UniREditBench: A Unified Reasoning-based Image Editing Benchmark

Feng Han[1,2*], Yibin Wang[1,2*], Chenglin Li[2,3], Zheming Liang[2], Dianyi Wang[1,2], Yang Jiao[1], Zhipeng Wei[4], Chao Gong[1], Cheng Jin[1,2], and Jiaqi Wang[2,5†]

[1] Fudan University [2] Shanghai Innovation Institute [3] Zhejiang University [4] UC Berkeley [5] JD.COM

Project Page: https://maplebb.github.io/UniREditBench/

Abstract. Recent advances in multimodal generative models have driven substantial improvements in image editing. However, current generative models still struggle with handling diverse and complex image editing tasks that require implicit reasoning, underscoring the need for a comprehensive benchmark to systematically assess their performance across various reasoning scenarios. Existing benchmarks primarily focus on single-object attribute transformation in realistic scenarios, which, while effective, encounter two key challenges: **(1)** they largely overlook multi-object interactions as well as game-world scenarios that involve human-defined rules, which are common in real-life applications; **(2)** they only rely on textual references to evaluate the generated images, potentially leading to systematic misjudgments, especially in complex reasoning scenarios. To this end, this work proposes **UniREditBench**, a unified benchmark for reasoning-based image editing evaluation. It comprises 2,700 meticulously curated samples, covering both real- and game-world scenarios across 8 primary dimensions and 18 sub-dimensions. To improve evaluation reliability, we introduce multimodal dual-reference evaluation, providing both textual and ground-truth image references for each sample assessment. Furthermore, we design an automated multi-scenario data synthesis pipeline and construct **UniREdit-Data-100K**, a large-scale synthetic dataset with high-quality chain-of-thought (CoT) reasoning annotations. We fine-tune Bagel on this dataset and develop **UniREdit-Bagel**, demonstrating substantial improvements in both in-domain and out-of-distribution settings. Through thorough benchmarking of both open-source and closed-source image editing models, we reveal their strengths and weaknesses across various aspects.

Keywords: Image Editing · Visual Reasoning · Generative Model

1 Introduction

Recent advances in multimodal generative models have led to remarkable improvements in instruction-conditioned image editing. Generative models [3, 41,

* Equal contribution. † Corresponding author.

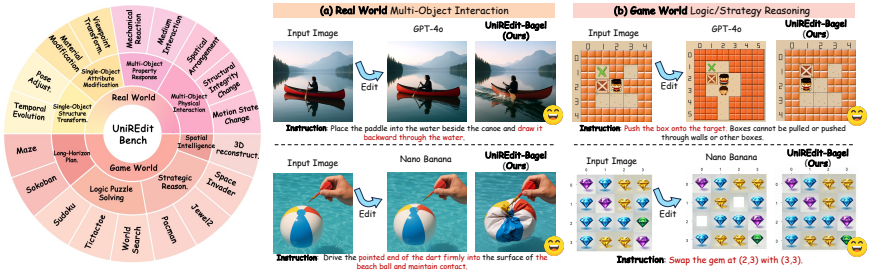

Fig. 1: UniREditBench covers both real-world and game-world reasoning scenarios across 8 primary dimensions and 18 sub-dimensions. We provide editing cases of (a) real-world multi-object interaction, and (b) game-world logical/strategy reasoning.

42, 45, 49, 52, 54], including Step1X-Edit [23], FLUX-Kontext [2], Bagel [6], Nano Banana [8], and GPT-4o [14], have demonstrated a powerful ability to understand diverse textual instructions and generate semantically consistent image edits. In parallel, reinforcement learning-based training strategies [21, 37, 38, 48] are continuously advancing, further enhancing the capabilities of image editing models. With these rapid developments, the need for a more comprehensive benchmark to evaluate model editing capabilities across different aspects has become increasingly essential. Early benchmarks [23, 51] focus on local details or global stylistic changes, e.g., style transfer, color alteration, and object removal. However, they fail to cover editing tasks that require models to perform implicit reasoning [10, 53], which are commonly used in real-life applications. As illustrated in Fig. 1, when editing instructions involving real-world or human-defined game rules, current models often generate results that lack physical plausibility. To this end, recent efforts have introduced reasoning-aware evaluation across temporal, spatial, and logical dimensions [56], and proposed a knowledge-grounded taxonomy assessing factual, conceptual, and procedural knowledge types [46].

Despite their effectiveness, these benchmarks still face two significant challenges: **(1)** they primarily focus on single-object attribute changes in realistic scenarios, neglecting multi-object interactions and game-world scenarios involving human-defined rules (see Tab. 1). This narrow scope restricts their ability to evaluate how effectively models generalize across a wider range of complex reasoning contexts. Additionally, **(2)** they mainly rely on textual references to evaluate the generated images [46, 56], which may lead to systematic misjudgments, especially in complex reasoning-based editing scenarios (see Fig. 2).

In this work, we posit that: **(1)** While current models exhibit proficiency in perceptual instruction following and simple reasoning editing settings (e.g., *Transform an intact apple to a bitten one*), they still struggle with complex reasoning-based image editing that necessitates the comprehension of multi-object interaction characteristics (e.g., *Draw the paddle backward through the water*) as well as logical constraints of puzzle and game scenarios (e.g., *Control the player and push the box to the target*), as illustrated in Fig. 1. **(2)** Relying on textual references in evaluating complex reasoning-based image editing tasks

Table 1: Comparison of reasoning-based image editing benchmarks. "S-Obj" and "M-Obj" denote single-object and multi-object editing, respectively. UniREditBench stands out for its dual-reference evaluation protocol as well as broader coverage of real- and game-world scenarios. See Appendix for Detailed analysis.

Benchmark	Size	▼ Evaluation Validity			▼ Real World Scenario								▼ Game World Scenario						
		Reference Image	Eval Ref. Modality (Text/Image)	Unified Dual-ref (Text+Image)	Attribute (S-Obj)	Temporal (S-Obj)	Pose (S-Obj)	Structure (M-Obj)	Spatial (M-Obj)	Motion (M-Obj)	Mechanic (M-Obj)	Medium (M-Obj)	Logical	Long-plan	Strategic	Spatial	Explicit Rules	Action Control	Intra-game Multi-task
SmartEdit [13]	219	All (219/219)	Image-only		✔				✔										
RISE [56]	360	Partial (70/360)	Text-major		✔	✔			✔				✔	✔					
KRIS [46]	1,267	Partial (50/1,267)	Text-major		✔	✔			✔			✔	✔	✔					
UniREditBench	**2,700**	**All (2,700/2,700)**	**Dual**	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

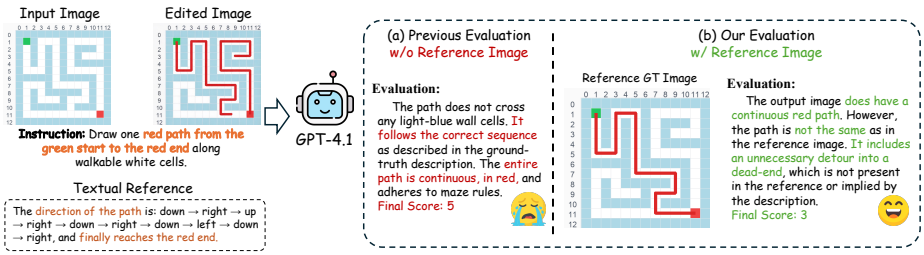

Fig. 2: Image editing evaluation comparison. Current text-reference-only evaluation potentially leads to misjudgments, while our dual-reference evaluation results in more reliable assessments.

often leads to unreliable judgments. As shown in Fig. 2 (a), the text-reference-only evaluator assigns an inflated score even when the edited image introduces an additional faulty path. Therefore, we intuitively believe that incorporating a ground-truth (GT) image as an additional visual reference can enable more precise evaluation.

To this end, this work proposes **UniREditBench**, a unified benchmark for reasoning-based image editing assessment with broader evaluation dimension coverage and robust evaluation pipeline. Specifically, **(1)** we adopt a scenario-to-category hierarchical dimension design, covering diverse reasoning types in both real-world and game-world scenarios (shown in Fig. 1): it includes 2,700 carefully curated samples organized across 8 primary dimensions and 18 sub-categories, e.g., *multi-object interaction* in real world, and *long-horizon game planning* in game world. Meanwhile, **(2)** as illustrated in Fig. 2, in contrast to existing work that relies solely on textual references for evaluation, we introduce additional reference GT images to facilitate direct visual comparison with the generated image. By utilizing the visual cues provided by the reference image, the evaluator is able to more accurately and reliably assess the alignment of the generated image with the given instruction, as shown in Fig. 2 (b). Furthermore, to ensure the diversity and reliability of samples in this benchmark, we design a **multi-scenario data synthesis pipeline**. Specifically, as shown in Fig. 4, **(a)** For *real-world* scenarios, we first handcraft a few reference text prompts, including the original image description, the editing instruction, and the textual reference of edited effect. These prompts are then scaled up using

the VLM. Finally, all resulting textual descriptions are directly used to generate pairs of original and edited image. **(b)** For *game-world* scenarios, we first design diverse game problems, and then use Python programs to generate image pairs, instructions, and textual references of edited effects, ensuring both logical and visual correctness in these rule-intensive scenarios [19, 31]. Ultimately, all data samples in UniREditBench undergo VLM-based filtering and human inspection to ensure their reliability and accuracy.

Based on our data synthesis pipeline, we also propose **UniREdit-Data-100K**, a comprehensive reasoning-based image editing dataset with high-quality chain-of-thought (CoT) reasoning annotations, consisting of detailed, step-by-step reasoning traces generated using VLM, as shown in Fig. 4. To validate its reliability and effectiveness, we fine-tune the Bagel [6] on this dataset, resulting in **UniREdit-Bagel**. Experimental results demonstrate that the fine-tuned model achieves substantial improvements on both UniREditBench and other out-of-distribution benchmarks [46,56]. Additionally, through comprehensive evaluation of both open- and closed-source editing models on our UniREditBench, we reveal their strengths and weaknesses across diverse reasoning-based scenarios.

Contributions: (1) We introduce UniREditBench, a unified benchmark for reasoning-based image editing that covers both real-world and game-world scenarios across 8 primary dimensions and 18 sub-dimensions, augmented with reference GT images to enable robust evaluation; (2) We design a multi-scenario data synthesis pipeline and develop UniREdit-Data-100K, a large-scale synthetic reasoning-based image editing dataset that includes high-quality CoT reasoning annotations. By fine-tuning the Bagel on this dataset, we develop UniREdit-Bagel and achieve substantial improvements, validating the effectiveness and reliability of our dataset; (3) Through comprehensive benchmarking of both open- and closed-source models, we systematically identify their strengths and weaknesses across diverse reasoning-based editing scenarios, offering valuable insights for advancing future models.

2 Related Work

Instruction-based Image Editing. Instruction-based image editing models aim to bridge semantic understanding of instructions with accurate visual manipulation. Traditional methods perform editing by altering the diffusion trajectory without requiring additional training, including partial denoising from intermediate SDE steps [24], cross-attention control [12,43], mask-guided blending [1,41,42], CLIP- or diffusion-guided manipulation [7,17], and latent inversion for fidelity preservation [16, 40]. Besides, several studies employ visual-language models (VLMs) to provide prompts, spatial priors, or synthetic supervision to guide a generative editing model [3,9,11,27,28,54,55]. Recent unified frameworks aim to use a single model for both image understanding and editing in a complementary direction [34, 45, 49]. For instance, Bagel [6] features a *think* mode that produces reasoning text prior to editing to enhance instruction fidelity and consistency. While effective, current methods still face challenges with complex

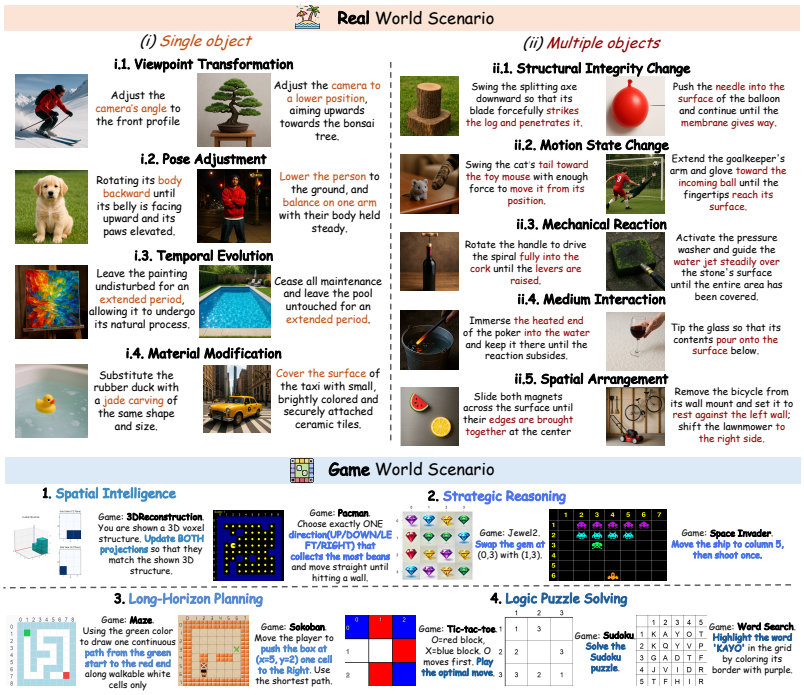

Fig. 3: Qualitative cases of evaluation dimensions in UniREditBench. We present qualitative examples for each dimension across both real-world and game-world scenarios.

reasoning-based editing, underscoring the need for comprehensive benchmarks to assess their performance across various reasoning scenarios.

Reasoning-based Benchmarks for Image Generation and Editing. In T2I generation, several benchmarks [20, 25, 29, 35] have been developed to assess the reasoning capabilities of models in generating images. For example, WISE [25] focuses on assessing models' world knowledge, such as cultural and physical understanding, while UniGenBench++ [35] unifies semantic generation evaluation, covering 10 primary dimensions and 27 sub-dimensions, such as logic reasoning, relational understanding, supporting multilingual and varying-length assessments. In image editing evaluation, recent reasoning-based benchmarks [15, 33, 46, 56] like RISEBench [56] aim to examine temporal, spatial, and logical editing capabilities of editing models. Besides, KRISBench [46] introduces a knowledge-grounded taxonomy covering factual, conceptual, and procedural types. However, these benchmarks primarily focus on single-object knowledge and attribute reasoning. We posit that extending evaluation to multi-object interactions and scenarios governed by human-defined rules is a crucial next step. As for image quality evaluation [18, 50], recent works like UnifiedReward [36, 39] adopt the "VLM-as-a-judge" paradigm, leveraging the powerful capabilities of VLMs to score and provide explanatory judgments. In image editing tasks, evaluation is more challenging because the evaluator needs to assess not only image quality but also understand complex editing instructions and final edited effects.

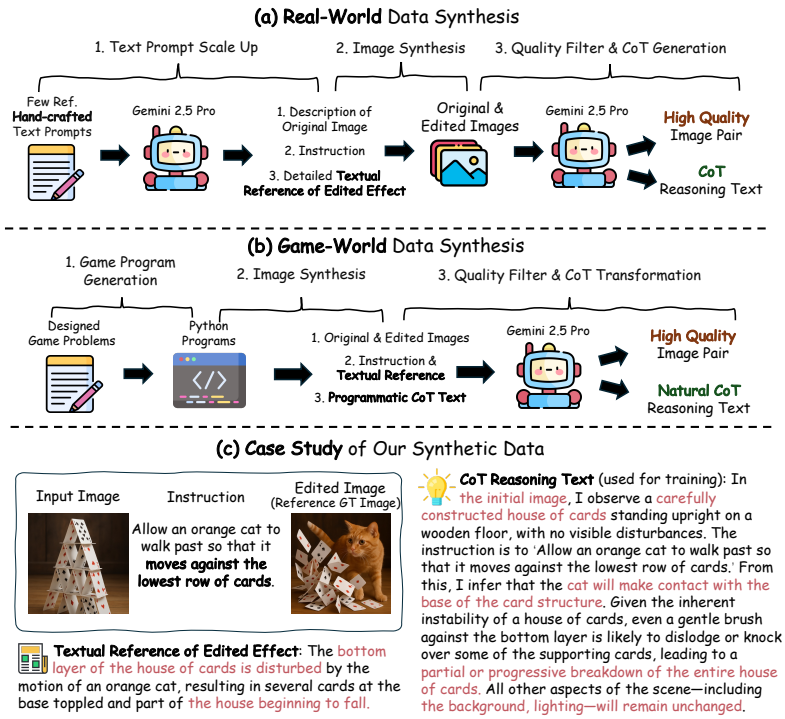

Fig. 4: Multi-scenario data synthesis pipeline. (a) Real-world data synthesis pipeline; (b) Game-world data synthesis pipeline; and (c) Case study of our synthesized data.

Most studies like RISEBench and KRISBench utilize the property model [14] to rate instruction following, temporal consistency, and image quality. Despite effectiveness, their evaluation relies mainly on textual references, which may lead to systematic misjudgments in complex reasoning tasks.

To this end, this work proposes UniREditBench, a unified reasoning-based image editing benchmark that spans a broad range of evaluation dimensions across real-world and game-world scenarios with multimodal dual-reference evaluation for more reliable and accurate assessments.

3 UniREditBench

3.1 Overview

With the rapid advancements in image editing models, existing benchmarks are gradually becoming less adequate to fully capture their comprehensive capabilities, particularly their reasoning-based editing abilities. Specifically, current benchmarks encounter two major challenges: **(1)** their evaluation primarily focuses on simple single-object attribute edits in real-world scenarios, neglecting complex multi-object interactions, as well as logical or strategic reasoning in

game-world scenarios, where explicit human-defined rules govern the outcomes (Tab. 1); **(2)** their evaluation predominantly relies on CLIP-based metrics or VLM-based evaluators with text-only references, which may offer insufficient or inaccurate assessments, particularly in complex reasoning-intensive editing scenarios (Fig. 2).

To this end, this work proposes **UniREditBench**, a unified reasoning-based image editing benchmark that covers a broad spectrum of reasoning dimensions in different scenarios. Compared with previous studies, this benchmark exhibits several key superiorities:

- **Broader scenario and reasoning dimension coverage.** It contains 2,700 high-quality samples organized into 8 primary reasoning dimensions and 18 sub-dimensions, spanning both real-world and game-world image editing tasks (Sec. 3.2).
- **Reliable dual-reference evaluation.** For each sample assessment, we introduce both the textual reference and ground-truth (GT) image reference. This multimodal reference enables VLM evaluators to perform direct and fine-grained comparisons at both the textual and visual levels with the generated images, leading to more reliable evaluation (Sec. 3.3).
- **Scalable multi-scenario data synthesis.** We propose an automatic data synthesis pipeline with distinct generation strategies tailored for real-world and game-world scenarios (Sec. 3.4).

3.2 Evaluation Dimensions

In real-life applications, image editing scenarios often involve diverse requirements spanning both real and game-world contexts, where complex contextual understanding and implicit reasoning capabilities are crucial for accurate image edits. Therefore, UniREditBench organizes reasoning-based image editing tasks into a scenario-to-category hierarchy framework. As illustrated in Fig. 1, it covers both real-world and game-world scenarios across 8 primary dimensions and 18 sub-categories, each representing a unique visual reasoning challenge with 150 human-inspected examples. We will elaborate on each dimension in the following.

Real-World Scenarios Real-world scenarios involve editing tasks that reflect the perceptual and interaction dynamics observed in natural environments. These tasks involve transformations of individual objects or complex interactions among multiple objects. To handle such tasks, models must capture the semantic, physical, and temporal characteristics of objects, as well as their relationships.

- **Single-Object Transformation** targets variations intrinsic to an individual object, including viewpoint and attribute changes that do not disrupt spatial relationships within the scene:
  - **Viewpoint Transformation:** Altering the perspective or viewing angle to exhibit alternative views of the same object (e.g., side, top-down, close-up).

  - **Pose Adjustment:** Modifying the articulation or positioning of an object's parts, such as limb configurations or postural shifts.
  - **Temporal Evolution:** Simulating natural progressions over time like aging, decay, or seasonal changes impacting the object's appearance.
  - **Material Modification:** Changing inherent surface or material properties (e.g., color, texture) while preserving geometry and location.
- **Multi-Object Interaction** involves mutual influences and state changes arising from the physical or spatial interactions among multiple objects:
  - **Structural Integrity Change:** Physical deformations resulting from forces or collisions.
  - **Motion State Change:** Dynamics induced by contact or force transmission leading to altered movement or posture.
  - **Mechanical Reaction:** State transitions caused by device operation or functional interactions.
  - **Medium Interaction:** Changes mediated by substances or environmental factors that affect appearance or state.
  - **Spatial Arrangement:** Reorganization or repositioning of multiple objects within the scene.

Game-World Scenarios Game-world scenarios consist of tasks within synthetic environments governed by human-defined rules, evaluating logical, strategic, spatial, and long-horizon reasoning capabilities. These tasks require models to plan, deduce, and act in accordance with the explicit rules that govern the environment.

- **Long-Horizon Planning** requires multi-step sequential reasoning to accomplish distant goals, exemplified by navigation or puzzle games such as Maze-solving and Sokoban.
- **Logical Puzzle Solving** involves constraint satisfaction and symbolic inference to produce valid solutions under formal rule sets, including Sudoku, Tic-tac-toe, and Word Search.
- **Strategic Reasoning** requires resource management, adversarial planning over time, captured by games like Pacman, Jewel2, and Space Invader.
- **Spatial Intelligence** focuses on geometric and topological reasoning within 3D environments, such as reconstructing spatial layouts in gaming contexts.

Representative examples are provided in Figs. 1 and 3 to illustrate the scope and diversity of evaluation dimensions, and highlight the complexity and variety of tasks in our benchmark.

3.3 Dual-Reference Evaluation

Evaluating reasoning-based image editing is intrinsically challenging due to the need for the evaluator to accurately understand the implicit reasoning intentions within the instruction. To achieve reliable and comprehensive assessments, we

introduce a VLM-based multi-dimensional scoring schema, leveraging both textual and visual evaluation references. Specifically, for each sample, this pipeline evaluates three core dimensions:

Instruction Following measures how accurately the generated image reflects the input instruction, focusing on whether the explicit effect of the edit is properly manifested. Here, the VLM compares the output image G against both the textual reference of edited effect R_t and the corresponding reference GT image R_i to verify compliance:

$$S_{\text{IF}} = \text{VLM}(O, I, G, R_i, R_t)$$

where O represents the original image, I denotes the editing instruction.

Visual Consistency assesses the preservation of image regions and attributes unrelated to the edit instruction, ensuring that changes are localized and do not inadvertently alter irrelevant scene elements. This criterion favors models capable of accurate, fine-grained editing rather than wholesale regeneration:

$$S_{\text{VC}} = \text{VLM}(O, I, G)$$

Visual Quality evaluates the realism of the generated output, checking for artifacts, distortions, and physical or logical implausibility in the final image:

$$S_{\text{VQ}} = \text{VLM}(G)$$

We choose GPT-4.1 [14] as the VLM evaluator. Each score S ranges from 1 to 5, following prior detailed scoring guidelines [46, 56]. Finally, the overall evaluation score aggregates these via weighted sum:

$$S_{\text{Overall}} = \alpha_1 S_{\text{IF}} + \alpha_2 S_{\text{VC}} + \alpha_3 S_{\text{VQ}},$$

where $\alpha_1 = 0.5$, $\alpha_2 = 0.3$, and $\alpha_3 = 0.2$. This setting prioritizes instruction following to emphasize the importance of adhering to the instruction, and also incorporates visual consistency and quality, ensuring that areas unrelated to the instruction are preserved and that the overall image quality is maintained. Details of the weight selection strategy are provided in the Appendix.

3.4 Multi-Scenario Data Synthesis

Given the distinct characteristics of real- and game-world contexts, we develop a specialized data generation process for each scenario, as illustrated in Fig. 4. Detailed elaboration of each data synthesis process is provided below.

For **Real-World Scenario**, we employ a "text-then-image" data generation strategy. Specifically, **(1)** this process begins with hand-crafted textual triples that describe the original image, the editing instruction, and the textual reference of edited effect (a reasoning-based narrative of the anticipated outcome). Next, we use the powerful VLM [5] to expand this initial set into a large corpus of text triples. Subsequently, **(2)** these curated textual triples are input to

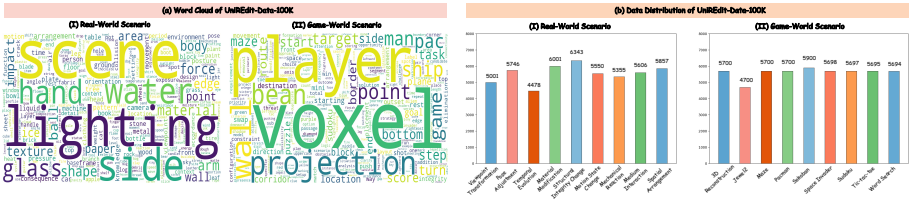

Fig. 5: Statistic visualization. We visualize (a) word clouds and (b) data distribution of our UniREdit-Data-100K.

GPT-4o [14] to synthesize the original and edited images in alignment with the described textual reference of edited effect. **(3)** Finally, in the quality filtering stage, VLM [5] is used to assess the generated images based on visual fidelity, instruction alignment, and potential hallucination risks. Additionally, it generates reasoning chain-of-thought (CoT) text for each qualified instance, ensuring the production of high-quality, reasoning-based image editing training data.

In **Game-World Scenario**, game states are inherently well-suited to be represented as structured reasoning-based editing data, where instructions can naturally be solved using Python code. Inspired by Game-RL [31], **(1)** we first design a diverse collection of game-based problems and develop corresponding Python programs tailored to each category. **(2)** Then, these programs automatically generate paired original and edited images, along with instructions, textual reference effects, and programmatic CoT reasoning traces. **(3)** To bridge the gap between programmatic and natural language CoT reasoning formats, we use VLMs to convert these reasoning traces into explanations that align with human inference patterns. Finally, quality filtering is applied to ensure the integrity and reliability of the data.

Overall, this multi-scenario data synthesis pipeline generates **UniREdit-Bench**, a unified reasoning-based image editing benchmark, and **UniREdit-Data-100K**, a large-scale dataset with high-quality CoT annotations. We detail this dataset in the next section.

4 UniREdit-Data-100K

To enhance the capability of current generative models on reasoning-driven image editing, we propose UniREdit-Data-100K, which contains 100,421 samples spanning 8 reasoning dimensions and 18 categories defined in Sec. 3.2.

4.1 Statistical Analysis

UniREdit-Data-100K is designed with an emphasis on balance and diversity, ensuring that each reasoning category contains over 4,000 instances to effectively support model training across a wide range of editing tasks. It is divided into two primary scenarios: (i) Real-World Scenario, which captures natural object attributes and complex multi-object interactions, and (ii) Game-World Scenario, presenting structured, rule-based editing challenges, such as puzzles and strategic planning games. We visualize the word cloud for both real-world and game-world

Table 2: In-domain quantitative comparisons on UniREditBench. *GPT-4.1* is used as the evaluator. Best scores are in **bold**.

Model	Real World Scenario					Game World Scenario					
	Attribute Modification	Structure Transform	Physical Interaction	Property Response	Avg.	Spatial Intelligence	Strategic Reason	Long-Horizon Plan	Logic Puzzle Solving	Avg.	Overall
Closed-source Models											
FLUX-Kontext-Pro	47.35	47.16	44.37	41.44	45.00	49.12	48.58	51.16	40.49	46.52	45.77
Seedream4.0	69.54	73.13	67.88	62.40	68.20	39.27	43.54	43.79	51.91	45.91	57.03
Wan2.5	74.66	69.74	63.51	62.87	67.23	63.73	47.13	55.00	54.93	52.67	61.36
Nano Banana	77.10	78.88	71.86	74.70	75.22	66.74	**56.11**	56.83	**64.91**	60.39	68.26
GPT-4o	**82.67**	**84.75**	**79.81**	**77.40**	**81.01**	**77.73**	51.44	**67.61**	63.79	**62.07**	**73.39**
Open-source Models											
MagicBrush	43.97	46.09	42.93	46.64	44.69	63.58	33.59	30.72	35.31	36.85	40.77
Omnigen2	55.47	58.27	51.44	50.70	53.69	70.28	27.50	36.25	24.31	33.14	43.41
Bagel	61.85	59.55	52.56	54.47	56.60	54.25	35.21	37.69	39.83	39.42	48.01
Lumina-DiMOO	52.84	52.31	50.31	50.83	51.44	61.23	36.96	39.57	53.09	45.61	48.54
Step1X-Edit	59.69	56.36	53.85	54.84	55.93	65.68	34.89	47.01	43.89	44.00	50.15
Bagel-Think	61.45	59.51	52.65	55.68	56.80	66.29	43.23	43.00	41.30	45.10	50.96
DreamOmni2	63.52	58.35	52.68	54.02	56.64	72.42	42.17	48.80	48.09	48.98	52.81
UniWorld-V2	72.96	69.37	63.65	61.69	66.55	49.27	40.00	53.41	37.53	43.19	54.87
Qwen-Image-Edit	75.68	73.03	70.59	64.67	70.95	56.73	36.63	48.80	37.68	41.92	56.52
UniREdit-Bagel (Ours)	**76.73**	**77.80**	**76.57**	**71.44**	**75.74**	**84.90**	**72.83**	**84.88**	**83.72**	**80.48**	**78.15**

subsets in Fig. 5 (a) and the detailed distribution of samples across different categories in Fig. 5 (b). These visualizations highlight the extensive vocabulary as well as the broad coverage across various categories of our dataset.

4.2 UniREdit-Bagel

To further validate the effectiveness of our dataset, we use it to fine-tune Bagel [6], a unified understanding and generative model. Specifically, each training sample consists of the input image O, an editing instruction I, a stepwise CoT text C that grounds the edit effects step by step, and the target edited image G. During training, the original image and instruction are first input into the model, which then generates a textual reasoning trace and synthesizes the edited image. We supervise both the textual reasoning trace and the visual edit. Formally, for reasoning text supervision, we minimize the negative log-likelihood:

$$\mathcal{L}_{\text{text}} = -\sum_{t}^{T} \log p_\theta\big(y_t \mid y_{<t}, O, I\big).$$

For image generation, we supervise the latent flow-matching loss [22] between the VAE latents of O and G, conditioned on (O, I, C):

$$\mathcal{L}_{\text{img}} = \mathbb{E}_{t\sim\mathcal{U}(0,1)} \left\| u_\theta(z_t, t\,;\, O, I, C) - u^\star(z_t, t) \right\|_2^2,$$

where u_θ is the learned time-conditioned velocity field on the latent path from z_O to z_G, and $u^\star$ is the target velocity. Finally, the overall objective is

$$\mathcal{L} = \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{img}}\mathcal{L}_{\text{img}}.$$

Under the influence of $\mathcal{L}_{\text{text}}$, the model enhances its reasoning ability through explicit CoT learning, which effectively guides the accurate image editing, while $\mathcal{L}_{\text{img}}$ improves both the correctness and fidelity of the edited image.

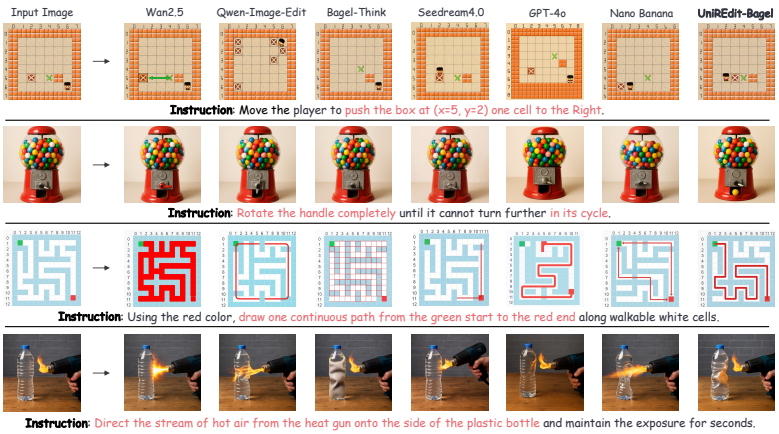

Fig. 6: Qualitative editing result comparison. Our UniREdit-Bagel demonstrates significant superiority in both instruction following and visual quality compared with state-of-the-art closed-source and open-source models.

5 Experiment

5.1 Implementation Details

Baselines. We benchmark *closed-source* models including: GPT-4o [14], Nano Banana [8], Gemini-2.0 [5], Seedream4.0 [26], Wan2.5 [32], and FLUX-Kontext-Pro [2], as well as *open-source* models including: Bagel [6], Qwen-Image-Edit [44], Step1X-Edit [23], FLUX.1-Kontext-dev [2], Emu2 [30], Omnigen2 [45], Omnigen [47], HiDream-Edit [4], MagicBrush [54], and AnyEdit [52].
Training and Evaluation. We train all Bagel [6] components, except the VAE, for 5,000 iterations on UniREdit-Data-100K using the Adam optimizer and a cosine learning-rate schedule with 500 warm-up steps, a peak learning rate of 2×10^{-5}, and a minimum learning rate of 10^{-6}. The loss weights are set to $\lambda_{\text{text}} = 2$ and $\lambda_{\text{img}} = 1$. During inference, we use the official inference settings provided by Bagel. To ensure fair comparisons with other baselines, we adopt the original inference configurations of these models.

5.2 Benchmarking Results on UniREditBench

As shown in Tab. 2, among **closed-source** models, GPT-4o achieves the highest overall performance across all scenarios, with Nano Banana performing comparably. Wan2.5 delivers balanced results on real-world tasks but lags on game scenarios that require strategic reasoning. Besides, Seedream4.0 is competitive on *structure transform* yet encounters challenges in game scenarios. Among **open-source** baselines, Qwen-Image-Edit performs strongly on real-world tasks such as *attribute modification* and *structure transform*. However, most models remain comparatively weak on game scenarios like *Strategic Reasoning*. **Overall**, compared with open-source methods, closed-source models, particularly GPT-4o,

Table 3: Out-of-distribution quantitative performance comparison on RISEBench [56]. *GPT-4.1* is used as the evaluator. Best scores are in **bold**.

Models	Temporal	Causal	Spatial	Logical	**Overall**
Closed-source Models					
Gemini-2.0-Flash-pre	10.6%	13.3%	11%	2.3%	9.4%
Seedream4.0	12.9%	12.2%	11.0%	7.1%	10.8%
Gemini-2.0-Flash-exp	8.2%	15.5%	23.0%	4.7%	13.3%
GPT-4o	**34.1%**	32.2%	**37.0%**	10.6%	28.9%
Nano Banana	25.9%	**47.8%**	**37.0%**	**18.8%**	**32.8%**
Open-source Models					
HiDream-Edit	0.0%	0.0%	0.0%	0.0%	0.0%
OmniGen	1.2%	1.0%	0.0%	1.2%	0.8%
Step1X-Edit	0.0%	2.2%	2%	3.5%	1.9%
Bagel	3.5%	4.4%	9.0%	5.9%	5.8%
FLUX.1-Kontext-Dev	2.3%	5.5%	13.0%	1.2%	5.8%
Qwen-Image-Edit	4.7%	10.0%	17.0%	2.4%	8.9%
Bagel-Think	4.7%	15.5%	14.0%	1.2%	9.2%
UniREdit-Bagel (Ours)	**22.4%**	**18.9%**	**21.0%**	**10.6%**	**18.3%**

maintain a clear advantage. While some open-source models are competitive on specific real-world tasks, they generally struggle with complex reasoning in game scenarios. Notably, only GPT-4o and Nano Banana achieve an average score greater than 60 on game scenarios, underscoring that this setting remains highly challenging and serves as a useful test for current models.

5.3 Comparison Results of UniREdit-Bagel

Quantitative. UniREdit-Bagel achieves the best overall performance among all closed- and open-source models on UniREditBench, surpassing the second-place GPT-4o by a substantial margin. The largest gains occur in game-world scenarios (+17.08), indicating exceptional capability of understanding and processing complex reasoning image editing tasks. In out-of-distribution performance comparison, UniREdit-Bagel achieves the strongest open-source results across all categories on RISEBench, shown in Tab. 3, improving upon the Bagel-Think baseline by 9.1 points and surpassing the closed-source Gemini-2.0-Flash-exp by 5.0 points. It also remains competitive with top closed-source models like Nano Banana, narrowing the gap between open- and closed-source models.

Qualitative. The qualitative results presented in Fig. 6 highlight the strengths of our UniREdit-Bagel across various tasks. Specifically, in Row 4, most models fail to reliably reproduce the physical effect induced by the heat. Although several baselines (e.g., Nano Banana and Qwen-Image-Edit) capture the heat-induced warping of a plastic bottle under sustained heat gun exposure, they fail to preserve the heat trace. Notably, UniREdit-Bagel not only renders the deformation accurately but also preserves the heat trace, offering superior visual consistency. Besides, in the Sokoban and Maze game settings (Rows 1 and 3), most baselines (e.g. Seedream4.0, Nano Banana, and Wan2.5) struggle with

Fig. 7: Mean Absolute Error (MAE) between human and VLM scores under different reference settings.

Table 4: MAE between human and VLM scores across Instruction Following (IF), Visual Consistency (VC), and Visual Quality (VQ).

Model	IF	VC	VQ	Avg.
Qwen-Image-Edit	0.73	0.62	0.47	0.61
Nano Banana	0.85	0.42	0.66	0.64
GPT-4o	0.42	0.68	0.22	0.44
UniREdit-Bagel	0.66	0.54	0.51	0.57
Avg.	0.67	0.57	0.47	

box pushing and maze completion. In contrast, UniREdit-Bagel excels in both fulfilling the instruction and maintaining the coherence of unrelated content.

5.4 Discussion and Analysis

VLM-as-Judge Assessment To assess the reliability of using VLMs as evaluators, we conducted a user study with six human experts on 200 randomly selected samples across four models (Qwen-Image-Edit, Nano Banana, GPT-4o, UniREdit-Bagel). We reported the Mean Absolute Error (MAE) between experts' scores and GPT-4.1 scores in Tab. 4. It is demonstrated that the average MAE of each score and model falls below 1, indicating that scores by GPT-4.1 are closely aligned with those predicted by human experts.

Ablation on Dual-Reference Evaluation Protocol We conduct an ablation study under different reference settings. Compared with settings that omit the image or the text reference (Fig. 7), our dual-reference protocol consistently achieves the lowest MAE between human and VLM scores across all expert score levels, validating the accuracy of our evaluation pipeline.

More Analysis We provide additional analyses in the Appendix, including **(1)** an ablation study of the score weighting parameters, showing that $(\alpha_1, \alpha_2, \alpha_3) = (0.50, 0.30, 0.20)$ produces scores most consistent with human experts and is adopted as the weighting scheme; **(2)** quantitative results on KRISBench, showing that training on UniREdit-Data-100K improves out-of-distribution generalization on other benchmarks; **(3)** consistent scoring trends between Qwen3VL-32B-Instruct and GPT-4.1; **(4)** an ablation study on game-world data, demonstrating transferability to real-world scenarios; **(5)** investigations into the role of chain-of-thought (CoT) reasoning; **(6)** data-scaling analysis, confirming consistent gains from 20K to 100K samples; and **(7)** additional qualitative comparisons and model-specific failure modes.

6 Conclusion

This paper presents **UniREditBench**, a unified reasoning-based benchmark for image editing with broad dimension coverage and a robust dual-reference evalu-

ation protocol. We further introduce a multi-scenario synthesis pipeline and release **UniREdit-Data-100K**, a large-scale dataset with high-quality CoT annotations. To demonstrate its effectiveness, we fine-tune Bagel on this dataset and obtain **UniREdit-Bagel**, which achieves substantial quantitative and qualitative gains. Comprehensive benchmarking of open-source and closed-source image editing models reveals their strengths and weaknesses across various aspects.

Acknowledgments

This work is supported by Shanghai Innovation Institute.

References

1. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: CVPR. pp. 18208–18218 (2022)
2. Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742
3. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR. pp. 18392–18402 (2023)
4. Cai, Q., Chen, J., Chen, Y., Li, Y., Long, F., Pan, Y., Qiu, Z., Zhang, Y., Gao, F., Xu, P., et al.: Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. arXiv preprint arXiv:2505.22705 (2025)
5. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
6. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
7. Gong, C., Li, D., Pan, Y., Chen, J., Yao, T., Mei, T.: Freeinpaint: Tuning-free prompt alignment and visual rationality enhancement in image inpainting. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 4239–4247 (2026)
8. Google: Introducing gemini 2.5 flash image, our state-of-the-art image model. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/ (2025)
9. Han, F., Jiao, Y., Chen, S., Xu, J., Chen, J., Jiang, Y.G.: Controlthinker: Unveiling latent semantics for controllable image generation through visual reasoning. arXiv preprint arXiv:2506.03596 (2025)
10. He, Q., Chen, X., Wang, C., Pan, Y., Hu, X., Gan, Z., Wang, Y., Wang, C., Li, X., Zhang, J.: Reasoning to edit: Hypothetical instruction-based image editing with visual reasoning. arXiv preprint arXiv:2507.01908 (2025)
11. He, R., Ma, K., Huang, L., Huang, S., Gao, J., Wei, X., Dai, J., Han, J., Liu, S.: Freeedit: Mask-free reference-based image editing with multi-modal instruction. arXiv preprint arXiv:2409.18071 (2024)

12. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
13. Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., Zhou, J., Dong, C., Huang, R., Zhang, R., et al.: Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In: CVPR. pp. 8362–8371 (2024)
14. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
15. Jia, B., Huang, W., Tang, Y., Qiao, J., Liao, J., Cao, S., Zhao, F., Feng, Z., Gu, Z., Yin, Z., et al.: Compbench: Benchmarking complex instruction-guided image editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1112–1122 (2026)
16. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: CVPR. pp. 6007–6017 (2023)
17. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: CVPR. pp. 2426–2435 (2022)
18. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. NeurIPS **36**, 36652–36663 (2023)
19. Li, C., Chen, Q., Li, Z., Tao, F., Zhang, Y.: Vcbench: A controllable benchmark for symbolic and abstract challenges in video cognition. arXiv preprint arXiv:2411.09105 (2024)
20. Li, O., Wang, Y., Hu, X., Huang, H., Chen, R., Ou, J., Tao, X., Wan, P., Qi, X., Feng, F.: Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? arXiv preprint arXiv:2509.03516 (2025)
21. Li, Z., Liu, Z., Zhang, Q., Lin, B., Yuan, S., Yan, Z., Ye, Y., Yu, W., Niu, Y., Yuan, L.: Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. arXiv preprint arXiv:2510.16888 (2025)
22. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
23. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
24. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
25. Niu, Y., Ning, M., Zheng, M., Jin, W., Lin, B., Jin, P., Liao, J., Feng, C., Ning, K., Zhu, B., et al.: Wise: A world knowledge-informed semantic evaluation for text-to-image generation. arXiv preprint arXiv:2503.07265 (2025)
26. Seed, B.: Seedream 4.0. `https://seed.bytedance.com/en/seedream4_0` (2025)
27. Shen, F., Jiang, X., He, X., Ye, H., Wang, C., Du, X., Li, Z., Tang, J.: Imagdressing-v1: Customizable virtual dressing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6795–6804 (2025)
28. Shen, F., Tang, J.: Imagpose: A unified conditional framework for pose-guided person generation. Advances in neural information processing systems **37**, 6246–6266 (2024)
29. Sun, K., Fang, R., Duan, C., Liu, X., Liu, X.: T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation. arXiv preprint arXiv:2508.17472 (2025)

30. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners. In: CVPR. pp. 14398–14409 (2024)
31. Tong, J., Tang, J., Li, H., Mou, Y., Zhang, M., Zhao, J., Wen, Y., Song, F., Zhan, J., Lu, Y., et al.: Code2logic: Game-code-driven data synthesis for enhancing vlms general reasoning. arXiv preprint arXiv:2505.13886 (2025)
32. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
33. Wang, C., Zhou, Y., Wang, Q., Wang, Z., Zhang, K.: Complexbench-edit: Benchmarking complex instruction-driven image editing via compositional dependencies. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 13391–13397 (2025)
34. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
35. Wang, Y., Li, Z., Zang, Y., Bu, J., Zhou, Y., Xin, Y., He, J., Wang, C., Lu, Q., Jin, C., et al.: Unigenbench++: A unified semantic evaluation benchmark for text-to-image generation. arXiv preprint arXiv:2510.18701 (2025)
36. Wang, Y., Li, Z., Zang, Y., Wang, C., Lu, Q., Jin, C., Wang, J.: Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. arXiv preprint arXiv:2505.03318 (2025)
37. Wang, Y., Li, Z., Zang, Y., Zhou, Y., Bu, J., Wang, C., Lu, Q., Jin, C., Wang, J.: Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. arXiv preprint arXiv:2508.20751 (2025)
38. Wang, Y., Tan, Z., Wang, J., Yang, X., Jin, C., Li, H.: Lift: Leveraging human feedback for text-to-video model alignment. arXiv preprint arXiv:2412.04814 (2024)
39. Wang, Y., Zang, Y., Li, H., Jin, C., Wang, J.: Unified reward model for multimodal understanding and generation. arXiv preprint arXiv:2503.05236 (2025)
40. Wang, Y., Zhang, W., Jin, C.: Magicface: Training-free universal-style human image customized synthesis. arXiv preprint arXiv:2408.07433 (2024)
41. Wang, Y., Zhang, W., Xu, H., Jin, C.: Dreamtext: High fidelity scene text synthesis. In: CVPR. pp. 28555–28563 (2025)
42. Wang, Y., Zhang, W., Zheng, J., Jin, C.: Primecomposer: Faster progressively combined diffusion for image composition with attention steering. In: ACM MM. pp. 10824–10832 (2024)
43. Wang, Y., Zhou, C., Xu, H.: Enhancing object coherence in layout-to-image synthesis. arXiv preprint arXiv:2311.10522 (2023)
44. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
45. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
46. Wu, Y., Li, Z., Hu, X., Ye, X., Zeng, X., Yu, G., Zhu, W., Schiele, B., Yang, M.H., Yang, X.: Kris-bench: Benchmarking next-level intelligent image editing models. arXiv preprint arXiv:2505.16707 (2025)
47. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. In: CVPR. pp. 13294–13304 (2025)
48. Xiao, Y., Song, L., Chen, Y., Luo, Y., Chen, Y., Gan, Y., Huang, W., Li, X., Qi, X., Shan, Y.: Mindomni: Unleashing reasoning generation in vision language models with rgpo. arXiv preprint arXiv:2505.13031 (2025)

49. Xin, Y., Qin, Q., Luo, S., Zhu, K., Yan, J., Tai, Y., Lei, J., Cao, Y., Wang, K., Wang, Y., et al.: Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. arXiv preprint arXiv:2510.06308 (2025)
50. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. NeurIPS **36**, 15903–15935 (2023)
51. Ye, Y., He, X., Li, Z., Lin, B., Yuan, S., Yan, Z., Hou, B., Yuan, L.: Imgedit: A unified image editing dataset and benchmark. arXiv preprint arXiv:2505.20275 (2025)
52. Yu, Q., Chow, W., Yue, Z., Pan, K., Wu, Y., Wan, X., Li, J., Tang, S., Zhang, H., Zhuang, Y.: Anyedit: Mastering unified high-quality image editing for any idea. In: CVPR. pp. 26125–26135 (2025)
53. Zhang, D., He, L., Yan, R., Shen, F., Tang, J.: R-genie: Reasoning-guided generative image editing. arXiv preprint arXiv:2505.17768 (2025)
54. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. NeurIPS **36**, 31428–31449 (2023)
55. Zhang, S., Yang, X., Feng, Y., Qin, C., Chen, C.C., Yu, N., Chen, Z., Wang, H., Savarese, S., Ermon, S., et al.: Hive: Harnessing human feedback for instructional visual editing. In: CVPR. pp. 9026–9036 (2024)
56. Zhao, X., Zhang, P., Tang, K., Zhu, X., Li, H., Chai, W., Zhang, Z., Xia, R., Zhai, G., Yan, J., et al.: Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. arXiv preprint arXiv:2504.02826 (2025)