WEAGENT-MMGENEDIT: A FULL-STACK RECIPE FOR MULTIMODAL AGENTIC IMAGE GENERATION AND EDITING

ABSTRACT

Image generation and editing models have advanced rapidly, yet remain unreliable when prompts require external world knowledge. Bounded and long-tail parametric knowledge prevents direct or reason-then-generate approaches from recovering the required facts and visual appearances. Existing agentic generation and editing methods mitigate this limitation with retrieval tools, yet remain constrained by insufficient visual verification, overloaded policy models, and weak integration of retrieved textual and visual evidence. To address these limitations, we present **WeAgent-MMGenEdit**, a full-stack recipe including a multimodal harness, a scalable data construction pipeline, a comprehensive benchmark, and post-training methods for the agent policy and image backend. We first introduce WeAgent-Harness, a multimodal runtime with persistent evidence management and dedicated verification and integration tools that organize retrieved multimodal evidence into a dense carrier. Upon this, we develop a scalable pipeline for prompt synthesis and agentic trajectory collection, yielding 23K supervised trajectories and 14.7K RL tasks with three-layer verifiable checklists. We further introduce WeBench-MMGenEdit, a bilingual benchmark covering both knowledge-intensive image generation and multi-image editing. Finally, a two-sided post-training recipe based on SFT and RL improves the agent policy and image backend. Together, WeAgent-MMGenEdit enables a 30B-total/3B-active policy to outperform similarly sized policy models and approach the performance of a 1T-parameter agent.

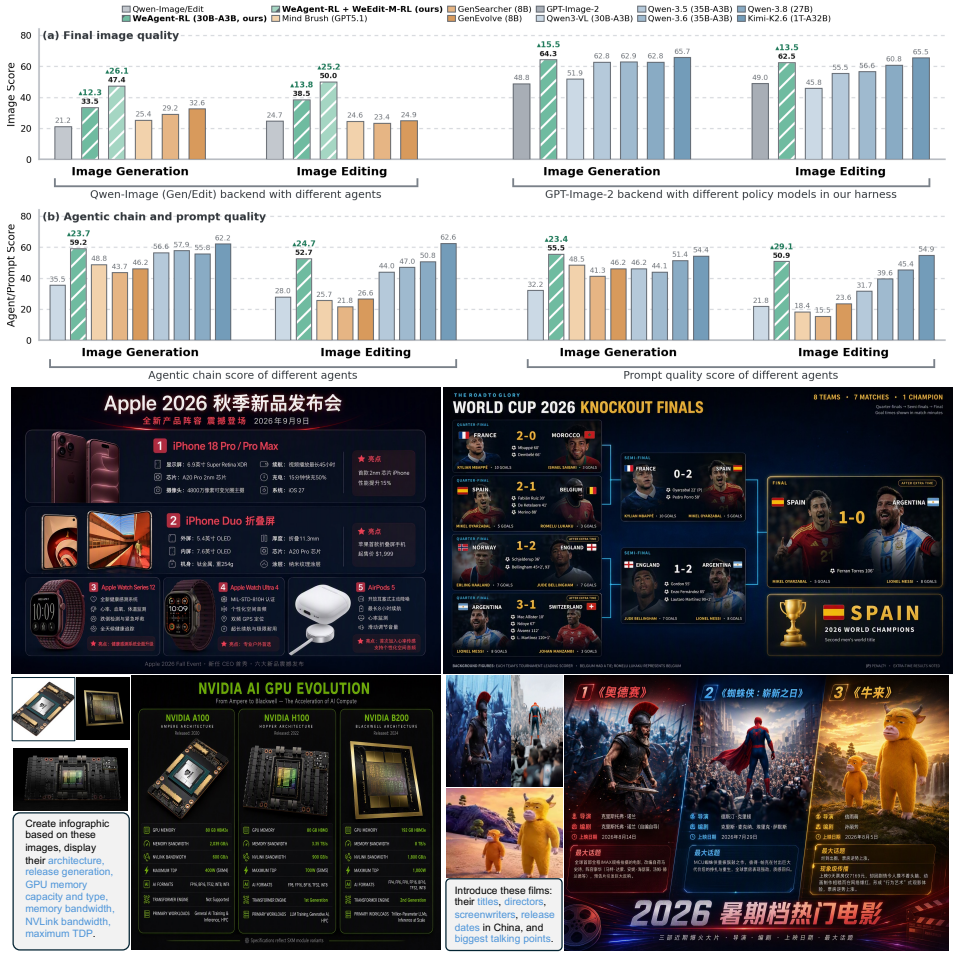

Figure 1: WeAgent-MMGenEdit outperforms similarly sized models in image generation and editing.

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.

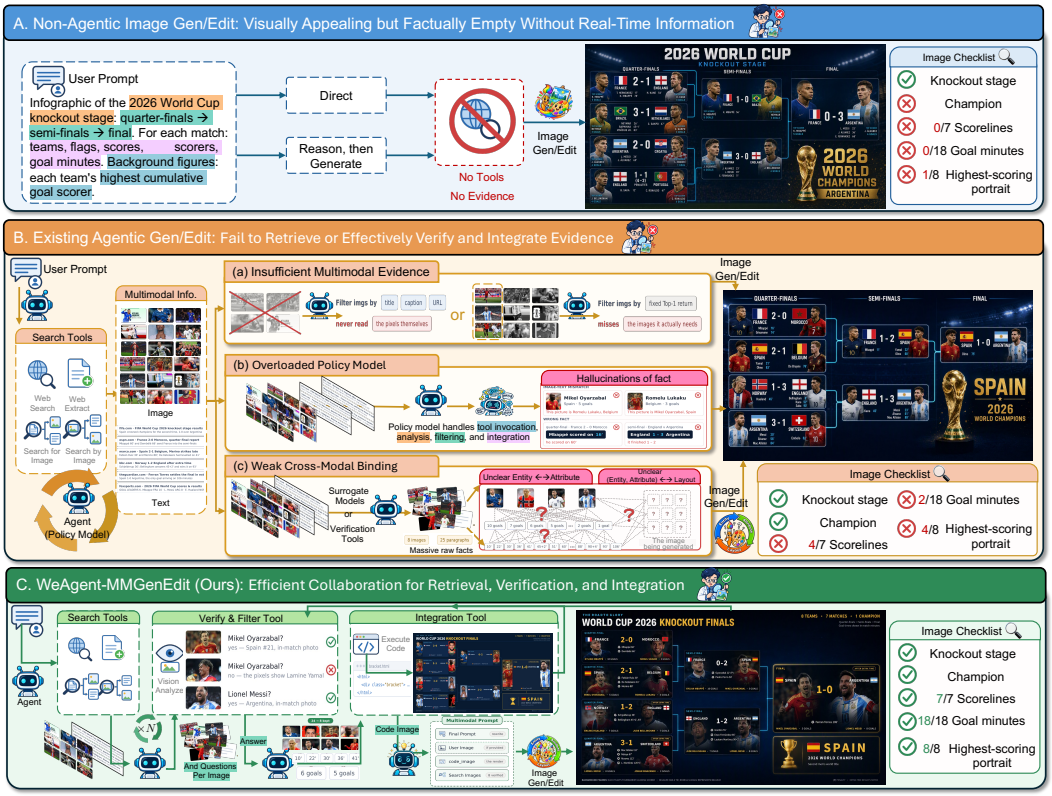

Figure 2: **Three paradigms for multimodal knowledge-intensive image generation and editing.**

1 INTRODUCTION

Recent advances in diffusion models (qwe, 2026.7; byt, 2026.7; Labs, 2026) have substantially improved the capabilities of image generation and editing systems. Leading proprietary models, including Gemini-3.1-Image (goo, 2026.2) and GPT-Image-2 (ope, 2026.4), can follow complex instructions, render legible text, preserve visual content, and support generation, editing, and multi-reference composition. Despite this progress, these systems remain largely closed-book: they perform well when all required content is provided in the prompt or reference images, but become unreliable when the task depends on external world knowledge.

Yet, an increasing range of real-world image generation and editing scenarios expose this limitation, as they require up-to-date or highly specialized information. Examples include updating an infographic with the latest rankings or financial results, reconstructing a recently completed tournament bracket, or editing an image to reflect the authentic appearance of a specific person, product, or place. Unlike conventional generation and editing, such multimodal knowledge-intensive image generation and editing requires models to acquire both textual knowledge and visual evidence beyond their parameters, and to faithfully transfer them into the generated pixels. As shown in Figure 2, direct generation is fundamentally constrained by training cutoffs and sparse long-tail knowledge. Moreover, although "reason-then-generate" paradigms (Sun et al., 2026; Kou et al., 2026; Jiang et al., 2026) use Vision-Language Models (VLMs) to reason over and enrich user prompts, they still rely on internalized parametric knowledge and therefore cannot access the up-to-date facts or authentic visual appearances required by these tasks.

To address this, agentic image generation and editing (Feng et al., 2026; He et al., 2026a; Chen et al., 2026b;a; Wang et al., 2026; Zhang et al., 2026d) has emerged as a promising direction. By equipping a policy model with search tools and an execution harness, these systems can dynamically acquire external textual and visual information from the open web. However, retrieval alone is not enough: the acquired evidence must still be reliably verified, filtered, and integrated before generation. As illustrated in Figure 2 B, existing systems still exhibit three limitations in this agentic process: **I) Insufficient multimodal verification**: visual candidates are often selected based on metadata (titles, captions, or URLs) or restricted to a small number of top-ranked results. As a result, the

policy may not explicitly inspect the pixels of all candidate images before selection, causing critical entities, attributes, or viewpoints to be missed. **II) Overloaded policy model**: a single policy is often responsible for both planning and evidence filtering, while simultaneously maintaining a growing context of retrieved images and passages. This increasing burden can dilute attention over long trajectories and lead to incorrect or conflated factual information. **III) Weak cross-modal binding**: even when correct facts and reference images are retrieved, they are often passed to the generator as loosely organized text and image attachments. The associations among entities, attributes, and spatial layouts are therefore left implicit, making it difficult for the generator to faithfully realize layout-dense outputs such as infographics, brackets, and timelines. Beyond these methodological limitations, an additional gap lies in evaluation. Existing agentic benchmarks (Feng et al., 2026; He et al., 2026a; Chen et al., 2026b), primarily focus on generation, leaving knowledge-intensive multi-image editing underexplored.

To this end, we propose **WeAgent-MMGenEdit**, a full-stack recipe integrating the agent harness, dataset, benchmark, and two-sided post-training: **I) WeAgent-Harness:** a multimodal runtime for retrieving, verifying, and integrating external evidence. It maintains persistent multimodal evidence and organizes verified textual and visual information into rendered carriers with explicit entity–attribute–layout bindings. **II) WeDataset-MMGenEdit:** a verifiable data pipeline for multi-hop generation and editing, with three-level checklists covering the agentic chain, generation input, and final image. It yields 23K SFT trajectories and 14.7K RL tasks. **III) WeBench-MMGenEdit:** a 300-case human-audited bilingual benchmark balanced across generation/editing and English/Chinese. It emphasizes knowledge-intensive multi-image editing and evaluates the agentic process, generation input, and final image separately. **IV) Two-sided Post-Training:** agent-side SFT and checklist-grounded RL improve evidence acquisition, verification, and integration, while image-side SFT and RL improve multi-reference ability and faithful evidence rendering.

As summarized in Figure 1, WeAgent-MMGenEdit consistently improves both the agentic process and final image quality across generation and editing. Within the same WeAgent-Harness, the WeAgent policy outperforms similarly sized policies and approaches the performance of a trillion-parameter agent with only about 3% of its total parameters. These gains transfer consistently across multiple image backends and, when combined with image-side post-training, enable WeAgent-MMGenEdit to substantially outperform existing open-source agentic generation and editing systems on both WeBench-MMGenEdit and public benchmarks.

2 Related Work

2.1 Image Generation and Editing

Recent image generation and editing models have substantially improved instruction following, text rendering, visual fidelity, and multi-reference conditioning (Rombach et al., 2022; Podell et al., 2024; Peebles & Xie, 2023; Esser et al., 2024; Wu et al., 2025a; Cao et al., 2025; ope, 2026.4; goo, 2026.2). Post-training through supervised adaptation and reinforcement learning further improves controllability, reference conditioning, and perceptual quality (Hu et al., 2022; Tan et al., 2025; Zhang et al., 2026a; Wu et al., 2025b; Zheng et al., 2025b; Liu et al., 2025). However, most systems remain *closed-book*, limiting their ability to handle generation and editing that require external factual.

2.2 Multimodal Agent and Harness

Language-model agents combine reasoning with external tools for iterative interaction (Schick et al., 2023; Yao et al., 2022; Qin et al., 2024), while multimodal agents further support visual observations and interaction with web and computer environments (Wu et al., 2023; Hu et al., 2024; He et al., 2024; Xie et al., 2024; Wu et al., 2026; Yang et al., 2025). Recent work also highlights the *agent harness* for managing tools, context, memory, and execution (Pan et al., 2026; Chen et al., 2026c; Meng et al., 2026; Huang et al., 2026; Shao et al., 2026). However, existing harnesses are largely general-purpose and provide limited support for visual verification and structured multimodal evidence integration.

2.3 Agentic Image Generation and Editing

Agentic image generation and editing combines visual synthesis with planning and tool use, ranging from generation/editing tool orchestration (Wang et al., 2024; Chen et al., 2025; Ren et al., 2026; He et al., 2026b) to knowledge-seeking systems that retrieve external textual and visual evidence (Son et al., 2025; He et al., 2026a; Feng et al., 2026; Chen et al., 2026a;b; Zhang et al., 2026d). However, existing systems remain limited in evidence verification and integration, agentic post-training, and knowledge-intensive multi-image editing. WeAgent-MMGenEdit addresses these gaps with explicit verification and integration, two-sided post-training, and a balanced benchmark.

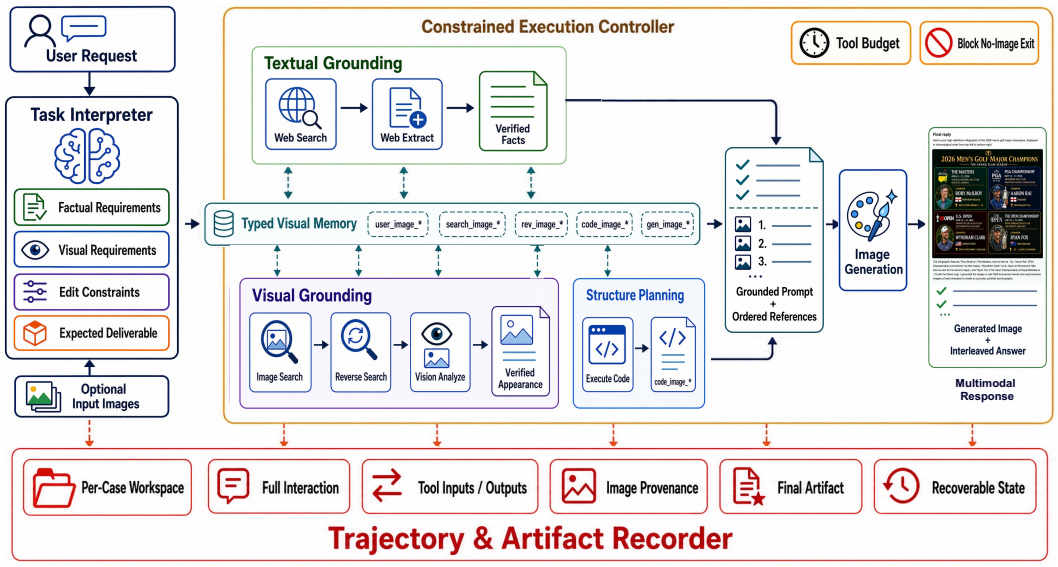

Figure 3: **WeAgent-Harness: A Multimodal Runtime for Retrieval, Verification, and Integration.**

3 WeAgent-Harness: Multimodal Agentic Harness

A harness defines an agent's observations, actions, and information persistence across turns. For multimodal knowledge-intensive image generation and editing, retrieval alone is insufficient; external evidence must also be reliably verified and structured for generation. We therefore introduce **WeAgent-Harness (WeHarness)**, a unified retrieve–verify–integrate–deliver runtime with persistent multimodal evidence management and constrained execution (Figure 3).

3.1 Agent Runtime and Multimodal Memory

Reasoning-and-Action Protocol. Each turn follows a think–act–observe paradigm (Yao et al., 2022): the policy produces a brief reasoning trace, issues one tool call or the final answer, and receives the tool output as the next observation. This design enables process-level supervision and RL while keeping verification and integration explicit. Each episode follows fixed turn and tool budgets, with constrained termination and automatic repair of unambiguous malformed tool calls. The same harness is used for inference and RL rollout.

Persistent Multimodal Evidence Store. Evidence is stored by reference rather than carried in context. User-provided, retrieved, rendered, and generated images are assigned stable identifiers with provenance in a per-trajectory workspace. These identifiers enable evidence to be accessed across tools only when needed, reducing long-horizon context overhead.

3.2 Retrieve-Verify-Integrate-Deliver Toolchain

Tool Interface. The policy interacts with seven tools: four for multimodal retrieval and three for verification, integration, and final image generation. Together, they implement the retrieve–verify–integrate–deliver workflow.

Multimodal Evidence Retrieval. `Web_Search` and `Web_Extract` retrieve task-relevant textual evidence, while `Search_for_Image` and `Search_by_Image` acquire visual evidence through text- and image-based search. Retrieved candidates are registered in the evidence store for subsequent verification and integration.

Explicit Visual Verification. `Vision_Analyze` selectively inspects registered images through focused image–question pairs, with multiple candidates verified in parallel. A separately deployed VLM (identical to the policy model) performs visual inspection without occupying the main agent context, distinguishing retrieved candidates from verified visual evidence.

Structured Evidence Integration. `Execute_Code` organizes verified textual and visual evidence into a rendered carrier, typically through an HTML layout. The carrier explicitly encodes entity–attribute–layout relations in pixels, providing structured multimodal evidence for final generation.

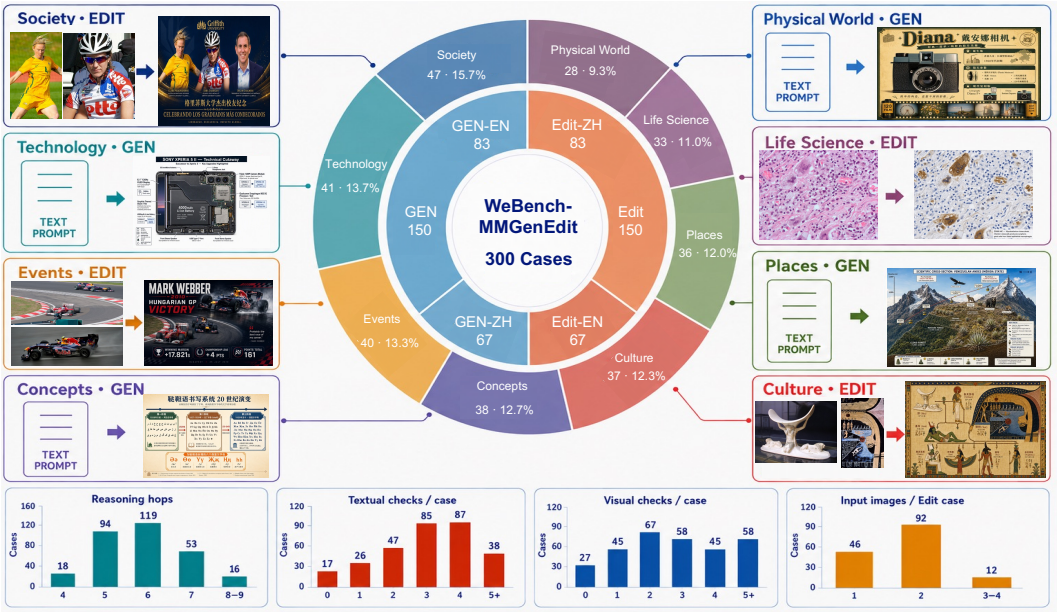

Figure 4: **Composition of WeBench-MMGenEdit.** The benchmark contains 300 human-audited cases, exactly balanced between generation and editing and between English and Chinese.

Unified Delivery. `Image_Generation` provides a unified interface for generation and multi-reference editing, conditioned on the final prompt and optional reference images. The image backend is decoupled from the policy, allowing different generators to share the same agent workflow.

4 WeAgent-MMGenEdit Dataset and Benchmark

We build **WeDataset-MMGenEdit**, a verifiable training corpus for multimodal knowledge-intensive generation and editing, and **WeBench-MMGenEdit**, a human-audited benchmark. Both share a three-level specification covering the agentic process, generation input, and final image.

4.1 WeDataset-MMGenEdit

Verifiable Task Construction. We construct a bilingual multimodal concept bank aligned with a knowledge graph, and synthesize tasks through single-page and multi-hop graph-based routes. Each task includes three-level checklists for the agentic process, generation input, and final image.

Trajectory Collection and Grading. Synthesized tasks are executed by an expert policy in WeAgent-Harness to collect complete agentic trajectories. Each trajectory is independently graded at the three levels: high-quality trajectories are retained for SFT, while valid but imperfect ones are routed to RL. After filtering, WeDataset-MMGenEdit contains approximately 23K SFT trajectories and 14.7K RL tasks. Further details on data construction, task synthesis, grading criteria, and filtering are provided in the supplementary material.

4.2 WeBench-MMGenEdit

Construction and Composition. WeBench-MMGenEdit contains 300 human-verified held-out tasks, balanced across generation/editing and English/Chinese, spanning eight domains. Tasks require multi-hop textual and visual retrieval; 84.0% include hidden visual references for evaluating visual knowledge, and 69.3% of editing cases involve multiple user-provided images.

Three-Level Evaluation. We independently evaluate: **(i) Agentic Chain Score** for evidence retrieval and integration; **(ii) Prompt Score** for textual and visual knowledge sufficiency; and **(iii) Image Output Score** for instruction following, knowledge accuracy, preservation, and visual quality. This three-level protocol separates agent-side failures from image-backend failures. Detailed benchmark construction, evaluation criteria, and metric definitions are provided in the Appendix.

5 Agentic Post-Training

We post-train the agent policy in two stages while keeping the image backend frozen. Agentic SFT learns the WeAgent-Harness interaction protocol from expert trajectories, followed by checklist-grounded RL to improve evidence acquisition, verification, and integration.

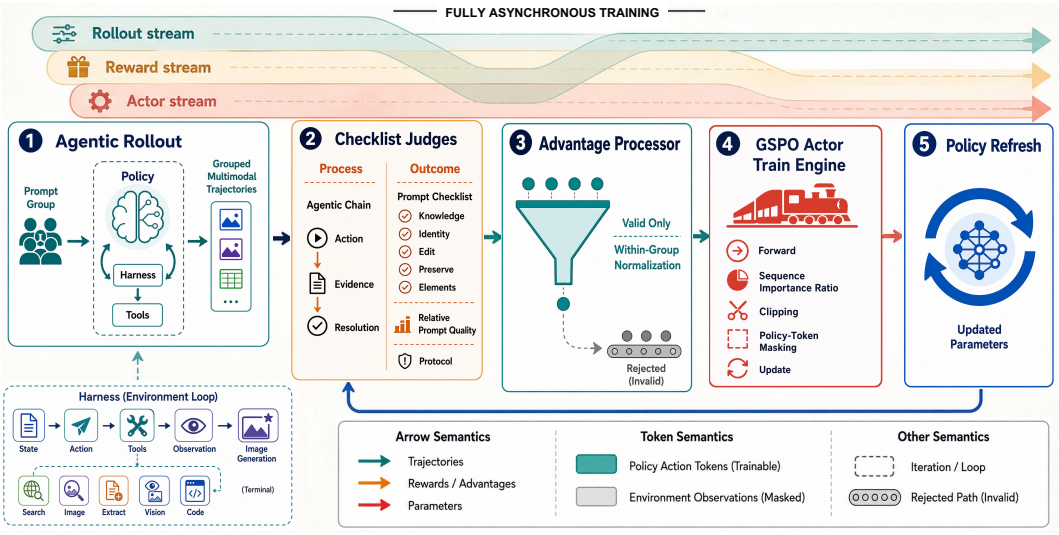

Figure 5: Asynchronously Checklist-grounded agentic RL.

Problem Formulation. Given a task $x = (u, \mathcal{I}_{\text{user}})$, the policy π_θ interacts with WeAgent-Harness through a multi-turn trajectory $\tau = (r_1, a_1, o_1, \ldots, r_T, a_T, o_T)$. The terminal `Image_Generation` call produces the generation input $c = (p_{\text{gen}}, \mathcal{I}_{\text{ref}})$. Agentic post-training optimizes the policy that produces (τ, c).

5.1 Stage I: Agentic Supervised Fine-Tuning

We train the policy on complete expert reasoning–action–observation trajectories collected with WeAgent-Harness. Given the serialized sequence $z = (z_1, \ldots, z_L)$, we apply next-token prediction only to policy-generated tokens: $\mathcal{L}_{\text{SFT}} = -\sum_{n=1}^{L} m_n \log \pi_\theta(z_n \mid z_{<n}, x)$, where $m_n = 1$ for policy-generated tokens and 0 otherwise. This teaches tool use, multimodal evidence handling, and generation-input construction while treating environment observations as context only.

5.2 Stage II: Checklist-Grounded Agentic RL

SFT learns successful interaction patterns but does not directly optimize evidence quality or prompt sufficiency. We therefore use the task-level checklists from Section 4.1 as RL supervision, jointly optimizing the agentic process and the multimodal input ultimately delivered to the image model.

Grouped Agentic Rollout. Starting from the SFT policy, we optimize with GSPO (Zheng et al., 2025a). Each prompt produces $G = 8$ trajectories in WeAgent-Harness, enabling group-relative optimization over complete agentic trajectories.

Checklist-Grounded Reward. Each trajectory is evaluated by isolated judges for the agentic process and final generation input. The reward combines generation-input sufficiency, process quality, relative quality, and protocol compliance: $R = 0.80\, S_{\text{input}} + 0.10\, S_{\text{agentic}} + 0.05\, S_{\text{relative}} + 0.05\, S_{\text{protocol}}$, where S_{input} emphasizes generation-input sufficiency and S_{agentic} measures checklist-grounded process quality. For each agentic checklist item, $S^{(i)}_{\text{item}} = 0.25\, s^{(i)}_{\text{act}} + 0.30\, s^{(i)}_{\text{evi}} + 0.45\, s^{(i)}_{\text{res}}$.

Robust Agentic RL. Live tool environments introduce failures unrelated to policy quality. We therefore exclude invalid trajectories from optimization and group normalization, and apply gradients only to policy-generated tokens.

Asynchronous Training. Rollout, reward evaluation, and actor optimization run asynchronously using Relax (Zhang et al., 2026c; Chu et al., 2026), overlapping expensive web interaction, judging, and policy updates. Further optimization and implementation details are provided in the Appendix.

6 Image Editing Post-Training

WeAgent-Harness produces heterogeneous visual references with different resolutions, aspect ratios, and semantic roles. We therefore post-train the image backend in two stages: SFT followed by Diffusion-NFT RL to improve multi-reference adherence and text clarity. The resulting models are denoted as **WeEdit-M-SFT** and **WeEdit-M-RL**.

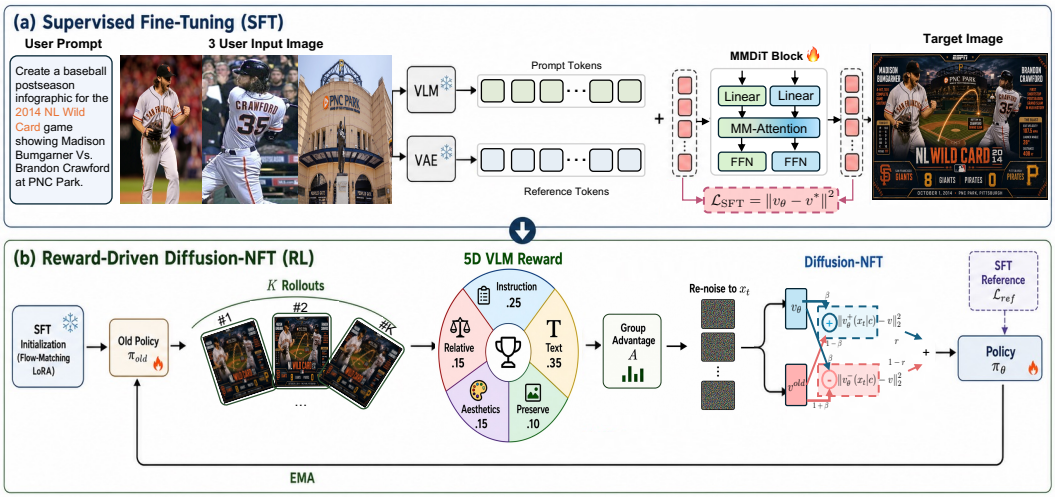

Figure 6: Two-stage image editing post-training.

6.1 Stage I: Multi-Reference Image Editing SFT

Heterogeneous Multi-Reference Conditioning. Each reference is encoded by a frozen VAE and vision–language encoder, with independent resolution buckets preserving heterogeneous aspect ratios. Their latent tokens are concatenated with per-segment shape information: $\mathbf{Z} = [\mathbf{Z}_y; \mathbf{Z}_{x_1}; \ldots; \mathbf{Z}_{x_n}]$, We train LoRA adapters (Hu et al., 2022) with a rectified-flow objective (Liu et al., 2023), applying supervision only to target-image tokens. The resulting **WeEdit-M-SFT** establishes stable multi-reference conditioning.

6.2 Stage II: Multi-Objective Image Editing RL

Grouped Candidate Rollout. Starting from WeEdit-M-SFT, we optimize the editor with Diffusion-NFT (Zheng et al., 2025b). For each multimodal condition, multiple candidates are sampled and compared within the same group, with challenging multi-reference cases prioritized for RL.

Multi-Objective Reward. Each candidate is evaluated along five dimensions: instruction accuracy, text accuracy, preservation, aesthetics, and relative quality. We derive continuous scores from judge probabilities and combine them as $R = 0.25\, s_{\text{instr}} + 0.35\, s_{\text{text}} + 0.10\, s_{\text{presv}} + 0.15\, s_{\text{aes}} + 0.15\, s_{\text{rel}}$. This decomposed reward avoids masking severe failures behind a single holistic score and directly targets multi-reference adherence, text clarity, preservation, and visual quality.

Diffusion-NFT Optimization. Rewards are normalized within each candidate group to obtain relative advantages. Diffusion-NFT converts these preferences into a flow-matching objective without backpropagating through the full sampling trajectory, reinforcing high-reward candidates while suppressing low-reward ones. To limit reward-induced drift, the policy is regularized toward the frozen WeEdit-M-SFT reference.

7 Experiment

We primarily evaluate on WeBench-MMGenEdit at three levels: **Agentic Chain Score** measures textual/visual knowledge retrieval (TKR/VKR) and multimodal evidence integration (MEI); **Prompt Score** measures whether the generation input contains sufficient textual and visual knowledge (TKS/VKS); and **Image Output Score** measures instruction adherence (IA), textual/visual knowledge accuracy (TKA/VKA), content preservation (CP), and visual quality (VQ). We initialize the agent policy from Qwen3-VL-30B-A3B (Bai et al., 2025) and the image backend from Qwen-Image-Edit-2509 (qwe, 2025), and post-train both with SFT followed by RL. All benchmark evaluations use GPT-5.5 as the judge model. Further details on training configurations, reward models, evaluation protocols, hyperparameters, evaluation on KnowGen (Feng et al., 2026) and Mind-Bench (He et al., 2026a) are provided in the supplementary material.

7.1 Main Results

WeAgent-Harness Provides the First Gain. Direct Gen/Edit suffers from low knowledge accuracy, while reason-then-generate offers limited gains as it cannot retrieve external evidence. In contrast,

Table 1: Quantitative Results on Image Generation and Editing.

Model	Policy Model	Gen/Edit Model	Harness	Image Generation					Image Editing					
				IA	TKA	VKA	VQ	W_Avg	IA	TKA	VKA	CP	VQ	W_Avg
Direct Gen/Edit														
GPT-Image-2	N/A	GPT-Image-2	N/A	63.20	33.83	42.18	85.13	48.79	53.86	23.67	47.95	81.06	82.33	48.99
Gemini3.1-Flash	N/A	Gemini3.1-Flash	N/A	67.18	46.50	45.70	81.41	55.12	56.35	31.49	54.60	75.08	77.72	51.79
Gemini3.1-Flash-Lite	N/A	Gemini3.1-Flash-Lite	N/A	65.77	46.22	43.84	79.06	53.81	53.07	30.52	47.95	66.71	76.27	48.51
Seedream-5.0-pro	N/A	Seedream-5.0-pro	N/A	58.07	29.91	35.56	80.01	43.58	54.09	26.37	47.72	77.34	78.18	48.60
FLUX2-dev	N/A	FLUX2-dev	N/A	34.87	4.58	22.99	55.07	22.69	32.22	2.66	22.20	51.16	50.54	24.99
Qwen-Image(Gen/Edit)	N/A	Qwen-Image(Gen/Edit)	N/A	33.18	3.25	21.60	53.11	21.25	31.65	1.57	24.39	52.41	48.79	24.74
Reason then Gen/Edit														
Reason+GPT-Image-2	Kimi-K2.6	GPT-Image-2	N/A	65.84	38.35	42.43	85.30	51.28	56.04	23.79	50.33	83.07	81.79	50.13
Reason+Gemini3.1-Flash-Lite	Kimi-K2.6	Gemini3.1-Flash-Lite	N/A	65.01	44.14	41.79	78.40	52.26	52.71	27.36	49.55	70.22	75.50	47.92
Open-Source Agentic Gen/Edit														
Mind Brush	GPT-5.1	Qwen-Image(Gen/Edit)	Mind Brush	34.53	7.99	29.10	52.07	25.35	30.67	3.58	23.11	46.25	50.80	24.64
GenSearcher	GenSearcher-8B	Qwen-Image(Gen/Edit)	GenSearcher	37.01	13.51	32.64	54.53	29.20	29.07	2.22	22.61	47.77	46.80	23.38
GenEvolve	GenEvolve-8B	Qwen-Image(Gen/Edit)	GenEvolve	40.27	19.91	33.83	55.73	32.63	28.87	10.03	20.79	39.02	50.07	24.93
Agentic Gen/Edit within WeAgent-Harness														
Opus-5 + GPT-Image-2	Claude-Opus-5	GPT-Image-2	WeHarness	**82.22**	**70.74**	**65.57**	**89.26**	**74.06**	**75.53**	**64.12**	**75.17**	**83.38**	**85.73**	**73.55**
GPT-5.6-Sol + GPT-Image-2	GPT-5.6-Sol	GPT-Image-2	WeHarness	80.80	67.28	64.99	88.93	72.17	72.87	55.31	73.34	82.50	86.33	69.41
Kimi-K2.6 + GPT-Image-2	Kimi-K2.6 (1T-A32B)	GPT-Image-2	WeHarness	75.53	58.93	56.86	88.01	65.73	68.13	50.48	68.78	79.79	85.20	65.50
Kimi-K2.6 + Gemini3.1-Flash-Lite	Kimi-K2.6 (1T-A32B)	Gemini3.1-Flash-Lite	WeHarness	71.67	56.51	53.42	81.40	62.32	62.35	49.15	61.26	73.17	78.21	60.30
Qwen-3.5 + GPT-Image-2	Qwen-3.5 (35B-A3B)	GPT-Image-2	WeHarness	72.58	54.98	54.98	86.67	62.75	58.36	39.24	53.01	75.38	81.78	55.46
Qwen-3.6 + GPT-Image-2	Qwen-3.6 (35B-A3B)	GPT-Image-2	WeHarness	73.49	58.57	54.61	86.59	62.91	59.26	40.26	56.31	75.45	82.62	56.58
Qwen-3.8 + GPT-Image-2	Qwen-3.8 (27B)	GPT-Image-2	WeHarness	72.36	57.71	54.86	86.11	62.81	64.10	47.63	56.28	78.71	82.25	60.79
Qwen3-VL + GPT-Image-2 (Baseline 1)	Qwen3-VL (30B-A3B)	GPT-Image-2	WeHarness	63.92	38.82	45.67	85.52	51.86	50.74	24.53	48.01	74.20	80.94	45.80
Qwen3-VL + Gemini3.1-Flash-Lite (Baseline 2)	Qwen3-VL (30B-A3B)	Gemini3.1-Flash-Lite	WeHarness	61.73	42.35	42.79	77.64	51.21	47.77	21.86	39.21	66.72	73.99	42.98
Qwen3-VL + Qwen-Image(Gen/Edit) (Baseline 3)	Qwen3-VL (30B-A3B)	Qwen-Image(Gen/Edit)	WeHarness	37.94	14.66	27.57	52.53	28.34	34.93	11.08	25.05	52.57	52.97	29.42
WeAgent-MMGenEdit (Ours)														
WeAgent-SFT + GPT-Image-2	WeAgent-SFT (30B-A3B)	GPT-Image-2	WeHarness	71.92	**54.41**	**53.20**	87.81	60.83	61.22	**42.66**	**57.71**	78.34	83.74	58.56
WeAgent-RL + GPT-Image-2	WeAgent-RL (30B-A3B)	GPT-Image-2	WeHarness	73.67	**59.64**	**56.87**	87.93	64.31	64.47	**48.25**	**62.96**	79.78	82.87	62.52
vs. Baseline 1	–	–	–	**+9.75**	**+20.82**	**+11.20**	**+2.41**	**+12.45**	**+13.73**	**+23.72**	**+14.95**	**+5.58**	**+1.93**	**+16.72**
WeAgent-SFT + Gemini3.1-Flash-Lite	WeAgent-SFT (30B-A3B)	Gemini3.1-Flash-Lite	WeHarness	67.72	**51.62**	**47.82**	78.39	57.50	55.68	**37.67**	**54.81**	67.72	76.22	53.08
WeAgent-RL + Gemini3.1-Flash-Lite	WeAgent-RL (30B-A3B)	Gemini3.1-Flash-Lite	WeHarness	69.86	**56.33**	**49.83**	78.75	60.08	56.58	**41.52**	**55.42**	68.44	74.73	54.43
vs. Baseline 2	–	–	–	**+8.13**	**+13.98**	**+7.04**	**+1.11**	**+8.87**	**+8.81**	**+19.66**	**+16.21**	**+1.72**	**+0.74**	**+11.45**
WeAgent-RL + Qwen-Image(Gen/Edit)	WeAgent-RL (30B-A3B)	Qwen-Image(Gen/Edit)	WeHarness	42.83	**23.23**	**29.03**	54.83	33.54	41.78	**22.39**	**40.53**	56.59	54.67	38.50
WeAgent-RL + WeEdit-M-SFT	WeAgent-RL (30B-A3B)	WeEdit-M-SFT	WeHarness	49.19	**27.24**	**35.41**	59.12	38.65	49.93	**31.92**	**47.78**	60.44	63.63	46.24
WeAgent-RL + WeEdit-M-RL	WeAgent-RL (30B-A3B)	WeEdit-M-RL	WeHarness	56.53	**39.13**	**43.47**	63.61	47.39	53.11	**37.03**	**47.27**	70.21	66.01	49.98
vs. Baseline 3	–	–	–	**+18.59**	**+24.47**	**+15.90**	**+11.08**	**+19.05**	**+18.18**	**+25.95**	**+22.22**	**+17.64**	**+13.04**	**+20.56**

WeAgent-Harness substantially improves knowledge accuracy, particularly VKA, through explicit visual verification and structured evidence integration.

Post-Training Further Improves Agentic Gen/Edit. Agentic SFT improves knowledge grounding, while RL further enhances evidence selection and integration, enabling WeAgent-RL to outperform similarly sized policies (Qwen Series) and existing agentic systems. Image-side SFT and RL further improve multi-reference image editing. Together, the fully post-trained stack improves W_Avg by 19.05/20.56 points over the untrained Qwen3-VL + Qwen-Image baseline on generation/editing.

Comparison of Agentic Chain and Prompt Scores. WeAgent-RL substantially improves both agentic-process and prompt-level scores over the untrained Qwen3-VL baseline, with particularly large gains in visual knowledge sufficiency (VKS: +25.65/+33.62 on generation/editing). It also outperforms existing open-source agentic systems and similarly sized policies across both levels. These results show that post-training improves not only multimodal evidence acquisition and integration, but also its reliable knowledge transfer to the image model.

Table 2: Comparison of Agentic Chain and Prompt Scores.

Model Name	Harness	Image Generation							Image Editing						
		Agentic Chain Score				Prompt Score			Agentic Chain Score				Prompt Score		
		TKR	VKR	MEI	Avg	TKS	VKS	Avg	TKR	VKR	MEI	Avg	TKS	VKS	Avg
Open-Source Agentic Gen/Edit															
Mind Brush (GPT-5.1)	Mind Brush	50.31	46.95	49.27	48.84	52.91	44.10	48.51	20.76	20.17	36.10	25.68	26.71	10.01	18.36
GenSearcher-8B	GenSearcher	45.41	44.17	41.54	43.71	43.89	38.73	41.31	18.66	19.13	27.54	21.78	22.55	8.42	15.49
GenEvolve-8B	GenEvolve	42.14	53.66	42.71	46.17	46.03	46.47	46.25	22.38	21.59	35.92	26.63	29.92	17.37	23.65
Agentic Gen/Edit within WeAgent-Harness															
Opus-5	WeHarness	68.04	64.22	69.14	67.13	74.28	59.31	66.80	65.46	67.61	70.40	67.82	69.76	72.11	70.94
GPT-5.6-Sol	WeHarness	70.26	64.58	71.23	68.69	71.28	59.17	65.23	63.90	66.57	70.35	66.94	63.72	61.84	62.78
Kimi-K2.6 (1T-A32B)	WeHarness	65.32	58.15	63.01	62.16	65.30	43.42	54.36	58.09	66.66	62.97	62.57	57.90	51.84	54.87
Qwen-3.5 (35B-A3B)	WeHarness	60.39	55.27	54.07	56.58	59.71	32.76	46.24	43.71	41.49	46.86	44.02	45.69	17.63	31.66
Qwen-3.6 (35B-A3B)	WeHarness	63.71	53.24	56.69	57.88	62.15	26.08	44.11	44.19	46.75	50.07	47.00	47.68	31.58	39.63
Qwen-3.8 (27B)	WeHarness	60.03	49.75	57.65	55.81	62.44	40.39	51.42	52.48	45.89	54.08	50.82	53.38	37.37	45.38
Qwen3-VL (30B-A3B) (Baseline)	WeHarness	39.24	31.14	36.07	35.48	41.98	22.37	32.17	23.41	25.20	35.47	28.03	29.65	13.95	21.80
WeAgent-SFT (30B-A3B)	WeHarness	58.29	55.37	54.83	56.16	58.55	47.48	53.02	47.99	44.76	51.93	48.23	48.24	28.79	38.52
WeAgent-RL (30B-A3B)	WeHarness	64.54	55.88	57.08	59.17	63.05	48.02	55.54	53.71	48.92	55.49	52.71	54.13	47.57	50.85
vs. Baseline	–	**+25.30**	**+24.74**	**+21.01**	**+23.68**	**+21.07**	**+25.65**	**+23.36**	**+30.30**	**+23.72**	**+20.02**	**+24.68**	**+24.48**	**+33.62**	**+29.05**

Qualitative Results. Figure 7 shows that direct image models often miss task-specific factual and visual details, while WeAgent-RL improves knowledge correctness through multimodal evidence retrieval, verification, and integration. These gains generalize across image backends and are further enhanced by image-side post-training, yielding more faithful generation and multi-image editing.

7.2 Ablation Study

Ablation Study. Table 3 shows that visual verification and structured evidence integration provide complementary gains beyond retrieval alone, with the full WeAgent-Harness achieving the best performance. On the agent side, SFT substantially improves performance, while RL further benefits



Figure 7: Qualitative comparison on knowledge-intensive generation and editing.

Table 3: Ablation study of WeAgent-MMGenEdit.

Ablation						Performance	
Search Tools	Verify Tools	Integrate Tools	Agent SFT	Agent RL	Process Reward	Gen. W-Avg	Edit. W-Avg
Harness Ablation							
✓	×	×	–	–	–	47.22	42.14
✓	✓	×	–	–	–	49.03	43.30
✓	×	✓	–	–	–	48.70	42.91
✓	✓	✓	–	–	–	51.86	45.80
Agent Post-Training Ablation							
✓	✓	✓	✓	×	–	60.83	58.56
✓	✓	✓	✓	✓	×	62.70	60.80
✓	✓	✓	✓	✓	✓	**64.31**	**62.52**

from checklist-grounded process rewards, reaching $64.31/62.52$ W-Avg on generation/editing. On the image side, multi-reference SFT improves conditioning on heterogeneous references, and subsequent RL further enhances their faithful execution, reaching $47.39/49.98$. Together, these results validate the contributions of WeAgent-Harness and our two-sided post-training recipe.

8 Conclusion

In this paper, we present WeAgent-MMGenEdit, a full-stack framework for multimodal knowledge-intensive image generation and editing that jointly addresses external evidence acquisition, multimodal verification and integration, agentic post-training, and multi-reference image generation and editing. We introduce WeAgent-Harness, a unified multimodal runtime that supports evidence retrieval, explicit visual verification, and structured evidence integration for final generation. We further construct a checklist-grounded data construction and evaluation pipeline, together with agent-side SFT and RL, to improve the acquisition and organization of external multimodal knowledge. On the image side, we introduce multi-reference SFT and reinforcement learning to better utilize heterogeneous visual references produced by agentic workflows. Additionally, we establish WeBench-MMGenEdit, a bilingual benchmark covering both knowledge-intensive generation and multi-image editing with process-level and output-level evaluation. Extensive experiments show that WeAgent-MMGenEdit improves diverse image backends, outperforms existing open-source agentic systems on WeBench-MMGenEdit and public benchmarks, and enables a 30B-A3B policy to surpass similarly sized models while approaching the performance of a 1T-parameter agent with only about 3% of the parameters.

9 Limitations

Due to inherent bottlenecks in image generation models, rendering very small text remains challenging. Furthermore, accurate knowledge acquisition relies on the availability and stability of retrieval tools.

10 AI USE STATEMENT

Generative AI models were used for task and data synthesis, agentic trajectory collection, and model-based evaluation as part of the proposed methodology. Generative AI models were also used for minor language polishing. All AI-assisted content was reviewed and verified by the authors.

REFERENCES

Qwen-image-edit-2509. https://huggingface.co/Qwen/Qwen-Image-Edit-2509, 2025.

Kimi-k2.6. https://huggingface.co/moonshotai/Kimi-K2.6, 2026.

Qwen3.5-35b-a3b. https://huggingface.co/Qwen/Qwen3.5-35B-A3B, 2026.2.

Gemini-3.1-flash-image. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-image, 2026.2.

Gemini-3.1-flash-lite-image. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-lite-image, 2026.3.

Qwen3.6-35b-a3b. https://huggingface.co/Qwen/Qwen3.6-35B-A3B, 2026.4.

Gpt-image-2. https://developers.openai.com/api/docs/models/gpt-image-2, 2026.4.

Gpt-5.6-sol. https://openai.com/zh-Hans-CN/index/previewing-gpt-5-6-sol/, 2026.6.

Seedream5.0. https://seed.bytedance.com/zh/blog/beyond-generation-it-understands-design-introducing-seedream-5-0-pro, 2026.7.

Claude-opus-5. https://www.anthropic.com/news/claude-opus-5, 2026.7.

Qwen-image-3.0. https://qwen.ai/blog?id=qwen-image-3.0, 2026.7.

Qwen3.8-27b. https://huggingface.co/Qwen/Qwen3.8-27B, 2026.8.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Chieh-Yun Chen, Min Shi, Gong Zhang, and Humphrey Shi. T2i-copilot: A training-free multi-agent text-to-image system for enhanced prompt interpretation and interactive generation. In *ICCV*, 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*, 2026a.

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. *arXiv preprint arXiv:2605.21605*, 2026b.

Tingyang Chen, Shuo Lu, Kang Zhao, Weicheng Meng, Hanlin Teng, Tianhao Li, Chao Li, Xule Liu, Jian Liang, Zhizhong Zhang, et al. Harnessx: A composable, adaptive, and evolvable agent harness foundry. *arXiv preprint arXiv:2606.14249*, 2026c.

Zheng Chu, Xiao Wang, Jack Hong, Huiming Fan, Yuqi Huang, Yue Yang, Guohai Xu, Chenxiao Zhao, Cheng Xiang, Shengchao Hu, et al. Redsearcher: A scalable and cost-efficient framework for long-horizon search agents. *arXiv preprint arXiv:2602.14234*, 2026.

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 *ICML*, 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.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models. In *ACL*, 2024.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *NeurIPS*, 2024.

Yue Huang, Wenjie Wang, Han Bao, Yuchen Ma, Xiaonan Luo, Yi Nian, Haomin Zhuang, Zheyuan Liu, Yue Zhao, and Xiangliang Zhang. Memoharness: Agent harnesses that learn from experience. *arXiv preprint arXiv:2607.14159*, 2026.

Kaixun Jiang, Yuzheng Wang, Junjie Zhou, Pandeng Li, Zhihang Liu, Chen-Wei Xie, Zhaoyu Chen, Yun Zheng, and Wenqiang Zhang. Genagent: Scaling text-to-image generation via agentic multimodal reasoning. *arXiv preprint arXiv:2601.18543*, 2026.

Siqi Kou, Jiachun Jin, Zetong Zhou, Ye Ma, Yugang Wang, Quan Chen, Peng Jiang, Xiao Yang, Jun Zhu, Kai Yu, et al. Think-then-generate: Reasoning-aware text-to-image diffusion with llm encoders. *arXiv preprint arXiv:2601.10332*, 2026.

Black Forest Labs. Flux2. https://bfl.ai/models/flux-2, 2026.

Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Feize Wu, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, et al. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025.

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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Qianyu Meng, Yanan Wang, Liyi Chen, Yihang Li, Wei Wu, Wenyuan Jiang, Qimeng Wang, Chengqiang Lu, Yan Gao, Yi Wu, et al. Agent harness for large language model agents: A survey. 2026.

Linyue Pan, Lexiao Zou, Shuo Guo, Jingchen Ni, and Hai-Tao Zheng. Natural-language agent harnesses. *arXiv preprint arXiv:2603.25723*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *CVPR*, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Shuai Shao, Kangning Zhang, Qingyao Li, Shijian Wang, Hao Wang, Wenxiang Jiao, Yuan Lu, Yi Guo, Weiwen Liu, and Weinan Zhang. Harness-r1: Learning to edit executable runtime harnesses from agent failure trajectories. *arXiv preprint arXiv:2608.02276*, 2026.

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.

Kaiyue Sun, Weiyang Jin, Chengqi Duan, Rongyao Fang, Xian Liu, Yuwei Niu, Chunwei Wang, Aoxue Li, and Xihui Liu. Universe: Empower unified generation with reasoning and knowledge. In *CVPR*, 2026.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *ICCV*, pp. 14940–14950, 2025.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Ziwei Chai, Y Charles, HS Che, Cheng Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Haozhe Wang, Weijia Feng, Jinpeng Yu, Che Liu, Ping Nie, Fangzhen Lin, Jiaming Liu, Ruihua Huang, Jimmy Lin, Wenhu Chen, et al. Search beyond what can be taught: Evolving the knowledge boundary in agentic visual generation. *arXiv preprint arXiv:2607.05382*, 2026.

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing. In *NeurIPS*, 2024.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

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

Keming Wu, Sicong Jiang, Max Ku, Ping Nie, Minghao Liu, and Wenhu Chen. Editreward: A human-aligned reward model for instruction-guided image editing. *arXiv preprint arXiv:2509.26346*, 2025b.

Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, and Klara Nahrstedt. Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. In *ICLR*, volume 2026, 2026.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *NeurIPS*, 2024.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. In *CVPR*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Hui Zhang, Dexiang Hong, Maoke Yang, Yutao Cheng, Zhao Zhang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatidesign: A unified multi-conditional diffusion transformer for creative graphic design. In *ICLR*, 2026a.

Hui Zhang, Juntao Liu, Zongkai Liu, Liqiang Niu, Fandong Meng, Zuxuan Wu, and Yu-Gang Jiang. Weedit: A dataset, benchmark and glyph-guided framework for text-centric image editing. *arXiv preprint arXiv:2603.11593*, 2026b.

Liujie Zhang, Benzhe Ning, Rui Yang, Xiaoyan Yu, Jiaxing Li, Lumeng Wu, Jia Liu, Minghao Li, Weihang Chen, Weiqi Hu, et al. Relax: An asynchronous reinforcement learning engine for omni-modal post-training at scale. *arXiv preprint arXiv:2604.11554*, 2026c.

Zekai Zhang, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiaoyue Chen, et al. Qwen-image-agent: Bridging the context gap in real-world image generation. *arXiv preprint arXiv:2606.26907*, 2026d.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025a.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025b.