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**[1], **Chao Gong**[1], **Cheng Jin**[1,2], **Jingjing Chen**[1†], **Jiaqi Wang**[2†]
[1]Fudan University, [2]Shanghai Innovation Institute, [3]Zhejiang University
Project Page: maplebb.github.io/UniREditBench

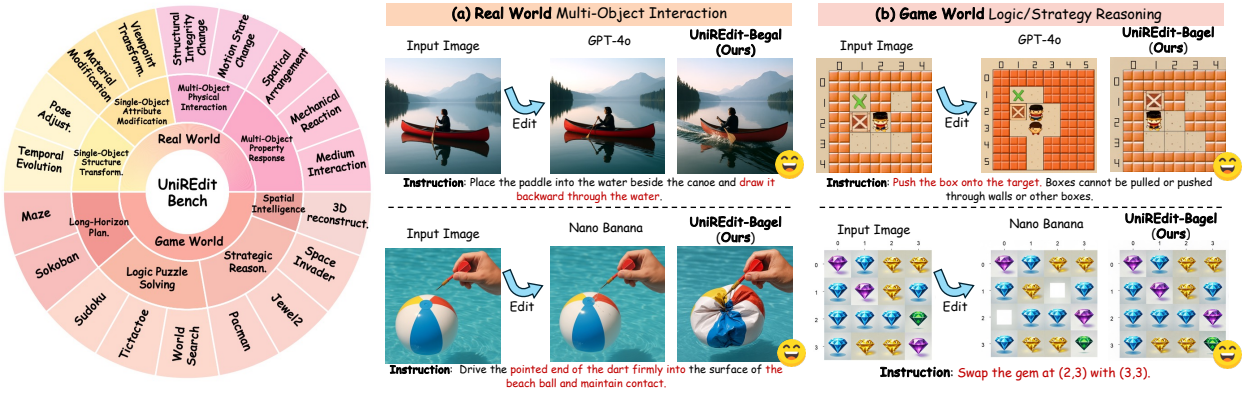

Figure 1. UniREditBench covers both real-world and game-world reasoning scenarios across 8 primary dimensions and 18 sub-dimensions. We provide qualitative editing cases of (a) real-world multi-object interaction, and (b) game-world logical/strategy reasoning.

Abstract

*Recent advances in multi-modal 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.*

1. Introduction

Recent advances in multimodal generative models have led to remarkable improvements in instruction-conditioned image editing. Generative models [3, 31, 36, 38, 42, 45, 48], including Step1X-Edit [19], FLUX-Kontext [2], Bagel [6], Nano-banana [7], and GPT-4o [12], have demonstrated a powerful ability to understand diverse textual instructions and generate semantically consistent image edits. In parallel, reinforcement learning-based training strategies [17, 29, 34, 41] are continuously advancing, further enhancing the capabilities of image editing models. With these rapid

*Equal contribution. †Corresponding author.

Table 1. Reasoning-based image editing benchmark comparison. Our UniREditBench excels in broader scenario and evaluation dimension coverage. "S-Obj" indicates single-object while "M-Obj" indicates multi-object.

Benchmark	Size	Reference Images	Real World Scenario							Game World Scenario			
			Attribute (S-Obj)	Temporal (S-Obj)	Pose (S-Obj)	Spatial (M-Obj)	Motion (M-Obj)	Mechanic (M-Obj)	Medium (M-Obj)	Logical	Long-planing	Strategic	Spatial
SmartEdit [11]	219	219	✔			✔							
RISE [50]	360	70	✔	✔		✔				✔	✔		
KRIS [39]	1,267	50	✔	✔		✔			✔	✔	✔		
UniREditBench	**2,700**	**2,700**	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

developments, the need for a more comprehensive benchmark to evaluate model editing capabilities across different aspects has become increasingly essential. Early benchmarks [19, 44] 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 [8, 47], 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 [50], and proposed a knowledge-grounded taxonomy assessing factual, conceptual, and procedural knowledge types [39].

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 reference to evaluate the generated images [39, 50], 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 solely on text-based references in evaluating complex reasoning-based image editing task often leads to unreliable judgments. As shown in Fig. 2 (a), the text-reference-only evaluator assigns an inflated score even 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 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. 3, **(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 resulted 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 reference of edited effects, ensuring both logical and visual correctness in these rule-intensive scenarios [16, 25]. Ultimately, all data samples in UniREditBench undergo VLM-based filtering and human inspection to ensure their reliability and accuracy.

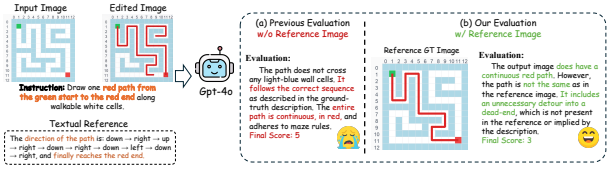

Figure 2. Image editing evaluation comparison. Current text-reference-only evaluation potentially leads to misjudging, while our dual-reference evaluation results in more reliable assessments.

Based on our data synthesis pipeline, we also propose **UniREdit-Data-100K**, a comprehensive reasoning-based

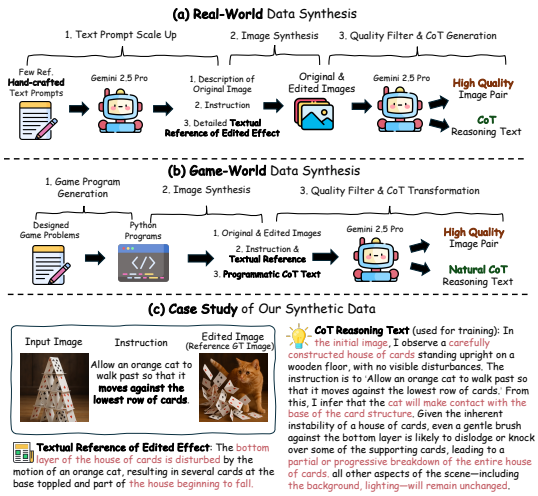

Figure 3. Multi-Scenario data synthesis pipeline. (a) Real-world data synthesis pipeline; (b) Game-world data synthesis pipeline; and (c) Case study of our synthesized data.

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. 3. 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 [39, 50]. 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.

Contribution: (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 [20], cross-attention control [10, 28], mask-guided blending [1, 31, 36], CLIP- or diffusion-guided manipulation [14], and latent inversion for fidelity preservation [13, 30]. Besides, several studies employ visual-language models (VLMs) to provide prompts, spatial priors, or synthetic supervision to guide a generative editing model [3, 9, 48, 49]. Recent unified frameworks aim to use a single model for both image understanding and editing in a complementary direction [27, 38, 42]. 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 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 have been developed to assess the reasoning capabilities of models in generating images. For example, WISE [21] focuses on assessing models' world knowledge, such as cultural and physical understanding, while UniGenBench++ [32] 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 like RISEBench [50] aim to examine temporal, spatial, and logical editing capabilities of editing models. Besides, KRIS-Bench [39] introduces a knowledge-grounded taxonomy covering factual, conceptual, and procedural types. However, these benchmarks primarily focus on single-object knowledge and attribute reasoning. We suppose that extending evaluation to multi-object interactions and scenarios governed by human-defined rules is a crucial next step. As for image quality evaluation [15, 43], recent works like UnifiedReward [33, 35] 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. Most studies like RISEBench and KRISBench utilize the property model [12], to rate instruction following, temporal consistency, and image quality. Despite effectiveness, their evaluation relies solely 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.

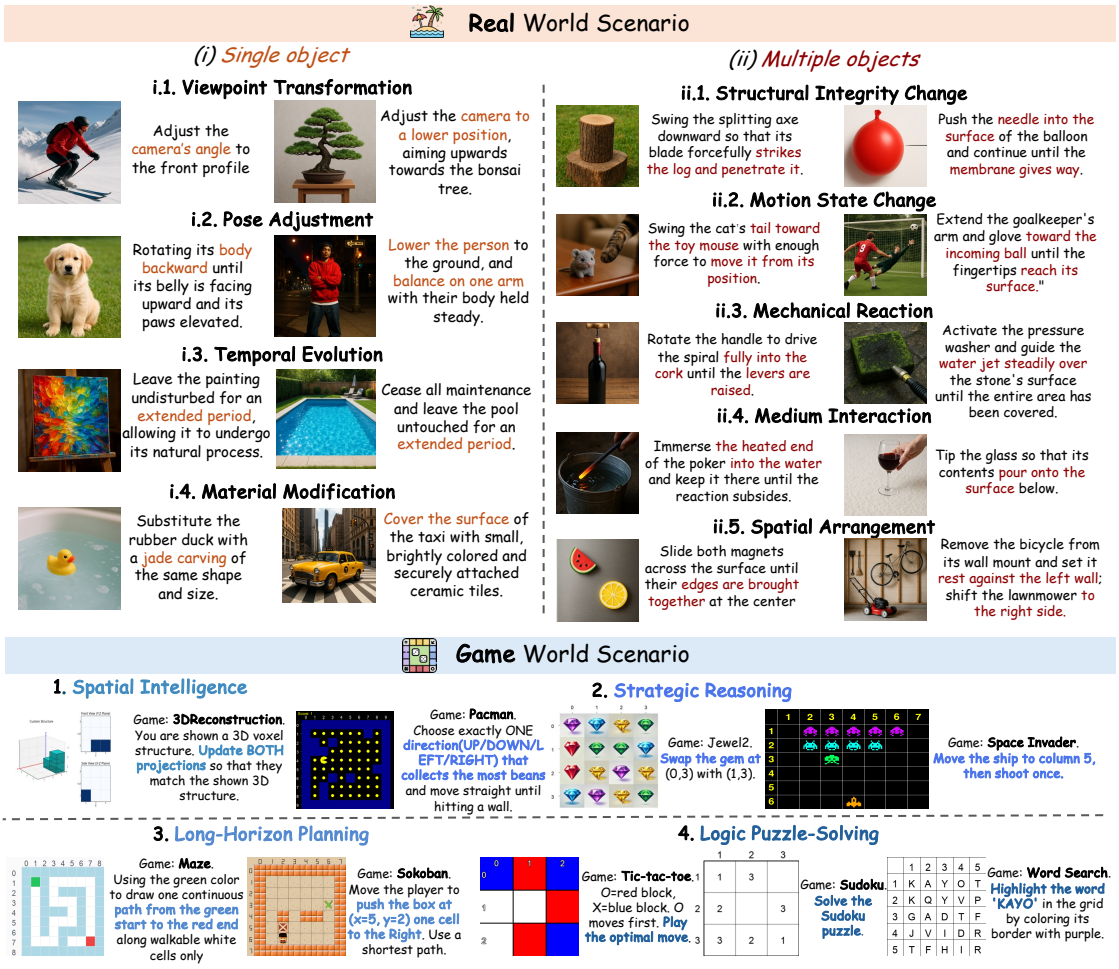

Figure 4. Qualitative cases of evaluation dimensions in UniREditBench. We present qualitative examples for each dimension across both real-world and game-world scenarios.

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 rely 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 design both the textual reference and ground-truth (GT) image reference. This multi-modal reference enables vision-language model (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-world 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.

3.2.1. Real-World Scenarios

Real-world scenarios involve editing tasks that reflect the perceptual and interaction dynamics commonly observed in natural environments. These tasks may 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.

1. **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.
   - **Attribute Modification:** Changing inherent surface or material properties (e.g., scolor, texture) while preserving geometry and location.
2. **Multi-Object Interaction** involves mutual influences and state changes arising from the physical or spatial interactions among multiple objects:
   - **Structural 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 Response:** State transitions caused by device operation or functional interactions.
   - **Medium Effects:** Changes mediated by substances or environmental factors that affect appearance or state.
   - **Spatial Rearrangement:** Reorganization or repositioning of multiple objects within the scene.

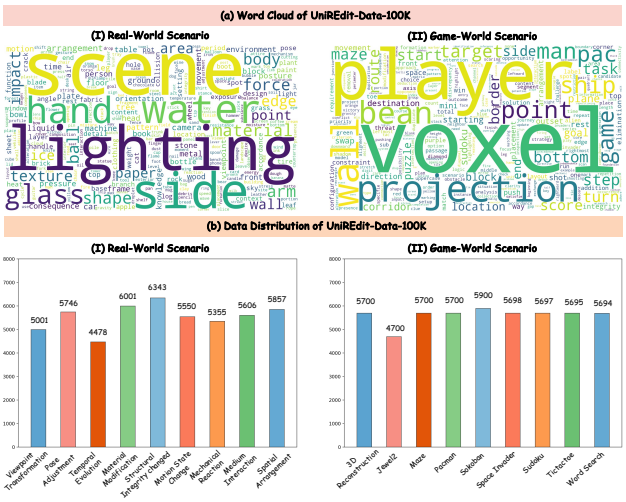

Figure 5. Statistic visualization. We visualize (a) Word clouds and (b) data distribition of our UniREdit-data-100K.

3.2.2. 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, modeled after 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 4 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 prompt. 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 eval-

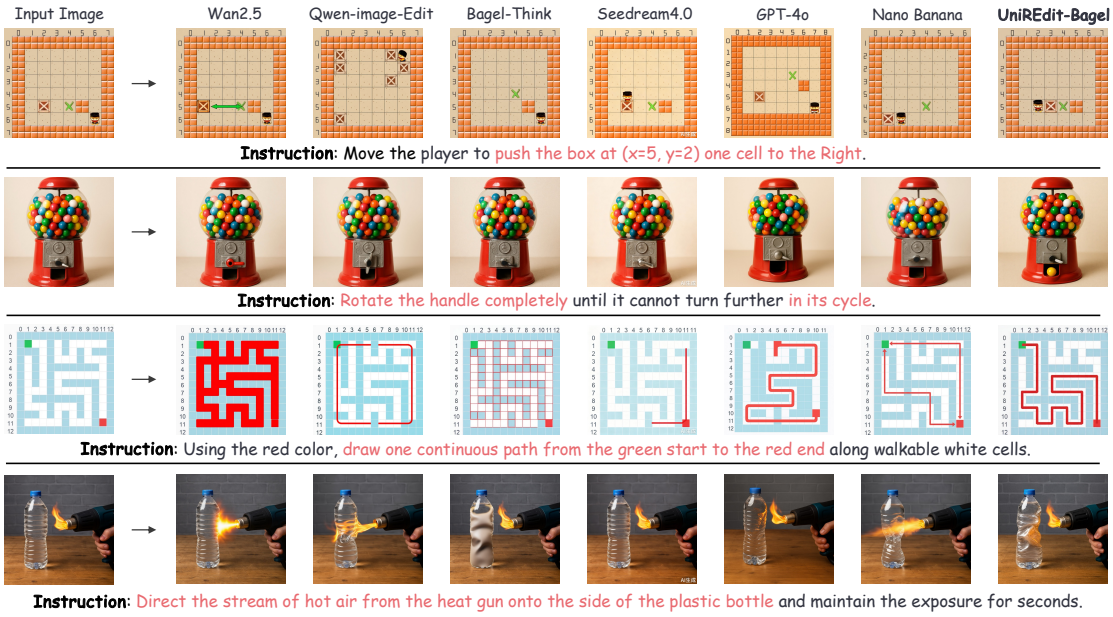

Figure 6. Qualitative editing result comparison. Our UniREdit-Bagel demonstrates significant superiority in both instruction following and visual quality compared with state-of-the-art closed-sourced and open-sourced models.

uates three core dimensions:

- **Instruction Following** measures how accurately the generated image reflects the input instruction, focusing on whether the explicit effect of the edits 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 and perceptual integrity 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 [12] as VLM evaluator. Each score S ranges from 1 to 5, following prior detailed scoring guidelines [39, 50]. 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 accurately adhering to the prompt's intent, and also incorporates visual consistency and quality, ensuring that areas unrelated to the instruction are preserved and that the overall image quality is maintained.

3.4. Multi-Scenario Data Synthesis

Given the distinct characteristics of real- and game-world contexts, we develop specialized data generation process for each scenario, as illustrated in Fig. 3. 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 GPT-4o [12] 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 us-

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					Overall
	Attribute Modification	Structure Transform	Physical Interaction	Property Response	Avg.	Spatial Intelligence	Strategic Reason	Long-Horizon Plan	Logic Puzzle Solving	Avg.	
Closed-source Models											
Flux-Kontext-Pro	45.68	45.65	42.85	40.17	43.59	43.54	44.03	49.80	40.07	44.36	43.72
Seedream4.0	69.54	73.13	67.88	62.40	68.24	39.27	43.54	43.79	51.91	44.63	57.05
Wan2.5	74.13	69.92	64.23	65.16	68.36	63.39	46.13	56.32	52.24	54.52	60.59
Nano Banana	77.54	78.45	72.41	77.39	76.45	66.26	**56.83**	56.35	**65.85**	61.32	68.38
GPT-4o	**83.27**	**83.93**	**80.43**	**77.62**	**81.31**	**78.82**	50.28	**66.02**	65.66	**65.19**	**71.64**
Open-source Models											
MagicBrush	43.10	46.24	43.88	47.70	45.23	63.43	33.77	30.83	35.17	40.80	40.98
Omnigen2	54.18	57.75	52.87	53.52	54.58	70.78	27.61	37.11	24.69	40.05	43.96
Step1X-Edit	59.50	57.37	56.07	56.75	57.42	62.90	34.17	44.63	44.03	46.43	50.12
Bagel-Think	61.27	59.83	53.95	57.11	58.04	65.65	43.46	43.82	40.18	48.28	51.25
Qwen-Image-Edit	75.19	73.16	71.80	67.54	71.92	57.03	34.80	47.37	37.70	44.22	56.46
UniREdit-Bagel (Ours)	**79.24**	**78.61**	**76.43**	**71.81**	**76.52**	**84.93**	**73.98**	**86.14**	**84.01**	**82.27**	**78.87**

ing Python code. Inspired by Game-RL [46], **(1)** we first design a diverse collection of game-based tasks and develop corresponding Python programs tailored to each category. **(2)** Then, these programs automatically generate paired original and edited images, along with instructions, 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 are applied to ensure the integrity and reliability of the data.

Overall, this multi-scenario data synthesis pipeline generates our **UniREditBench**, a unified reasoning-based image editing benchmark, and **UniREdit-Data-100K**, a large-scale synthetic dataset with high-quality CoT annotations. We will provide a detailed elaboration of this dataset in the following 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 scenario types: (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 subsets in Fig. 5 (a) and the detailed distribution of samples across different categories in Fig. 5 (b). These visualizations highlight the extensive vocabulary of our dataset that captures the diverse visual attributes, as well as its broad coverage across various categories.

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 [18] 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 generation, while $\mathcal{L}_{\text{img}}$ improves both the correctness and fidelity of the edited image.

5. Experiment

5.1. Implementation Details

Baselines. We benchmark *closed-source* models including: GPT-4o [12], Nano Banana [7], Gemini-2.0 [5], Seedream4.0 [23], Wan 2.5 [26], and Flux-Kontext-Pro [2], as well as *open-source* models including: Bagel [6], Qwen-Image-Edit [37], Step1X-Edit [19], Flux.1-Kontext-dev [2], Emu2 [24], Omnigen2 [38], Omnigen [40], HiDream-Edit [4], MagicBrush [48], and AnyEdit [45].

Training and Evaluation. We train all components of Bagel [6] except the VAE model for 5,000 iterations using the Adam optimizer and the cosine learning rate scheduler on UniREdit-Data-100K. The scheduler includes 500 warm-up steps, with the peak learning rate of 2e-5 and minimum learning rate of 1e-6. 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 average performance across all scenarios, with Nano Banana showing comparable capability. Wan2.5 delivers balanced results on real-world tasks but lags on game scenarios that require strategic reasoning. Besides, Seedream4.0 performs reasonably well on the *Structure Transform* dimension yet encounters challenges in game scenarios. Among **open-source** baselines, Qwen-Image-Edit performs strongly on real-world tasks such as *Attribute Modify* 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, 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 OOD performance comparison, UniREdit-Bagel achieves the strongest open-source results across all four 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 and GPT-4o, narrowing the gap between open- and closed-source models.

Table 3. Out-of-distribution quantitative performance comparison on RISEBench [50]. *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%
Seedream-4.0	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%**

Qualitative. The qualitative results presented in Fig. 6 highlight the strengths of our UniREdit-Bagel across various tasks. Specifically, in Fig. 6 (Row 4), most models fail to reliably reproduce the physical heat effect. Although several baselines, such as Nano Banana and Qwen-Image-Edit, successfully 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), Seedream-4.0, Nano Banana, and Wan-2.5 generally preserve instruction-irrelevant content but struggle with instruction-specific objectives. In contrast, UniREdit-Bagel excels in both fulfilling the instruction and maintaining the coherence of unrelated content.

6. Conclusion

This paper introduces **UniREditBench**, a unified reasoning-based benchmark for image editing with broader evaluation dimension coverage and a reliable dual-reference evaluation pipeline. Additionally, we design a multi-scenario data synthesis pipeline and release **UniREdit-Data-100K**, a large-scale dataset with high-quality chain-of-thought (CoT) annotations. To demonstrate its effectiveness, we fine-tune Bagel on this dataset, resulting in **UniREdit-Bagel**, which achieves significant improvements both quantitatively and qualitatively. Through comprehensive benchmarking of both open-source and closed-source image editing models, we highlight their strengths and weaknesses across various aspects.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, pages 18208–18218, 2022. 3

[2] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 1, 8

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 1, 3

[4] Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025. 8

[5] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, 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, 8

[6] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 1, 3, 7, 8

[7] 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. 1, 8

[8] Qingdong He, Xueqin Chen, Chaoyi Wang, Yanjie Pan, Xiaobin Hu, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xiangtai Li, and Jiangning Zhang. Reasoning to edit: Hypothetical instruction-based image editing with visual reasoning. *arXiv preprint arXiv:2507.01908*, 2025. 2

[9] Runze He, Kai Ma, Linjiang Huang, Shaofei Huang, Jialin Gao, Xiaoming Wei, Jiao Dai, Jizhong Han, and Si Liu. Freeedit: Mask-free reference-based image editing with multi-modal instruction. *arXiv preprint arXiv:2409.18071*, 2024. 3

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[11] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *CVPR*, pages 8362–8371, 2024. 2

[12] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 3, 6, 8

[13] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, pages 6007–6017, 2023. 3

[14] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, pages 2426–2435, 2022. 3

[15] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *NeurIPS*, 36:36652–36663, 2023. 3

[16] Chenglin Li, Qianglong Chen, Zhi Li, Feng Tao, and Yin Zhang. Vcbench: A controllable benchmark for symbolic and abstract challenges in video cognition. *arXiv preprint arXiv:2411.09105*, 2024. 2

[17] Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, and Li Yuan. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025. 1

[18] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 7

[19] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 1, 2, 8

[20] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3

[21] Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025. 3

[22] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020. 1

[23] ByteDance Seed. Seedream 4.0. https://seed.bytedance.com/en/seedream4_0, 2025. 8

[24] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, pages 14398–14409, 2024. 8

[25] Jingqi Tong, Jixin Tang, Hangcheng Li, Yurong Mou, Ming Zhang, Jun Zhao, Yanbo Wen, Fan Song, Jiahao Zhan, Yuyang Lu, et al. Code2logic: Game-code-driven data synthesis for enhancing vlms general reasoning. *arXiv preprint arXiv:2505.13886*, 2025. 2

[26] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 8

[27] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3

[28] Yibin Wang, Changhai Zhou, and Honghui Xu. Enhancing object coherence in layout-to-image synthesis. *arXiv preprint arXiv:2311.10522*, 2023. 3

[29] Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*, 2024. 1

[30] Yibin Wang, Weizhong Zhang, and Cheng Jin. Magicface: Training-free universal-style human image customized synthesis. *arXiv preprint arXiv:2408.07433*, 2024. 3

[31] Yibin Wang, Weizhong Zhang, Jianwei Zheng, and Cheng Jin. Primecomposer: Faster progressively combined diffusion for image composition with attention steering. In *ACM MM*, pages 10824–10832, 2024. 1, 3

[32] Yibin Wang, Zhimin Li, Yuhang Zang, Jiazi Bu, Yujie Zhou, Yi Xin, Junjun He, Chunyu Wang, Qinglin Lu, Cheng Jin, et al. Unigenbench++: A unified semantic evaluation benchmark for text-to-image generation. *arXiv preprint arXiv:2510.18701*, 2025. 3

[33] Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318*, 2025. 3

[34] Yibin Wang, Zhimin Li, Yuhang Zang, Yujie Zhou, Jiazi Bu, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. Pref-grpo: Pairwise preference reward-based grpo for stable text-to-image reinforcement learning. *arXiv preprint arXiv:2508.20751*, 2025. 1

[35] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025. 3

[36] Yibin Wang, Weizhong Zhang, Honghui Xu, and Cheng Jin. Dreamtext: High fidelity scene text synthesis. In *CVPR*, pages 28555–28563, 2025. 1, 3

[37] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 8

[38] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 1, 3, 8

[39] Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang Yu, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. Kris-bench: Benchmarking next-level intelligent image editing models. *arXiv preprint arXiv:2505.16707*, 2025. 2, 3, 6, 1

[40] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *CVPR*, pages 13294–13304, 2025. 8

[41] Yicheng Xiao, Lin Song, Yukang Chen, Yingmin Luo, Yuxin Chen, Yukang Gan, Wei Huang, Xiu Li, Xiaojuan Qi, and Ying Shan. Mindomni: Unleashing reasoning generation in vision language models with rgpo. *arXiv preprint arXiv:2505.13031*, 2025. 1

[42] Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025. 1, 3

[43] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *NeurIPS*, 36:15903–15935, 2023. 3

[44] 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. *arXiv preprint arXiv:2505.20275*, 2025. 2

[45] 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 *CVPR*, pages 26125–26135, 2025. 1, 8

[46] Dazhi Zhan, Xin Liu, Wei Bai, Wei Li, Shize Guo, and Zhisong Pan. Game-rl: Generating adversarial malware examples against api call based detection via reinforcement learning. *TDSC*, 2025. 7

[47] Dong Zhang, Lingfeng He, Rui Yan, Fei Shen, and Jinhui Tang. R-genie: Reasoning-guided generative image editing. *arXiv preprint arXiv:2505.17768*, 2025. 2

[48] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *NeurIPS*, 36:31428–31449, 2023. 1, 3, 8

[49] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. In *CVPR*, pages 9026–9036, 2024. 3

[50] Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, et al. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *arXiv preprint arXiv:2504.02826*, 2025. 2, 3, 6, 8

UniREditBench: A Unified Reasoning-based Image Editing Benchmark

Supplementary Material

A. Data Filtering

We design a comprehensive, multi-stage pipeline that performs data filtering, i.e., *instruction de-duplication*, *quality filtering*, and *human inspection*, to remove redundancy and low-quality data. Detailed elaboration of the filtering pipeline is provided below.

A.1. Instruction De-duplication

During the first stage of the real-world scenario, text prompt for image editing are sampled from the Gemini-2.5-Pro, which may potentially introduce repeated or near-duplicate entries. We remove redundancy along two aspects: exact matches and semantic similarity.

- **Exact-Match Deduplication**: We first normalize the *Original Image Description* by converting it to lowercase and removing punctuation. Afterward, we extract the set of words from the normalized text. If two samples contain identical word sets, they are considered duplicates, as the descriptions are effectively the same. These duplicate samples are then filtered out to ensure data diversity.
- **Semantic-Similarity Deduplication**: We use a sentence-transformers [22] model to extract sentence embeddings for both the *Original Image Description* and *Edit Instruction*. We then compute the pairwise similarity between these embeddings. If the similarity score exceeds a threshold of 0.7 for either description and instruction, the samples are deemed semantically redundant and are filtered out to enhance dataset diversity.

These complementary exact-match and semantic filters improve dataset diversity by eliminating both literal and paraphrastic duplicates.

A.2. Quality Filtering

To ensure the quality of both the generated text and images, we evaluate and filter them across six key dimensions: text hallucination, instruction adherence, content preservation, visual quality, image hallucination, and CoT quality. Scores for each dimension are assigned by the Gemini-2.5-Pro model on a 1–5 scale. Only samples that achieve the maximum score across all dimensions are retained.

- **Text Hallucination**: We evaluate the textual reference for hallucinated content, defined as entities or visual effects that are not mentioned in the Instruction or that cannot be plausibly induced by the given Instruction.
- **Instruction Following**: We compare the edited image with the textual reference to assess whether the generated visual changes accurately reflect the specified effects. Samples that demonstrate poor adherence to the instructions and text reference are discarded.
- **Content Preservation**: We assess whether regions unrelated to the edit instruction, such as the background, remain consistent between the original and edited images, ensuring stability in unaffected areas.
- **Visual Quality**: We assess whether the generated images meet fundamental quality standards, specifically by ensuring they are free from artifacts or degradation.
- **Image Hallucination**: We examine the edited images for any unintended additions or alterations beyond the specified textual reference, such as the appearance of additional objects.
- **CoT Quality**: We evaluate the correctness of the chain-of-thought (CoT) reasoning text, focusing on whether the analysis of the original image and instruction is logical and sound.

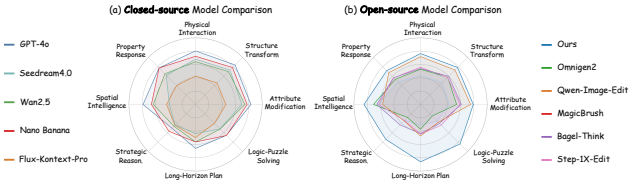

Figure 7. Benchmarking result visualization. (a) Closed-source model comparison; (b) Open-source model comparison.

Table 4. **Quantitative comparisons on KRIS [39]**. *GPT-4.1* is used as the evaluator. Best scores are in **bold** while second-best is underlined.

Model	Attribute Perception	Spatial Perception	Temporal Prediction	Social Science	Natural Science	Logical Reasoning	Instruction Decompose	**Overall Score**
Closed-source Models								
Doubao	70.92	59.17	40.58	65.50	61.19	47.75	60.58	60.70
Step 3ϕ vision	69.67	61.08	63.25	66.88	60.88	49.06	54.92	61.43
Gemini-2.0	66.33	63.33	63.92	68.19	56.94	54.13	71.67	62.41
GPT-4o	**83.17**	**79.08**	**68.25**	**85.50**	**80.06**	**71.56**	**85.08**	**80.09**
Open-source Models								
MagicBrush	53.92	39.58	-	42.94	38.06	30.00	23.08	37.15
AnyEdit	47.67	45.17	-	38.56	42.94	36.56	26.92	38.55
Emu2	51.50	48.83	22.17	34.69	38.44	24.81	45.00	39.70
Step1X-Edit	55.50	51.75	-	44.69	49.06	40.88	22.75	43.29
Bagel-Think	69.27	67.58	-	65.00	62.11	47.33	49.22	60.77
UniREdit-Bagel (Ours)	**71.75**	**71.00**	-	**69.20**	**65.99**	**59.91**	**51.55**	**65.45**

A.3. Human Inspection

In addition to our automated filtering pipeline, we perform a final manual check of each data instance. To facilitate this, we developed two web-based interfaces and enlisted eight expert annotators to carry out two-stage filtering process:

- **Initial Filtering**: Annotators remove samples with extremely erroneous textual references or substandard generated images.

Table 5. Detailed 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									Overall
	Attribute Modification		Structure Transform		Physical Interaction			Property Response		Spatial Intelligence	Strategic Reason			Logic Puzzle Solving			Long-Horizon Plan		
	Viewpoint Transformation	Material Modification	Pose Adjustment	Temporal Evolution	Structural Integrity Change	Motion State Change	Spatial Arrangement	Mechanical Reaction	Medium Interaction	3D Reconstruction	Space Invader	Jewel2	Pacman	Word Search	Tictactoe	Sudoku	Maze	Sokoban	
Closed-source Models																			
Flux-Kontext-Pro	34.75	56.62	54.95	36.36	42.15	42.52	43.89	38.01	42.33	43.54	47.45	34.09	50.55	34.88	58.44	26.88	46.70	52.90	43.72
Seedream4.0	64.77	74.32	80.19	66.07	59.48	64.68	79.48	61.58	63.21	39.27	46.28	43.00	41.34	48.55	38.43	68.75	54.15	33.43	57.05
Wan2.5	72.97	75.30	79.97	59.86	64.51	66.56	61.60	63.67	66.66	63.39	44.67	53.73	40.00	58.54	–	45.94	65.17	47.47	60.59
Nano Banana	75.37	79.72	85.55	71.35	70.65	73.36	73.22	**79.44**	75.34	66.26	**61.43**	**54.65**	**54.40**	**64.65**	40.92	**91.99**	62.65	**50.05**	68.38
GPT-4o	**83.55**	**82.98**	**92.16**	**75.70**	**76.10**	**76.32**	**88.88**	78.97	**76.28**	**78.82**	58.88	44.95	47.02	64.55	**63.80**	68.62	**82.27**	49.77	**71.64**
Open-source Models																			
MagicBrush	35.65	50.55	48.34	44.13	45.07	47.83	38.73	47.88	47.52	63.43	44.74	26.26	30.30	34.12	31.05	40.33	32.65	29.00	40.98
Omnigen2	50.32	58.03	67.73	47.77	49.90	57.63	51.09	55.05	51.98	70.78	51.17	1.50	30.17	23.07	46.17	4.85	39.63	34.58	43.96
Step-1X-Edit	51.82	67.18	63.29	51.45	54.98	62.18	51.03	55.72	57.78	62.90	33.62	35.40	33.48	43.02	49.92	39.17	54.53	34.73	50.12
Bagel-Think	58.38	64.15	63.38	56.27	54.40	58.63	48.82	56.47	57.75	65.65	47.30	42.12	40.97	47.80	40.35	32.40	48.83	38.80	51.25
Qwen-Image-Edit	72.08	**78.31**	81.75	64.57	68.78	69.51	77.12	67.83	67.25	57.03	36.58	37.79	30.02	48.47	33.00	31.63	60.02	34.71	56.46
UniREdit-Bagel (Ours)	**84.43**	74.06	**85.13**	**72.08**	**74.31**	**71.95**	**83.05**	**71.17**	**72.45**	**84.93**	**86.93**	**61.25**	**73.76**	**87.48**	**70.65**	**93.90**	**97.73**	**74.55**	**78.87**

- **Manual Correction**: Annotators make refinements to the textual reference effect that are only slightly incorrect, ensuring alignment and accuracy.

Two web interfaces are shown in Figs 11 and 12.

B. Detailed Benchmarking Results

We provide detailed benchmarking results on our UniREditBench for each category in Tab 5.

C. More Quantitative Results

We provide more quantitative out-of-distribution performance comparisons on KRISBench in Tab. 4.

D. More Qualitative Comparison Results

We provide additional qualitative comparisons on UniREditBench in Fig. 8 and 9, and comparisons on RISEBench in Fig. 10.

E. Ethical statement

In this work, we affirm our commitment to ethical research practices and responsible innovation. To the best of our knowledge, this study does not involve any data, methodologies, or applications that raise ethical concerns. All experiments and analyses were conducted in compliance with established ethical guidelines, ensuring the integrity and transparency of our research process.

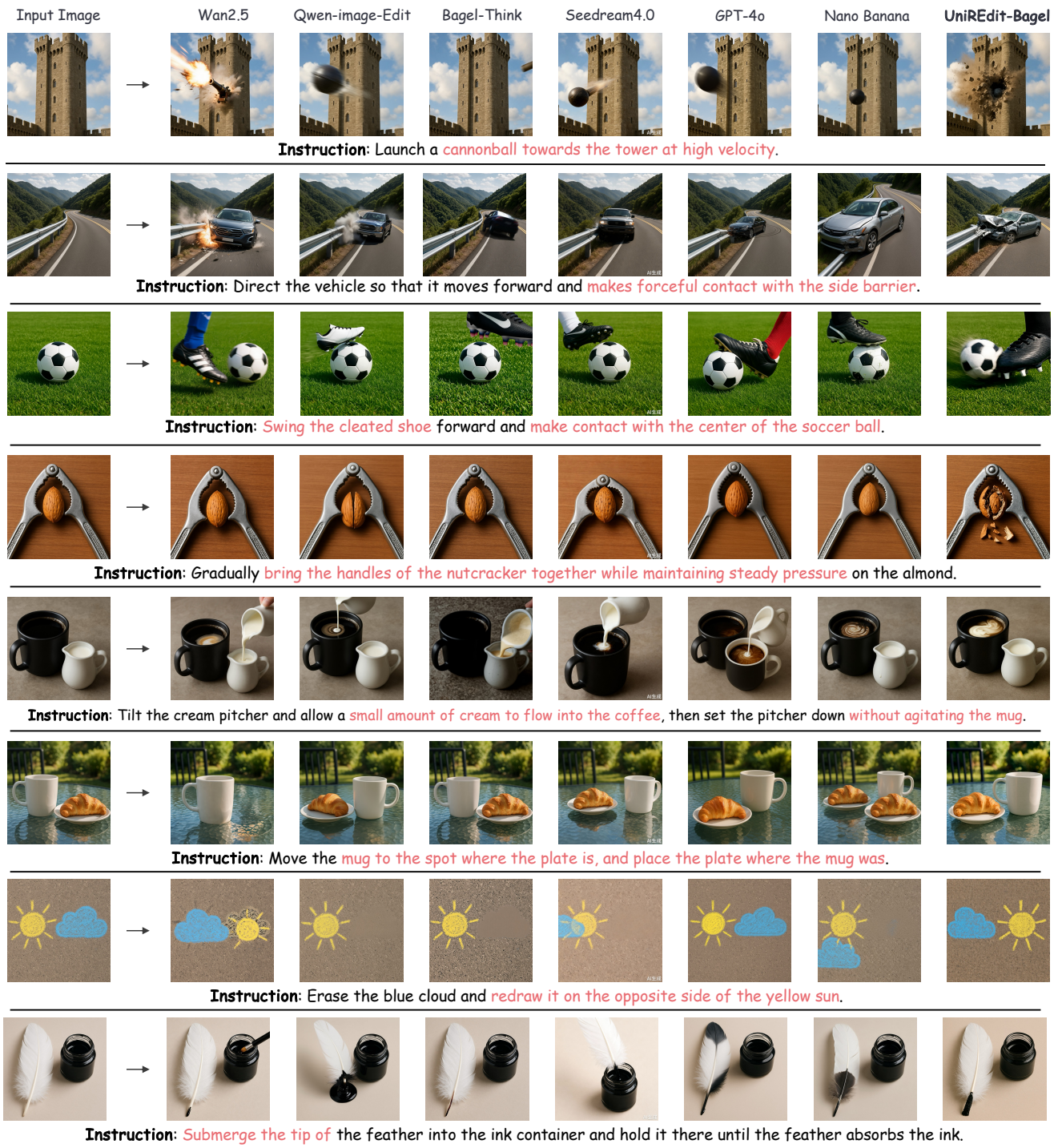

Figure 8. Qualitative editing result comparison on UniREditBench. Our UniREdit-Bagel demonstrates significant superiority in both instruction following and visual quality compared with state-of-the-art closed-sourced and open-sourced models.

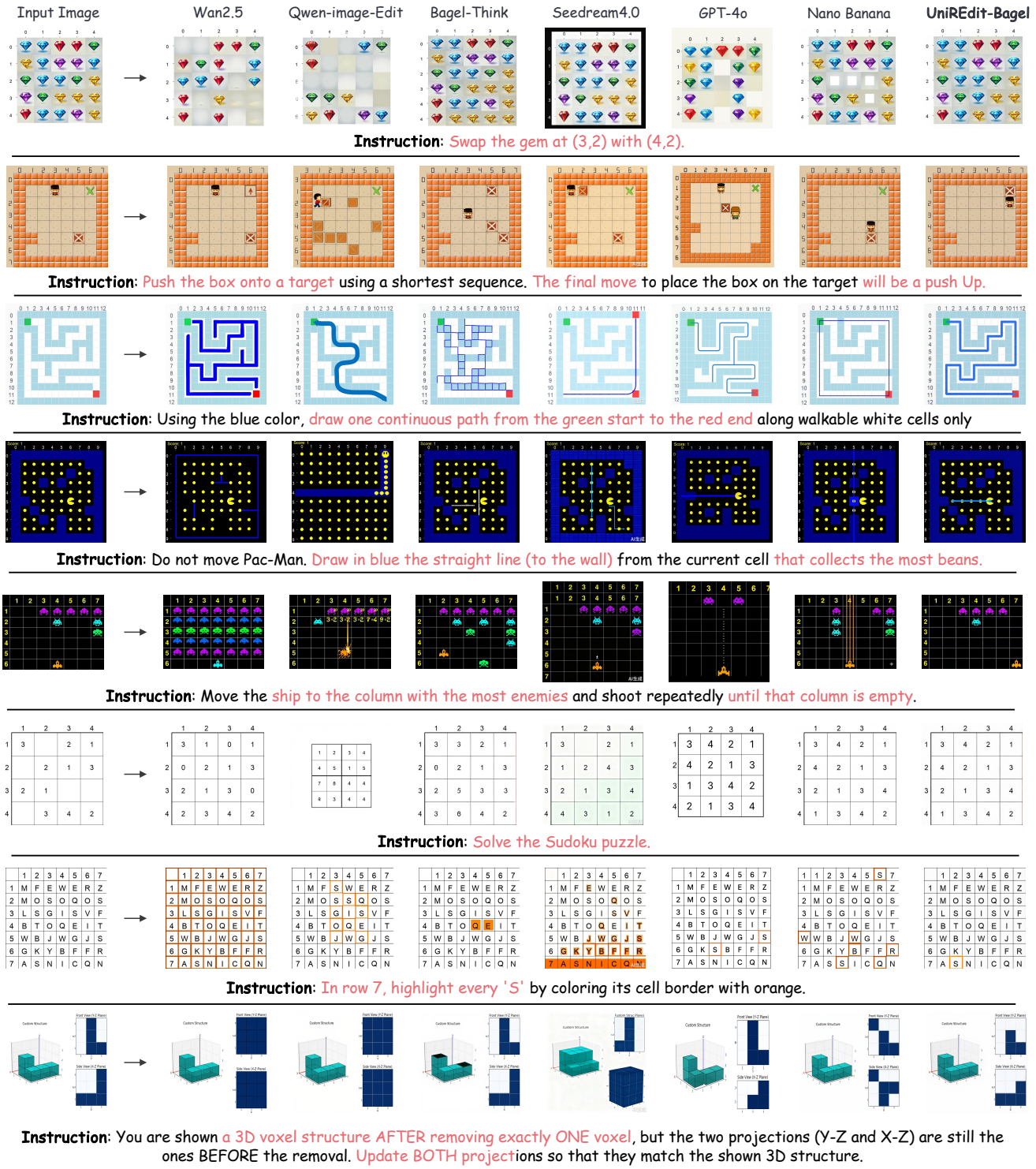

Figure 9. Qualitative editing result comparison on UniREditBench. Our UniREdit-Bagel demonstrates significant superiority in both instruction following and visual quality compared with state-of-the-art closed-sourced and open-sourced models.

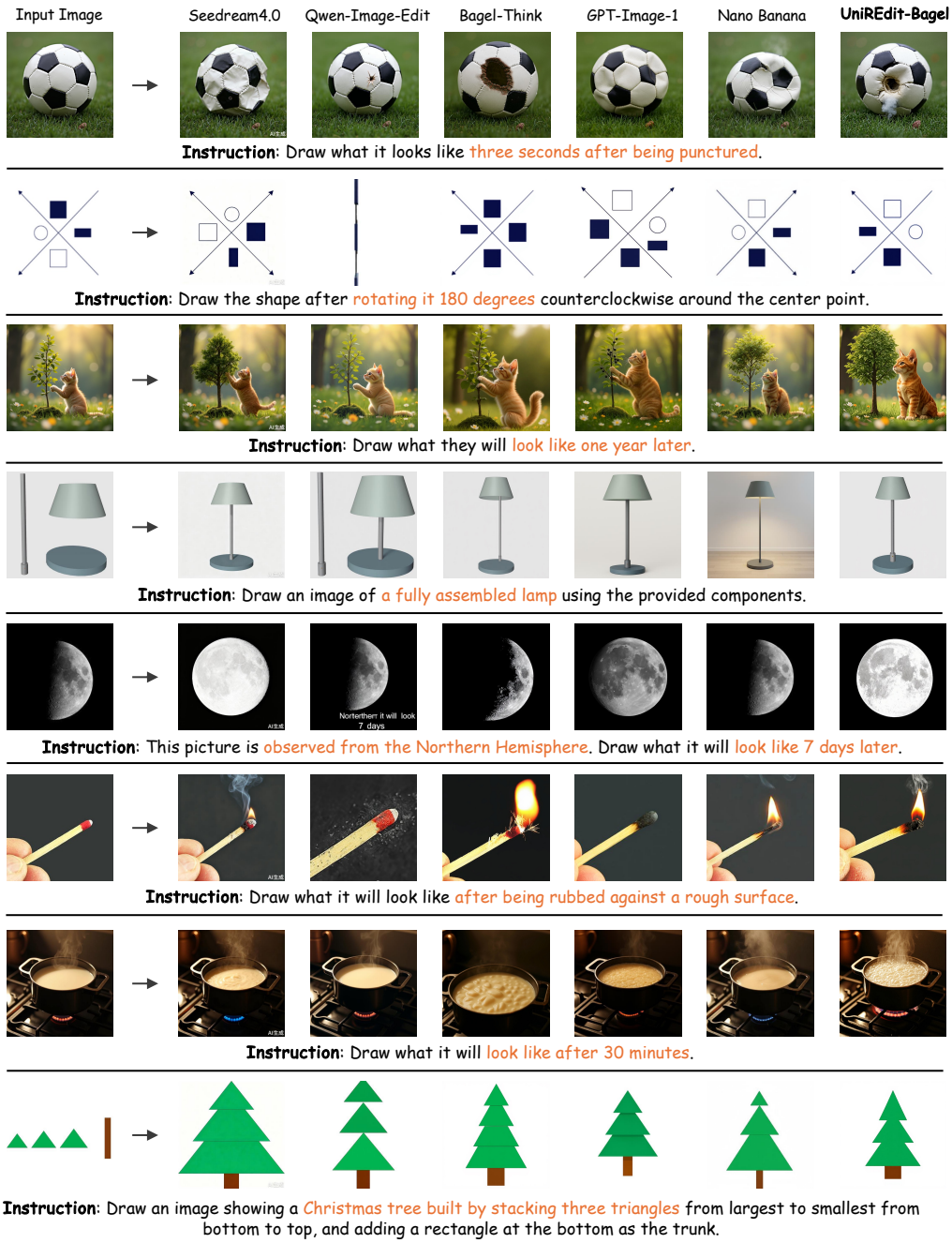

Figure 10. Qualitative editing result comparison on RISEBench. Our UniREdit-Bagel demonstrates significant superiority in both instruction following and visual quality compared with state-of-the-art closed-sourced and open-sourced models.

Initial Filtering

Class: `Structural Integrity Change`

Before Image | After Image

Instruction: Release an arrow so that it travels directly toward the apple and continues its motion through the point where the apple is positioned.

Reference Effect: The apple remains on the fence post but now has a clean, straight hole passing through its center, created where the arrow entered and exited. Splintered edges of skin and some apple flesh are visible around the entry and exit points. The arrow is seen continuing on its flight path, just beyond the apple, with no significant fragmentation or debris in the air.

✅ Yes ❌ No

Figure 11. Web interface of the initial filtering stage.

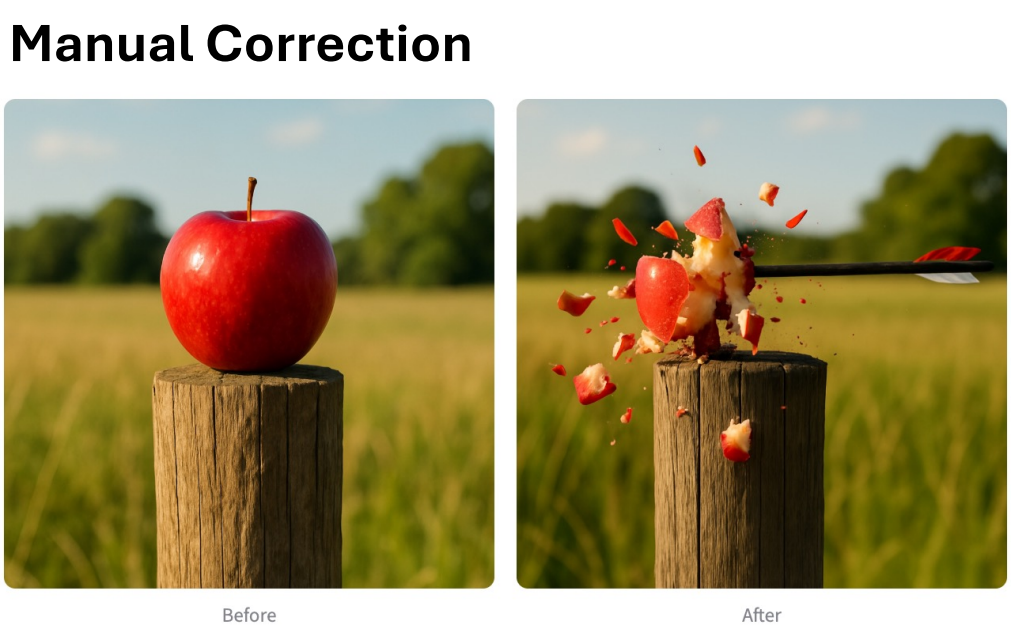

Figure 12. Web interface of the manual correction stage.