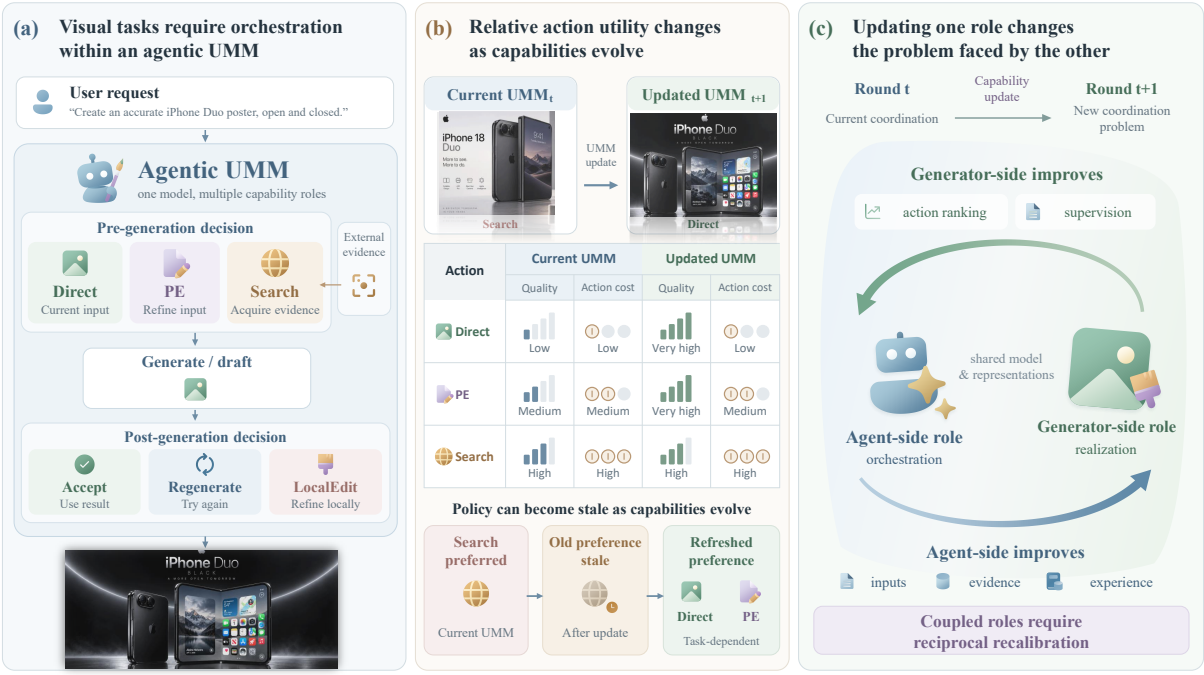

Figure 1 Reciprocal learning in agentic visual generation. (a) Agent orchestration and generator realization are roles within a unified multimodal model (UMM); PE denotes prompt enhancement. (b) Generator learning changes action utility. (c) Agent learning changes inputs and evidence for later curricula. Images and quality levels are illustrative.

SearchGen already teaches the generator and then recalibrates its reasoner (Wang et al., 2026), while GenEvolve distills contrasting trajectory experience into an evolving agent (Chen et al., 2026b). Building on these advances, we study three linked decisions: *Who should learn? Which actions remain useful after an update? What should the updated roles learn next?*

We introduce **RSIgen**, a framework for reciprocal self-improvement. Dynamic Responsibility Routing (DRR) uses matched interventions on evidence, inputs, and supported actions to select learning recipients and skill subtypes and construct compatible targets.

After generator learning, RSIgen repeats affected action comparisons with the agent, branch inputs, evidence, evaluator, and image seeds fixed. These comparisons refresh supervision according to updated relative action utility across acquisition, input construction, stopping, and supported repair. The agent then learns with the updated generator fixed, and the retained roles construct the next curriculum.

Across BAGEL, Qwen-Image, and Wan-Image-7B, RSIgen improves AgentGen-Bench Full Overall9 by 3.2–3.3 points over SearchGen-3 + costs, our three-cycle SearchGen adaptation with action costs. Blind evaluation on 200 real-user requests yields pooled tie-adjusted preferences of 70.0% for task completion and 64.9% for visual quality against this baseline.

Our contributions, with mechanism studies on Wan, are:

- **Diagnosis-guided allocation:** distinguish execution choices from learning-recipient allocation and construct compatible targets, with gains at matched role-level training exposure.
- **Supervision refresh:** re-evaluate tested pre- and post-generation actions after generator learning, improving adaptation over current-generator on-policy learning at a shared total budget.
- **Reciprocal curriculum analysis:** establish the benefit of updated curriculum sources through controls that freeze those sources while keeping recipient learning active.

2 Related Work

Agentic visual generation. Retrieval-augmented generation addresses knowledge and reference requirements through external information. Re-Imagen conditions synthesis on retrieved image–text pairs, while retrieval-augmented diffusion uses retrieved image embeddings (Chen et al., 2022; Rombach et al., 2022). Agentic systems decide how to acquire and use this information: Gen-Searcher learns search decisions with a frozen generator and image/text feedback; ToolArtist learns native multimodal tool use; Qwen-Image-Agent organizes planning, search, and memory; and Unify-Agent jointly supervises textual trajectories and image generation (Feng et al., 2026; Zhao et al., 2026; Zhang et al., 2026b; Chen et al., 2026a). RSIgen studies how experience from this process should train different capabilities and how action supervision should change when the generator also learns.

Experience reuse and generator-dependent adaptation. SearchGen trains the generator with Diffusion-DPO and recalibrates the reasoner through positive-advantage rejection finetuning with the updated generator (Wang et al., 2026). Generator-dependent adaptation is therefore already part of its design. RSIgen adds matched diagnosis for recipient-specific targets and explicit re-evaluation of acquisition and repair supervision. SearchGen-3 + costs is our local three-cycle extension of the original single DPO–RFT pass, so both systems undergo repeated adaptation. GenEvolve extracts structured experience from contrasting same-request trajectories and supplies it to a privileged teacher for token-level distillation (Chen et al., 2026b). RSIgen instead asks which role should learn from a validated experience and when its action labels should be reconsidered.

Reciprocal multimodal self-improvement. HermesFlow aligns cross-task preferences; UniRL and UniCorn construct multimodal learning tasks from generated experience; SUDER introduces dual self-rewards; and Ask, Solve, Generate alternates role-specific updates (Yang et al., 2025; Mao et al., 2025; Han et al., 2026a; Hong et al., 2025; Thawkar et al., 2026). Internal Gap revisits previously underutilized samples as capabilities improve, while Dreaming in Flow combines flow-level feedback with visual replay grounding (Han et al., 2026b; Hao et al., 2026). RSIgen couples the roles through curriculum construction and action supervision: generator learning changes useful agent decisions, and agent learning changes the experience available for later generator training. Its construction-source controls evaluate the latter direction while recipients continue learning.

Optimization and feedback. Flow matching, DPO, Diffusion-DPO, and GRPO provide the underlying optimization tools (Lipman et al., 2022; Rafailov et al., 2023; Wallace et al., 2023; Shao et al., 2024). Flow-GRPO, T2I-R1, UniGRPO, and UniRL-Zero apply outcome-driven optimization to generation or multimodal reasoning (Liu et al., 2025; Jiang et al., 2025; Liu et al., 2026; Wang et al., 2025a). Understanding-based rewards provide feedback for multimodal learning (Pan et al., 2026; Wang et al., 2025b; Jin et al., 2025). RSIgen uses completed-image rewards and trajectory costs for RL, with probes selecting curricula and validated trajectories providing action supervision. COMA assigns credit within policy optimization, whereas DRR assigns compatible learning targets before optimization (Foerster et al., 2018). SPIN and self-rewarding models also use model-generated supervision (Chen et al., 2024; Yuan et al., 2024). RSIgen updates capability roles under a fixed outer protocol, rather than modifying agent implementations or improvement mechanisms (Zhang et al., 2025, 2026a).

Scope of the comparison. SRUM freezes its understanding evaluator while training generation, and Endogenous Reprompting freezes generation while learning evaluation and reprompting (Jin et al., 2025; Tang et al., 2026). GEMS and Mind-Brush broaden agentic generation through memory, skills, search, and reasoning (He et al., 2026b,a). DAgger addresses the distribution

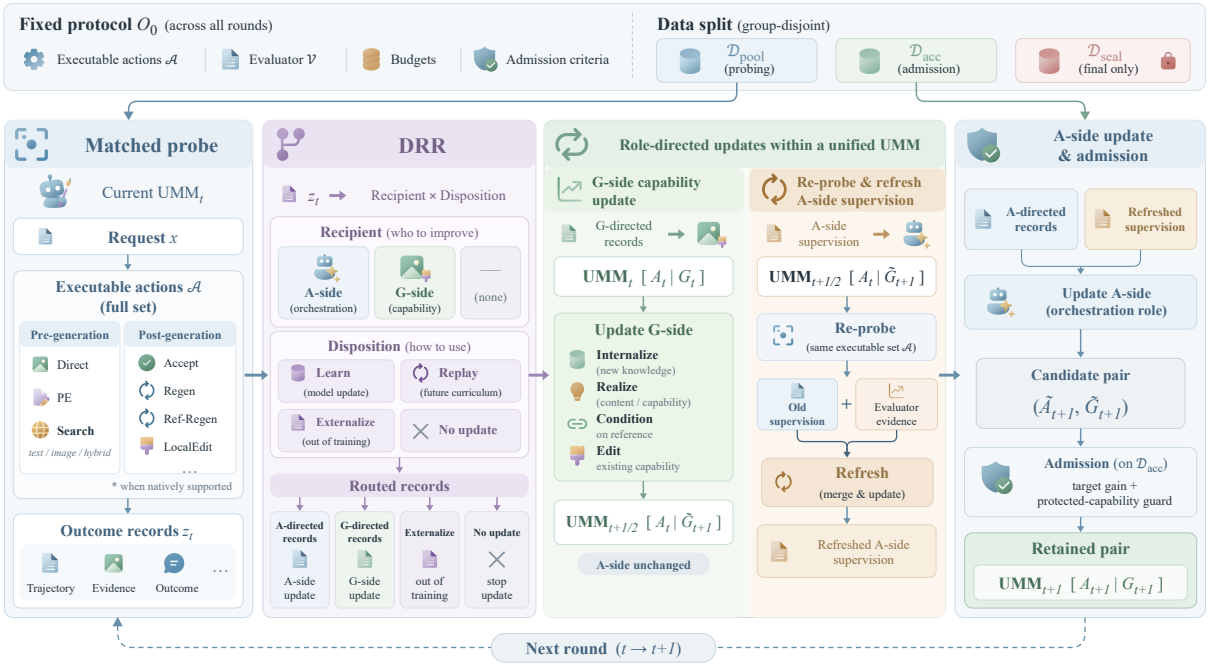

Figure 2 RSIgen's learning procedure. DRR compiles matched probes into typed curricula. Generator learning precedes re-probing and agent learning; retained roles construct the next curriculum. Pool, admission, and sealed evaluation data have separate roles. UMM denotes the unified model's understanding/orchestration and generation roles. Evaluator feedback and capability memory support training.

induced by a changing policy (Ross et al., 2011); RSIgen additionally considers changing action utility induced by generator learning.

3 Task-Centric Reciprocal Self-Improvement

RSIgen links learning-recipient allocation, supervision refresh, and recursive curriculum construction under a fixed outer protocol (Figure 2).

3.1 Task, capability roles, and outcome rewards

At round t, $S_t = (A_t, G_t, M_t)$ comprises the native understanding/agent and generation roles of a unified multimodal backbone, plus outer-loop capability memory. Each update changes its recipient while holding the other role's effective function fixed. Parameters follow the backbone; Wan uses full-parameter updates to the trained role. For request x,

$$x \xrightarrow{A_t:\text{Think}} \text{decision} \xrightarrow{A_t:\text{Action}} (p, e) \xrightarrow{G_t} I \xrightarrow{\mathcal{E}_0} y. \tag{1}$$

The agent supplies prompt p and optional evidence/reference e; $\mathcal{E}_0$ scores the completed image. The agent observes images and tool results, but evaluator scores and M_t remain training-side resources.

Pre-generation actions are Direct, PromptEnhance, TextSearch, ImageSearch, and HybridSearch. Post-generation actions are accept, regenerate, reference-conditioned regenerate, and LocalEdit where natively supported. Conditioning and editing actions require native support.

Both RL stages use completed-image outcomes. With request-applicable checks $\mathcal{J}(x)$ and normalized weights w_j,

$$R_G(I, x, e) = \sum_{j \in \mathcal{J}(x)} w_j v_j(I, x, e) - \lambda_{\text{viol}} c_{\text{viol}}(I, x, e), \tag{2}$$

$$R_A(\tau, x) = R_G(I_{\text{final}}, x, e_{\text{final}}) - \lambda_s C_s(\tau) - \lambda_g C_g(\tau) - \lambda_l C_l(\tau). \tag{3}$$

Figure 3 Schematic routing decisions. (a) Holding evidence and generator fixed, a corrected layout repairs a chronology error and yields an agent-construction candidate. (b) A module mismatch under checked input yields a generator-realization candidate when a compatible target is available. (c) Learning, externalization, and deferral depend on the dependency's properties and branch outcomes.

Checks v_j evaluate image requirements; c_{viol} penalizes predefined violations. Factual/reference targets are anchored independently of agent-selected evidence. C_s counts searches, C_g generations/edits, and C_l horizon-normalized decisions. Probe scores construct curricula; reference/KL regularization belongs to optimization.

The objective is held-out quality $Q(A, G)$ under training/inference budgets and a protected-capability guard. Protocol Ω_0 fixes actions, compilation, update stages, rewards, and admission. Pool $\mathcal{D}_{\text{pool}}$, admission $\mathcal{D}_{\text{acc}}$, and sealed evaluation $\mathcal{D}_{\text{seal}}$ remain source/request/artifact-group disjoint, including derived curricula and replay. Terminal reward alone cannot allocate learning; matched probes guide this decision.

3.2 Who should learn: diagnosis and curriculum compilation

Budgeted probes cover anchors, changed capability regions, uncertain comparisons, and under-covered requests. Matched interventions compare Direct, prompt enhancement, search modalities, Oracle evidence, and supported repairs:

$$q_t(x, a) = \frac{1}{K} \sum_{k=1}^{K} s(I_t^{(k)}(x, a), x), \tag{4}$$

where s scores the completed image. Comparisons fix the request, peer versions, evaluator, unaffected inputs, and image seeds. Each branch's retrieval artifacts are frozen for replay. Empty searches retain their outcomes and costs; unexecuted branches remain unmeasured.

Input checks validate request fidelity, evidence coverage, entity/reference binding, and interface compatibility. Oracle evidence is validated independently of image reward and serves diagnosis, not executable agent labels. Failed checks trigger reconstruction or no-update.

Dynamic Responsibility Routing (DRR) maps the versioned probe record z_t to a recipient, skill

subtype, and disposition: learn, replay, externalize, or no-update (Figure 3; Table 1). Oracle repair of inadequate retrieval suggests agent acquisition; verified prompt corrections suggest agent construction. Residual failure under valid, requirement-complete inputs suggests generator realization. These are operational learning hypotheses, not unique causal identifications.

Successful search also offers a learning use: Direct failure on stable reusable knowledge with validated Search success and no material Oracle gain suggests generator internalization. Volatile facts and instance-specific references retain external dependencies. Refresh tests whether learning supports no-search supervision.

For opportunity $o = (x, z_t)$, let $d_{o,r}$ be a compatible input–target pair for recipient/subtype $r \in \mathcal{R}(o)$. DRR selects $\rho(o) \in \mathcal{R}(o)$ or no-update and compiles $\mathcal{C}_r = \{d_{o,r} : \rho(o) = r\}$. Generator SFT requires a validated admissible image or successful generation target. Pool expansion independently validates each added example's inputs, subtype, and target.

In the poster example, agent construction pairs the request, evidence, and original input with the executed correction. Internalization pairs an evidence-free student request with a validated assisted image:

$$(p^{\mathrm{ctx}}, e^*) \longrightarrow I^*, \qquad G : (p^{\mathrm{D}}, \varnothing) \longrightarrow I^*. \tag{5}$$

The student preserves the request and entity identity but excludes retrieved facts and descriptions. Reference-following records retain their conditions. Generator curricula cover internalization, realization, reference following, and supported editing. Compatible flow generators use SFT/conditional flow matching and then Flow-GRPO for open-ended verifiable skills; internalization uses SFT only.

Routing precedence and admissible targets. DRR resolves upstream input failures before considering downstream realization. Regressions replay validated targets for their original recipient and subtype. Volatile dependencies remain external, while unresolved or mixed diagnoses are deferred. Table 1 summarizes the resulting routes.

Table 1 From probe outcomes to typed curricula. A route is admitted only when its inputs and training target are valid for the recipient. Oracle serves diagnosis and is excluded from executable agent labels.

Probe pattern	Recipient/subtype	Compiled training object
Oracle repairs inadequate retrieval	Agent: acquisition	Valid query and evidence-selection trajectory
Corrected input repairs the output	Agent: construction	Executed corrected prompt or condition
Search repairs a stable-knowledge failure	Generator: internalization	Evidence-free prompt and validated assisted image
Failure persists with sufficient valid input	Generator: realization	Compatible rendering, reference, or edit pair
Previously mastered capability regresses	Original recipient/replay	Validated prior target
Volatile dependency or unresolved diagnosis	Externalize or no-update	Retained access or further probes

3.3 Which actions remain useful: post-generator refresh

Post-generator probes cover affected regions, uncertain comparisons, and anchors with the agent, branch inputs, evidence, source images, continuation, evaluator, and seeds fixed. On common tested executable branches $\mathcal{A}_x^P$, empirical terminal values define a tolerance-aware preferred set:

$$\begin{aligned} \widehat{V}_t(x, a) &= \frac{1}{K} \sum_{k=1}^{K} R_A(\tau_{t,k}^a, x), \\ \widehat{\mathcal{B}}_t^{\epsilon}(x) &= \left\{ a \in \mathcal{A}_x^P : \widehat{V}_t(x, a) \geq \max_{b \in \mathcal{A}_x^P} \widehat{V}_t(x, b) - \epsilon \right\}. \end{aligned} \tag{6}$$

Here $\tau^a_{t,k}$ forces branch a with the common continuation. Replacing G_t with $\widetilde{G}_{t+1}$ under matched conditions defines $\widehat{\mathcal{B}}^{\epsilon}_{t+1/2}$. Old supervision is *stale* when its action becomes invalid or materially inferior under the updated generator. A preferred-set change need not invalidate a near-optimal old label; ambiguous comparisons can remain unlabeled. Rankings depend on differential action improvement.

An internalization-based no-search target additionally requires evidence-free success on reserved new uses of the same knowledge unit, a small remaining Search–Direct gap, and preserved assisted success. Failure of both branches never licenses this target. Refreshed supervision covers search, query/reference selection, prompt construction, stopping, and supported repair. The agent learns from validated executed trajectories through SFT, then GRPO with the updated generator fixed.

The effect of a generator update on rankings depends on differential gains. Define $\gamma_t(x,a) = \widehat{V}_{t+1/2}(x,a) - \widehat{V}_t(x,a)$. Then

$$\begin{aligned}\widehat{V}_{t+1/2}(x,a) - \widehat{V}_{t+1/2}(x,b) = \widehat{V}_t(x,a) - \widehat{V}_t(x,b) \\ + \gamma_t(x,a) - \gamma_t(x,b).\end{aligned} \tag{7}$$

Equal gains preserve the pairwise difference. A preferred set can change while an old label remains competitive, distinguishing an action shift from label staleness. Task success is evaluated separately from the relative ranking of branches.

3.4 What the roles learn next: recursive construction and admission

Writing $z_{t+1/2}$ for post-generator probes, an update attempt is

$$\begin{aligned}\widetilde{G}_{t+1} &= \mathcal{U}_G(G_t; \mathcal{C}_{A\to G}(z_t, M_t, \mathcal{D}_{\text{pool}})), \\ \widetilde{A}_{t+1} &= \mathcal{U}_A(A_t; \mathcal{C}_{G\to A}(z_{t+1/2}, M_t, \mathcal{D}_{\text{pool}}), \widetilde{G}_{t+1}).\end{aligned} \tag{8}$$

Here $G \to A$ denotes supervision compiled from generator behavior. Memory stores versioned outcomes for diagnosis, expansion, and replay. After admission, retained roles construct new experience through the fixed collector $\mathcal{P}_{\Omega_0}$:

$$z_{t+1} = \mathcal{P}_{\Omega_0}(A_{t+1}, G_{t+1}, M_{t+1}; \mathcal{D}_{\text{pool}}) \longrightarrow \mathcal{C}^{t+1}_{A\to G} \longrightarrow \widetilde{G}_{t+2}. \tag{9}$$

These *construction peers* supply updated evidence, inputs, and outcomes for subsequent curricula. Source controls freeze them while recipient learning and current-partner RL remain active, testing their combined contribution to content, target validity/yield, and supervision freshness (Section 5.5).

Admission compares target quality and protected generation/understanding capabilities with the retained state on identical held-out subsets, separate from curricula and rewards. Rejected attempts retain the state and consume their budget. Final evaluation compares retained endpoints with initialization and matched controls.

Protected-capability admission. For each candidate $\widetilde{S}_{t+1}$, let $\mathcal{D}^G_{\text{protect},t}$ and $\mathcal{D}^A_{\text{protect},t}$ be held-out generation and understanding subsets. Candidate and retained components are evaluated on identical items and settings within the attempt. For protected capability c, the guard is

$$\Delta_{c,t} = Q_{c,t}(\widetilde{S}_{t+1}) - Q_{c,t}(S_t) \geq -\epsilon_c, \qquad \text{for every protected } c. \tag{10}$$

Here $Q_{c,t}$ evaluates the relevant component: generation checks keep raw prompts fixed without agent assistance, and understanding checks invoke the agent's native understanding role without generation. Admission also requires the target-quality criterion. Protected items do not enter curriculum construction, training rewards, or the sealed final evaluation. The comparator is the current retained state S_t, and rejection leaves that state unchanged.

The tolerance ϵ_c applies locally to the protected subset at each admitted update, relative to the current retained state. Successive accepted steps can accumulate their permitted losses.

4 Experimental Setup

4.1 Backbones, benchmark, and comparison methods

We evaluate BAGEL, Qwen-Image, and Wan-Image-7B, with mechanism studies on Wan. Qwen-Image uses Qwen2.5-VL as its condition encoder (Wu et al., 2025); Wan uses full-parameter updates to the trained role. Within each backbone, trained comparisons share initialization, tools, resource allowances, corresponding seeds, and three update attempts. Direct and prompt enhancement provide untrained inference references.

Agentic initialization is the starting tool-using pair. Agent-only adaptation follows Gen-Searcher with a fixed generator and RSIgen's completed-image outcome and trajectory-cost contract; the original Gen-Searcher also uses final-text feedback (Feng et al., 2026). SearchGen-3 + costs repeats Diffusion-DPO and positive-advantage rejection finetuning for three cycles (Wang et al., 2026). Joint outcome RL trains both roles from terminal outcomes. Uniform reciprocal learning shares RSIgen's role-specific optimization stages and refresh, while changing curriculum allocation. Uniform and failure-locus comparisons evaluate the curriculum package; one-sided updates test which roles need to learn.

Benchmark and metrics. AgentGen-Bench contains 751 requests. We follow its released Overall9 aggregation and evaluator (Wang et al., 2026); the original paper called it SearchGen-Bench. Overall9 averages Checklist, Rubric, Prompt Faithfulness, Image Quality, Text Rendering, AI Naturalness, Composition, Physical Plausibility, and Visual Reference, excluding Text Reference. NoSearch (NS) and Search-Intensive (SI) are the benchmark strata; Full is the global mean over each model's scored coverage. Applicable metric denominators differ, so averaging displayed components or nominal-count weighting of rounded NS/SI scores need not reproduce Full. Public rows have model-dependent coverage: Nano Banana Pro reports 738/751 scored requests and GPT-Image-2 718/751, whereas the published BAGEL and Qwen-Image rows cover 751/751. For example, the former's published Full 65.0 differs from 64.9 obtained by nominal-count weighting of its rounded stratum scores. Evaluation references are distinct from runtime retrieval inputs and training targets. Dataset and public-result definitions follow the benchmark release and project results.

4.2 Matched mechanism studies

The allocation study uses 4,000 opportunities and matches role-level exposure: 1,600 agent items with 3.2M SFT and 2.4M RL tokens, and 2,000 generator items with 4,000 SFT and 7,200 RL image presentations. This comparison is separate from the original curriculum-package ablation.

The action-refresh diagnostic uses 200 held-out Test-mini requests, separate from training, tuning, checkpoint selection, and the human-evaluation requests. It compares (G_0, G_1), (G_1, G_2), and (G_2, G_3) under fixed prompts, evidence, action packages, and paired image seeds, with 200 canonical post-states per stage. Action probing uses eight images per action/version. Stale-label and refresh continuation share agent initialization, training requests, the updated rollout generator, and realized token/training-compute budgets. A separate Stage-2 comparison additionally includes current-generator on-policy learning and matches preparation, rollout, and optimization within a common 100-unit budget: $c = 100C/B_{\text{adapt}}$, so one unit is 1% of the single-stage budget.

Construction-source controls keep recipient learning and current-partner RL active, while freezing the agent, generator, or both as curriculum sources. Every arm uses 1,800 agent and 2,200 generator records per attempt, totaling 5,400 and 6,600. Separate role evaluation uses 200 requests and four image seeds per request/state; agent evaluation fixes G_0, and generator evaluation distinguishes direct and assisted inputs.

4.3 Human judgments, statistical units, and resource accounting

Human evaluation compares RSIgen with SearchGen-3 + costs on the same 200 real-user requests across all three backbones. Three independent blind ratings per pair assess task completion and visual quality, giving 600 judgments per backbone and 1,800 pooled judgments per dimension. Preference credits a tie as half a win. Repeated ratings and reuse across backbones do not increase the number of distinct requests.

Main benchmark results aggregate multiple training seeds. The continuation diagnostic uses 10,000 paired request-cluster bootstrap replicates, retaining all stages, matched branches, and policy arms for each sampled request. Conditional metrics recompute their numerators and denominators in each resample before the equal-weight stage mean. The resulting percentile 95% intervals quantify request variation conditional on checkpoints and frozen evidence, separately from variation across training runs. Image seeds are generation replicates; sequential stages describe progression. Searches and generations/edits are counted per request, with different calls potentially having different compute and latency costs. Failed calls and rejected attempts consume their assigned budgets.

5 Results and Analysis

5.1 Benchmark quality and within-backbone comparisons

RSIgen reaches AgentGen-Bench Full scores of 39.9, 50.3, and 47.5 on BAGEL, Qwen-Image, and Wan-Image-7B (Table 2). Matched training comparisons in Table 3 show gains of 3.2, 3.2, and 3.3 points over SearchGen-3 + costs, and 3.7, 3.6, and 3.7 over uniform reciprocal learning. RSIgen also exceeds joint outcome RL on every backbone. Improvements over SearchGen-3 + costs in Table 3 are larger on SI (3.4–3.6 points) than NS (0.9–2.2).

Table 2 AgentGen-Bench overview (0–100). NS/SI: NoSearch/Search-Intensive; Full: Full-slice Overall9. Colors mark best and second-best within each group. Public rows have model-dependent coverage; matched training comparisons appear in Table 3.

	Checklist		Visual Ref.		Image Quality		Overall9		
Model	NS	SI	NS	SI	NS	SI	NS	SI	Full
Commercial Models									
GPT-Image-2	**84.8**	**74.4**	**70.9**	**70.4**	**73.4**	**77.4**	**78.9**	**75.5**	**76.0**
Grok-Imagine-2.0-Low	83.0	73.1	68.7	68.3	72.5	76.9	76.9	74.9	75.2
Qwen-Image-3-Max	79.1	65.8	64.0	59.6	70.3	71.1	73.7	67.6	68.4
Nano Banana Pro	78.4	60.7	63.0	55.4	70.7	68.1	73.3	63.6	65.0
Qwen-Image-2	76.2	49.9	60.3	42.0	70.2	62.5	71.1	54.1	56.4
Generation Only									
Mage-Flow	**64.2**	23.0	**47.1**	16.2	**64.7**	**38.7**	**61.2**	**28.4**	**32.8**
Flux.2-Klein-9B	63.7	**23.8**	41.1	**16.7**	63.2	36.1	58.5	26.9	31.1
Flux.2-Klein-4B	55.8	20.1	35.3	12.7	59.7	33.0	52.3	23.9	27.7
Unified MLLMs									
SenseNova-U1	68.1	29.3	51.1	21.0	64.3	39.5	62.8	31.6	35.7
OmniGen2	49.1	16.4	30.6	9.5	56.4	29.2	47.4	20.4	24.0
Show-o2	32.7	8.1	19.9	3.4	47.2	30.2	34.4	17.5	19.8
BAGEL	51.5	18.7	32.1	14.2	56.2	30.4	49.1	22.2	25.8
BAGEL + RSIgen	63.4	36.4	45.6	27.4	61.4	43.3	60.4	36.8	39.9
Qwen-Image	74.9	25.0	60.1	17.7	69.2	37.9	70.0	28.6	34.1
Qwen-Image + RSIgen	**77.2**	**47.3**	**63.4**	**37.4**	**70.7**	**52.3**	**73.2**	**46.8**	**50.3**
Wan-Image-7B	65.3	22.3	49.3	16.3	65.1	37.2	62.3	27.4	32.1
Wan-Image-7B + RSIgen	70.4	44.3	55.4	35.4	67.2	50.2	67.3	44.4	47.5

Table 4 breaks down the three backbones and their RSIgen variants by capability. On SI, Checklist improves by 17.7–22.3 points and Text Rendering by 26.2–30.2, compared with 12.9–14.4 for Image

Table 3 Matched within-backbone comparisons on AgentGen-Bench. Trained methods share initialization and resource allowances within each backbone; scores aggregate multiple training seeds. NS/SI denote NoSearch/Search-Intensive, and Full is the global Overall9 mean. Direct and prompt enhancement are untrained Wan references; dashes indicate results not reported in this comparison.

	BAGEL			Qwen-Image			Wan-Image-7B		
Method	NS	SI	Full	NS	SI	Full	NS	SI	Full
Direct	–	–	–	–	–	–	62.3	27.4	32.1
Prompt enhancement	–	–	–	–	–	–	64.1	31.2	35.6
Agentic initialization	54.2	25.4	29.3	70.6	35.3	40.0	64.3	33.4	37.5
A-only / Gen-Searcher adapt.	56.3	29.4	33.0	71.2	39.4	43.6	65.2	36.3	40.2
SearchGen-3 + costs	58.2	33.4	36.7	72.3	43.2	47.1	66.2	40.8	44.2
Joint outcome RL	57.6	32.2	35.6	71.7	42.4	46.3	65.7	39.8	43.3
Uniform reciprocal	58.6	32.8	36.2	72.1	42.8	46.7	66.3	40.3	43.8
RSIgen	**60.4**	**36.8**	**39.9**	**73.2**	**46.8**	**50.3**	**67.3**	**44.4**	**47.5**

Quality. Improvements therefore span task requirements and text rendering as well as visual quality. These before/after backbone comparisons complement the matched training controls in Table 3; component scores use the benchmark's applicable denominators.

Table 4 Component scores for the three backbones and their RSIgen variants. Ch: Checklist; Ru: Rubric; Pr: Prompt Faithfulness; IQ: Image Quality; Tx: Text Rendering; Nat: AI Naturalness; Co: Composition; Ph: Physical Plausibility; VR: Visual Reference. Overall9 appears in Table 2.

Model	Ch	Ru	Pr	IQ	Tx	Nat	Co	Ph	VR
A. NoSearch (NS)									
BAGEL	51.5	49.2	37.7	56.2	16.1	48.8	62.8	58.5	32.1
BAGEL + RSIgen	63.4	62.1	58.3	61.4	50.6	60.2	68.4	73.2	45.6
Qwen-Image	74.9	72.0	65.5	69.2	61.1	61.7	77.0	78.6	60.1
Qwen-Image + RSIgen	77.2	76.1	72.4	70.7	68.4	69.2	80.1	81.3	63.4
Wan-Image-7B	65.3	63.2	57.4	65.1	54.3	57.2	73.4	75.1	49.3
Wan-Image-7B + RSIgen	70.4	69.2	65.3	67.2	60.4	63.2	76.1	75.4	55.4
B. Search-Intensive (SI)									
BAGEL	18.7	17.7	12.4	30.4	2.2	25.0	32.4	38.1	14.2
BAGEL + RSIgen	36.4	35.2	31.4	43.3	29.5	37.2	44.3	44.8	27.4
Qwen-Image	25.0	23.8	17.4	37.9	9.2	29.9	41.0	45.5	17.7
Qwen-Image + RSIgen	47.3	46.2	42.4	52.3	39.4	45.2	55.1	55.7	37.4
Wan-Image-7B	22.3	21.2	15.4	37.2	10.3	30.2	41.3	51.2	16.3
Wan-Image-7B + RSIgen	44.3	43.2	39.4	50.2	36.5	43.2	53.4	53.2	35.4

5.2 Human evaluation on real-user requests

RSIgen receives more wins than losses for both dimensions on every backbone (Table 5). Tie-adjusted task-completion preference ranges from 68.5% to 71.4%, and visual-quality preference from 63.5% to 66.1%. The pooled scores are 70.0% and 64.9%, respectively.

The full counts distinguish raw wins from tie-adjusted preference: task completion gives $100(972 + 577/2)/1800 = 70.0\%$, versus a raw win proportion of 54.0%; visual quality gives $100(828 + 679/2)/1800 = 64.9\%$, versus 46.0% raw wins. Unknown judgments are zero throughout. Equal judgment counts make pooled preferences equal to the unrounded, equally weighted backbone averages. These are descriptive point estimates on 200 distinct requests, without request-cluster confidence intervals.

5.3 Diagnostic allocation and quality–tool-use tradeoffs

Table 6 tests the Wan curriculum package and updated roles. Agent-only and generator-only updates reach 40.2 and 41.7, compared with RSIgen's 47.5. Uniform reciprocal and failure-

Table 5 Blind human evaluation against SearchGen-3 + costs on the same 200 real-user requests. Each backbone has 600 judgments per dimension (three ratings per pair); Overall pools 1,800 judgments on those same requests. W/T/L is from RSIgen's perspective; preference is $100(W + 0.5T)/(W + T + L)$. Unknown judgments are zero.

	Task completion		Visual quality	
Backbone	W / T / L	Pref.	W / T / L	Pref.
BAGEL	312 / 198 / 90	68.5%	264 / 234 / 102	63.5%
Qwen-Image	324 / 194 / 82	70.2%	276 / 228 / 96	65.0%
Wan-Image-7B	336 / 185 / 79	71.4%	288 / 217 / 95	66.1%
Overall	972 / 577 / 251	**70.0%**	828 / 679 / 293	**64.9%**

locus-only routing reach 43.8 and 44.1. Removing post-generator re-probing lowers Full to 45.0; removing capability memory or pool expansion lowers it to 45.9 or 45.3. These comparisons evaluate the combined curriculum recipe, including which opportunities and compatible targets reach each role.

Table 6 Wan ablations on AgentGen-Bench. Full measures quality; S and G/E count inference searches and generations/edits per request.

Variant	Full ↑	S	G/E
Full RSIgen	47.5	**1.16**	**1.30**
A-only updates	40.2	1.59	1.45
G-only updates	41.7	1.76	1.51
Uniform reciprocal	43.8	1.41	1.42
Failure-locus-only routing	44.1	1.51	1.44
No post-G re-probe	45.0	1.53	1.39
No capability memory	45.9	1.31	1.36
No full-pool expansion	45.3	1.29	1.35
No trajectory costs	**48.0**	1.86	1.56

Allocation under matched role exposure. The separate allocation study holds agent and generator training volumes fixed (Section 4). In Table 7, DRR reaches Full 47.4, gaining 1.4 over compatible-uniform (46.0) and 1.1 over failure-locus (46.3). Searches and generations/edits decrease from 1.29/1.37 to 1.21/1.32 relative to compatible-uniform. Thus the gain persists at matched role-level item, token, and image exposure.

Table 7 Wan allocation comparison with matched role-level training exposure. Δ is relative to compatible-uniform. These results are separate from the curriculum-package ablations.

Strategy	Full ↑	Δ Full	Searches ↓	Gen./edits ↓
Compatible-uniform	46.0	0.0	1.29	1.37
Failure-locus	46.3	+0.3	1.27	1.36
DRR	**47.4**	**+1.4**	**1.21**	**1.32**

All arms use 4,000 opportunities. Agent: 1,600 training entries, 3.20M SFT tokens and 2.40M RL tokens. Generator: 2,000 entries, 4,000 SFT and 7,200 RL image presentations. Calls are per request.

Inference quality and tool use. The no-cost variant reaches Full 48.0 versus RSIgen's 47.5, with searches rising from 1.16 to 1.86 and generations/edits from 1.30 to 1.56 per request: 60.3% and 20.0% increases from the displayed averages. Trajectory costs trade 0.5 quality points for fewer inference calls. RSIgen also achieves higher quality with slightly fewer calls than SearchGen-3 + costs across all three backbones (Table 8). Call counts measure tool use rather than wall-clock latency.

Table 8 Inference-call averages for SearchGen-3 + costs and RSIgen. Counts charge search and generation/edit calls; different calls can have different latency and compute costs.

Backbone	Method	Full	Searches	Generations/edits
BAGEL	SearchGen-3 + costs	36.7	1.21	1.32
BAGEL	RSIgen	**39.9**	**1.18**	**1.31**
Qwen-Image	SearchGen-3 + costs	47.1	1.21	1.31
Qwen-Image	RSIgen	**50.3**	**1.13**	**1.28**
Wan-Image-7B	SearchGen-3 + costs	44.2	1.26	1.33
Wan-Image-7B	RSIgen	**47.5**	**1.16**	**1.30**

5.4 Action utility shifts and post-generator adaptation

Matched branches and evaluated populations. The diagnostic distinguishes changes in action utility from the policy's response to new supervision. Old/new generator comparisons fix the agent, enhanced prompts, search evidence, references, source images, continuation, evaluator, and paired seeds. Search branches retain their artifacts and logical call costs, including when retrieval yields no usable evidence. A canonical post-state uses the same source image for both checkpoints and continuation arms. Accept returns that image; regeneration and reference-conditioned regeneration use the fixed input packages; LocalEdit is included only when the localized repair is identifiable and supported by both checkpoints. Edit ineligibility removes the action, while retaining the state in the other post-action metrics.

There are 200 evaluated pre-generation requests and 200 post-states per stage. Across 600 stage–request observations, text and image search yield no usable evidence in 162 (27.0%) and 204 (34.0%) cases, with zero infrastructure failures after retries. LocalEdit is eligible for 126, 121, and 116 states across the stages. Empty retrieval outcomes remain included in action values. End-to-end Mini O9 evaluates the separate complete policy workflow, including its chosen queries and induced state distribution.

Generator learning changes relative action utility. Table 9A reports preference shifts in 22.2% of pre-generation and 20.8% of post-generation decisions; old labels are stale in 17.0% and 15.7%. A shift changes the preferred action set, whereas staleness makes the old supervision action invalid or materially inferior. Pre-generation shift counts are 49, 44, and 40, and stale counts are 38, 34, and 30; post-generation shift counts are 45, 42, and 38, and stale counts are 34, 31, and 29. Headline rates average the stage proportions.

Search-to-Direct accounts for $69/133 = 51.9\%$ of pre-generation shifts; Regenerate-to-LocalEdit accounts for $49/125 = 39.2\%$ of post-generation shifts (Table 10). Accept's image quality is checkpoint-invariant, so shifts involving Accept reflect changes in the relative benefit of intervention.

Measured branch examples. Two numerical cases, separate from the visual requests in Figure 7, illustrate differential action gains. In one, Direct quality rises from 0.38 to 0.75 while Search remains at 0.82–0.83, narrowing the Search–Direct gap from 0.44 to 0.08. In a repair case, LocalEdit's quality-minus-incremental-cost value rises from 0.54 to 0.77 and overtakes regeneration (0.62 to 0.64); Accept remains at 0.52 and reference-conditioned regeneration changes from 0.59 to 0.64.

Refreshed supervision improves continuation. Stale-label and refresh arms share initialization, requests, the updated rollout generator, and realized training compute. Refresh reduces pre/post regret from 5.37/4.60 to 3.30/3.00 and increases three-stage Mini O9 from 47.37 to 50.20. Table 9B also reports complementary alignment measures: Pre Track conditions on pre-generation shift cases, Post Agree uses all evaluated post-states, and LocalEdit precision/recall condition on

Table 9 Wan action refresh on Test-mini. A: branch changes over 200 requests and 200 canonical post-states per stage, with 95% intervals in brackets. B: matched continuation with the updated generator; the final column gives refresh-minus-stale 95% intervals. All estimates average three stages. Intervals use 10,000 paired request-cluster bootstrap replicates and condition on checkpoints and frozen evidence.

A. Action changes after generator updates

Decision	Evaluated per stage	Preference shift (%)	Stale labels (%)
Pre-generation	200	22.2 [18.1, 26.5]	17.0 [13.4, 20.9]
Post-generation	200	20.8 [16.9, 25.1]	15.7 [12.3, 19.3]

B. Policy adaptation and paired differences

Metric	Old A_j	Stale labels	Refresh	Δ: 95% CI
Pre regret ↓	7.63	5.37	**3.30**	[−2.60, −1.60]
Post regret ↓	6.10	4.60	**3.00**	[−2.00, −1.20]
Mini O9 ↑	44.77	47.37	**50.20**	[1.80, 3.90]
Pre Track (%) ↑	34.0	52.9	**78.2**	[18.00, 32.00]
Post Agree (%) ↑	67.3	75.1	**83.4**	[5.00, 12.00]
LocalEdit precision (%) ↑	62.6	72.7	**83.8**	[7.00, 16.00]
LocalEdit recall (%) ↑	49.5	61.4	**76.3**	[10.00, 20.00]

Regret uses reward units; percentage-metric differences use percentage points. Pre Track conditions on pre-generation shifts, Post Agree uses all post-states, and LocalEdit precision/recall use eligible states.

Table 10 Grouped action transitions pooled across stages. Shares divide by 133 pre-generation or 125 post-generation shifts. SearchModeSwitch includes switches involving HybridSearch; Repair groups admissible repair actions.

Phase	Transition	Count	Share (%)
Pre	Search→Direct	69	51.9
Pre	Direct→Search	15	11.3
Pre	SearchModeSwitch	24	18.0
Pre	Search→PromptEnhance	13	9.8
Pre	PromptEnhance→Search	12	9.0
Post	Regenerate→LocalEdit	49	39.2
Post	ReferenceRegenerate→LocalEdit	21	16.8
Post	LocalEdit→Regenerate	16	12.8
Post	Accept→Repair	18	14.4
Post	Repair→Accept	21	16.8

eligible states. The direction is consistent across all three stages: refresh improves Mini O9 over stale continuation by 2.8, 2.9, and 2.8 points, and Pre Track by 25.9, 25.8, and 24.0 percentage points, respectively.

Regret averages the best finite-bank action value minus the selected action value in the reward units of Eq. (3). Pre Track and Post Agree measure preference membership under the updated generator. Empty conditional denominators are undefined. Finite-sample maximization can bias absolute regret upward; for refresh-minus-stale contrasts on the same action bank, the common maximum cancels, leaving the difference between their selected-action values. Table 9 gives paired request-cluster 95% intervals, including [1.8, 3.9] for the Mini O9 gain, conditional on the checkpoints and frozen evidence.

Comparison with current-generator on-policy learning. The separate Stage-2 comparison charges preparation as well as rollout and optimization within the common 100-unit budget (Table 11). Refresh reaches Mini O9 50.0, versus on-policy learning's 49.1 and stale supervision's 47.2. Pre/post regret decreases from 4.0/3.7 to 3.5/3.2 relative to on-policy learning. These Stage-2 values differ from the preceding three-stage continuation means and should be interpreted under their own matched-budget comparison.

Figure 4 Wan curriculum-source controls on AgentGen-Bench Full. Recipients keep learning with current-partner RL while the named construction sources are frozen.

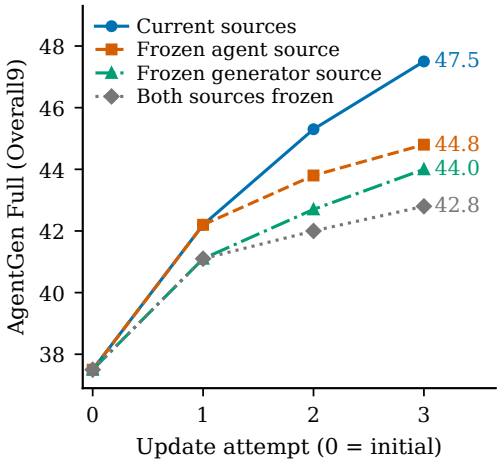

Table 11 Wan Stage-2 adaptation at a matched total budget of 100, including preparation, rollout, and optimization. On-policy uses the current generator; Δ O9 is relative to this arm. One cost unit is 1% of the common single-stage budget.

Metric	Stale	Current-G on-policy	Refresh
Quality and action selection			
Mini O9 ↑	47.2	49.1	**50.0**
Δ O9	-1.9	0.0	**+0.9**
Pre regret	5.40	4.00	**3.50**
Post regret	4.60	3.70	**3.20**
Adaptation cost			
Preparation	5	0	20
Rollout	65	74	55
Optimization	30	26	25
Total	100	100	100

5.5 Updated construction sources and recursive curricula

Controls retain recipient learning. Freeze-A uses A_0 to construct curriculum evidence and inputs; Freeze-G uses G_0 for curriculum-side outcomes, including post-generator diagnosis. Freeze-both fixes both sources, while the full arm uses current sources. Trainable recipients continue learning in every arm, and agent RL uses its current updated generator. Source opportunities, retrieval snapshots, compiler, optimization stages, and resource caps define the matched comparison. Each arm derives its memory, targets, and scores from its designated sources. With common randomness, current and Freeze-A share the first attempt because both initially use A_0.

Current sources reach 47.5 on Wan, compared with 42.8 under dual freezing. Freezing only the agent gives 44.8; freezing only the generator gives 44.0. Successive quality gains with current sources diminish from 4.7 to 3.1 and 2.2 points across attempts (Figure 4). These controls measure the combined contribution of curriculum content, target validity/yield, and supervision freshness.

Usable curriculum yield at fixed training sizes. Each source arm considers 2,000 requests per phase, yielding 4,000 generator-target candidates before the generator update and 3,000 agent-target candidates afterward. Validity is the valid-to-candidate ratio. Cross-attempt construction totals count target entries, separately from selected training records and repeated optimization exposure. New, refreshed, and existing-use records partition the selected curriculum; existing use includes replay and first use of previously available targets.

Across three attempts, current sources yield 5,580/7,480 valid agent/generator targets, compared with 4,430/6,740 under dual freezing, with 1,950 versus zero refreshed agent targets. Every arm trains on 5,400/6,600 agent/generator items (Table 12). Quality therefore improves alongside validity and freshness at fixed selected training sizes, while content and construction yield change together.

5.6 Role trajectories and curriculum evolution

Performance under specified partners and inputs. Figure 5 evaluates four retained states on 200 requests with four image seeds per request/state. Agent evaluation fixes G_0; generator scores distinguish direct and evidence-assisted inputs. These readouts are separate from system Full. Agent quality rises from 38.6 to 41.9 at S_2, then declines to 41.6; assisted/direct generator quality rises from 40.9/33.0 to 45.9/38.4. Generator improvement thus coexists with a nonmonotonic fixed-partner agent trajectory.

Table 12 Wan curriculum construction and supervision changes across the four source controls. Counts sum target entries across three attempts. Valid and new-valid counts describe construction; refreshed counts describe updated targets actually used for training.

	Agent curricula			Generator curricula		
Source	Valid	New valid	Refreshed	Valid	New valid	Refreshed
Current	5,580	2,660	1,950	7,480	4,140	970
Freeze-A	5,280	2,560	1,660	7,060	3,820	790
Freeze-G	4,690	2,270	400	6,960	3,680	680
Freeze-both	4,430	2,120	0	6,740	3,540	570

Each arm uses 5,400 agent and 6,600 generator training entries over the three attempts. Repeated optimization of an entry contributes to exposure, not to these entry counts.

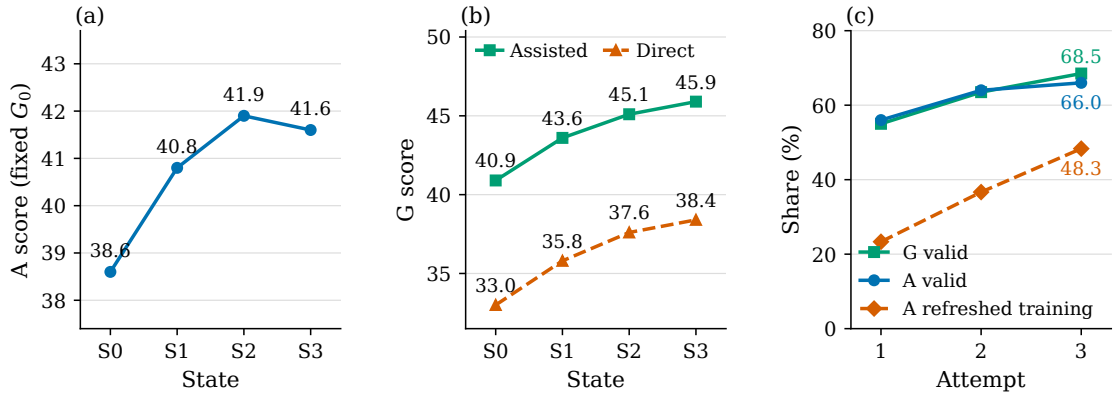

Figure 5 Wan role performance and curriculum validity across retained states. Evaluation uses 200 requests and four image seeds per request/state; agent evaluation fixes G_0. Role scores and curriculum units are defined in Section 5.6.

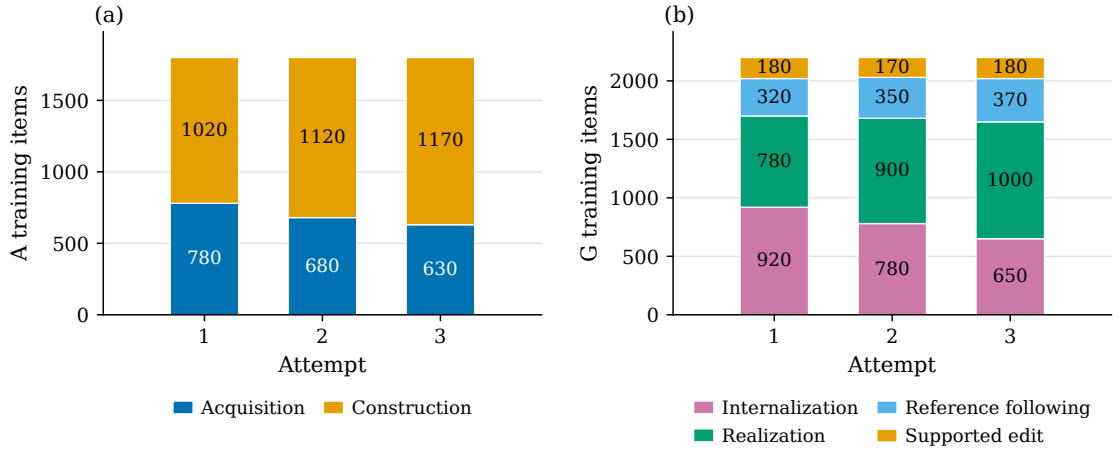

Figure 6 Current-source training curricula across three Wan update attempts. Each bar counts selected items by recipient subtype, with totals of 1,800 agent and 2,200 generator items per attempt. Replay is included in its subtype; candidate counts and repeated optimization exposures use separate units.

Validity, freshness, and skill composition. For current sources, agent/generator construction validity rises from 56.0%/55.0% to 66.0%/68.5%. Refreshed agent supervision grows from $420/1800 = 23.3\%$ to $870/1800 = 48.3\%$, while newly valid generator/agent targets decline from 1,500/1,080 to 1,180/650. Validity and freshness improve within fixed training sizes even as fewer targets become newly valid.

Figure 6 shows the selected records by skill. Agent construction grows from 1,020 to 1,170 while acquisition falls from 780 to 630; generator realization grows from 780 to 1,000 while internalization falls from 920 to 650. Subtype and new/refreshed/existing-use categories are

separate partitions of the same records, with fixed totals of 1,800 agent and 2,200 generator entries per attempt.

5.7 Qualitative results

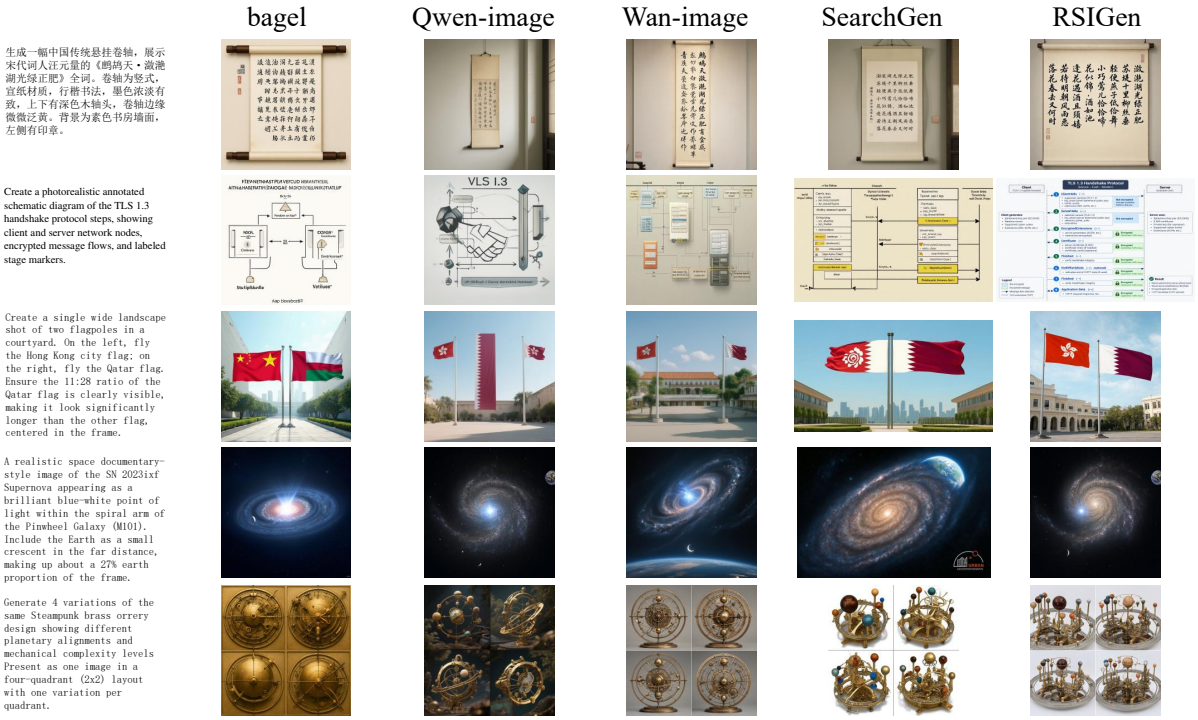

Figure 7 Five-request comparison with complete prompts at left and BAGEL, Qwen-Image, Wan-Image, SearchGen, and RSIgen outputs in columns. Table 3 reports aggregate comparisons.

Figure 7 compares five requests with their complete evaluation prompts and 25 model outputs: calligraphy, TLS 1.3, flags, astronomy, and coordinated design variations. RSIgen separates client/server columns and labels protocol stages in the TLS example, and renders four orrery designs in a consistent grid. The other requests illustrate text, entity appearance, and scene-composition requirements. These examples complement the aggregate benchmark and human judgments.

Figure 8 shows additional visual examples, including object depiction, sequential character scenes, product presentation, and stylized characters.

6 Discussion and Limitations

What the controls establish. The original uniform and failure-locus comparisons evaluate the diagnostic curriculum package. The matched-exposure allocation study isolates its benefit at equal role-level training volume; the matched-budget Stage-2 study compares refresh with ordinary current-generator adaptation. Construction-source freezes keep recipient learning active and measure the contribution of updated peers to curriculum content, validity/yield, and freshness. Their effects should be interpreted jointly rather than as a unique decomposition of causal responsibility.

For Wan, the descriptive difference-in-differences of the final source-control endpoints is $47.5 - 44.8 - 44.0 + 42.8 = 1.5$. This contrast summarizes combined content and target-yield effects; the primary result is the 4.7-point advantage of current over jointly frozen sources. Role trajectories also distinguish improving system quality from monotonic gains of every component under a fixed partner.

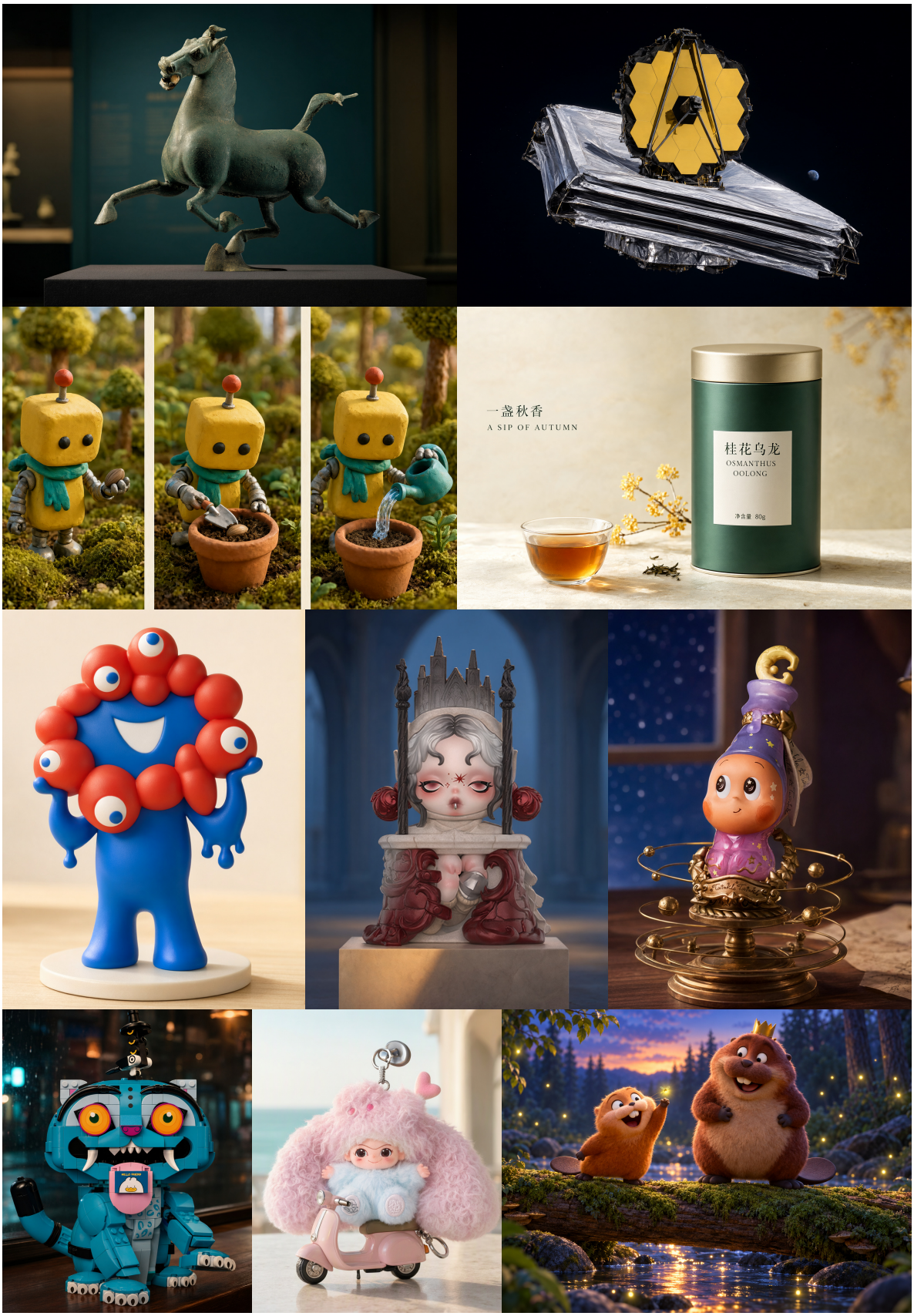

Figure 8 Additional RSIgen visual examples, including object depiction, sequential character scenes, product presentation, and stylized character designs.

Scope and uncertainty. Mechanism studies use Wan, whereas the benchmark improvements are evaluated across three backbones. Action comparisons use finite shared probe banks; their intervals measure request variation conditional on the trained checkpoints and frozen evidence. Human preferences are descriptive summaries of the evaluated real-user sample. Routing is an operational learning hypothesis rather than unique causal identification, and source controls change content and target yield together. The protected-capability tolerance is local to each accepted update; permitted losses can accumulate across successive steps.

Underlying generation capability. Current agentic visual generation systems, whether built from separate understanding and generation models or a unified model, often understand task requirements more reliably than they can realize them in generated images. RSIgen uses this capability asymmetry to improve collaboration between understanding and generation, translating stronger understanding into more effective supervision and better system-level performance. Nevertheless, the extent of further improvement remains constrained by the generative capabilities of the underlying models. Larger advances therefore also depend on improving these models' ability to translate understood requirements into faithful visual outputs.

7 Conclusion

RSIgen connects diagnosis-guided learning allocation, supervision refresh, and reciprocal curriculum construction. It improves AgentGen-Bench quality across three backbones, with complementary human preferences on 200 real-user requests. Wan controls show allocation and refresh gains at matched exposure and adaptation budget, while source controls connect recursive improvement to target validity and supervision freshness at fixed curriculum sizes. The role trajectories highlight the importance of adapting coordination as capabilities change. Section 6 discusses the scope and limitations.

References

Shuang Chen, Quanxin Shou, Hangting Chen, Yucheng Zhou, Kaituo Feng, Wenbo Hu, Yi-Fan Zhang, Yunlong Lin, Wenxuan Huang, Mingyang Song, Dasen Dai, Bolin Jiang, Manyuan Zhang, Shi-Xue Zhang, Zhengkai Jiang, Lucas Wang, Zhao Zhong, Yu Cheng, and Nanyun Peng. Unify-agent: A unified multimodal agent for world-grounded image synthesis, 2026a. https://arxiv.org/abs/2603.29620. arXiv:2603.29620; bibliographic record retrieved 2026-09-08.

Sixiang Chen, Zhaohu Xing, Tian Ye, Xinyu Geng, Yunlong Lin, Jianyu Lai, Xuanhua He, Fuxiang Zhai, Jialin Gao, and Lei Zhu. Genevolve: Self-evolving image generation agents via tool-orchestrated visual experience distillation, 2026b. https://arxiv.org/abs/2605.21605. Metadata verified against arXiv on 2026-09-19.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. Re-imagen: Retrieval-augmented text-to-image generator, 2022. https://arxiv.org/abs/2209.14491. arXiv:2209.14491; bibliographic record retrieved 2026-09-07.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models, 2024. https://arxiv.org/abs/2401.01335. arXiv:2401.01335; bibliographic record retrieved 2026-09-07.

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, 2026. https://arxiv.org/abs/2603.28767. arXiv:2603.28767; bibliographic record retrieved 2026-09-06.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11794. https://ojs.aaai.org/index.php/AAAI/article/view/11794.

Ruiyan Han, Zhen Fang, XinYu Sun, Yuchen Ma, Ziheng Wang, Yu Zeng, Zehui Chen, Lin Chen, Wenxuan Huang, Wei-Jie Xu, Yi Cao, and Feng Zhao. Unicorn: Towards self-improving unified multimodal models

through self-generated supervision, 2026a. https://arxiv.org/abs/2601.03193. arXiv:2601.03193; bibliographic record retrieved 2026-09-06.

Yujin Han, Hao Chen, Andi Han, Zhiheng Wang, Xinyu Liu, Yingya Zhang, Shiwei Zhang, and Difan Zou. Turning internal gap into self-improvement: Promoting the generation-understanding unification in MLLMs. In *International Conference on Learning Representations*, 2026b. https://proceedings.iclr.cc/paper_files/paper/2026/hash/dceda5883d3223231bef121cc8da2d61-Abstract-Conference.html. Preprint arXiv:2507.16663.

Ke Hao, Yuanzhi Liang, Tingxi Chen, Rui Li, Haibin Huang, Chi Zhang, Yun Gu, and Xuelong Li. Dreaming in flow: Generative grounding feedback for self-evolving unified multimodal models, 2026. https://arxiv.org/abs/2609.08282.

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, 2026a. https://arxiv.org/abs/2602.01756. Metadata verified against arXiv on 2026-09-19.

Zefeng He, Siyuan Huang, Xiaoye Qu, Yafu Li, Tong Zhu, Yu Cheng, and Yang Yang. Gems: Agent-native multimodal generation with memory and skills, 2026b. https://arxiv.org/abs/2603.28088. Metadata verified against arXiv on 2026-09-19.

Jixiang Hong, Yiran Zhang, Guanzhong Wang, Yi Liu, Ji-Rong Wen, and Rui Yan. SUDER: Self-improving unified large multimodal models for understanding and generation with dual self-rewards, 2025. https://arxiv.org/abs/2506.07963.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot, 2025. https://arxiv.org/abs/2505.00703. arXiv:2505.00703; bibliographic record retrieved 2026-09-06.

Weiyang Jin, Yuwei Niu, Jiaqi Liao, Chengqi Duan, Aoxue Li, Shenghua Gao, and Xihui Liu. SRUM: Fine-grained self-rewarding for unified multimodal models, 2025. https://arxiv.org/abs/2510.12784.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2022. https://arxiv.org/abs/2210.02747. arXiv:2210.02747; bibliographic record retrieved 2026-09-07.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl, 2025. https://arxiv.org/abs/2505.05470. arXiv:2505.05470; bibliographic record retrieved 2026-09-06.

Jie Liu, Zilyu Ye, Linxiao Yuan, Shenhan Zhu, Yu Gao, Jie Wu, Kunchang Li, Xionghui Wang, Xiaonan Nie, Weilin Huang, and Wanli Ouyang. Unigrpo: Unified policy optimization for reasoning-driven visual generation, 2026. https://arxiv.org/abs/2603.23500. arXiv:2603.23500; bibliographic record retrieved 2026-09-06.

Weijia Mao, Zhenheng Yang, and Mike Zheng Shou. Unirl: Self-improving unified multimodal models via supervised and reinforcement learning, 2025. https://arxiv.org/abs/2505.23380. arXiv:2505.23380; bibliographic record retrieved 2026-09-06.

Jiadong Pan, Liang Li, Yuxin Peng, Yu-Ming Tang, Shuohuan Wang, Yu Sun, Hua Wu, Qingming Huang, and Haifeng Wang. Learning to generate via understanding: Understanding-driven intrinsic rewarding for unified multimodal models, 2026. https://arxiv.org/abs/2603.06043. arXiv:2603.06043; bibliographic record retrieved 2026-09-06.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023. https://arxiv.org/abs/2305.18290. arXiv:2305.18290; bibliographic record retrieved 2026-09-07.

Robin Rombach, Andreas Blattmann, and Björn Ommer. Text-guided synthesis of artistic images with retrieval-augmented diffusion models, 2022. https://arxiv.org/abs/2207.13038. arXiv:2207.13038; bibliographic record retrieved 2026-09-07.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 627–635. PMLR, 2011. https://proceedings.mlr.press/v15/ross11a.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning

in open language models, 2024. https://arxiv.org/abs/2402.03300. arXiv:2402.03300; bibliographic record retrieved 2026-09-07.

Zhenchen Tang, Songlin Yang, Zichuan Wang, Bo Peng, Yang Li, Beibei Dong, and Jing Dong. Endogenous reprompting: Self-evolving cognitive alignment for unified multimodal models, 2026. https://arxiv.org/abs/2601.20305.

Ritesh Thawkar, Shravan Venkatraman, Omkar Thawakar, Abdelrahman Shaker, Fahad Khan, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. Ask, solve, generate: Self-evolving unified multimodal understanding and generation via self-consistency rewards, 2026. https://arxiv.org/abs/2606.27376. arXiv:2606.27376; bibliographic record retrieved 2026-09-06.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 2023. https://arxiv.org/abs/2311.12908. arXiv:2311.12908; bibliographic record retrieved 2026-09-07.

Fu-Yun Wang, Han Zhang, Michael Gharbi, Hongsheng Li, and Taesung Park. Unirl-zero: Reinforcement learning on unified models with joint language model and diffusion model experts, 2025a. https://arxiv.org/abs/2510.17937. arXiv:2510.17937; bibliographic record retrieved 2026-09-06.

Haozhe Wang, Weijia Feng, Jinpeng Yu, Che Liu, Ping Nie, Fangzhen Lin, Jiaming Liu, Ruihua Huang, Jimmy Lin, Wenhu Chen, and Cong Wei. Search beyond what can be taught: Evolving the knowledge boundary in agentic visual generation, 2026. https://arxiv.org/abs/2607.05382. arXiv:2607.05382; bibliographic record retrieved 2026-09-06.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation, 2025b. https://arxiv.org/abs/2503.05236. arXiv:2503.05236; bibliographic record retrieved 2026-09-06.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. https://arxiv.org/abs/2508.02324. arXiv:2508.02324; bibliographic record retrieved 2026-09-08.

Ling Yang, Xinchen Zhang, Ye Tian, Shiyi Zhang, Chenming Shang, Minghao Xu, Wentao Zhang, and Bin Cui. HermesFlow: Seamlessly closing the gap in multimodal understanding and generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2082. https://proceedings.neurips.cc/paper_files/paper/2025/hash/59d4e18a60490b9ed9913f3be2b14839-Abstract-Conference.html.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models, 2024. https://arxiv.org/abs/2401.10020. arXiv:2401.10020; bibliographic record retrieved 2026-09-07.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin godel machine: Open-ended evolution of self-improving agents, 2025. https://arxiv.org/abs/2505.22954. arXiv:2505.22954; bibliographic record retrieved 2026-09-06.

Jenny Zhang, Bingchen Zhao, Wannan Yang, Jakob Foerster, Jeff Clune, Minqi Jiang, Sam Devlin, and Tatiana Shavrina. Hyperagents, 2026a. https://arxiv.org/abs/2603.19461. arXiv:2603.19461; bibliographic record retrieved 2026-09-06.

Zekai Zhang, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiaoyue Chen, Xiao Xu, Yan Shu, Yanran Zhang, Yixian Xu, Yuxiang Chen, Zhendong Wang, Zihao Liu, Zikai Zhou, Huishuai Zhang, Dongyan Zhao, and Chenfei Wu. Qwen-image-agent: Bridging the context gap in real-world image generation, 2026b. https://arxiv.org/abs/2606.26907. arXiv:2606.26907; bibliographic record retrieved 2026-09-08.

Jiahao Zhao, Xiaomin Yu, Zhongxiang Sun, Fengwei Teng, Chengwei Qin, Xiaobin Hu, Jun Xu, and Shuicheng Yan. Toolartist: Tool-using unified multimodal models for agentic image generation, 2026. https://arxiv.org/abs/2608.04436. arXiv:2608.04436; bibliographic record retrieved 2026-09-06.