FROM EVALUATION TO EVOLUTION: BENCHMARKING IMAGE EDITING AND EVOLVING IMAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world image editing requires coordinating multiple transformations with preservation and reference constraints, yet aggregate scores can obscure unmet requirements. Useful evaluation must therefore reveal local failures and translate them into actionable priorities for subsequent model improvement. We introduce WEGENEDITBENCH, a benchmark of 2,740 instances and 148 task types spanning single-image, text-intensive, and multi-reference editing. Shared, pre-generated checklists decompose each request into atomic requirements and connect scores and deductions to visual evidence across edit execution, preservation, visual quality, and task-specific correctness. To translate diagnosis into training priorities, WEGENEVOSTUDIO combines failure analysis, retrieval-grounded data construction, supervised adaptation, and re-evaluation in a unified multi-agent framework. Among 15 evaluated configurations, GPT-Image-2 leads single-image editing with an Overall score of 4.076, while Seedream-5.0-Pro leads text-intensive editing at 4.291. Across four generation and editing benchmarks, adaptation yields task-dependent gains; on their report-selected Top-5 single-image targets, FLUX.2 [dev] improves Overall from 1.083 to 3.517, while Qwen-Image-Edit-2511 gains 13.2%. Together, our benchmark and framework provide a foundation for developing more reliable image models through interpretable evaluation and targeted, feedback-driven learning. The benchmark and code will be publicly released.



Figure 1: From evaluation to evolution. Aggregate scores can obscure unmet editing requirements and leave training priorities unspecified (left). WEGENEDITBENCH exposes these requirements through sample-specific checklists, scores, and rationales; WEGENEVOSTUDIO turns the resulting evidence into targeted data construction, model updates, and re-evaluation (right). Only selected checklist items are shown.

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.

1 INTRODUCTION

Rapid advances in text-to-image and unified multimodal models are expanding natural-language image editing. Proprietary systems such as GPT-Image-2 and Nano Banana Pro, alongside open-weight models such as Qwen-Image-Edit and FLUX.2 [dev], demonstrate strong instruction understanding and precise editing (OpenAI, 2026; Raisinghani, 2025; Qwen Team, 2025a; Black Forest Labs, 2025). Practical requests increasingly couple multiple transformations with preservation and reference constraints, such as replacing a product background while retaining packaging text and appearance. Evaluation must check whether these requirements are jointly satisfied: visual plausibility or a successful salient edit does not establish that the complete request has been fulfilled.

Existing benchmarks advance editing evaluation along complementary directions. GEdit and ImgEdit cover common operations, with ImgEdit also assessing interactive editing (Liu et al., 2025; Ye et al., 2025b); Complex-Edit, GEditBench v2, UnicBench, and ConceptEdit extend evaluation to compositional operations, open-ended requests, reasoning, and fine-grained concepts, respectively (Yang et al., 2025; Jiang et al., 2026; Ye et al., 2025a; Cui et al., 2026). Evaluation protocols are also becoming more granular: ComplexBench-Edit scores individual instructions and compositional dependencies (Wang et al., 2025), Qwen-Image-Bench introduces hierarchical criteria for text-to-image generation (Li et al., 2026), and Mind-Bench uses checklists for knowledge- and reasoning-driven generation (He et al., 2026). Building on these advances, general image editing requires sample-specific criteria grounded in source images to verify requested changes, preservation conditions, and reference assignments. Dimension-level summaries alone can obscure local omissions and constraint violations, leaving specific unmet requirements unidentified.

To this end, we introduce WEGENEDITBENCH, comprising 2,740 samples and 148 task types across single-image, text-intensive, and multi-reference editing. Real-world requests guide automated construction and human calibration. Pre-generated checklists shared across models decompose each sample into atomic requirements under Edit Goal (EG), Preservation (PR), Visual Quality (VQ), and Domain-specific Capability (DS). A fixed VLM judge records item-level rationales and deductions before assigning dimensional scores, making omitted edits, unintended changes, and incorrect reference binding traceable to specific requirements and visual evidence.

Benchmarking is only the starting point of the next model-improvement cycle. Data-centric development and feedback optimization motivate using evaluation to improve training data and model behavior, as exemplified by DataPerf and DDPO (Mazumder et al., 2023; Black et al., 2024). For image editing, OmniEdit and AnyEdit improve supervision through specialist models and broader task coverage (Wei et al., 2025; Yu et al., 2025), while ConceptEdit scales concepts and supervision density (Cui et al., 2026). DataEvolver explores construction-time feedback and downstream-validation-driven data iteration (Yan et al., 2026; Zhang et al., 2026); IterComp and JarvisEvo jointly refine generators or editors and their evaluators (Zhang et al., 2025; Lin et al., 2026). These efforts leave a complementary planning challenge for general editing: translating a target model's unmet requirements into subsequent data needs and revising construction priorities as its capabilities change. Category averages identify weak tasks but do not, by themselves, specify recurring missed constraints or the supervision needed to address them.

To connect diagnosis with learning, we introduce WEGENEVOSTUDIO, a unified multi-agent framework. The Diagnosis Agent uses category statistics, case-level rationales, and visual verification to formulate task-level data requirements; the Construction Agent uses these requirements to retrieve concepts and new source images and construct targeted editing instructions. The resulting supervision updates the model, whose re-evaluation informs subsequent training priorities. Experiments across four editing and generation benchmarks reveal requirement-level capability gaps and targeted adaptation gains alongside cross-task trade-offs. Our main contributions are:

- We introduce WEGENEDITBENCH, with 2,740 samples and 148 task types spanning single-image, text-intensive, and multi-reference editing. Its broad coverage of complex practical tasks and sample-specific atomic checklists connect scores to requirement-level rationales and visual evidence.
- We propose WEGENEVOSTUDIO, a unified multi-agent framework that connects benchmark failure diagnosis and training-goal planning to retrieval-grounded data construction and supervised model updates, with re-evaluation guiding subsequent evolution.

- We systematically evaluate 15 open- and closed-source editing configurations to identify gaps in execution, preservation, and reference handling, and conduct model-evolution experiments with Qwen-Image and FLUX.2 [dev] demonstrating targeted gains alongside cross-task trade-offs.

2 Related Work

Image Editing Benchmarks. Image editing benchmarks assess instruction following, preservation, and visual quality (Liu et al., 2025; Ye et al., 2025b;a). EditBench evaluates text-guided inpainting (Wang et al., 2023); MagicBrush provides annotated single- and multi-turn edits (Zhang et al., 2023). GEdit covers common operations (Liu et al., 2025), while ImgEdit introduces task-specific criteria and multi-turn evaluation (Ye et al., 2025b). Concise requests, however, underrepresent interacting edits and preservation constraints. Recent benchmarks examine reasoning in UnicBench (Ye et al., 2025a), open-ended requests in GEditBench v2 (Jiang et al., 2026), and compositional difficulty through operation counts in Complex-Edit (Yang et al., 2025). ComplexBench-Edit scores instructions and their dependencies (Wang et al., 2025). Yet these scores can obscure failed targets, attributes, or constraints; jointly verifying them across single-image, text-intensive, and multi-reference scenarios remains underexplored.

Agentic Multimodal Data Construction. Closing capability gaps requires diverse, reliable supervision (Zhao et al., 2024; Cui et al., 2026). InstructPix2Pix synthesizes editing pairs with language and diffusion models (Brooks et al., 2023). OmniEdit combines specialist supervision with quality-based sampling (Wei et al., 2025). AnyEdit diversifies tasks through adaptive synthesis and filtering (Yu et al., 2025). These pipelines largely follow predefined task inventories and quality criteria. Agentic approaches add adaptability: Gen-n-Val couples prompt generation with visual validation (Huang et al., 2026); ScaleEditor coordinates source expansion, editing, and verification through hierarchical agents (Chen et al., 2026a). For text-rich generation, DataEvolver revises retrieval and synthesis using rejection feedback (Yan et al., 2026). Construction quality, however, does not specify which supervision a target model needs to address recurring failures.

Feedback-Driven Learning. Feedback supports refinement and adaptation (Madaan et al., 2023; Shinn et al., 2023; Black et al., 2024). Self-Refine revises outputs through self-critique (Madaan et al., 2023); Reflexion retains verbal feedback for subsequent decisions (Shinn et al., 2023). Neither updates model parameters. DDPO instead optimizes diffusion policies against rewards (Black et al., 2024). Recent methods also adapt feedback: IterComp updates compositional reward models and generators (Zhang et al., 2025); JarvisEvo co-optimizes editing and evaluation policies (Lin et al., 2026). Visual DataEvolver combines sample correction with validation-driven data expansion for object rotation (Zhang et al., 2026). A planning challenge remains: translating unmet editing requirements into subsequent training tasks with traceable links between evidence and supervision.

3 WeGenEditBench

3.1 Real-World Scenarios and Task Coverage

To assess joint fulfillment of real-world editing requirements, WeGenEditBench spans three complementary settings: **single-image editing**, **text-intensive editing**, and **multi-reference editing** (Figure 2, left). Its taxonomy derives from over 16,000 manually collected requests and application examples from public internet sources, model documentation, and GitHub projects. Single-image tasks cover local, portrait, aesthetic, and reasoning-oriented edits; text-intensive tasks cover content, typography, and layout; multi-reference tasks test binding identity, product, style, and scene attributes to designated sources. Automated construction with human calibration yields 2,740 instances and 148 task types across 19 setting-specific families. Types count subset–label pairs, not disjoint operations; this organization connects application contexts to practical preservation, layout, and reference constraints.

3.2 Grounded Composition and Instruction Complexity

To keep instructions actionable, we compose compatible operations around visible subjects in source images, explicitly specifying preservation, layout, and reference constraints while excluding unsup-

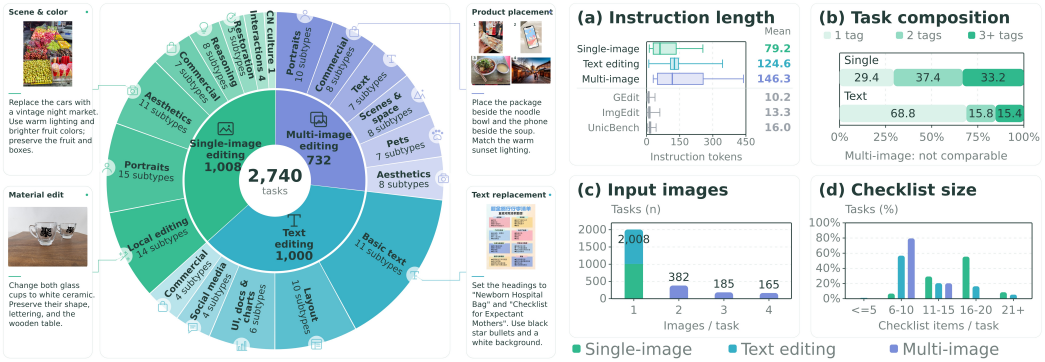

Figure 2: Taxonomy and characteristics of WEGENEDITBENCH. Left: settings, task families, examples, instance counts (inner), and subtypes (outer). Right: distributions of instruction length (a), tag multiplicity (b), input images (c), and checklist items (d). Colors indicate settings; lengths preserve source languages.

ported or contradictory combinations. Human review calibrates these compositions so that complexity reflects coordinated, plausible requirements (Appendix A.2).

Figure 2(a–c) characterizes the resulting instances. Mean instruction lengths are 79.2, 124.6, and 146.3 tokens for the three subsets, versus 10.2–16.0 for GEdit, ImgEdit, and UnicBench in this comparison. Multiple task labels occur in 70.6% of single-image and 31.2% of text-intensive instances, while multi-reference tasks use two to four images with distinct roles. These demands probe the limits of editing under interacting transformations, fine-grained preservation, and reference binding, exposing bottlenecks in satisfying multiple requirements simultaneously.

3.3 EVIDENCE-GROUNDED EVALUATION AND FAILURE DIAGNOSIS

To expose failures hidden by aggregate scores, we assess four dimensions, where applicable: **Edit Goal (EG)**, **Preservation (PR)**, **Visual Quality (VQ)**, and **Domain-specific Capability (DS)**, measuring execution, preservation, rendering quality, and task-specific correctness. From each instruction and its source images, we derive independently verifiable requirements specifying the target, attribute or location, and expected state. Text editing separates text-side and image-side checks; multi-reference editing binds attributes to designated sources. Checklists are generated before output evaluation, fixed within each evaluation version, and shared across compared models. Figure 2(d) shows their differing sizes: single-image checklists most often contain 16–20 items, versus 6–10 for text-intensive and multi-reference editing.

Holistic dimensional judgments can overemphasize visually salient edits while overlooking missed targets or unintended changes (Figure 3). Our fixed VLM judge records rationales before assigning item-level deductions, then assigns 0–5 dimensional scores based on each item's importance under the scoring rubric. In the illustrated case, silver boots coexist with an unchanged ornament: PR=VQ=5, yet separate checks expose the missed target and material inconsistency (EG=2, DS=0). Judgments thus remain traceable to specific unmet requirements and visual evidence. Three trained experts randomly audit 10% of cases in each benchmark; checklist corrections trigger re-evaluation of all compared models on affected cases using the same revised checklist.

Subset-specific aggregation. Single-image editing weights EG, PR, and VQ and subtracts the mean penalty over applicable DS task types T:

$$S_{\text{single}} = \text{clip}_{[0,5]}\left(0.6S_{\text{EG}} + 0.2S_{\text{PR}} + 0.2S_{\text{VQ}} - \frac{1}{|T|}\sum_{t\in T}\frac{5 - S_{\text{DS},t}}{4}\right). \tag{1}$$

The penalty is zero when T is empty. Text-intensive editing scores both sides separately, combining them when image-side requirements exist; otherwise, only S_{text} is used. Multi-reference editing combines goal fulfillment, reference-attribute preservation, and visual quality without DS score:

$$S_{\text{text-edit}} = 0.75S_{\text{text}} + 0.25S_{\text{image}}, \qquad S_{\text{multi}} = 0.4S_{\text{EG}} + 0.4S_{\text{PR}} + 0.2S_{\text{VQ}}. \tag{2}$$

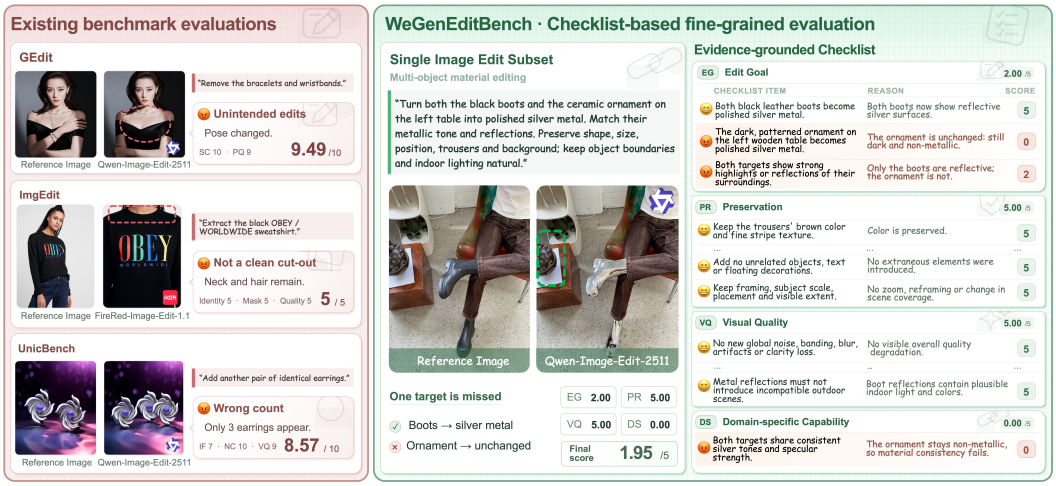

Figure 3: From scores to requirement-level diagnosis. Left: selected cases receive high scores despite visible errors. Right: WEGENEDITBENCH identifies an omitted edit through checklist scores and rationales. Cases and score scales differ across benchmarks.

Each subset's Overall averages valid instance scores, counting each instance once. Rubrics, human auditing, and invalid-response handling appear in Appendix B; the retained item-level evidence supports capability diagnosis in WEGENEVOSTUDIO.

4 WEGENEVOSTUDIO

4.1 UNIFIED EVALUATION-TO-LEARNING FRAMEWORK

To turn the unmet requirements in Section 3 into learning objectives, we introduce WEGENEVOSTUDIO, a unified framework centered on a shared, model-specific data prescription (Figure 4). For checkpoint θ_t, evaluation records $\mathcal{E}_t$ contain instructions, source/output images, scores, and available rationales. Diagnosis translates recurring failures into prescriptions $\mathcal{T}_t$, guiding construction of new tasks $\mathcal{X}_t$ and subsequent supervised adaptation:

$$\theta_t \xrightarrow{\text{evaluate}} \mathcal{E}_t \xrightarrow{\text{diagnose}} \mathcal{T}_t \xrightarrow{\text{construct}} \mathcal{X}_t \xrightarrow{\text{supervise/train}} \widetilde{\theta}_{t+1}. \tag{3}$$

To carry diagnostic evidence across stages, an agentic harness links checkpoints, evaluation records, prescriptions, and data revisions. In this shared context, an LLM delegates diagnosis and planning to the **Diagnosis Agent**, and retrieval-grounded synthesis to the **Construction Agent**. Scoped tools and validated resource interfaces preserve the evidence behind construction decisions. Validation selects the next checkpoint, whose re-evaluation refreshes the prescription so that learning goals respond to persistent or newly exposed weaknesses. Implementation details and execution contracts appear in Appendix C.1.

4.2 FROM DIAGNOSIS TO TARGETED DATA PLANS

Category averages locate weak areas but leave recurring failure conditions unexplained. The Diagnosis Agent combines score-based category selection with case-level review, using requirement-level reasons when available. With a compatible anchor model, roughly 60% of slots prioritize the largest relative deficits and the remainder low absolute scores, with deduplication; otherwise, selection uses the target model's own statistics. Within each category, it contrasts representative or severe failures with stronger cases across diagnostic dimensions. VLM-assisted review checks instructions and source/output images for omitted operations, affected objects, and unintended changes. Cross-case synthesis yields a capability profile Φ_t linking recurring patterns to supporting cases and conditions; hypotheses about training deficiencies remain distinct from observed failures.

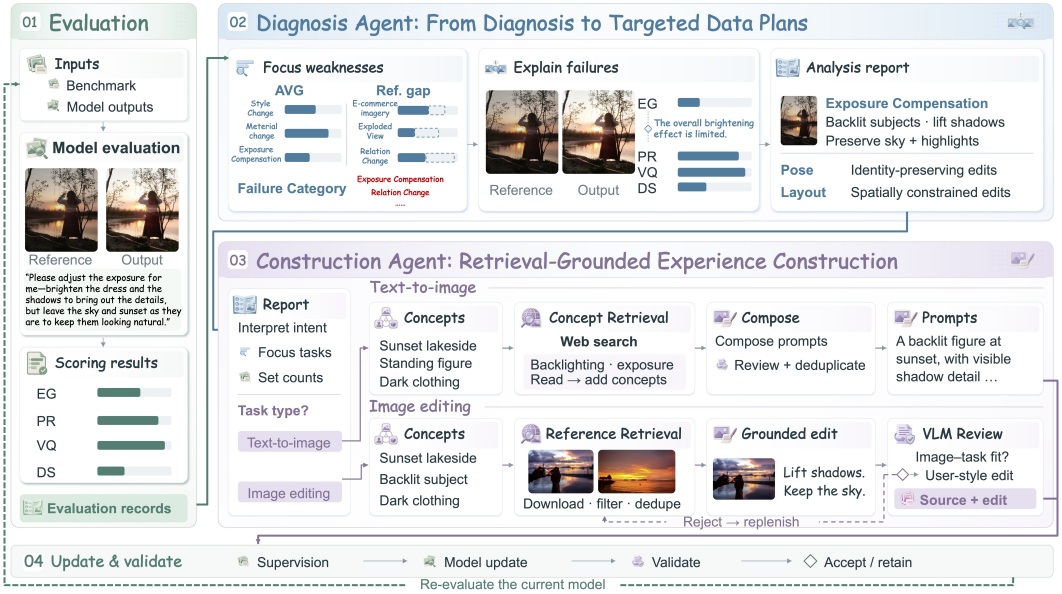

Figure 4: WEGENEVOSTUDIO translates evaluation evidence into targeted training data. The Diagnosis Agent selects weaknesses and forms data prescriptions; the Construction Agent retrieves concepts or source images to synthesize generation and editing tasks. Supervised updates and re-evaluation close the loop. The exposure-editing example and score bars illustrate the workflow.

To make these observations reusable for data collection, the prescription $\mathcal{T}_t$ identifies weak task types and, when evidence is available, summarizes source conditions and capability requirements. Supporting cases justify the diagnosis within the report; the construction interface receives task labels and optional category notes. For example, preservation failures during background replacement motivate diverse sources with clearly visible products and explicit product-preservation requirements. These conditions guide acquisition and synthesis without reconstructing the failed benchmark scene. Figure 7 details the diagnosis procedure.

4.3 RETRIEVAL-GROUNDED EXPERIENCE CONSTRUCTION

To address diagnosed needs in new visual contexts, the Construction Agent maps selected tasks to our taxonomy and allocates examples across categories. All benchmarks use a shared generator without importing their construction templates. Category-specific queries retrieve editing sources from a library or the web, followed by format, resolution, and duplicate checks. For text-to-image generation, web-text research expands task concepts to reduce reliance on entities the LLM frequently proposes. Feedback sets capability priorities; retrieval supplies visual instances for training.

To avoid infeasible or weakly targeted instructions, each generation call grounds the selected task and category note in a retrieved source image, not an evaluated input–output example. It specifies visible targets, intended changes, and preservation conditions, rejecting incompatible sources or forced combinations. VLM review and rewriting produce varied, feasible tasks while retaining the capability focus. Labels and provenance link tasks to the originating report (Figure 8); the supervision pipeline then supplies training targets. This is benchmark-informed task-level adaptation; instance-reuse boundaries and the scope of reported generalization are specified in Appendix C.4.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Benchmark evaluation. We evaluate 15 configurations: GPT-Image-2 (OpenAI, 2026), Nano-Banana-Pro (Raisinghani, 2025), Qwen-Image-3.0 (Qwen Team, 2026), and Seedream-5.0-Pro (ByteDance Seed, 2026) are closed-source; open-source configurations comprise HunyuanImage-3.0-Instruct-Distil (Tencent Hunyuan Foundation Model Team, 2025) and HunyuanImage-3.0-

Table 1: Quantitative comparison on WEGENEDITBENCH-SINGLEIMAGEEDIT. Bold and underlined scores denote the best and second best.

Model Name	EG ↑	PR ↑	VQ ↑	DS ↑	Overall ↑
Closed-source Models					
GPT-Image-2	**4.298**	4.049	**4.463**	<u>4.182</u>	**4.076**
Nano-Banana-Pro	4.159	3.752	4.370	4.050	3.882
Qwen-Image-3.0	4.120	3.469	4.331	4.126	3.813
Seedream-5.0-Pro	<u>4.207</u>	<u>4.084</u>	<u>4.434</u>	**4.206**	<u>4.029</u>
Open-source Models					
HunyuanImage-3.0-Instruct-Distil	4.020	**4.175**	4.280	3.952	3.841
HunyuanImage-3.0-Instruct-Distil-CoT	4.071	4.065	4.265	4.008	3.862
FireRed-Image-Edit-1.1	4.091	3.550	4.227	3.902	3.736
JoyAI-Image-Edit	4.042	3.547	4.169	3.794	3.667
Qwen-Image-Edit-2511	3.963	3.567	4.095	3.708	3.588
Boogu-Image-0.1-Edit	3.850	3.718	4.092	3.768	3.564
SenseNova-U1-8B-MoT-CoT	3.948	3.636	3.932	3.605	3.535
SenseNova-U1-8B-MoT	4.030	3.408	3.917	3.674	3.552
LongCat-Image-Edit	3.761	3.439	4.003	3.643	3.410
HiDream-O1-Image	3.158	2.942	3.855	3.231	2.816
FLUX.2 [dev]	2.056	3.343	3.393	1.975	1.828

Instruct-Distil-CoT, FireRed-Image-Edit-1.1 (FireRed Team, 2026), JoyAI-Image-Edit (Song et al., 2026), Qwen-Image-Edit-2511 (Qwen Team, 2025a), Boogu-Image-0.1-Edit (Chen et al., 2026b), SenseNova-U1-8B-MoT and SenseNova-U1-8B-MoT-CoT (Diao et al., 2026), LongCat-Image-Edit (Meituan LongCat Team et al., 2025), HiDream-O1-Image (Cai et al., 2026), and FLUX.2 [dev] (Black Forest Labs, 2025). All 15 cover single-image and text-intensive editing, and six cover multi-reference editing. All core benchmark scores (Tables 1–3) use `gemini-3.1-pro-preview` with shared checklists and the subset-specific protocol in Section 3.3.

Model evolution. We adapt Qwen-Image-2512 for generation, Qwen-Image-Edit-2511 for editing, and FLUX.2 [dev] for both, with separate rank-16 LoRA adapters per backbone–benchmark pair (Hu et al., 2022; Qwen Team, 2025b;a; Black Forest Labs, 2025). Each pair uses 25,000 training examples, evenly allocated to Top-$K = 5$ selected categories, on 32 Cambricon MLU590 accelerators with batch size 1 per device. Experiments cover WeGenBench, GEdit, ImgEdit (WeChatCV, 2026; Liu et al., 2025; Ye et al., 2025b), and WEGENEDITBENCH-SINGLEIMAGEEDIT. GEdit/ImgEdit guide selection through category scores; WeGenBench adds native rationales, while WEGENEDITBENCH supplies checklist-level evidence. All use a shared constructor with new source instances. Settings and comparison qualifications appear in Appendix D.

5.2 SCENARIO-SPECIFIC CAPABILITIES AND FINE-GRAINED DIAGNOSIS

Scenario-dependent strengths. Tables 1–3 show that leadership depends on the editing scenario. GPT-Image-2 leads single-image and multi-reference Overall (4.076 and 4.138), but Seedream-5.0-Pro surpasses it on text-intensive editing (4.291 versus 4.132). Seedream leads all four text-side dimensions, whereas GPT-Image-2 retains stronger image-side EG (4.604 versus 4.362). The reversal therefore reflects textual execution, preservation, and presentation rather than uniformly stronger visual editing. Among open-source configurations, HunyuanImage-3.0-Instruct-Distil-CoT leads single-image and text-intensive Overall (3.862 and 3.882). These results motivate evaluating complementary scenarios: single-image leadership alone does not establish superiority on text-aware or reference-conditioned requests.

Text and reference bottlenecks. Separating requirements further identifies where competitive models remain unreliable. JoyAI-Image-Edit achieves image-side EG of 4.565 on text-intensive tasks, but text-side EG is only 3.900: completing the visual change does not ensure correct text execution. For multi-reference editing, HunyuanImage-3.0-Instruct-Distil is the strongest evaluated open-source model (Overall 3.327), yet its EG and PR (3.279 and 3.493) trail GPT-Image-2 (4.358 and 4.134). This profile exposes gaps in both requested composition and fidelity to assigned reference attributes. The scenario and dimension splits distinguish these bottlenecks from general visual quality, giving more specific capability boundaries than a single ranking.

Table 2: Quantitative comparison on WEGENEDITBENCH-TEXTEDIT. Bold and underlined scores denote the best and second best.

Model Name	Text-side				Image-side				Overall ↑
	EG ↑	PR ↑	VQ ↑	DS ↑	EG ↑	PR ↑	VQ ↑	DS ↑	
Closed-source Models									
GPT-Image-2	4.289	4.031	<u>4.329</u>	<u>4.399</u>	**4.604**	3.919	<u>4.421</u>	**4.452**	<u>4.132</u>
Nano-Banana-Pro	4.295	<u>4.092</u>	4.310	4.322	4.396	<u>4.208</u>	4.232	4.306	4.116
Qwen-Image-3.0	<u>4.367</u>	3.732	<u>4.329</u>	4.385	4.509	3.869	4.406	4.414	4.109
Seedream-5.0-Pro	**4.415**	**4.327**	**4.416**	**4.504**	4.362	4.207	**4.455**	<u>4.429</u>	**4.291**
Open-source Models									
SenseNova-U1-8B-MoT-CoT	4.039	3.880	3.844	3.943	4.340	4.092	3.935	4.090	3.762
HunyuanImage-3.0-Instruct-Distil	3.914	3.801	3.917	4.133	4.361	4.177	4.109	4.214	3.742
HunyuanImage-3.0-Instruct-Distil-CoT	4.090	3.807	4.102	4.188	4.422	**4.258**	4.194	4.323	3.882
SenseNova-U1-8B-MoT	4.013	3.809	3.902	3.924	4.408	3.830	3.895	4.099	3.733
JoyAI-Image-Edit	3.900	3.617	3.751	3.975	<u>4.565</u>	3.925	4.129	4.287	3.630
FireRed-Image-Edit-1.1	3.712	3.308	3.646	3.890	4.497	3.622	3.932	4.245	3.423
Qwen-Image-Edit-2511	3.658	3.583	3.655	3.820	4.384	3.830	3.969	4.112	3.429
Boogu-Image-0.1-Edit	3.842	3.180	3.732	3.915	4.412	3.459	3.898	4.197	3.482
LongCat-Image-Edit	3.027	3.560	3.385	3.320	4.187	3.639	3.789	3.949	2.901
HiDream-O1-Image	2.893	2.541	3.213	3.145	3.568	3.051	3.483	3.539	2.502
FLUX.2 [dev]	2.183	3.313	3.640	2.384	2.724	3.331	3.570	2.782	2.140

Table 3: Quantitative comparison on WEGENEDITBENCH-MULTIIMAGEEDIT. Bold and underlined scores denote the best and second best.

Model Name	EG ↑	PR ↑	VQ ↑	Overall ↑
Closed-source Models				
GPT-Image-2	**4.358**	**4.134**	<u>3.707</u>	**4.138**
Nano-Banana-Pro	4.031	3.993	3.340	3.877
Seedream-5.0-Pro	<u>4.270</u>	<u>4.093</u>	**3.750**	<u>4.095</u>
Open-source Models				
HunyuanImage-3.0-Instruct-Distil	3.279	3.493	3.091	3.327
FireRed-Image-Edit-1.1	2.714	2.977	3.055	2.887
Qwen-Image-Edit-2511	2.589	2.884	3.093	2.808

Beyond aggregate success. Strong aggregate results still leave individual requirements unresolved. Of GPT-Image-2's 644 single-image outputs scoring at least 4, 315 contain a checklist item scored at most 2 (48.9%). In multi-reference editing, 108 GPT-Image-2 outputs share identical dimensional and Overall scores, yet 44 contain such an item and 64 do not. On 280 shared tasks with both text-side and image-side requirements, only 25.0% of GPT-Image-2 and 27.1% of Seedream outputs score at least 4 on every applicable item on both sides. Thus high scores or success on one component do not establish joint fulfillment. Checklist records identify the missed requirement and its rationale; analysis and validation scope appear in Appendix B.4.

5.3 FEEDBACK-DRIVEN ADAPTATION AND CAPABILITY EVOLUTION

Adaptation across feedback formats. Table 4 tests WEGENEVOSTUDIO with native benchmark feedback, without requiring WEGENEDITBENCH checklists. On report-selected Top-5 unions, Qwen/FLUX.2 gain 0.968/0.253 Final on WeGenBench; on WEGENEDITBENCH-SINGLEIMAGEEDIT, Qwen gains 0.373 Overall and FLUX.2 improves from 1.083 to 3.517. Score-guided GEdit adaptation improves FLUX.2 Overall by 3.944 on targets and 3.009 overall (Appendix D.3). These results support the framework's utility across feedback formats and generation/editing tasks. Matched ablations remain necessary to isolate the added benefit of fine-grained reasons.

Changes in editing behavior. Figure 5 shows concrete behavioral changes after Qwen adaptation: raising an arm instead of placing the hand at the hip, producing an exploded coffee-machine arrangement instead of a cutaway, and preserving white shoelaces during recoloring. These cases illustrate more faithful operation selection and attribute preservation. Additional generation and editing comparisons appear in Appendix G.4.

Table 4: Quantitative comparison on report-selected Top-5 categories across four benchmarks. Target sets differ by backbone; green/red marks positive/negative Δ.

Model	T2I	Image editing											
	WeGen Bench	WeGenEditBench-SingleImageEdit					GEdit			ImgEdit			
	F ↑	EG ↑	PR ↑	VQ ↑	DS ↑	F ↑	SC ↑	PQ ↑	O ↑	I ↑	Q ↑	D ↑	F ↑
Qwen-Image	7.700	3.244	2.646	3.500	3.565	2.834	8.515	7.131	7.511	4.621	4.374	4.400	4.465
+ WeGenEvoStudio	8.668	3.705	2.650	4.105	3.437	3.208	8.172	7.241	7.476	4.621	4.416	4.450	4.496
Δ	**+0.968**	**+0.461**	**+0.003**	**+0.605**	**-0.128**	**+0.373**	**-0.343**	**+0.109**	**-0.035**	+0.000	**+0.042**	**+0.051**	**+0.031**
FLUX.2 [dev]	7.726	1.550	3.178	2.780	0.809	1.083	3.425	7.592	3.465	1.915	2.626	2.873	2.471
+ WeGenEvoStudio	7.979	3.767	3.865	4.414	3.380	3.517	7.949	8.032	7.409	3.821	3.722	3.955	3.833
Δ	**+0.253**	**+2.217**	**+0.687**	**+1.634**	**+2.571**	**+2.434**	**+4.524**	**+0.440**	**+3.944**	**+1.906**	**+1.096**	**+1.082**	**+1.362**



Figure 5: Qualitative editing comparisons of Qwen-Image-Edit-2511 and its WeGenEvoStudio adapters on GEdit, ImgEdit, and WeGenEditBench.

Capability trade-offs. Benefits remain model- and task-dependent: Qwen loses 0.035 GEdit Overall and 0.128 single-image DS on the selected categories. Targeted Overall gains therefore do not imply improvement in every dimension. These are benchmark-informed adaptation results; ImgEdit remains descriptive because inference settings are not fully matched.

6 Conclusion

We presented WeGenEditBench and WeGenEvoStudio to connect fine-grained evaluation with targeted model improvement. Composite requests require feedback that separates visual plausibility from complete requirement fulfillment and identifies what to learn next. Shared checklists expose unmet requirements, while the agentic framework turns diagnosis into training data and supervised updates. GPT-Image-2 leads single-image editing at 4.076 Overall, whereas Seedream-5.0-Pro leads text-intensive editing at 4.291. On their report-selected Top-5 single-image targets, FLUX.2 [dev] improves Overall from 1.083 to 3.517, while Qwen-Image-Edit-2511 gains 13.2%. These gains and capability trade-offs support interpretable evaluation as actionable training feedback. Extending this connection across editing scenarios can guide the development of more reliable and adaptable image models.

AI USE STATEMENT

Generative AI tools were used primarily to polish manuscript text and assist with writing portions of the experimental code, including English editing and LaTeX formatting. The research ideas, experimental approach, and framework design were independently developed by the author team. The proposed system itself employs language and vision-language models for task construction, checklist-based evaluation, failure diagnosis, and training-data construction, as described in Sections 3 and 4. The authors are responsible for the reported methods, analyses, and conclusions.

ETHICS STATEMENT

This work studies the evaluation and improvement of image-editing models. Such models can modify identities, depicted events, and textual content, creating risks of misleading imagery and privacy violations. The benchmark draws on public internet sources and application examples; public accessibility does not by itself establish redistribution rights. Use and release of source images must respect applicable licenses, privacy, and intellectual property. Our evaluation focuses on editing capability and requirement satisfaction; it does not constitute a comprehensive safety assessment.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 describe benchmark construction, scoring, and the evaluation-to-learning framework, while Section 5.1 specifies the evaluated models and adaptation settings. Appendix A details data construction and coverage; Appendix B specifies evaluation, aggregation, and human auditing; and Appendix C describes the two agents and their interfaces. Appendix D documents run-specific settings and comparison qualifications. Additional quantitative results, the complete taxonomy, and qualitative examples appear in Appendices F and G.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training Diffusion Models with Reinforcement Learning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.13301.

Black Forest Labs. FLUX.2 [dev]: Official Model Card. Hugging Face, 2025. URL https://huggingface.co/black-forest-labs/FLUX.2-dev. Accessed September 18, 2026.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023. URL https://arxiv.org/abs/2211.09800.

ByteDance Seed. Seedream 5.0 Pro. Official model announcement, 2026. URL https://seed.bytedance.com/en/seedream5_0_pro.

Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Yingwei Pan, Yi Peng, Zhaofan Qiu, Kai Yu, Yiheng Zhang, Hao Ai, Siying Bai, Yang Chen, Zhihui Chen, Fengbin Gao, Ying Guo, Dong Li, Zhen Shen, Leilei Shi, Jing Wang, Siyu Wang, Yimeng Wang, Rui Zheng, Ting Yao, and Tao Mei. HiDream-O1-Image: A Natively Unified Image Generative Foundation Model with Pixel-level Unified Transformer. *arXiv preprint arXiv:2605.11061*, 2026. URL https://arxiv.org/abs/2605.11061.

Guanzhou Chen, Erfei Cui, Changyao Tian, Danni Yang, Ganlin Yang, Yu Qiao, Hongsheng Li, Gen Luo, and Hongjie Zhang. ScaleEdit-12M: Scaling Open-Source Image Editing Data Generation via Multi-Agent Framework, 2026a. URL https://arxiv.org/abs/2603.20644.

Guoxuan Chen, Chufeng Xiao, Haoran Yang, Siyue Xie, Binxiao Huang, Ming Zhang, Cheuk Him Chau, Xinyu Fu, Yingzhao Lian, Tom S. Y. Li, Jintao Lin, Bowen Dong, Zian Qian, Yuhao Liu, Yuxuan Hu, Weikang Shi, Bin Zou, Bowen Zheng, Haoxuan Che, Chang Chen, Yuyang He, Heyang Sun, Tianyu Huang, Chong Hou Choi, Cheng Gong, Han Shi, Haoli Bai, Xihui Liu, Hongsheng Li, Qifeng Chen, Chao Huang, Rui Liu, and Chenyang Lei. Boogu-Image-0.1: Boosting Open Agentic Multimodal Generation via Understanding under a Minimal Budget. *arXiv preprint arXiv:2607.13125*, 2026b. URL https://arxiv.org/abs/2607.13125.

Long Cui, Xiaoqian Liu, Qi Qin, Yi Xin, Tao Lin, Jianguo Li, and Linfeng Zhang. Unlocking the Potential of Image Editing via Concept Scaling and Dense Supervision, 2026. URL https://arxiv.org/abs/2608.16812.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. SenseNova-U1: Unifying Multimodal Understanding and Generation with NEO-unify Architecture. *arXiv preprint arXiv:2605.12500*, 2026. URL https://arxiv.org/abs/2605.12500.

FireRed Team. FireRed-Image-Edit-1.1. Official Hugging Face model card, 2026. URL https://huggingface.co/FireRedTeam/FireRed-Image-Edit-1.1. Accessed September 23, 2026.

Feng Han, Yibin Wang, Chenglin Li, Zheming Liang, Dianyi Wang, Yang Jiao, Zhipeng Wei, Chao Gong, Cheng Jin, and Jiaqi Wang. UniREditBench: A Unified Reasoning-Based Image Editing Benchmark. In *Computer Vision – ECCV 2026*, pp. 287–304. Springer, 2026. doi: 10.1007/978-3-032-37447-9_16. URL https://link.springer.com/chapter/10.1007/978-3-032-37447-9_16.

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, 2026. URL https://arxiv.org/abs/2602.01756.

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 *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jing-En Huang, I-Sheng Fang, Tzuhsuan Huang, Yu-Lun Liu, Chih-Yu Wang, and Jun-Cheng Chen. Gen-n-Val: Agentic Image Data Generation and Validation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8786–8795, 2026. URL https://arxiv.org/abs/2506.04676.

Zhangqi Jiang, Zheng Sun, Xianfang Zeng, Yufeng Yang, Xuanyang Zhang, Yongliang Wu, Wei Cheng, Gang Yu, Xu Yang, and Bihan Wen. GEditBench v2: A Human-Aligned Benchmark for General Image Editing, 2026. URL https://arxiv.org/abs/2603.28547.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. VIEScore: Towards Explainable Metrics for Conditional Image Synthesis Evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.663. URL https://aclanthology.org/2024.acl-long.663/.

Niantong Li, Guangzheng Hu, Weixu Qiao, Ying Ba, Qichen Hong, Shijun Shen, Jinlin Wang, Fan Zhou, Jianye Kang, Xin Shang, Ziyi He, Wei Wang, Dalin Li, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiao Xu, Xiaoyue Chen, Yuxiang Chen, Yan Shu, Yanran Zhang, Yilei Chen, Yixian Xu, Zekai Zhang, Zhendong Wang, Zihao Liu, Zikai Zhou, Hongzhu Shi, Yi Wang, Bing Zhao, Hu Wei, Lin Qu, and Chenfei Wu. Qwen-Image-Bench: From Generation to Creation in Text-to-Image Evaluation, 2026. URL https://arxiv.org/abs/2605.28091.

Yunlong Lin, Linqing Wang, Kunjie Lin, Zixu Lin, Kaixiong Gong, Wenbo Li, Bin Lin, Zhenxi Li, Shiyi Zhang, Yuyang Peng, Wenxun Dai, Xinghao Ding, Chunyu Wang, and Qinglin Lu. JarvisEvo: Towards a Self-Evolving Photo Editing Agent with Synergistic Editor-Evaluator Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27291–27302, 2026. URL https://arxiv.org/abs/2511.23002.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1X-Edit: A Practical Framework for General Image Editing, 2025. URL https://arxiv.org/abs/2504.17761.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://doi.org/10.52202/075280-2019.

Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Will Cukierski, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Raje, Max Bartolo, Evan Sabri Eyuboglu, Amirata Ghorbani, Emmett Goodman, Addison Howard, Oana Inel, Tariq Kane, Christine R. Kirkpatrick, D. Sculley, Tzu-Sheng Kuo, Jonas W. Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Ce Zhang, James Y. Zou, Carole-Jean Wu, Cody Coleman, Andrew Y. Ng, Peter Mattson, and Vijay Janapa Reddi. DataPerf: Benchmarks for Data-Centric AI Development. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0235. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/112db88215e25b3ae2750e9eefcded94-Abstract-Datasets_and_Benchmarks.html.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, Xunliang Cai, Yayong Guan, and Jie Hu. LongCat-Image Technical Report. *arXiv preprint arXiv:2512.07584*, 2025. URL https://arxiv.org/abs/2512.07584.

OpenAI. GPT-Image-2. OpenAI model documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-image-2. Accessed September 23, 2026.

Qwen Team. Qwen-Image-Edit-2511: Official Model Card. Hugging Face, 2025a. URL https://huggingface.co/Qwen/Qwen-Image-Edit-2511. Accessed September 18, 2026.

Qwen Team. Qwen-Image-2512. Official Hugging Face model card, 2025b. URL https://huggingface.co/Qwen/Qwen-Image-2512. Accessed September 23, 2026.

Qwen Team. Qwen-Image-3.0. Official model announcement, 2026. URL https://qwen.ai/blog?id=qwen-image-3.0. Accessed September 23, 2026.

Naina Raisinghani. Introducing Nano Banana Pro. Google DeepMind, 2025. URL https://blog.google/innovation-and-ai/products/nano-banana-pro/.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://doi.org/10.52202/075280-0377.

Lin Song, Wenbo Li, Guoqing Ma, Wei Tang, Bo Wang, Yuan Zhang, Yijun Yang, Yicheng Xiao, Jianhui Liu, Yanbing Zhang, Guohui Zhang, Wenhu Zhang, Hang Xu, Nan Jiang, Xin Han, Haoze Sun, Maoquan Zhang, Haoyang Huang, and Nan Duan. JoyAI-Image: Awaking Spatial Intelligence in Unified Multimodal Understanding and Generation. *arXiv preprint arXiv:2605.04128*, 2026. URL https://arxiv.org/abs/2605.04128.

Super Intelligence Team, Xiaohongshu Inc. FireRed-Image-Edit-1.0 Technical Report, 2026. URL https://arxiv.org/abs/2602.13344.

Tencent Hunyuan Foundation Model Team. HunyuanImage 3.0 Technical Report. *arXiv preprint arXiv:2509.23951*, 2025. URL https://arxiv.org/abs/2509.23951.

Chenglin Wang, Yucheng Zhou, Qianning Wang, Zhe Wang, and Kai Zhang. ComplexBench-Edit: Benchmarking Complex Instruction-Driven Image Editing via Compositional Dependencies. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. doi: 10.1145/3746027.3758301. URL https://arxiv.org/abs/2506.12830.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. Imagen Editor and EditBench: Advancing and Evaluating Text-Guided Image Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18359–18369, 2023. URL https://arxiv.org/abs/2212.06909.

WeChatCV. WeGenBench: Official Benchmark Repository. GitHub, 2026. URL https://github.com/WeChatCV/WeGenBench. Repository accessed September 18, 2026; year denotes access year.

Cong Wei, Zheyang Xiong, Weiming Ren, Xinrun Du, Ge Zhang, and Wenhu Chen. OmniEdit: Building Image Editing Generalist Models Through Specialist Supervision. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.07199.

Siyu Yan, Yizhen Gao, Yilin Wang, Dongxing Mao, and Alex Jinpeng Wang. DataEvolver: Self-Evolving Multi-Agent Data Construction for Text-Rich Image Generation, 2026. URL https://arxiv.org/abs/2606.31537.

Siwei Yang, Mude Hui, Bingchen Zhao, Yuyin Zhou, Nataniel Ruiz, and Cihang Xie. Complex-Edit: CoT-Like Instruction Generation for Complexity-Controllable Image Editing Benchmark, 2025. URL https://arxiv.org/abs/2504.13143.

Keming Ye, Zhipeng Huang, Canmiao Fu, Qingyang Liu, Jiani Cai, Zheqi Lv, Chen Li, Jing Lyu, Zhou Zhao, and Shengyu Zhang. UnicEdit-10M: A Dataset and Benchmark Breaking the Scale-Quality Barrier via Unified Verification for Reasoning-Enriched Edits, 2025a. URL https://arxiv.org/abs/2512.02790.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. ImgEdit: A Unified Image Editing Dataset and Benchmark. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://arxiv.org/abs/2505.20275.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. AnyEdit: Mastering Unified High-Quality Image Editing for Any Idea. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26125–26135, 2025. URL https://arxiv.org/abs/2411.15738.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. MagicBrush: A Manually Annotated Dataset for Instruction-Guided Image Editing. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.10012.

Qisong Zhang, Wenzhuo Wu, Zhuangzhuang Jia, Yunhao Yang, Huayu Zhang, Xianghao Zang, Zhixiang He, Zhongjiang He, Kongming Liang, and Zhanyu Ma. DataEvolver: Let Your Data Build and Improve Itself via Goal-Driven Loop Agents, 2026. URL https://arxiv.org/abs/2605.01789.

Xinchen Zhang, Ling Yang, Guohao Li, Yaqi Cai, Jiake Xie, Yong Tang, Yujiu Yang, Mengdi Wang, and Bin Cui. IterComp: Iterative Composition-Aware Feedback Learning from Model Gallery for Text-to-Image Generation. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.07171.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. UltraEdit: Instruction-based Fine-Grained Image Editing at Scale. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2407.05282.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, Hua Yang, Xue Yang, and Haodong Duan. Envisioning Beyond the Pixels: Benchmarking Reasoning-Informed Visual Editing, 2025. URL https://arxiv.org/abs/2504.02826.

Supplementary Material

CONTENTS OF THE APPENDIX

A Benchmark Construction and Coverage . . . 15
A.1 Application-Grounded Taxonomy . . . 15
A.2 Source-Grounded Curation . . . 15
A.3 Instruction and Record Consistency . . . 15
A.4 Coverage and Evaluation Resolution . . . 16
A.5 Task-Specific Strengths and Bottlenecks . . . 16
B Evaluation Protocol and Diagnostic Evidence . . . 17
B.1 Shared Requirements and Item Judgments . . . 17
B.2 Scoring, Applicability, and Aggregation . . . 17
B.3 Human Audit and Shared Revisions . . . 17
B.4 Diagnostic Scope and Statistical Conventions . . . 18
C Agentic Diagnosis and Data Construction . . . 18
C.1 Harness and Resource Contracts . . . 18
C.2 Diagnosis Agent: Evidence to Data Prescriptions . . . 18
C.3 Construction Agent: Retrieval-Grounded Task Synthesis . . . 19
C.4 Task-Level Feedback and Instance Boundaries . . . 20
D Experimental Protocol and Extended Analysis . . . 20
D.1 Shared Training Settings and Evaluation Scope . . . 20
D.2 Updates and Evaluation Records . . . 21
D.3 Targeted Gains and Broader Behavior . . . 21
D.4 Web Search and Concept-Specific Effects . . . 21
E Limitations and Future Directions . . . 21
F Supporting Results and Complete Taxonomy . . . 22
F.1 Quantitative Comparison and Diagnostic Tables . . . 22
F.2 Primary-Task Quantitative Comparisons . . . 25
F.3 Complete Task Taxonomy . . . 34
G Qualitative Case Studies . . . 38
G.1 WeGenEditBench-SingleImageEdit . . . 38
G.2 WeGenEditBench-TextEdit . . . 44
G.3 WeGenEditBench-MultiImageEdit . . . 50
G.4 Qualitative Evidence of Model Adaptation . . . 54

A BENCHMARK CONSTRUCTION AND COVERAGE

A.1 APPLICATION-GROUNDED TAXONOMY

WEGENEDITBENCH contains 2,740 instances: 1,008 single-image, 1,000 text-intensive, and 732 multi-reference tasks. Over 16,000 manually collected application requests and examples inform the taxonomy; automated construction and human refinement produce the evaluation records. The subsets contain 8/5/6 first-level families and 65/35/48 second-level types. The 148 types count subset–label pairs, not disjoint operations. Single-image and text-intensive instances may carry multiple labels; multi-reference instances have one primary label and two to four ordered inputs (382/185/165 cases). Table 5 and Appendix F.3 give counts and definitions. Instruction lengths use `tiktoken==0.12.0`, `o200k_base`, and `encode_ordinary` on the stored instruction without translation, trimming, or chat templates. The subset means are 79.2/124.6/146.3 tokens. Figure 2 additionally reports task-label multiplicity, input counts, and checklist sizes; multi-reference primary labels are excluded from the multiplicity comparison. Original-language lengths describe specification detail rather than a controlled cross-benchmark difficulty scale.

Tables 33–35 retain all 148 task types and their definitions. Family counts use original first-level annotations; task rows give operational definitions. Single/text records have 2,154/1,468 memberships, while all 732 multi-reference records have one primary label. Overlap denotes coverage, not extra instances. Single-image record 1006 carries an additional aesthetic-family annotation, explaining 320 original family members versus 319 reconstructed from child labels.

A.2 SOURCE-GROUNDED CURATION

Figure 6 shows five stages. **Visual Asset Curation** filters public images for format, resolution, clarity, and usable subjects or text. **Semantic Grounding** describes visible entities, OCR content, spatial relations, and reference roles. **Compositional Task Design** combines compatible labels with preservation, layout, and reference conditions, skipping unsupported targets. **Checklist Specification** converts the selected instruction and sources into observable requirements before output scoring. **Human Validation** checks feasibility, labels, image–instruction alignment, and checklist consistency; ambiguous or contradictory requirements are revised.

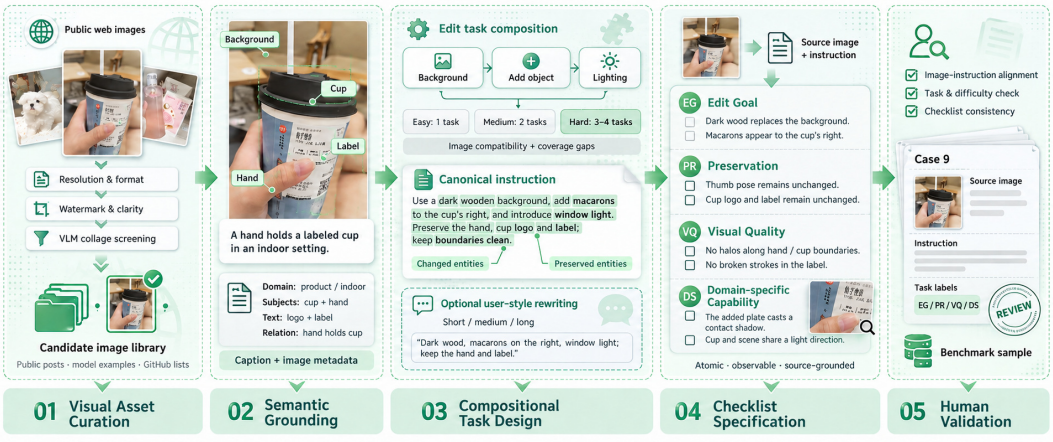

Figure 6: Benchmark curation from visual assets to human-validated instructions and checklists. The five stages are illustrated with abbreviated text from case 9.

A.3 INSTRUCTION AND RECORD CONSISTENCY

Single-image and text-intensive generation uses proposal bands of one, two, or three to four task labels; text tasks may add at most one/two visual-edit labels in the medium/high bands. These are construction controls, not calibrated difficulty levels. Short, medium, and long expressions can share

a canonical intention, but only the selected expression is evaluated: unselected variants cannot add checklist obligations. Exports retain IDs, ordered sources, selected instructions, labels, checklists, and provenance. Text records separate the two sides; reference checks identify their assigned source.

A.4 Coverage and Evaluation Resolution

Tables 6 and 7 separate coverage from retained judgment granularity. GEdit/GEditBench v2 collect practical requests, ImgEdit includes multi-turn interaction, and Complex-Edit/ComplexBench-Edit examine composition and dependencies (Liu et al., 2025; Jiang et al., 2026; Ye et al., 2025b; Yang et al., 2025; Wang et al., 2025). UnicBench adds instance-specific reasoning points, while REDEdit, RISEBench, and UniREditBench specialize text or reasoning assessment (Ye et al., 2025a; Super Intelligence Team, Xiaohongshu Inc., 2026; Zhao et al., 2025; Han et al., 2026). WeGenEditBench combines single-image, detailed text/layout, and multi-reference tasks with shared, sample-specific checks. Even one instruction can yield separate attribute, preservation, location, and source-binding judgments. This improves diagnostic resolution and auditability; superior human-aligned accuracy requires a same-output controlled comparison.

Comparison conventions. Inputs exclude judge-only ground-truth images, and taxonomies retain their native granularity. ImgEdit statistics use its 737-case Basic release, excluding UGE and multi-turn sequences. Token ranges use `o200k_base`: Chinese for GEdit, UnicBench, and ours; English for ImgEdit. Dashes indicate unmeasured ranges. UnicBench supplies instance-specific reasoning points for 250 of 1,100 cases. Multi-dimensional evaluation includes non-VLM metrics. The protocol comparison describes retained judgments, not independently verified human alignment or every possible free-form explanation.

A.5 Task-Specific Strengths and Bottlenecks

Tables 14–32 report all 19 task families. Single/text categories overlap; averages use valid, applicable scores and per-case Overall. Valid-case sets differ across models, so small gaps are descriptive. Bold and underline rank the displayed scores across all models, including ties.

Competitive preservation does not imply complete execution. HunyuanImage-3.0-Instruct-Distil leads PR in six of eight single-image families. Its CoT variant ranks second overall on restoration (4.070), above Seedream-5.0-Pro (4.044); the base variant nearly matches Nano-Banana-Pro on cultural editing (3.957 versus 3.965). Yet on action/interaction editing, its leading PR of 3.975 accompanies EG of only 3.640, versus GPT-Image-2's 4.376. Open models can thus approach closed models on specific tasks or preservation, while still struggling to execute the requested transformation.

Text bottlenecks depend on the application. On posters/layout, JoyAI-Image-Edit nearly matches GPT-Image-2 in image-side EG (4.552 versus 4.567), but trails in text-side EG (3.534 versus 4.144). Strong visual editing therefore does not establish textual correctness. Hunyuan's CoT variant reaches 4.084 Overall on social-media editing, 0.143 below GPT-Image-2. For interfaces/documents/charts, however, the strongest open model, SenseNova-U1-8B-MoT-CoT, reaches only 3.724 versus Seedream's 4.392. The separate task and side scores locate this gap more precisely than a single text-editing average.

Multi-reference gaps are especially pronounced for text. Hunyuan leads the evaluated open models in all six multi-reference families, yet its Overall gap to the strongest closed model ranges from 0.580 on pets to 1.112 on text. Multi-reference text is its lowest-scoring family (2.878), with EG at 2.700; FireRed and Qwen-Image-Edit-2511 fall below 1.9 EG. These results identify reference-conditioned text editing as a shared weakness, although they do not by themselves separate reference-binding errors from text-rendering failures.

B Evaluation Protocol and Diagnostic Evidence

B.1 Shared Requirements and Item Judgments

All three WeGenEditBench subsets use Gemini 3.1 Pro Preview (`gemini-3.1-pro-preview`) as the judge. Checklists are generated from the selected instruction and source images before scoring and shared by all compared models within an evaluation version. Each item identifies an observable target, attribute, location, or reference assignment under EG/PR/VQ/DS. The judge receives the instruction, ordered sources, output, and checklist, and records a deduction $\delta_{i,j} \in [0, 5]$, score $5 - \delta_{i,j}$, and visual rationale for each item before assigning dimensional scores. Dimensional scores follow their rubrics and need not equal item means. Literal text and source indices remain explicit; invented requirements and redundant restatements are excluded.

Edit Goal (EG). EG measures whether each requested transformation is applied to the intended target with the specified attribute, quantity, and spatial relation. It covers object operations, requested text changes, and reference-conditioned composition. For a request to turn both boots and an ornament into silver metal, changing only the boots leaves an editing goal unmet.

Preservation (PR). PR measures whether content and attributes outside the requested changes remain faithful to the source. Checks cover protected identities, object geometry, existing text, layout, and background, as applicable. In multi-reference editing, PR checks the attributes assigned to each source. Preservation is judged relative to the instruction; a requested change is not itself a preservation error.

Visual Quality (VQ). VQ measures rendering quality and visual coherence, including clarity, structural integrity, natural boundaries, and freedom from blur, artifacts, or implausible seams. Text-side checks examine legibility and glyph integrity; image-side checks assess the surrounding visual content. A visually polished output may therefore receive high VQ while failing EG or PR requirements.

Domain-specific Capability (DS). DS measures correctness under task-specific criteria beyond generic execution, preservation, and rendering quality. Applicable checks address material behavior, physical or geometric relations, typographic hierarchy, or cultural semantics. In the metal-editing example, consistent metallic tones and reflections across both targets form a joint requirement: making one object reflective does not establish material consistency across the pair.

B.2 Scoring, Applicability, and Aggregation

For single-image instance i, $B_i = 0.6S_{\mathrm{EG},i} + 0.2S_{\mathrm{PR},i} + 0.2S_{\mathrm{VQ},i}$ and $p_{i,t} = (5 - S_{\mathrm{DS},i,t})/4$. Equation 1 subtracts the mean applicable-task penalty and clips to $[0, 5]$; task-conditioned summaries use the corresponding penalty. Empty DS sets incur no penalty and have no DS score. Empty EG/PR/VQ checklists receive 5; omitted dimension scores in compact responses fall back to item means. Missing required fields or mismatched item counts invalidate the response. Stored scores retain four decimals. Text editing evaluates EG/PR/VQ/DS separately on the text and image sides, combining their scores with weights 0.75/0.25 only when image-side requirements exist; otherwise the text-side score is final. Multi-reference EG/PR/VQ correspond to native IF/RF/VQ and use weights 0.4/0.4/0.2 (Equation 2); no separate DS is annotated. Subset Overall averages valid instance scores once per instance, including unsuccessful edits. Inapplicable criteria and invalid judge responses are excluded from their respective denominators; task-label overlap does not duplicate instances in Overall. The weights define an operational rubric, not a human-calibrated preference model. High aggregate scores alongside low item scores demonstrate retained diagnostic detail, not superior accuracy over an independently evaluated holistic judge.

B.3 Human Audit and Shared Revisions

Three trained experts randomly audit 10% of cases in each benchmark, checking whether requirements follow from the shown inputs and deductions are supported by outputs. Problematic checklists

are corrected first; every compared model is re-evaluated on affected cases using the same revised checklist. Each version fixes the judge, prompt, input processing, checklist, and invalid-response policy. A score–item discrepancy alone is not proof of model failure: numerical aggregation, unsupported requirements, and judge inconsistency must be distinguished. The audit corrects identified problems but does not independently validate every judgment or quantify accuracy gains over holistic evaluation. Inter-rater agreement and error rates cannot be inferred from this corrective audit.

B.4 Diagnostic Scope and Statistical Conventions

The audit covers 15,214 model–task and 194,836 item records over 2,740 tasks; 262 inapplicable items are excluded. Single/text analyses include Qwen-Image-Edit-2511, FireRed-Image-Edit-1.1, HunyuanImage-3.0-Instruct-Distil-CoT, GPT-Image-2, Seedream-5.0-Pro, and JoyAI-Image-Edit. Multi-reference comparisons use Qwen, FireRed, GPT, and Seedream; an additional Hunyuan run is audited separately because its configuration mapping is unconfirmed. High-scoring outputs have Overall ≥ 4; severe items score ≤ 2. A side passes only if every applicable item scores ≥ 4; empty image-side lists are not successes. Each output contributes once per diagnostic event. The six-model shared text set contains 920 cases, including 280 dual-side tasks. Proportion intervals use 2,000 source-image-cluster bootstrap replicates; multi-reference clustering uses ordered reference tuples, with first-source sensitivity checks. Shared non-first references can leave residual dependence. Diagnostic analyses use the item-level judgments, dimensional scores, and subset-specific Overall defined in Section 3.3. Reproducing stored Overall verifies arithmetic, not semantic correctness. Main leaderboard and adaptation tables report point estimates; these diagnostic intervals do not quantify ranking uncertainty or training-seed variability. Tables 8 and 9 report the diagnostic outcomes.

Recorded diagnostic outcomes. Table 8 conditions on each model's high-scoring outputs rather than a shared ranking set. Within the single-image Overall ≥ 4 group, fixing 16 applicable items gives 50/104 severe-item cases (48.1%, 95% CI 38.4–57.7%). Exact multi-reference score vector $(5, 5, 3, 4.6)$ maps to both severe-item and no-severe-item outcomes: 44 with and 64 without severe items for GPT, and 30 with and 50 without for Seedream. At item threshold ≤ 1, neither has events among multi-reference outputs with Overall ≥ 4.5. Table 9 uses the same 280 dual-side tasks. More requirements mechanically reduce joint success, so these rates establish neither a causal complexity effect nor independently verified failure rates.

C Agentic Diagnosis and Data Construction

C.1 Harness and Resource Contracts

The canonical registry exposes 68 tools (aliases excluded): benchmarking 12, analysis 17, construction 10, training 6, evolution 4, discovery 1, and system operations 18. The orchestration LLM instantiates Diagnosis and Construction Agents, with defaults of four concurrent children and one delegation level. Each child has isolated context and a tool allowlist. Schema validation and committed resource references connect evaluation records, reports, datasets, training jobs, and checkpoints. Visual evidence is stored separately from original judgments and can be reused through content fingerprints.

Compatibility with native benchmark feedback. WeGenEvoStudio accepts benchmark-specific evaluation records through adapters, with category scores supporting task prioritization and native rationales or checklist items providing additional evidence when available. The shared interface preserves each benchmark's metric definitions and score direction; it does not require external benchmarks to adopt EG/PR/VQ/DS or our checklist format. This separates compatibility with an evaluation protocol from the granularity of the evidence it provides.

C.2 Diagnosis Agent: Evidence to Data Prescriptions

Figure 7 summarizes diagnosis. Adapters normalize category axes, score scales, and score direction. With a compatible anchor, selection allocates approximately 60% of slots to relative deficits

and 40% to absolute weaknesses, with at least one absolute slot. Duplicate selections are removed and reciprocal-rank ordering fills remaining slots; absent relative deficits, selection uses absolute scores. This is slot allocation, not score interpolation. Selected categories bound case review: weak/successful contrasts distinguish recurring symptoms from isolated outputs. Instructions, sources, outputs, and available reasons remain linked to case IDs.

Diagnosis at the available feedback granularity. In our experiments, GEdit and ImgEdit guide task selection through category scores; WeGenBench additionally contributes its own rationales, without our checklist; WEGENEDITBENCH provides requirement-level scores and reasons. Category scores identify where to allocate data, while available explanations and visual review help specify what should change or remain intact. Missing item-level feedback is not treated as an existing checklist judgment. All routes produce task-level prescriptions, allowing the framework to operate with coarse feedback and incorporate finer evidence when supplied. Failure groups retain affected IDs and score losses; visual inspection checks the reported defect. Reports distinguish observations, interpretations, causal hypotheses, and recommendations. A data prescription names the capability to reinforce and supported source conditions, constraints, and acceptance criteria; reference and coverage checks precede commitment. Plan defaults (categories/cases per category/visual cap/LLM cap) are quick 5/1/0/6, standard 10/3/30/31, and comprehensive up to 100/5/200/201. Explicit run settings override these defaults; evolution selects $K = 5$. These budgets describe software options, not every experimental trajectory.

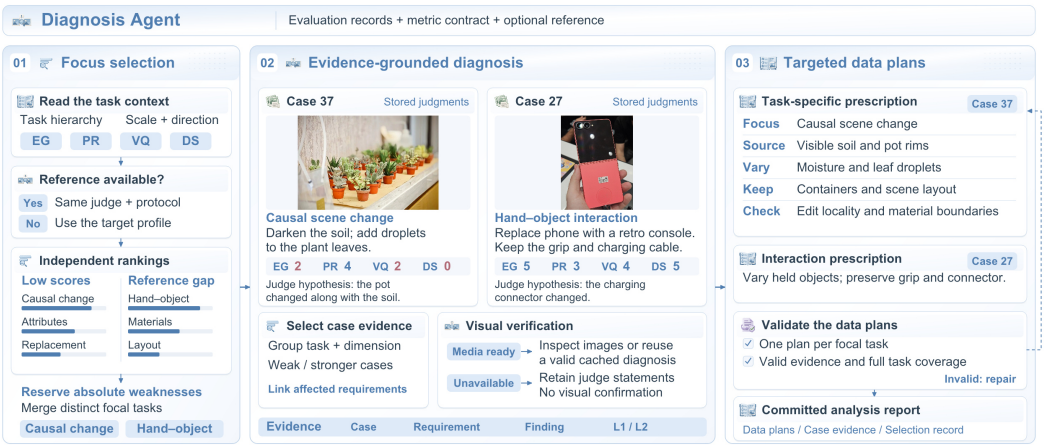

Figure 7: Diagnosis Agent: category selection, visual review, failure grouping, and data prescriptions. Cases 37/27 use stored Qwen judgments (EG/PR/VQ/DS: 2/4/2/0 and 5/3/4/5). Source illustrations, ranking bars, and prescriptions are schematic; failure descriptions are judge hypotheses.

C.3 CONSTRUCTION AGENT: RETRIEVAL-GROUNDED TASK SYNTHESIS

Figure 8 shows task alignment, retrieval, grounded instruction generation, and review. Selected labels map to the reusable taxonomy and optional category notes; the instruction budget is allocated across these tasks. Editing queries seek suitable unedited content, varying subjects and contexts. Acquired images pass format/resolution checks and registry-level exact-duplicate handling, with optional perceptual filtering. The generator receives an accepted source, metadata, task, and notes, then specifies feasible changes, protected content, and spatial conditions. Compatible secondary operations enrich requests; unsuitable sources are skipped. VLM review checks category alignment and feasibility before rewriting canonical instructions into user-style requests. Accepted records retain both versions, source hashes, labels, filtering outcomes, and report/dataset identifiers. The downstream supervision pipeline supplies targets. For text-to-image generation, keyword expansion and optional web-text research provide diverse concepts and defining visual attributes without an editing source. The implemented adaptation branches cover generation and single-image editing.

A shared construction interface. Across feedback formats, the Construction Agent receives selected task labels, requested counts, and optional category notes. Score-guided plans specify task priorities; richer evidence can refine source conditions and preservation constraints. Both use the same retrieval, grounded generation, and review workflow. Thus, external benchmarks can guide data construction without supplying our sample-specific checklists or their own instruction-generation templates.

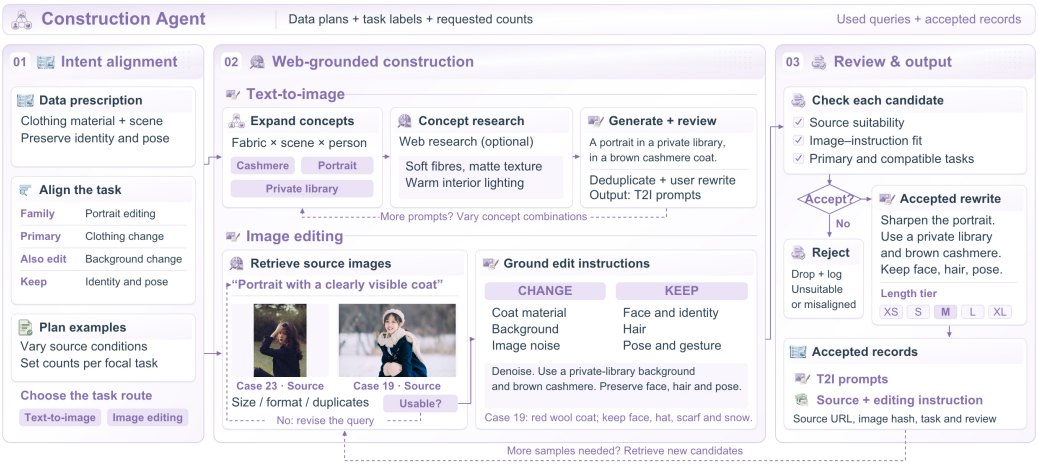

Figure 8: Construction Agent: task alignment, concept research or image retrieval, grounded instruction generation, and review. Cases 23/19 illustrate the interface. Accepted records retain source and instruction provenance; target supervision and training follow downstream.

C.4 Task-Level Feedback and Instance Boundaries

Cases justify capability selection within the diagnosis report; construction consumes task labels and optional category notes. Per-image calls do not receive evaluated input–output pairs as demonstrations or import external benchmarks' generation templates. Retrieved content determines the new scene, while feedback sets capability priorities. This is benchmark-informed task-level adaptation. New sources and registry deduplication do not certify independence from all evaluation images or semantic near-duplicates; content-overlap audits and a sealed holdout are needed for stronger generalization claims.

D Experimental Protocol and Extended Analysis

D.1 Shared Training Settings and Evaluation Scope

The benchmark comparison evaluates 15 configurations on single-image/text editing and six on multi-reference editing. Evolution uses Qwen-Image-2512 for generation, Qwen-Image-Edit-2511 for editing, and FLUX.2 [dev] for both, with a separate rank-16 LoRA adapter per backbone–benchmark pair. The main evolution experiments each use 25,000 training examples, with 5,000 per selected task for $K = 5$. The per-device training batch size is 1. The feedback adapters and shared construction interface in Appendix C.1 support category scores from GEdit/ImgEdit, native rationales from WeGenBench, and checklist evidence from WeGenEditBench. These experiments test the utility of the framework across native feedback formats. They do not isolate the contributions of fine-grained reasons or agent orchestration; equal-budget feedback variants on the same benchmark remain necessary for that comparison. The training schedule is 782 steps for Qwen's single-image adapter and 2,346 steps for the other adapters. FLUX.2 uses the corrected `bytelevel_v1` implementation. Report-selected case counts for Qwen/FLUX.2 are 277/328 (WeGenBench), 540/473 (GEdit), 433/425 (ImgEdit), and 82/78 (single-image). Full evaluation covers 1,000/1,000, 1,212/1,210, and 737/736 cases on the first three benchmarks. Sets are dedupli-

cated unions and differ by backbone. GEdit averages $O_i = \sqrt{\mathrm{SC}_i \mathrm{PQ}_i}$ per case; ImgEdit's detail score D differs from our DS. Deltas use unrounded means.

Hardware and external services. All local experiments, including model training and inference, run on a 32-card Cambricon MLU590 system. Closed-source image generation, API-based VLM evaluation, and web retrieval are accessed through external APIs; their provider-side hardware is outside this local configuration. We report hardware and training schedules, but do not report end-to-end API expenditure or accelerator-hours.

D.2 Updates and Evaluation Records

Constructed tasks are paired with target supervision into D_t; the backend updates the parent checkpoint and records the data revision and training recipe λ_t:

$$\widetilde{\theta}_{t+1} = \mathrm{Train}(\theta_t, D_t; \lambda_t). \tag{4}$$

Re-evaluation links the updated model to revised capability priorities:

$$\theta_t \longrightarrow \Phi_t \longrightarrow \mathcal{T}_t \longrightarrow \mathcal{X}_t \longrightarrow D_t \longrightarrow \theta_{t+1} \longrightarrow \Phi_{t+1}. \tag{5}$$

Run records retain parentage, supervision, filtering, inference settings, and evaluation versions. Validation supports target and retention checks; diagnosed benchmark cases are not an independent holdout. FLUX.2 GEdit uses GPT-4.1 VIEScore (Ku et al., 2024) on 473 target and 1,210 full-benchmark common-valid pairs, without mixing diagnosis-judge scores. ImgEdit's reported inference configurations remain unmatched: FLUX.2 uses Diffusers/guidance 4.0 before adaptation and DiffSynth/embedded guidance 2.5 afterward. Its deltas therefore include inference changes. Qwen's ImgEdit training coverage also differs from its five-category diagnostic scope. The WeGenEditBench single-image comparison shares cases and judge identity, but exact generation-setting equivalence remains unverified.

D.3 Targeted Gains and Broader Behavior

Table 10 separates targeted gains from full-benchmark changes: Qwen gains 0.670 WeGenBench F but loses 0.190 GEdit Overall; FLUX.2 gains 0.146 and 3.009, respectively. All five FLUX.2 GEdit targets improve, with larger SC than PQ gains (Table 11); motion change has the largest Overall increase (+5.295). Selected Qwen examples in Figure 5 show better operation choice and local preservation. Together, the results support task-dependent improvements, not uniform gains. FLUX.2 WeGenBench originally assigns minimum scores to two target and five full-benchmark parsing failures. Restricting to parseable pairs yields 326/995 cases and gains of 0.230/0.124 rather than 0.253/0.146 (Table 12). Both use the named Qwen3.8-Flash-Next judge and deduction prompt; historical serving-version identity is unverified. Faulty text-encoding runs are excluded from all gains.

D.4 Web Search and Concept-Specific Effects

Qwen-Image-2512 search-enabled and search-disabled adapters each train on 25,000 examples for 2,346 steps using 32 Cambricon MLU590 accelerators and a per-device batch size of 1. Both receive one Qwen3.8-Flash-Next evaluation pass. Table 13 keeps the original 277-case Top-5 aggregate while displaying four category columns. Search changes target/full F by +0.018/+0.021, improving counterintuitive concepts, symbols, and IP but reducing the animal score. With matched training-data size and step schedules, these heterogeneous changes indicate concept-dependent retrieval effects; the small aggregate gains do not establish statistical significance without uncertainty estimates.

E Limitations and Future Directions

Evaluation reliability. Explicit checklists and shared human revisions improve auditability, but requirements and VLM judgments can remain wrong. Independent annotations and matched holistic-versus-checklist evaluations are needed to quantify accuracy and localization gains. Diagnostic rates describe recorded judgments; the separate 18-case exploratory review and exclusion of 11 disputed items do not validate all outputs.

Coverage and generalization. Collected requests are not a probability sample of user demand, and label/checklist counts are imperfect complexity measures. Separate benchmark-specific adapters and adaptive task selection limit unseen-task generalization claims. Equal-budget random, score-only, and full-diagnosis comparisons, overlap audits, and sealed tests remain necessary to isolate feedback contributions and adaptive overfitting.

Further extensions. Future work can extend adaptation to text-intensive and multi-reference editing and add retention constraints. Reproduction requires versioned sources, judges, filtering records, supervision, and checkpoints; source accessibility alone does not establish redistribution rights.

F Supporting Results and Complete Taxonomy

F.1 Quantitative Comparison and Diagnostic Tables

Table 5: Statistics of the three source files. Length is measured in tokens using `o200k_base`. Types are subset-qualified; multi-reference primary labels are not comparable to multi-label task counts.

Statistic	Single-image	Text-intensive	Multi-reference
Instances	1,008	1,000	732
Task families	8	5	6
Task types	65	35	48
Mean instruction length	79.2	124.6	146.3
Median instruction length	60.5	126	116.5
Instruction length range	12–255	10–344	30–439
Mean task labels	2.14	1.47	—
Multi-label instances (%)	70.6	31.2	—
Mean input images	1.00	1.00	2.70

Table 6: Quantitative comparison on image-editing benchmarks: task coverage. Conventions are specified in Appendix A.4.

Benchmark	**Taxonomy**	**Tokens**	**Task emphasis**	**Inputs**
GEdit-Bench (2025)	11 categories	2–38	General editing	1
ImgEdit (2025b)	9 Basic types	4–60	Single/multi-turn	1
REDEdit (2026)	15 categories	—	General/text	1
RISEBench (2025)	4 families / 16 types	—	Implicit reasoning	1
UnicBench (2025a)	4 families / 22 types	6–42	General/reasoning	1
UniREditBench (2026)	8 dimensions / 18 subtypes	—	Real/game reasoning	1
Complex-Edit (2025)	9 groups / 24 operations	—	1–8 operations	1
ComplexBench-Edit (2025)	3 dependency levels	—	Parallel/chained edits	1
GEditBench v2 (2026)	22 tasks + open set	—	Hybrid/open-set	1
WeGenEditBench	**19 families / 148 types**	**10–439**	**Single/text/multi**	**1–4**

Table 7: Comparison of evaluation protocols and diagnostic outputs. Conventions are specified in Appendix A.4.

Benchmark	Multi-dim.	Verification criteria	Retained judgment unit
GEdit-Bench	Yes	General rubric	SC/PQ scores; explanations in VIEScore
ImgEdit	Yes	Task-specific rubrics	Dimension scores + brief reason
REDEdit	Yes	Task/text criteria; OCR	Dimension scores; text-error metrics
RISEBench	Yes	Reasoning guidelines	Reasoning, consistency, plausibility
UnicBench	Yes	Instance reasoning points	Dimension scores + reason
UniREditBench	Yes	Text + GT image references	Reference-aided dimension scores
Complex-Edit	Yes	Detailed IF/IP/PQ rubrics	Scores per evaluation dimension
ComplexBench-Edit	Yes	Edit-type criteria; chain rules	Per-instruction scores; region consistency
GEditBench v2	Yes	IF/VQ/VC comparisons	Pairwise dimensional preferences
WeGenEditBench	**Yes**	**Shared instance checklist**	**Per-requirement score, deduction, and rationale**

Table 8: GPT-Image-2 outputs with at least one recorded item score ≤ 2, conditional on the specified Overall threshold. Counts refer to model outputs; intervals are 95% cluster-bootstrap intervals.

Subset	Overall	Count	Rate	95% CI
Single-image	≥ 4	315/644	48.9%	[45.1, 52.8]%
Text-intensive	≥ 4	171/666	25.7%	[22.4, 29.0]%
Multi-reference	≥ 4	119/503	23.7%	[19.9, 27.2]%
Single-image	≥ 4.5	101/323	31.3%	[26.5, 36.6]%

Table 9: Joint satisfaction of text-side and image-side requirements on 280 shared tasks. A side passes only when every applicable item scores at least 4. Counts within each row sum to 280.

Model	Both sides	Text only	Image only	Neither
GPT-Image-2	70 (25.0%)	29	77	104
Seedream-5.0-Pro	76 (27.1%)	49	66	89

Table 10: Full-benchmark results for the adapters in Table 4, under the same protocol qualifications.

Model	WeGenBench	GEdit			ImgEdit			
	F ↑	SC ↑	PQ ↑	O ↑	I ↑	Q ↑	D ↑	F ↑
Qwen-Image	8.170	8.774	7.676	7.962	4.704	4.438	4.490	4.544
+ WeGenEvoStudio	8.840	8.338	7.597	7.772	4.688	4.474	4.476	4.546
Δ	**+0.670**	**-0.436**	**-0.079**	**-0.190**	**-0.016**	**+0.035**	**-0.014**	**+0.002**
FLUX.2 [dev]	8.315	4.388	7.621	4.395	2.323	3.124	3.207	2.885
+ WeGenEvoStudio	8.461	7.785	8.132	7.404	4.170	4.099	4.190	4.153
Δ	**+0.146**	**+3.398**	**+0.511**	**+3.009**	**+1.846**	**+0.976**	**+0.984**	**+1.269**

Table 11: FLUX.2 [dev] results after 2,346 training steps on the five report-selected GEdit categories, evaluated with GPT-4.1 VIEScore. Scores refer to the adapted model, and Δ denotes the change from its paired baseline.

Category	SC	PQ	O	Δ SC	Δ PQ	Δ O
`motion_change`	8.114	8.430	7.907	**+5.747**	**+0.848**	**+5.295**
`subject-remove`	7.272	8.316	6.911	**+3.991**	**+0.351**	**+3.706**
`color_alter`	9.275	7.812	8.308	**+5.463**	**+0.050**	**+4.371**
`material_alter`	7.625	7.600	7.002	**+4.438**	**+0.575**	**+3.678**
`subject-replace`	7.817	7.933	7.226	**+3.658**	**+0.425**	**+3.173**

Table 12: Sensitivity of FLUX.2 WeGenBench Final scores to judge-response parsing failures. Both comparison sets retain the same baseline and adapted checkpoints.

Scope	**Evaluation policy**	**Before**	**After**	Δ
Report Top-5	Original fail-soft	7.726	7.979	**+0.253**
Report Top-5	Both responses parseable	7.767	7.997	**+0.230**
Full benchmark	Original fail-soft	8.315	8.461	**+0.146**
Full benchmark	Both responses parseable	8.352	8.475	**+0.124**

Table 13: Web-search comparison on WeGenBench using Qwen-Image-2512. All scores are F ↑ (1–10). Δ is search-enabled minus search-disabled. Top-5 includes all five selected categories (277 cases), although only four category columns are displayed.

Setting	Counter-intuitive	Icons / symbols	Animals	IP	Top-5	Full
Baseline	7.150	7.540	8.000	7.810	7.700	8.170
Without web search	8.450	8.440	**8.903**	8.506	8.650	8.819
With web search	**8.700**	**8.760**	8.611	**8.671**	**8.668**	**8.840**
Δ	**+0.250**	**+0.320**	**-0.292**	**+0.165**	**+0.018**	**+0.021**

F.2 PRIMARY-TASK QUANTITATIVE COMPARISONS

F.2.1 SINGLE-IMAGE EDITING

Table 14: Quantitative comparison on WEGENEDITBENCH-SINGLEIMAGEEDIT: Basic Local Editing. Bold and underline denote the best and second-best scores.

Model Name	EG↑	PR↑	VQ↑	DS↑	**Overall↑**
Closed-source Models					
GPT-Image-2	**4.424**	3.966	**4.445**	<u>4.225</u>	**4.143**
Nano-Banana-Pro	4.317	3.640	4.402	4.056	3.963
Qwen-Image-3.0	4.228	3.517	4.359	4.125	3.893
Seedream-5.0-Pro	<u>4.328</u>	3.985	<u>4.424</u>	**4.251**	<u>4.091</u>
Open-source Models					
HunyuanImage-3.0-Instruct-Distil	4.101	**4.147**	4.271	4.002	3.895
HunyuanImage-3.0-Instruct-Distil-CoT	4.197	<u>4.030</u>	4.239	4.042	3.933
FireRed-Image-Edit-1.1	4.171	3.496	4.224	3.934	3.780
JoyAI-Image-Edit	4.118	3.520	4.172	3.869	3.727
Qwen-Image-Edit-2511	4.043	3.495	4.101	3.761	3.636
Boogu-Image-0.1-Edit	4.000	3.626	4.094	3.763	3.635
SenseNova-U1-8B-MoT-CoT	4.071	3.610	3.943	3.632	3.614
SenseNova-U1-8B-MoT	4.147	3.317	3.948	3.723	3.622
LongCat-Image-Edit	3.938	3.371	4.011	3.670	3.513
HiDream-O1-Image	3.286	2.962	3.896	3.199	2.897
FLUX.2 [dev]	1.984	3.311	3.307	1.753	1.707

Table 15: Quantitative comparison on WEGENEDITBENCH-SINGLEIMAGEEDIT: Aesthetic and Photographic Editing. Bold and underline denote the best and second-best scores.

Model Name	EG↑	PR↑	VQ↑	DS↑	**Overall↑**
Closed-source Models					
GPT-Image-2	**4.096**	4.068	**4.494**	**4.215**	**3.974**
Nano-Banana-Pro	3.983	3.880	4.424	4.010	3.803
Qwen-Image-3.0	3.916	3.485	4.324	4.149	3.699
Seedream-5.0-Pro	4.035	<u>4.163</u>	<u>4.441</u>	<u>4.189</u>	<u>3.939</u>
Open-source Models					
HunyuanImage-3.0-Instruct-Distil	3.917	**4.234**	4.357	4.018	3.823
HunyuanImage-3.0-Instruct-Distil-CoT	3.985	4.093	4.319	4.098	3.848
FireRed-Image-Edit-1.1	<u>4.038</u>	3.670	4.268	4.012	3.764
JoyAI-Image-Edit	4.016	3.539	4.187	3.824	3.661
Qwen-Image-Edit-2511	3.893	3.720	4.075	3.768	3.587
Boogu-Image-0.1-Edit	3.778	3.720	4.064	3.806	3.525
SenseNova-U1-8B-MoT-CoT	3.779	3.651	3.953	3.628	3.445
SenseNova-U1-8B-MoT	3.963	3.495	3.911	3.761	3.549
LongCat-Image-Edit	3.562	3.509	4.087	3.749	3.344
HiDream-O1-Image	3.184	3.010	3.849	3.399	2.885
FLUX.2 [dev]	2.215	3.298	3.628	2.302	2.044

Table 16: Quantitative comparison on WEGENEDITBENCH-SINGLEIMAGEEDIT: Portrait Editing. Bold and underline denote the best and second-best scores.

Model Name	EG↑	PR↑	VQ↑	DS↑	**Overall↑**
Closed-source Models					
GPT-Image-2	**4.443**	<u>4.156</u>	**4.464**	**4.328**	**4.222**
Nano-Banana-Pro	<u>4.390</u>	3.696	4.350	4.226	4.050
Qwen-Image-3.0	4.332	3.451	4.337	4.279	3.977
Seedream-5.0-Pro	4.376	4.125	<u>4.402</u>	<u>4.325</u>	<u>4.162</u>
Open-source Models					
HunyuanImage-3.0-Instruct-Distil	4.272	**4.231**	4.306	4.071	4.038
HunyuanImage-3.0-Instruct-Distil-CoT	4.193	4.095	4.305	4.162	3.986
FireRed-Image-Edit-1.1	4.192	3.506	4.186	4.030	3.811
JoyAI-Image-Edit	4.195	3.505	4.177	3.880	3.774
Qwen-Image-Edit-2511	4.263	3.449	4.186	3.937	3.819
Boogu-Image-0.1-Edit	4.094	3.618	4.161	3.945	3.749
SenseNova-U1-8B-MoT-CoT	4.240	3.556	4.017	3.756	3.748
SenseNova-U1-8B-MoT	4.276	3.359	4.016	3.834	3.749
LongCat-Image-Edit	4.096	3.461	4.120	3.949	3.711
HiDream-O1-Image	3.496	2.793	3.975	3.589	3.100
FLUX.2 [dev]	1.995	3.341	3.352	1.990	1.785

Table 17: Quantitative comparison on WEGENEDITBENCH-SINGLEIMAGEEDIT: Reasoning-based Editing. Bold and underline denote the best and second-best scores.

Model Name	EG↑	PR↑	VQ↑	DS↑	**Overall↑**
Closed-source Models					
GPT-Image-2	**4.228**	4.134	**4.433**	<u>4.143</u>	**4.036**
Nano-Banana-Pro	4.138	3.745	4.423	3.992	3.864
Qwen-Image-3.0	4.153	3.627	4.385	3.923	3.825
Seedream-5.0-Pro	<u>4.155</u>	4.074	<u>4.429</u>	**4.184**	<u>3.990</u>
Open-source Models					
HunyuanImage-3.0-Instruct-Distil	3.822	**4.205**	4.346	3.846	3.715
HunyuanImage-3.0-Instruct-Distil-CoT	3.845	<u>4.155</u>	4.302	4.021	3.754
FireRed-Image-Edit-1.1	3.907	3.698	4.249	3.790	3.631
JoyAI-Image-Edit	4.026	3.582	4.213	3.722	3.655
Qwen-Image-Edit-2511	3.758	3.585	4.114	3.573	3.438
Boogu-Image-0.1-Edit	3.763	3.834	4.153	3.687	3.527
SenseNova-U1-8B-MoT-CoT	4.057	3.723	4.012	3.595	3.631
SenseNova-U1-8B-MoT	4.029	3.357	4.080	3.729	3.589
LongCat-Image-Edit	3.586	3.507	3.966	3.593	3.303
HiDream-O1-Image	3.120	3.055	3.862	3.171	2.802
FLUX.2 [dev]	1.773	3.405	3.288	1.738	1.592

Table 18: Quantitative comparison on WEGENEDITBENCH-SINGLEIMAGEEDIT: Product and Commercial Editing. Bold and underline denote the best and second-best scores.

Model Name	EG↑	PR↑	VQ↑	DS↑	**Overall↑**
Closed-source Models					
GPT-Image-2	4.117	3.372	<u>4.314</u>	4.125	3.789
Nano-Banana-Pro	**4.223**	3.269	4.234	4.088	<u>3.806</u>
Qwen-Image-3.0	<u>4.190</u>	3.216	4.230	<u>4.153</u>	3.792
Seedream-5.0-Pro	4.003	<u>3.806</u>	**4.471**	**4.264**	**3.874**
Open-source Models					
HunyuanImage-3.0-Instruct-Distil	3.976	**3.838**	4.189	3.901	3.716
HunyuanImage-3.0-Instruct-Distil-CoT	3.912	3.460	4.042	3.790	3.556
FireRed-Image-Edit-1.1	4.002	2.866	4.036	3.701	3.457
JoyAI-Image-Edit	3.926	3.226	4.135	3.706	3.505
Qwen-Image-Edit-2511	3.881	2.865	3.999	3.451	3.321
Boogu-Image-0.1-Edit	3.792	3.241	3.939	3.551	3.349
SenseNova-U1-8B-MoT-CoT	3.647	3.336	3.733	3.523	3.242
SenseNova-U1-8B-MoT	3.742	2.837	3.598	3.425	3.141
LongCat-Image-Edit	3.464	2.505	3.486	3.127	2.848
HiDream-O1-Image	2.346	2.726	3.653	2.482	2.070
FLUX.2 [dev]	2.366	3.422	3.389	2.242	2.099

Table 19: Quantitative comparison on WEGENEDITBENCH-SINGLEIMAGEEDIT: Restoration and Low-level Enhancement. Bold and underline denote the best and second-best scores.

Model Name	EG↑	PR↑	VQ↑	DS↑	**Overall↑**
Closed-source Models					
GPT-Image-2	**4.347**	**4.425**	**4.565**	**4.462**	**4.272**
Nano-Banana-Pro	3.612	4.003	<u>4.550</u>	<u>4.392</u>	3.726
Qwen-Image-3.0	3.672	3.950	4.313	4.290	3.678
Seedream-5.0-Pro	4.117	4.202	4.465	4.362	4.044
Open-source Models					
HunyuanImage-3.0-Instruct-Distil	4.002	<u>4.304</u>	4.328	4.218	3.932
HunyuanImage-3.0-Instruct-Distil-CoT	<u>4.224</u>	4.300	4.444	4.146	<u>4.070</u>
FireRed-Image-Edit-1.1	4.083	3.897	4.323	3.982	3.839
JoyAI-Image-Edit	4.060	3.843	4.215	3.998	3.797
Qwen-Image-Edit-2511	3.392	4.261	4.157	3.974	3.462
Boogu-Image-0.1-Edit	3.322	4.010	4.211	3.835	3.346
SenseNova-U1-8B-MoT-CoT	3.583	4.027	4.057	4.066	3.536
SenseNova-U1-8B-MoT	3.429	4.006	4.115	3.835	3.390
LongCat-Image-Edit	3.230	3.988	4.194	3.923	3.308
HiDream-O1-Image	3.260	3.149	3.681	3.747	3.009
FLUX.2 [dev]	2.678	3.620	3.749	3.233	2.639

Table 20: Quantitative comparison on WEGENEDITBENCH-SINGLEIMAGEEDIT: Action and Interaction Editing. Bold and underline denote the best and second-best scores.

Model Name	EG↑	PR↑	VQ↑	DS↑	**Overall↑**
Closed-source Models					
GPT-Image-2	**4.376**	3.662	**4.364**	<u>4.009</u>	**3.983**
Nano-Banana-Pro	<u>4.353</u>	3.317	4.190	3.762	3.804
Qwen-Image-3.0	4.190	3.047	4.198	3.910	3.691
Seedream-5.0-Pro	4.225	3.613	<u>4.312</u>	**4.047**	<u>3.882</u>
Open-source Models					
HunyuanImage-3.0-Instruct-Distil	3.640	**3.975**	4.133	3.338	3.390
HunyuanImage-3.0-Instruct-Distil-CoT	4.011	<u>3.751</u>	4.060	3.777	3.663
FireRed-Image-Edit-1.1	4.013	3.000	4.202	3.763	3.539
JoyAI-Image-Edit	4.092	2.951	4.150	3.824	3.582
Qwen-Image-Edit-2511	4.045	2.941	4.013	3.797	3.517
Boogu-Image-0.1-Edit	3.880	3.338	3.979	3.653	3.455
SenseNova-U1-8B-MoT-CoT	3.873	3.106	3.976	3.414	3.344
SenseNova-U1-8B-MoT	4.054	2.741	3.927	3.295	3.340
LongCat-Image-Edit	4.133	2.954	3.980	3.504	3.492
HiDream-O1-Image	3.334	2.746	3.749	3.072	2.817
FLUX.2 [dev]	1.781	3.221	3.180	1.130	1.392

Table 21: Quantitative comparison on WEGENEDITBENCH-SINGLEIMAGEEDIT: Traditional Chinese Cultural Editing. Bold and underline denote the best and second-best scores.

Model Name	EG↑	PR↑	VQ↑	DS↑	**Overall↑**
Closed-source Models					
GPT-Image-2	4.450	4.105	3.802	3.671	3.919
Nano-Banana-Pro	**4.566**	3.703	<u>3.960</u>	3.774	<u>3.965</u>
Qwen-Image-3.0	<u>4.496</u>	3.264	3.850	3.880	3.840
Seedream-5.0-Pro	4.410	**4.340**	**4.097**	**4.082**	**4.104**
Open-source Models					
HunyuanImage-3.0-Instruct-Distil	4.320	<u>4.311</u>	3.939	3.862	3.957
HunyuanImage-3.0-Instruct-Distil-CoT	4.284	4.187	3.803	3.677	3.838
FireRed-Image-Edit-1.1	4.233	3.818	3.764	3.457	3.670
JoyAI-Image-Edit	4.431	3.636	3.896	3.574	3.808
Qwen-Image-Edit-2511	4.255	3.875	3.873	3.254	3.666
Boogu-Image-0.1-Edit	4.236	4.029	3.702	<u>3.982</u>	3.833
SenseNova-U1-8B-MoT-CoT	4.203	3.883	3.467	3.090	3.514
SenseNova-U1-8B-MoT	4.216	3.796	3.352	3.538	3.594
LongCat-Image-Edit	3.918	3.835	3.491	2.787	3.263
HiDream-O1-Image	3.365	2.914	3.446	2.875	2.760
FLUX.2 [dev]	1.815	3.118	2.727	0.698	1.183

F.2.2 TEXT-INTENSIVE EDITING

Table 22: Quantitative comparison on WEGENEDITBENCH-TEXTEDIT: Basic Text Editing. Bold and underline denote the best and second-best scores.

Model Name	Text-side				Image-side				Overall ↑
	EG ↑	PR ↑	VQ ↑	DS ↑	EG ↑	PR ↑	VQ ↑	DS ↑	
Closed-source Models									
GPT-Image-2	4.249	4.050	4.308	4.356	**4.594**	3.924	4.409	**4.446**	4.109
Nano-Banana-Pro	4.317	4.129	4.303	4.305	4.392	4.220	4.227	4.295	4.141
Qwen-Image-3.0	4.350	3.692	4.289	4.395	4.507	3.887	4.408	4.410	4.096
Seedream-5.0-Pro	**4.366**	**4.327**	**4.412**	**4.475**	4.352	4.218	**4.447**	4.428	**4.260**
Open-source Models									
SenseNova-U1-8B-MoT-CoT	4.022	3.950	3.832	3.965	4.352	4.105	3.941	4.089	3.790
HunyuanImage-3.0-Instruct-Distil	3.917	3.868	3.894	4.120	4.362	4.174	4.111	4.220	3.776
HunyuanImage-3.0-Instruct-Distil-CoT	4.112	3.914	4.084	4.177	4.418	**4.261**	4.195	4.324	3.929
SenseNova-U1-8B-MoT	4.029	3.890	3.931	3.962	4.415	3.840	3.892	4.094	3.792
JoyAI-Image-Edit	3.934	3.716	3.754	4.024	4.582	3.920	4.129	4.275	3.710
FireRed-Image-Edit-1.1	3.785	3.444	3.673	3.972	4.498	3.624	3.930	4.237	3.549
Qwen-Image-Edit-2511	3.727	3.639	3.683	3.888	4.383	3.829	3.962	4.108	3.532
Boogu-Image-0.1-Edit	3.876	3.330	3.786	3.968	4.432	3.470	3.902	4.202	3.578
LongCat-Image-Edit	3.044	3.702	3.432	3.315	4.195	3.655	3.787	3.958	2.994
HiDream-O1-Image	2.864	2.584	3.237	3.121	3.554	3.052	3.477	3.535	2.513
FLUX.2 [dev]	2.205	3.429	3.677	2.391	2.717	3.339	3.566	2.776	2.214

Table 23: Quantitative comparison on WEGENEDITBENCH-TEXTEDIT: Posters and Layout. Bold and underline denote the best and second-best scores.

Model Name	Text-side				Image-side				Overall ↑
	EG ↑	PR ↑	VQ ↑	DS ↑	EG ↑	PR ↑	VQ ↑	DS ↑	
Closed-source Models									
GPT-Image-2	4.144	3.817	4.372	4.439	**4.567**	3.908	4.483	**4.471**	4.038
Nano-Banana-Pro	4.032	4.032	4.326	4.177	4.304	4.208	4.216	4.268	3.913
Qwen-Image-3.0	4.257	3.509	**4.477**	4.289	4.528	3.854	4.463	4.390	4.015
Seedream-5.0-Pro	**4.288**	**4.152**	4.316	**4.496**	4.260	4.187	**4.528**	4.407	**4.152**
Open-source Models									
SenseNova-U1-8B-MoT-CoT	3.715	3.795	3.846	3.817	4.264	4.200	4.000	4.032	3.532
HunyuanImage-3.0-Instruct-Distil	3.593	3.685	3.466	3.922	4.280	4.080	4.024	4.160	3.424
HunyuanImage-3.0-Instruct-Distil-CoT	3.860	3.735	3.900	4.099	4.360	**4.256**	4.224	4.340	3.695
SenseNova-U1-8B-MoT	3.697	3.671	3.688	3.807	4.304	3.992	4.056	4.184	3.471
JoyAI-Image-Edit	3.534	3.557	3.733	3.900	4.552	4.152	4.200	4.324	3.428
FireRed-Image-Edit-1.1	3.462	3.215	3.480	3.862	4.456	3.632	3.960	4.312	3.251
Qwen-Image-Edit-2511	3.299	3.425	3.348	3.631	4.352	3.840	3.944	4.140	3.115
Boogu-Image-0.1-Edit	3.633	3.260	3.715	3.901	4.464	3.376	3.960	4.180	3.389
LongCat-Image-Edit	2.643	3.388	3.122	3.235	4.112	3.704	3.840	3.900	2.624
HiDream-O1-Image	2.914	2.708	3.299	3.282	3.440	3.272	3.528	3.620	2.589
FLUX.2 [dev]	1.995	3.457	3.561	2.235	2.744	3.440	3.520	2.816	1.995

Table 24: Quantitative comparison on WEGENEDITBENCH-TEXTEDIT: Interfaces, Documents, and Charts. Bold and underline denote the best and second-best scores.

Model Name	Text-side				Image-side				Overall ↑
	EG ↑	PR ↑	VQ ↑	DS ↑	EG ↑	PR ↑	VQ ↑	DS ↑	
Closed-source Models									
GPT-Image-2	4.341	<u>4.120</u>	4.329	4.423	**4.639**	<u>4.139</u>	**4.500**	**4.583**	<u>4.187</u>
Nano-Banana-Pro	4.260	3.949	<u>4.333</u>	<u>4.434</u>	<u>4.595</u>	**4.351**	4.351	4.351	4.103
Qwen-Image-3.0	<u>4.385</u>	3.856	4.299	4.297	**4.639**	3.833	4.167	4.500	4.106
Seedream-5.0-Pro	**4.535**	**4.371**	**4.435**	**4.576**	4.541	4.027	<u>4.486</u>	<u>4.568</u>	**4.392**
Open-source Models									
SenseNova-U1-8B-MoT-CoT	4.040	3.780	3.791	3.934	4.432	3.730	3.649	4.243	3.724
HunyuanImage-3.0-Instruct-Distil	3.655	3.605	3.831	4.044	4.270	4.027	4.000	4.027	3.505
HunyuanImage-3.0-Instruct-Distil-CoT	3.842	3.520	4.000	4.248	4.378	3.919	3.811	4.257	3.670
SenseNova-U1-8B-MoT	3.938	3.751	3.780	3.926	4.405	3.595	3.595	4.108	3.646
JoyAI-Image-Edit	3.678	3.379	3.452	3.787	4.351	3.811	3.649	4.338	3.336
FireRed-Image-Edit-1.1	3.446	3.017	3.390	3.566	4.459	3.216	3.405	3.878	3.059
Qwen-Image-Edit-2511	3.497	3.407	3.503	3.752	4.243	3.432	3.486	4.041	3.236
Boogu-Image-0.1-Edit	3.599	2.621	3.226	3.641	4.297	3.243	3.459	3.973	3.051
LongCat-Image-Edit	2.836	3.203	3.028	3.168	4.027	3.216	3.243	3.959	2.588
HiDream-O1-Image	2.582	2.249	2.836	2.875	3.703	2.649	2.973	3.108	2.119
FLUX.2 [dev]	2.254	3.073	3.706	2.605	2.568	3.243	3.649	2.554	2.159

Table 25: Quantitative comparison on WEGENEDITBENCH-TEXTEDIT: Commercial Text–Image Editing. Bold and underline denote the best and second-best scores.

Model Name	Text-side				Image-side				Overall ↑
	EG ↑	PR ↑	VQ ↑	DS ↑	EG ↑	PR ↑	VQ ↑	DS ↑	
Closed-source Models									
GPT-Image-2	<u>4.448</u>	3.676	4.295	4.423	**4.782**	3.818	**4.473**	**4.518**	<u>4.160</u>
Nano-Banana-Pro	4.382	<u>4.027</u>	4.282	4.408	4.508	3.847	4.220	<u>4.432</u>	4.139
Qwen-Image-3.0	4.418	3.455	**4.391**	<u>4.515</u>	4.508	3.763	4.322	4.356	4.139
Seedream-5.0-Pro	**4.472**	**4.241**	<u>4.380</u>	**4.578**	4.474	3.895	<u>4.351</u>	4.368	**4.296**
Open-source Models									
SenseNova-U1-8B-MoT-CoT	4.045	3.573	3.800	3.933	4.373	4.034	4.068	4.102	3.700
HunyuanImage-3.0-Instruct-Distil	4.083	3.780	4.083	4.270	4.458	<u>4.051</u>	4.169	4.347	3.906
HunyuanImage-3.0-Instruct-Distil-CoT	4.211	3.642	4.183	4.259	4.576	**4.254**	4.220	4.314	3.960
SenseNova-U1-8B-MoT	4.164	3.618	3.873	3.927	4.492	3.678	3.814	3.992	3.767
JoyAI-Image-Edit	4.191	3.609	3.845	4.102	4.576	3.763	4.271	4.203	3.835
FireRed-Image-Edit-1.1	3.745	3.382	3.600	4.037	<u>4.610</u>	3.390	3.949	4.347	3.500
Qwen-Image-Edit-2511	3.745	3.527	3.682	3.942	4.508	3.763	4.068	4.127	3.516
Boogu-Image-0.1-Edit	3.955	3.236	3.773	4.194	4.254	3.119	3.898	4.331	3.623
LongCat-Image-Edit	3.255	3.436	3.564	3.705	4.424	3.492	4.034	3.890	3.147
HiDream-O1-Image	3.227	3.000	3.545	3.641	3.661	2.915	3.390	3.475	2.955
FLUX.2 [dev]	1.964	3.418	3.555	2.154	3.051	3.237	3.542	2.873	1.967

Table 26: Quantitative comparison on WEGENEDITBENCH-TEXTEDIT: Social-media Text–Image Editing. Bold and underline denote the best and second-best scores.

Model Name	Text-side				Image-side				Overall ↑
	EG ↑	PR ↑	VQ ↑	DS ↑	EG ↑	PR ↑	VQ ↑	DS ↑	
Closed-source Models									
GPT-Image-2	**4.457**	3.837	<u>4.511</u>	4.454	**4.571**	3.839	4.304	4.348	**4.227**
Nano-Banana-Pro	4.274	4.105	4.274	<u>4.541</u>	4.228	<u>4.404</u>	4.018	4.351	4.127
Qwen-Image-3.0	<u>4.448</u>	3.615	**4.521**	**4.549**	4.534	3.810	**4.483**	**4.466**	<u>4.209</u>
Seedream-5.0-Pro	4.333	**4.312**	4.387	4.449	4.175	<u>4.404</u>	<u>4.368</u>	<u>4.395</u>	4.195
Open-source Models									
SenseNova-U1-8B-MoT-CoT	4.042	3.990	4.177	4.021	4.345	4.069	3.879	4.164	3.858
HunyuanImage-3.0-Instruct-Distil	3.906	4.073	3.875	4.311	4.259	4.345	4.121	4.241	3.824
HunyuanImage-3.0-Instruct-Distil-CoT	4.260	<u>4.219</u>	4.260	4.207	4.379	**4.534**	4.293	4.241	4.084
SenseNova-U1-8B-MoT	3.979	3.917	4.125	4.046	4.534	3.759	3.828	4.121	3.790
JoyAI-Image-Edit	3.844	3.938	4.073	4.182	<u>4.552</u>	3.741	4.190	4.250	3.780
FireRed-Image-Edit-1.1	3.469	3.417	3.719	4.100	4.448	3.690	4.207	4.267	3.404
Qwen-Image-Edit-2511	3.510	3.792	3.750	3.953	4.379	3.966	4.241	4.190	3.454
Boogu-Image-0.1-Edit	3.896	3.646	4.052	3.982	4.328	3.828	4.034	4.207	3.682
LongCat-Image-Edit	3.417	3.646	3.542	3.886	4.172	3.655	3.810	4.069	3.295
HiDream-O1-Image	3.240	2.948	3.688	3.675	3.638	2.793	3.603	3.414	2.984
FLUX.2 [dev]	1.968	3.421	3.653	2.163	2.684	3.421	3.561	2.816	1.965

F.2.3 MULTI-REFERENCE EDITING

Table 27: Quantitative comparison on WEGENEDITBENCH-MULTIIMAGEEDIT: Portrait Editing. Bold and underline denote the best and second-best scores.

Model Name	EG↑	PR↑	VQ↑	**Overall↑**
Closed-source Models				
GPT-Image-2	<u>4.315</u>	<u>4.107</u>	<u>3.658</u>	<u>4.101</u>
Nano-Banana-Pro	4.284	**4.122**	3.236	4.009
Seedream-5.0-Pro	**4.374**	4.094	**3.835**	**4.154**
Open-source Models				
HunyuanImage-3.0-Instruct-Distil	3.496	3.656	3.216	3.504
FireRed-Image-Edit-1.1	2.980	2.987	3.107	3.008
Qwen-Image-Edit-2511	2.740	2.887	3.160	2.883

Table 28: Quantitative comparison on WEGENEDITBENCH-MULTIIMAGEEDIT: Product and Commercial Editing. Bold and underline denote the best and second-best scores.

Model Name	EG↑	PR↑	VQ↑	**Overall↑**
Closed-source Models				
GPT-Image-2	<u>4.220</u>	**4.068**	<u>3.339</u>	<u>3.983</u>
Nano-Banana-Pro	3.795	3.889	3.043	3.682
Seedream-5.0-Pro	**4.250**	<u>4.056</u>	**3.361**	**3.994**
Open-source Models				
HunyuanImage-3.0-Instruct-Distil	3.198	3.089	2.861	3.087
FireRed-Image-Edit-1.1	2.442	2.775	2.858	2.658
Qwen-Image-Edit-2511	2.550	2.767	2.875	2.702

Table 29: Quantitative comparison on WEGENEDITBENCH-MULTIIMAGEEDIT: Text Editing. Bold and underline denote the best and second-best scores.

Model Name	EG↑	PR↑	VQ↑	**Overall↑**
Closed-source Models				
GPT-Image-2	**4.147**	**3.983**	<u>3.690</u>	**3.990**
Nano-Banana-Pro	3.664	3.779	3.319	3.641
Seedream-5.0-Pro	<u>4.018</u>	<u>3.936</u>	**3.697**	<u>3.921</u>
Open-source Models				
HunyuanImage-3.0-Instruct-Distil	2.700	3.130	2.730	2.878
FireRed-Image-Edit-1.1	1.839	2.703	2.695	2.356
Qwen-Image-Edit-2511	1.873	2.508	2.754	2.303

Table 30: Quantitative comparison on WEGENEDITBENCH-MULTIIMAGEEDIT: Scene and Spatial Editing. Bold and underline denote the best and second-best scores.

Model Name	EG↑	PR↑	VQ↑	**Overall**↑
Closed-source Models				
GPT-Image-2	**4.453**	**4.222**	**3.795**	**4.229**
Nano-Banana-Pro	4.060	3.966	3.368	3.884
Seedream-5.0-Pro	<u>4.272</u>	<u>4.123</u>	<u>3.658</u>	<u>4.089</u>
Open-source Models				
HunyuanImage-3.0-Instruct-Distil	3.209	3.549	2.978	3.299
FireRed-Image-Edit-1.1	3.026	3.316	3.043	3.145
Qwen-Image-Edit-2511	2.880	3.179	3.145	3.053

Table 31: Quantitative comparison on WEGENEDITBENCH-MULTIIMAGEEDIT: Pet Editing. Bold and underline denote the best and second-best scores.

Model Name	EG↑	PR↑	VQ↑	**Overall**↑
Closed-source Models				
GPT-Image-2	**4.579**	**4.211**	<u>3.833</u>	**4.282**
Nano-Banana-Pro	4.357	4.018	3.384	4.027
Seedream-5.0-Pro	<u>4.463</u>	<u>4.176</u>	**3.926**	<u>4.241</u>
Open-source Models				
HunyuanImage-3.0-Instruct-Distil	3.787	3.734	3.468	3.702
FireRed-Image-Edit-1.1	2.912	2.895	3.316	2.986
Qwen-Image-Edit-2511	2.816	2.974	3.421	3.000

Table 32: Quantitative comparison on WEGENEDITBENCH-MULTIIMAGEEDIT: Aesthetic and Photographic Editing. Bold and underline denote the best and second-best scores.

Model Name	EG↑	PR↑	VQ↑	**Overall**↑
Closed-source Models				
GPT-Image-2	**4.455**	**4.223**	<u>3.955</u>	**4.263**
Nano-Banana-Pro	3.955	4.153	3.739	3.991
Seedream-5.0-Pro	<u>4.216</u>	<u>4.171</u>	**4.000**	<u>4.155</u>
Open-source Models				
HunyuanImage-3.0-Instruct-Distil	3.255	3.798	3.287	3.479
FireRed-Image-Edit-1.1	3.044	3.195	3.319	3.159
Qwen-Image-Edit-2511	2.646	3.000	3.204	2.899

F.3 Complete Task Taxonomy

F.3.1 Single-Image Task Taxonomy

Table 33: Single-Image Task Taxonomy. Family headings give unique covered instances; rows define the second-level tasks. The subset has 1,008 instances and 65 task labels.

Second-level task	Operational definition
Basic local editing	631 unique instances
Object addition	Insert a specified object with correct location, scale, and occlusion.
Object removal	Remove a target, reconstruct its background, and preserve surrounding content.
Object replacement	Replace a target while respecting its scene role.
Object count change	Change the specified objects to the requested count.
Local attribute editing	Change target color or attributes within the requested scope.
Material replacement	Change target material while retaining geometry and identity.
Motion-state editing	Change motion state and depict requested motion cues.
Texture-detail editing	Modify surface patterns or details while preserving carrier structure.
Size or shape adjustment	Adjust target proportions or shape with coherent surrounding geometry.
Background replacement	Replace the background while preserving foreground identity and boundaries.
Subject extraction	Extract the target with clean boundaries and the requested background.
Image outpainting	Extend the canvas with compatible scene content.
Regional repainting	Regenerate a specified region while preserving content outside it.
Political-symbol editing	Insert or modify requested political symbols within the source scene.
Portrait editing	309 unique instances
Portrait retouching	Retouch facial appearance while preserving identity and natural structure.
Makeup editing	Modify makeup while preserving facial identity and anatomy.
Hairstyle change	Change hairstyle while retaining other required subject attributes.
Hair-color change	Recolor hair while preserving its shape, face, and surroundings.
Facial-expression editing	Change facial expression while preserving identity and facial structure.
Clothing-style editing	Restyle clothing from one input while preserving the person.
Pose editing	Change body or limb configuration with plausible anatomy.
Formal or professional portrait	Present a portrait in the requested formal or professional style.
Portrait or role-play styling	Style a portrait for a specified role, character, or concept.
Themed portrait styling	Coordinate attire, props, and setting for a portrait theme.
Photo-print or album rendering	Render source content as a photograph, instant print, or album.
Animalized or creative face editing	Integrate requested animal-like or creative facial transformations.
Age editing	Change apparent age while preserving identity cues and scene content.
Surreal portrait editing	Add surreal elements while preserving the intended portrait composition.
Named-character editing	Introduce or evoke a specified recognizable character or persona.
Aesthetic and photographic editing	320 unique instances
Global stylization	Change global style while preserving requested semantic content.
Color-tone adjustment	Adjust palette or grading while preserving required scene details.
Relighting	Modify illumination, highlights, and shadows to match requested lighting.
Depth-of-field editing	Adjust focus and background blur around the intended subject.
Framing or zoom adjustment	Change framing or shot scale while preserving scene relationships.
Photographic-view refinement	Refine viewpoint or composition while maintaining scene plausibility.
Subject-view rotation	Rotate the target view while preserving recognizable structure.
Material-detail enhancement	Enhance texture or gloss while preserving material identity.
Sketch-to-image conversion	Render a sketch into an image consistent with its structure.
Drawing colorization or rendering	Color or render artwork while retaining contours and composition.
Fisheye selfie rendering	Create a selfie composition with coherent fisheye distortion.
Product and commercial editing	117 unique instances
White-background product image	Place a product on white while preserving its details.
Transparent-background product image	Extract a product onto transparency with clean contours.
Product-in-scene rendering	Restage a product in context while preserving its appearance.
Studio product photography	Create studio product photography with controlled lighting and materials.
Exploded product view	Separate or expose product components in the requested explanatory arrangement.
Creative product advertising	Create a stylized advertisement or object-inspired product arrangement.

Continued on next page.

Table 33 continued.

Second-level task	Operational definition
Dieline-to-package rendering	Render a flat packaging layout as a coherent 3D package.
Action and interaction editing	64 unique instances
Person–person interaction	Edit interpersonal interactions while preserving roles and contact relationships.
Sports or dynamic action	Depict a requested dynamic action with coherent body configuration.
Person–scene interaction	Edit person–scene interactions, including support and contact.
Hand–object interaction	Change a held object or grasp with plausible finger contact.
Reasoning-intensive editing	164 unique instances
Spatial-reasoning editing	Execute edits requiring spatial relations or relative-position reasoning.
Coordinated multi-object editing	Coordinate changes to multiple objects under a shared objective.
Relational editing	Change a relationship between entities, not just their appearance.
Physical-state editing	Change physical state and depict its visible consequences.
Causal-scene editing	Depict an event and its visually coherent consequences.
Idiom-grounded editing	Translate colloquial or idiomatic intent into a visual edit.
Composite-operation editing	Coordinate multiple transformations and their interacting constraints.
Negative or preservation constraints	Honor explicit prohibitions or invariances while executing edits.
Restoration and low-level enhancement	67 unique instances
Old-photo restoration	Repair old-photo degradation while retaining original subject content.
Deblurring	Reduce blur while preserving structure and avoiding unsupported details.
Denoising	Suppress noise while preserving meaningful texture and edges.
Exposure correction	Correct exposure while maintaining coherent color and contrast.
Super-resolution enhancement	Enhance resolution and detail while preserving image content.
Traditional Chinese cultural editing	33 unique instances
Traditional Chinese cultural editing	Edit requested traditional cultural motifs, attire, objects, or styles.

F.3.2 TEXT-INTENSIVE TASK TAXONOMY

Table 34: Text-Intensive Task Taxonomy. Family headings give unique covered instances; rows define the second-level tasks. The subset has 1,000 instances and 35 task labels.

Second-level task	Operational definition
Basic text editing	678 unique instances
Text-content replacement	Replace specified wording while preserving placement and surrounding content.
Text insertion	Insert wording without obscuring protected visual information.
Text removal	Remove text, reconstruct its carrier, and preserve non-target content.
Typography preservation	Change wording while preserving specified typography and arrangement.
Typography-attribute editing	Change typography while preserving wording and required structure.
Multilingual text replacement	Replace text in another language and adapt its layout.
Perspective-aware text editing	Edit surface text with correct perspective, curvature, lighting, and material.
Logo or symbol replacement	Replace a logo, icon, or symbol within its carrier.
Text-effect design	Apply decorative text effects while preserving character legibility.
Watermark insertion or removal	Add or remove a watermark while preserving required scene content.
Typographical correction	Correct erroneous characters while preserving surrounding text and presentation.
Interfaces, documents, and charts	177 unique instances
Interface-copy editing	Edit interface copy while preserving components, spacing, and alignment.
Table or document editing	Edit document fields or cells while retaining unrelated content and rules.
Chart-label editing	Edit chart labels while preserving the plot unless otherwise requested.
Visual-question completion	Solve a depicted question and write the answer where specified.
Quantity or price comparison	Compare visible quantities and mark or write the requested result.
Interface mockup design	Design an interface with specified copy, components, and hierarchy.

Continued on next page.

Table 34 continued.

Second-level task	Operational definition
Commercial text–image design	110 unique instances
Product-advertising poster	Create a product poster with specified copy, layout, and product fidelity.
Packaging-text editing	Edit packaging text while preserving geometry and non-target information.
Product-feature labels	Add product labels or callouts without obscuring protected details.
E-commerce detail module	Organize product information into an illustrated detail-page module.
Posters and visual layouts	221 unique instances
Xiaohongshu-style cover	Create a social cover with prominent titles and a clear subject.
Event or course poster	Organize event or course information into a poster hierarchy.
Seasonal marketing poster	Compose seasonal imagery and promotional copy.
Magazine-cover design	Arrange cover titles and supporting text around the source subject.
Infographic design	Organize information into visual sections, labels, and explanations.
Timeline graphic	Arrange events or stages sequentially with associated text and imagery.
Map or world-layout labeling	Create or edit place names, regions, legends, or map annotations.
Before–after comparison layout	Present requested states in a labeled comparison layout.
Poster-layout rearrangement	Rearrange poster hierarchy while preserving key visuals and copy.
Format or platform adaptation	Adapt format while retaining important content and readability.
Social-media graphics	96 unique instances
Venue or recommendation graphic	Create scene-grounded recommendation graphics with titles, tags, or highlights.
Outfit-description graphic	Add outfit labels or commentary without obscuring clothing.
Meme or sticker design	Combine a subject with wording matching its expression.
Pet portrait or identity card	Create a pet portrait card with names or descriptive text.

F.3.3 Multi-Reference Task Taxonomy

Table 35: Multi-Reference Task Taxonomy. Family headings give unique covered instances; rows define the second-level tasks. The subset has 732 instances and 48 task labels.

Second-level task	Operational definition
Portrait editing	150 unique instances
Face transfer	Transfer facial identity while preserving required body, pose, and scene.
Reference-based clothing transfer	Transfer reference clothing while preserving the target person's identity.
Hairstyle transfer	Transfer hairstyle without substituting unrelated identity attributes.
Expression transfer	Transfer expression while preserving the target person's identity.
Jewelry try-on	Fit reference jewelry with correct scale, contact, and occlusion.
Makeup and styling transfer	Transfer specified reference makeup or styling to the target.
Multi-person interaction	Compose referenced people with correct identities, roles, and interactions.
Person–scene composition	Place a referenced person with coherent pose, contact, and perspective.
Person–pet interaction	Compose a person and pet with correct reference assignments.
Surreal portrait composition	Combine portraits and surreal elements with correct subject assignments.
Product and commercial editing	120 unique instances
Artwork-to-packaging transfer	Transfer artwork or text onto packaging with correct folds and perspective.
Character-themed commercial display	Integrate referenced characters or branding into a commercial display.
Product–scene composition	Compose a reference product while preserving identity and physical placement.
Model–product–scene composition	Bind person, product, and scene to their designated references.
Dieline-guided package rendering	Render flat artwork and reference information as a 3D package.
Image-to-device display transfer	Insert imagery into a device display with correct crop and perspective.
Branding transfer	Apply reference branding while preserving its visual identity.
Multi-product composition	Arrange referenced products with correct identities and relationships.
Scene and spatial editing	117 unique instances
Scene-element addition	Add reference elements at requested target-scene locations.
Scene-element replacement	Replace scene elements with reference content while preserving surroundings.

Continued on next page.

Table 35 continued.

Second-level task	Operational definition
Reference-based background replacement	Transfer a reference background while retaining the required foreground.
Scene-viewpoint transformation	Reconstruct a scene from a reference-specified viewpoint.
Scene-style transfer	Transfer scene style while preserving target content and structure.
Weather or season transfer	Transfer weather or season while preserving identities and layout.
Scene-lighting or time transfer	Transfer reference lighting or time-of-day appearance.
Plan-to-scene rendering	Render a scene consistent with a reference map or floor plan.
Aesthetic and photographic editing	113 unique instances
Reference-style transfer	Transfer reference style while preserving requested target content.
Reference-guided line-art coloring	Color line art from references while respecting its structure.
Reference-material transfer	Transfer reference material appearance onto a target object.
Reference-guided relighting	Transfer illumination with plausible lighting on the target.
Photographic-viewpoint transfer	Transfer photographic viewpoint while preserving required subject details.
Composition transfer	Transfer arrangement or framing while preserving target subjects.
Color-palette transfer	Transfer a reference palette while preserving required content.
Lens or focus-effect transfer	Transfer focus or lens effects while preserving focal content.
Pet editing	114 unique instances
Pet portrait styling	Style a pet portrait while preserving the pet's identity.
Pet-pattern transfer	Transfer reference markings to the specified pet region.
Pet clothing transfer	Transfer pet clothing with coherent fit and preserved identity.
Pet-to-toy rendering	Render a pet as a reference-guided toy retaining identity cues.
Pet–scene composition	Compose a pet and scene with coherent scale, lighting, and contact.
Pet-expression transfer	Transfer pet expression while preserving identity and anatomy.
Pet-pose transfer	Transfer pet pose while preserving identity.
Text editing	118 unique instances
Full-image watermark transfer	Transfer a watermark with the requested coverage and appearance.
Interface or screen-text transfer	Transfer interface or display text with correct perspective and alignment.
Reference-guided text–image layout	Rearrange text and images using designated layout references.
Reference-guided text rendering	Render reference-guided text while preserving correct characters.
Reference-based meme generation	Combine reference content and wording into an expressive meme.
Multilingual text transfer	Transfer or translate reference text and adapt its presentation.
Typeface-reference transfer	Transfer reference typography while retaining requested wording.

G QUALITATIVE CASE STUDIES

Sections G.1–G.3 present 32 benchmark cases in the order WeGenEditBench-SingleImageEdit (12), WeGenEditBench-TextEdit (12), and WeGenEditBench-MultiImageEdit (8), with at most two panels per page. Each panel includes the shown instruction, source/reference images, and outputs from the labeled models. They illustrate visible behaviors; population-level conclusions use the quantitative analyses above. Multi-reference inputs retain their ordering so that identity, appearance, and position can be checked against the intended source.

G.1 WEGENEDITBENCH-SINGLEIMAGEEDIT: SINGLE-IMAGE EDITING CASES

Inspect the requested transformation together with non-target objects, local geometry, and surrounding appearance.



Figure 9: WeGenEditBench-SingleImageEdit: object addition (case 350). The panel shows the input image(s), editing instruction, and outputs from the labeled models.



Figure 10: WeGenEditBench-SingleImageEdit: object removal (case 110). The panel shows the input image(s), editing instruction, and outputs from the labeled models.

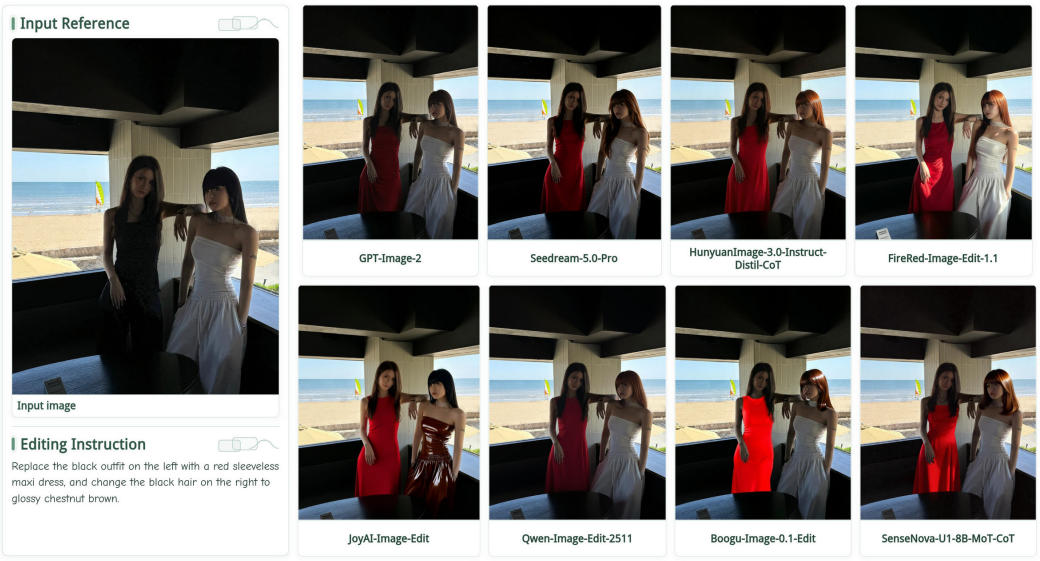

Figure 11: WeGenEditBench-SingleImageEdit: clothing-style editing (case 700). The panel shows the input image(s), editing instruction, and outputs from the labeled models.

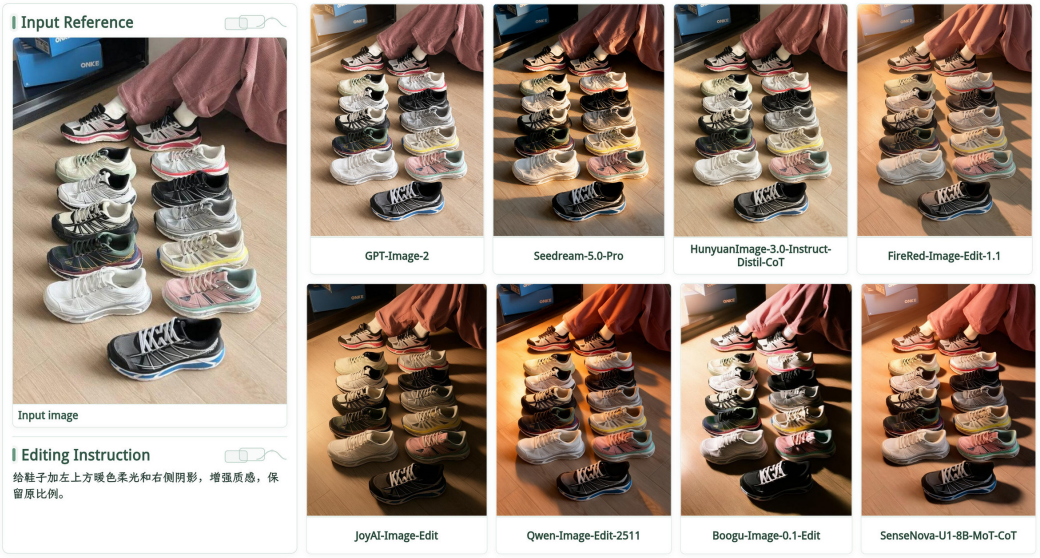

Figure 12: WeGenEditBench-SingleImageEdit: lighting redesign (case 118). The panel shows the input image(s), editing instruction, and outputs from the labeled models.



Figure 13: WeGenEditBench-SingleImageEdit: exploded product view (case 359). The panel shows the input image(s), editing instruction, and outputs from the labeled models.



Figure 14: WeGenEditBench-SingleImageEdit: dieline-to-package rendering (case 495). The panel shows the input image(s), editing instruction, and outputs from the labeled models.

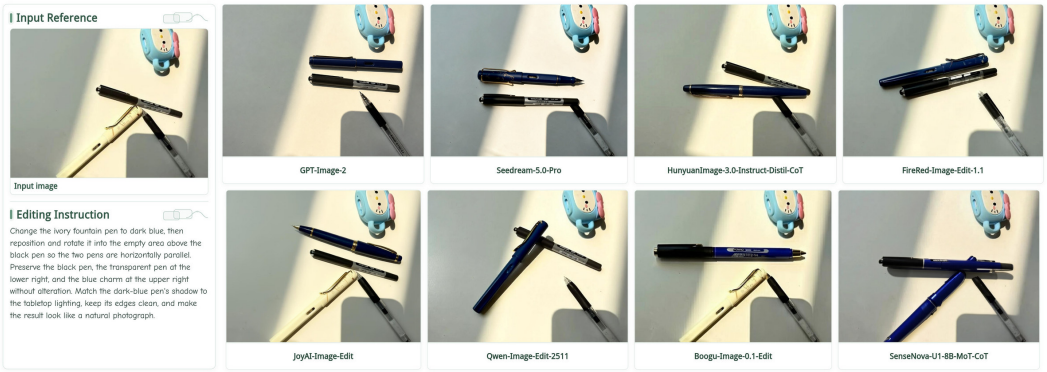

Figure 15: WeGenEditBench-SingleImageEdit: spatial-reasoning editing (case 142). The panel shows the input image(s), editing instruction, and outputs from the labeled models.

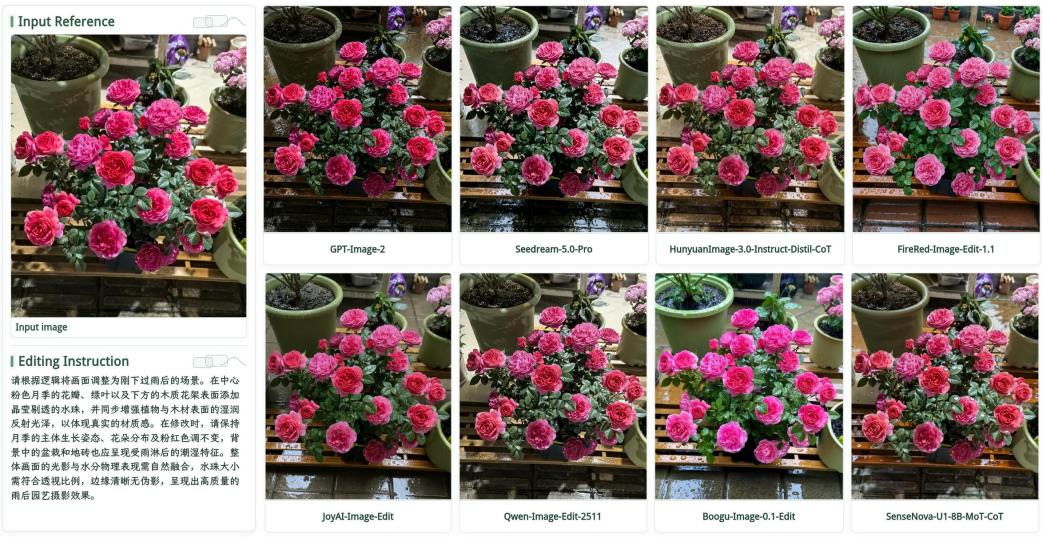

Figure 16: WeGenEditBench-SingleImageEdit: causal-scene editing (case 272). The panel shows the input image(s), editing instruction, and outputs from the labeled models.



Figure 17: WeGenEditBench-SingleImageEdit: old-photo restoration (case 356). The panel shows the input image(s), editing instruction, and outputs from the labeled models.



Figure 18: WeGenEditBench-SingleImageEdit: drawing colorization (case 241). The panel shows the input image(s), editing instruction, and outputs from the labeled models.

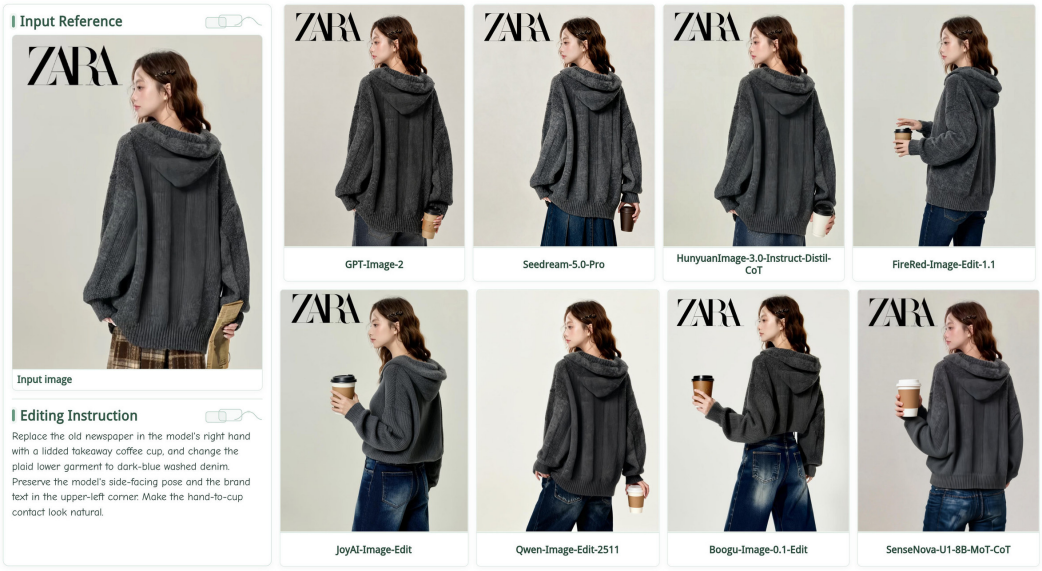

Figure 19: WeGenEditBench-SingleImageEdit: hand–object interaction (case 661). The panel shows the input image(s), editing instruction, and outputs from the labeled models.

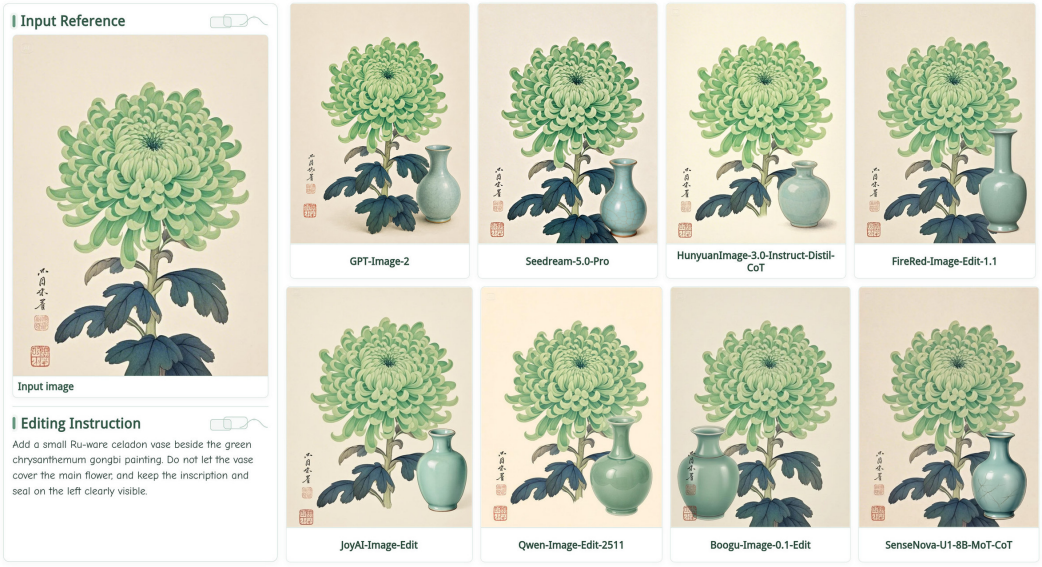

Figure 20: WeGenEditBench-SingleImageEdit: chinese traditional culture editing (case 550). The panel shows the input image(s), editing instruction, and outputs from the labeled models.

G.2 WeGenEditBench-TextEdit: Text-Intensive Editing Cases

Compare literal wording, typography, carrier geometry, and unaffected visual content separately.

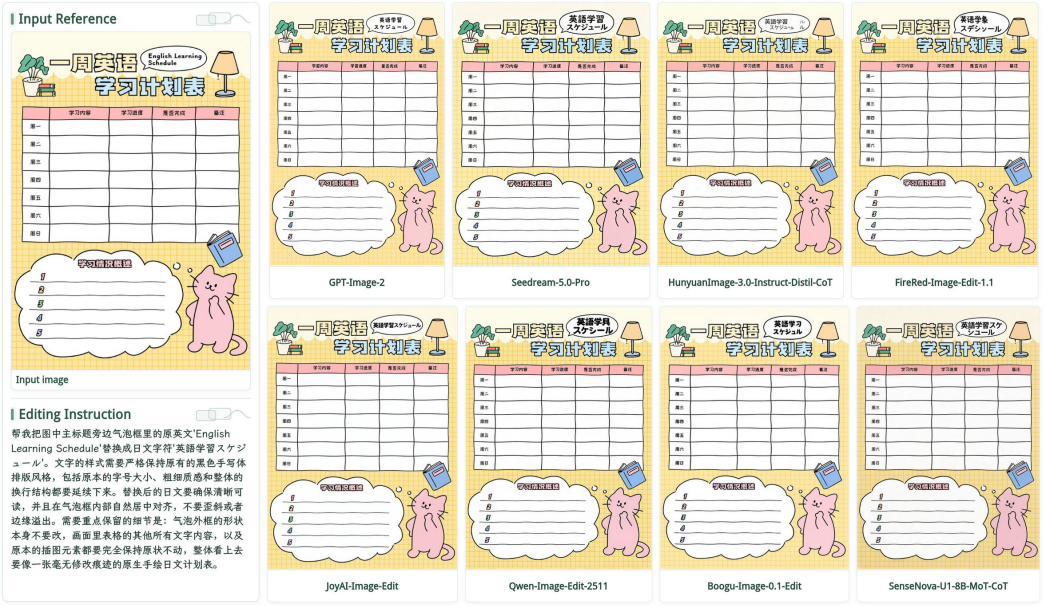

Figure 21: WeGenEditBench-TextEdit: multilingual text replacement (case 418). The panel shows the input image(s), editing instruction, and outputs from the labeled models.



Figure 22: WeGenEditBench-TextEdit: perspective-aware text editing (case 541). The panel shows the input image(s), editing instruction, and outputs from the labeled models.



Figure 23: WeGenEditBench-TextEdit: text removal (case 650). The panel shows the input image(s), editing instruction, and outputs from the labeled models.

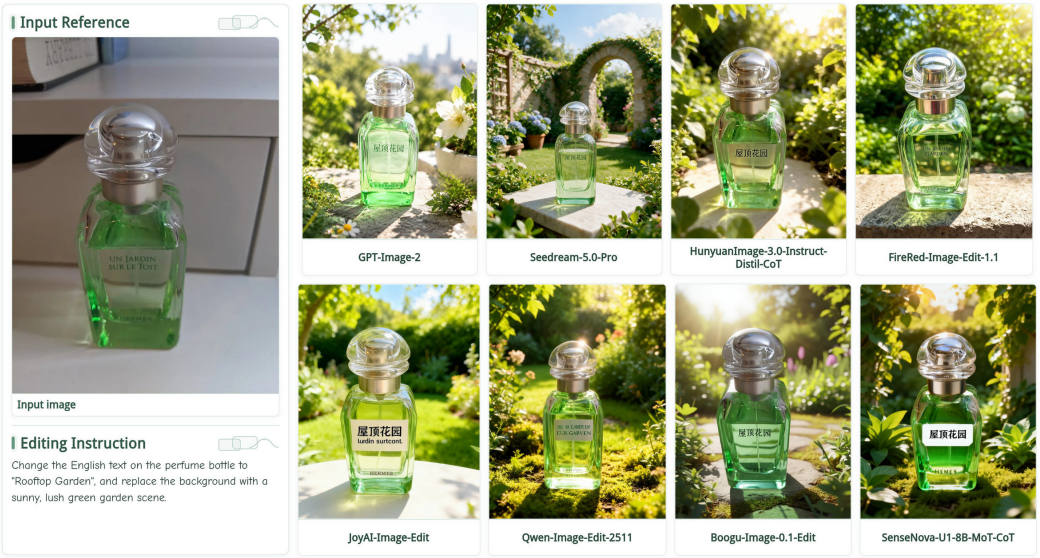

Figure 24: WeGenEditBench-TextEdit: product-packaging text editing (case 399). The panel shows the input image(s), editing instruction, and outputs from the labeled models.



Figure 25: WeGenEditBench-TextEdit: product-feature labels (case 502). The panel shows the input image(s), editing instruction, and outputs from the labeled models.



Figure 26: WeGenEditBench-TextEdit: poster layout reorganization (case 111). The panel shows the input image(s), editing instruction, and outputs from the labeled models.



Figure 27: WeGenEditBench-TextEdit: magazine cover (case 287). The panel shows the input image(s), editing instruction, and outputs from the labeled models.



Figure 28: WeGenEditBench-TextEdit: multi-format and multi-platform adaptation (case 980). The panel shows the input image(s), editing instruction, and outputs from the labeled models.

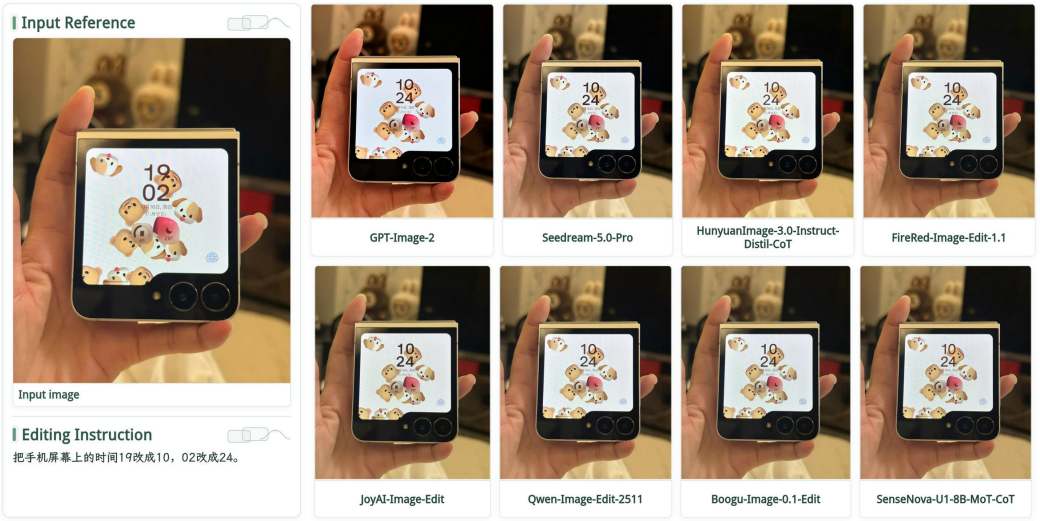

Figure 29: WeGenEditBench-TextEdit: interface and screenshot text editing (case 446). The panel shows the input image(s), editing instruction, and outputs from the labeled models.

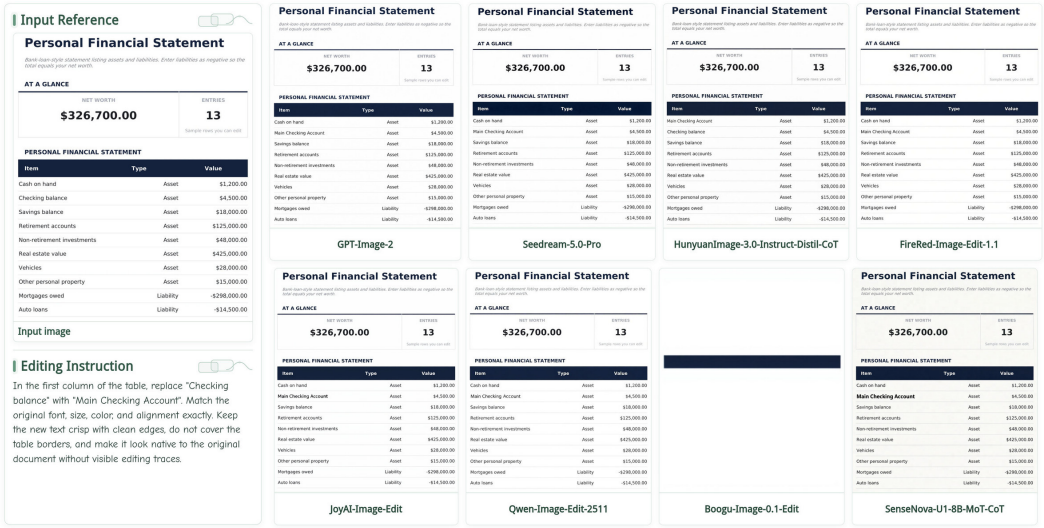

Figure 30: WeGenEditBench-TextEdit: local table and document editing (case 291). The panel shows the input image(s), editing instruction, and outputs from the labeled models.



Figure 31: WeGenEditBench-TextEdit: meme and sticker design (case 27). The panel shows the input image(s), editing instruction, and outputs from the labeled models.



Figure 32: WeGenEditBench-TextEdit: pet identification card (case 794). The panel shows the input image(s), editing instruction, and outputs from the labeled models.

G.3 WeGenEditBench-MultiImageEdit: Multi-Reference Editing Cases

Check whether each transferred identity, attribute, pose, or scene element is bound to its assigned reference, alongside the requested composition.



Figure 33: WeGenEditBench-MultiImageEdit: interpersonal action transfer (case 164). The panel shows the input image(s), editing instruction, and outputs from the labeled models.



Figure 34: WeGenEditBench-MultiImageEdit: model–product–scene composition (case 279). The panel shows the input image(s), editing instruction, and outputs from the labeled models.

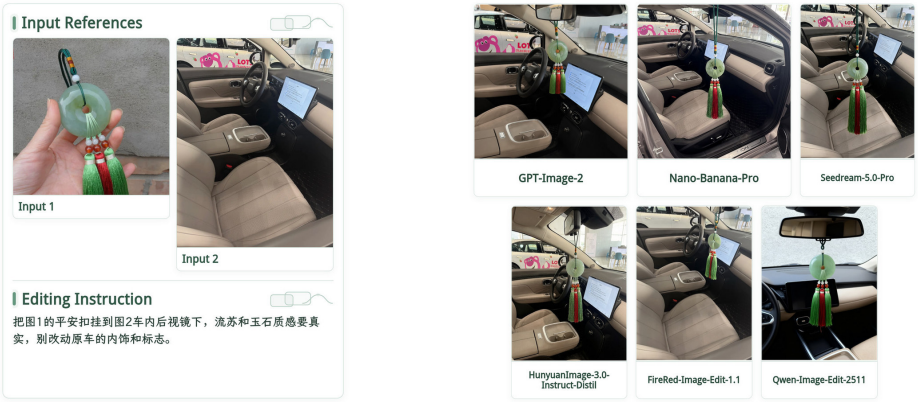

Figure 35: WeGenEditBench-MultiImageEdit: multi-product composition (case 573). The panel shows the input image(s), editing instruction, and outputs from the labeled models.

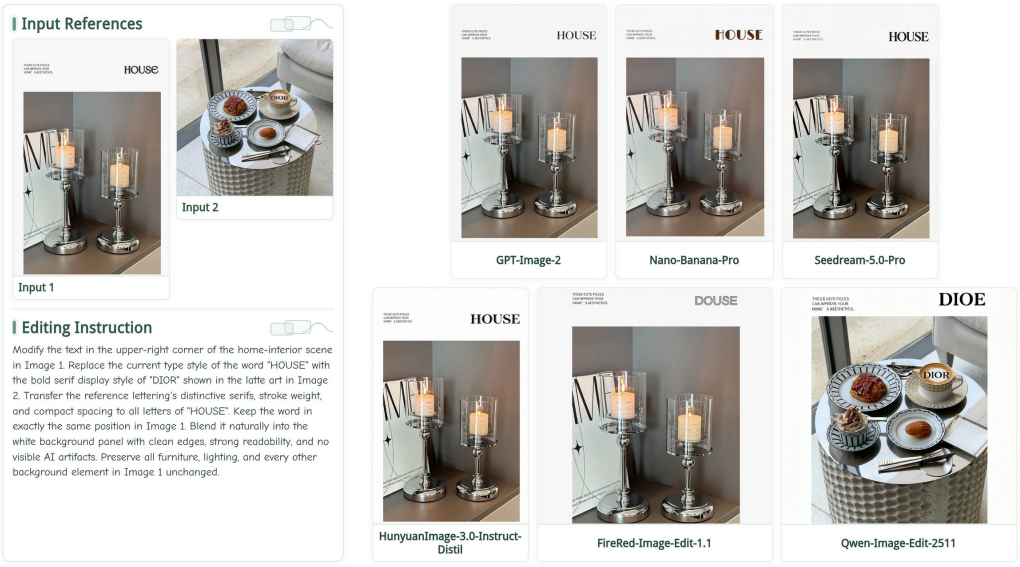

Figure 36: WeGenEditBench-MultiImageEdit: font-reference transfer (case 439). The panel shows the input image(s), editing instruction, and outputs from the labeled models.



Figure 37: WeGenEditBench-MultiImageEdit: scene lighting and time-of-day transfer (case 173). The panel shows the input image(s), editing instruction, and outputs from the labeled models.

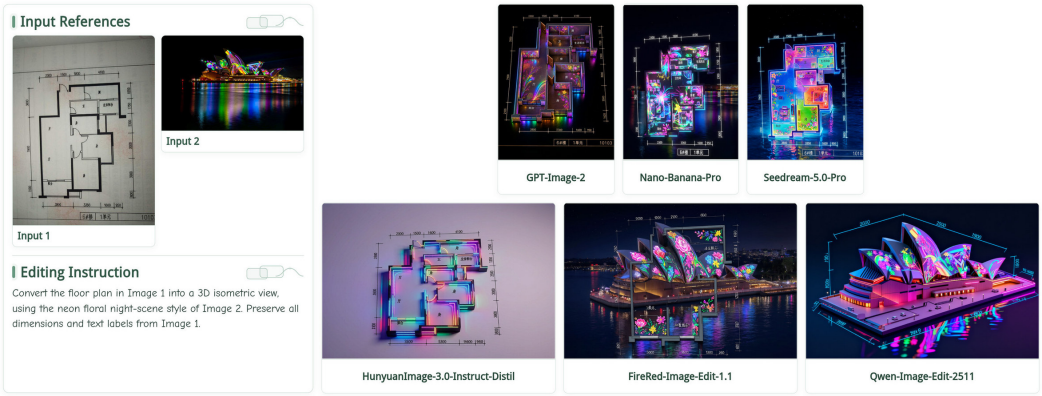

Figure 38: WeGenEditBench-MultiImageEdit: map or floor-plan to scene (case 674). The panel shows the input image(s), editing instruction, and outputs from the labeled models.



Figure 39: WeGenEditBench-MultiImageEdit: pet-to-toy transformation (case 277). The panel shows the input image(s), editing instruction, and outputs from the labeled models.

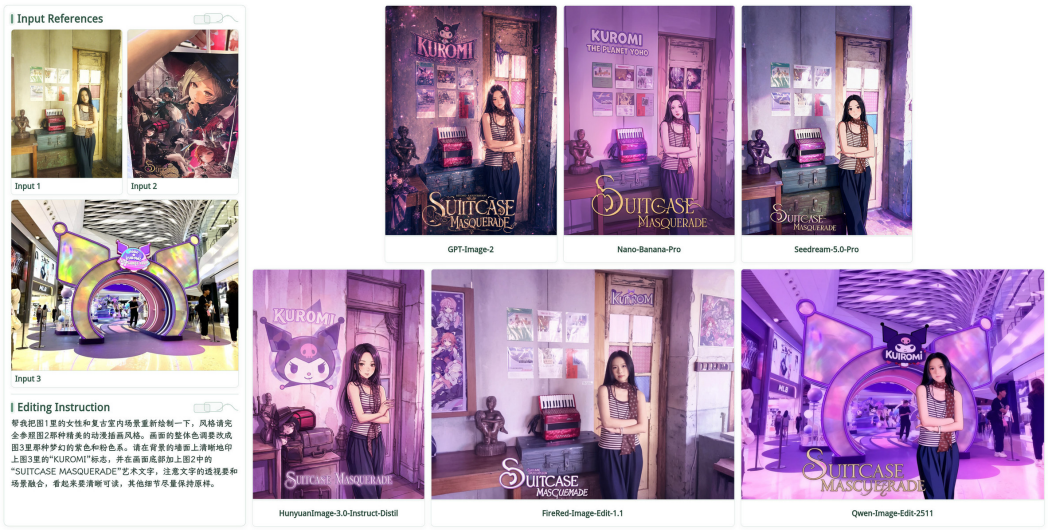

Figure 40: WeGenEditBench-MultiImageEdit: style transfer (case 203). The panel shows the input image(s), editing instruction, and outputs from the labeled models.

G.4 Qualitative Evidence of Model Adaptation



Figure 41: Qualitative text-to-image comparisons before and after WeGenEvoStudio adaptation. Each triplet shows the prompt, baseline, and adapted output, with FLUX.2 [dev] on the left and Qwen-Image-2512 on the right.

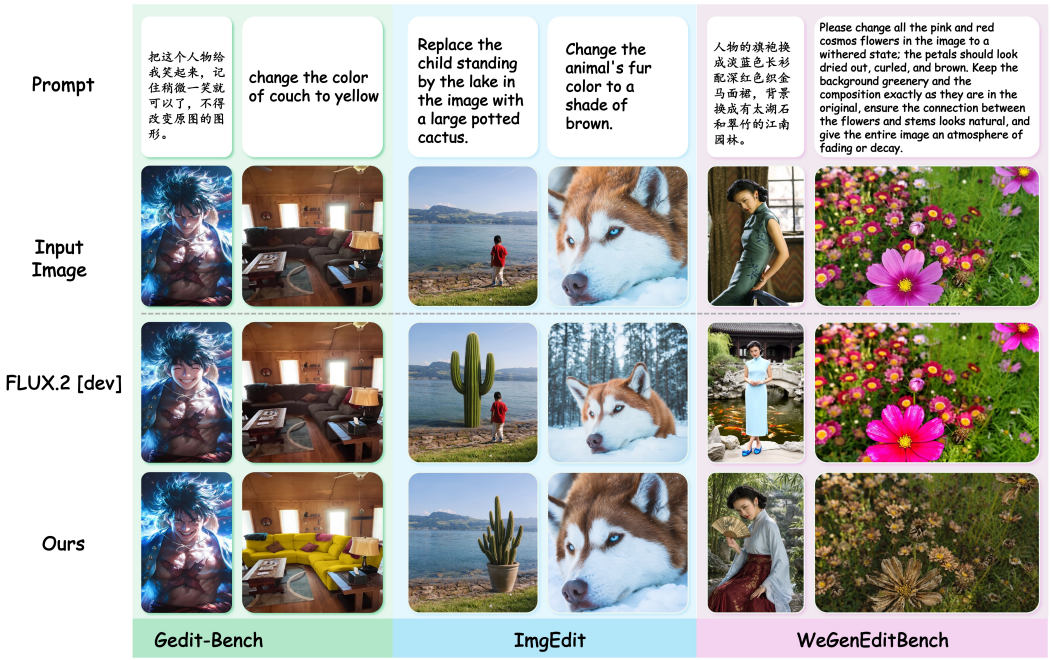

Figure 42: Additional editing comparisons for FLUX.2 [dev], following Figure 5. Examples cover expression and color changes, object replacement, clothing and scene composition, and flower withering. ImgEdit comparisons retain the inference-setting qualification in Appendix D.2.