WORLDEDIT: TOWARDS OPEN-WORLD IMAGE EDITING WITH A KNOWLEDGE-INFORMED BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in image editing models have demonstrated remarkable capabilities in executing explicit instructions, such as attribute manipulation, style transfer, and pose synthesis. However, these models often face challenges when dealing with implicit editing instructions, which describe the cause of a visual change without explicitly detailing the resulting outcome. These limitations arise because existing models rely on uniform editing strategies that are not equipped to handle the complex world knowledge and reasoning required for implicit instructions. To address this gap, we introduce **WorldEdit**, a dataset specifically designed to enable world-driven image editing. WorldEdit consists of high-quality editing samples, guided by paraphrased instructions that align with real-world causal logic. Furthermore, we provide **WorldEdit-Test** for evaluating the existing model's performance on causal editing scenarios. With WorldEdit, we use a two-stage training framework for fine-tuning models like Bagel, integrating with a causal verification reward. Our results show that the proposed dataset and methods significantly narrow the gap with GPT-4o and Nano-Banana, demonstrating competitive performance not only in instruction following but also in knowledge plausibility, where many open-source systems typically struggle. See Project Page.

1 INTRODUCTION

In recent years, image editing models (Simsar et al., 2024; Zhu et al., 2025; Simsar et al., 2025) have made remarkable progress, demonstrating excellent performance on tasks with explicit instructions—such as attribute modification (Cao et al., 2023; Xu et al., 2023), style transfer (Chung et al., 2024; Wang et al., 2023), and pose synthesis (Yin et al., 2025; Shen et al., 2023). However, as illustrated in Figure 1, when confronted with implicit editing instructions, which only provide the cause of a visual change without explicitly describing the resulting visual outcome, most existing models still exhibit significant limitations in editing quality.

An intuitive workaround is to paraphrase implicit instructions into more explicit editing prompts. Yet, as shown in Figure 5, we observe that even when using paraphrasing to convey editing intent to pre-trained generative models, the editing results of most models are still quite poor. On the one hand, the visual outcomes implied by such instructions are often highly complex and require accurate world knowledge to realize. For instance, the instruction "a water balloon hits a cactus" entails visual effects (*e.g.*, splashing trajectories of water droplets) that must adhere to physical laws and object interaction logic. How to generate high-quality textual prompts based on such knowledge to describe the visual outcome in detail remains an open problem (Deng et al., 2025).

On the other hand, even giving paraphrased instructions, many visual expressions remain challenging for pre-trained generative models to follow and render, like the single-sided structure of a Möbius strip or scattering pattern of collapsed building blocks in Figure 5. This reveals a critical limitation in the generalization ability of existing models, which is closely tied to the "input-dependency" of editing instructions. For conventional explicit instructions (*e.g.*, remove the object from the image), the correlation for instructions with the input image content is relatively low, like remove, which leads to consistent visual change logic across different objects and scenes (i.e., plausible removal of the target region and background completion). Thus, models can maintain good performance even on unseen scenes. In contrast, world knowledge-driven implicit instructions are highly input-contingent. As shown in Figure 1, applying the same action "hit a cactus" to balls of different materials results in

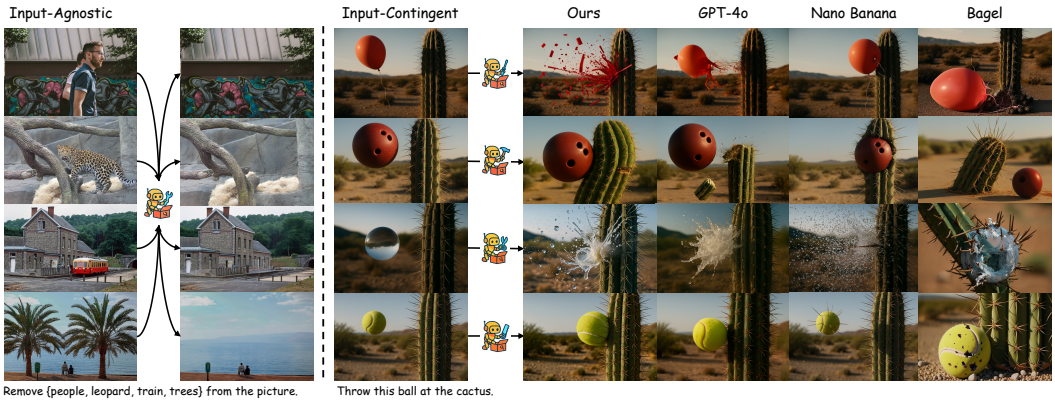

Figure 1: Unlike traditional image editing (*left*), which adopts a uniform editing strategy for different editing objects, world editing (*right*) needs to take into account the nature of the editing objects in the real world and produce editing results that conform to causal logic.

vastly different visual outcomes (*e.g.*, degree of deformation, reaction of the cactus) due to variations in physical properties (mass, elasticity, surface material, etc). This strong instruction–input–result coupling imposes far greater demands on model generalization than traditional tasks.

Thanks to the development of unified models (Wang et al., 2024; Zhou et al., 2024; Wang et al., 2025a; Deng et al., 2025), which are based on large-scale pre-trained models capable of handling both comprehension and generation tasks, these models can not only rewrite editing instructions based on input images but, more importantly, they implicitly capture the causal logic and visual relationships underlying world editing instructions through exposure to large-scale pre-training data. They introduce a promising path toward world image editing: transferring the world knowledge embedded in unified models into the editing process. Thus, an image dataset specifically designed for world-driven editing to stimulate unified models to leverage their powerful reasoning capabilities for improving image editing is required. Although recent efforts such as AnyEdit (Yu et al., 2025) have collected large-scale and diverse editing data, the vast majority still consist of traditional explicit instructions, with truly world knowledge-driven editing samples remaining scarce. KRIS-Bench (Wu et al., 2025) and RISEBench (Zhao et al., 2025) have proposed corresponding benchmark sets, but these cannot provide supervision for training and are limited in scale.

To address this gap, this paper proposes a comprehensive dataset **WorldEdit**. Using images from publicly available real-world segmentation datasets as original inputs, we employ instruction decomposition and multi-step editing via GPT-4o (Hurst et al., 2024) to generate edited images that conform to world knowledge logic. To ensure data quality, a two-stage filtering strategy—comprising instruction verification pre-filtering and image assessment post-filtering—is designed. The pre-filtering stage eliminates editing instructions with ambiguous causal logic through world knowledge consistency checks, while the post-filtering stage removes generated images that are visually unrealistic or violate world knowledge through visual plausibility assessment. Through this pipeline, the WorldEdit dataset ultimately contains 11k high-quality editing samples.

Finally, to validate the effectiveness of WorldEdit, we conducted a two-stage training procedure based on Bagel (Deng et al., 2025). In the first stage, we perform supervised fine-tuning with paraphrased instructions, enabling the model to better interpret implicit editing commands and align them with corresponding visual transformations. In the second stage, we further refine the model through reinforcement learning with a composite reward function that explicitly accounts for reasoning quality, visual fidelity, and causal consistency. This reward structure not only encourages the model to produce visually plausible and instruction-aligned images, but also grounds its generative process in interpretable reasoning traces and causal verification. Our main contributions are summarized as follows:

- We introduce **WorldEdit**, along with a challenging benchmark **WorldEdit-Test**, specifically designed to capture *cause-driven visual transformations* with world knowledge.
- With WorldEdit, we fine-tune Bagel in a two-stage process and introduce an *inversion-based causal verification reward* to better align generative behaviors with real-world causal logic.

- Our approach achieves **state-of-the-art** performance among open-source models on the WorldEdit-Test, demonstrating superior instruction generation and following capability.

2 Related Work

2.1 Text-based Image Editing

Image editing aims to modify image content according to given instructions while preserving both the consistency and naturalness of the edited output. Initially, diffusion models (Rombach et al., 2022; Esser et al., 2024), due to their remarkable image generation capabilities, were adapted for image editing by altering diffusion trajectories (Hertz et al., 2022; Tumanyan et al., 2023). Subsequent research incorporating masks (Avrahami et al., 2022; 2023), and multi-reference images (Kumari et al., 2023; Ruiz et al., 2023) has significantly enhanced the controllability of image editing. However, strong controllability does not necessarily imply intelligence. Since most diffusion-based methods rely on relatively small text encoders, they struggle to handle complex and fine-grained editing instructions that require reasoning. Recently, researchers have begun to explore the use of unified models for image editing. These models extend the generative capabilities of large language models (LLMs) to the visual domain, enabling cross-modal generation and understanding. Recent approaches such as Chameleon (Team, 2024), Emu3 (Wang et al., 2024), and Selftok (Wang et al., 2025a) adopt a unified next-token prediction paradigm by discretizing images, while Transfusion (Zhou et al., 2024) and Show-o (Xie et al., 2024) integrate image diffusion with autoregressive text prediction within a single framework. Compared to diffusion-based approaches, these models demonstrate improved performance in instruction following and semantic understanding. Furthermore, several commercial systems, such as GPT-4o (Hurst et al., 2024) and Gemini-2.0-Flash (Comanici et al., 2025), have exhibited impressive reasoning-based image editing capabilities, suggesting that unified LMMs provide a promising direction for image editing. However, they still fall short when confronted with instructions that implicitly require world knowledge. This limitation arises because, although unified models possess strong comprehension and generation abilities, it remains an open question whether these abilities can be mutually reinforced, and how their pretrained knowledge and reasoning skills can be effectively leveraged to advance image editing.

2.2 Benchmarks for Text-based Image Editing

Collecting high-quality image editing data is inherently challenging, as it requires strong consistency between pre- and post-edit images. Recent datasets such as InstructPix2Pix (Brooks et al., 2023), ImgEdit (Ye et al., 2025), AnyEdit (Yu et al., 2025), SEED-X (Ge et al., 2024) and UniReal (Chen et al., 2024) employing strategies like synthetic generation and video sampling to curate large-scale, high-quality editing pairs. However, these efforts primarily emphasize task complexity, instruction diversity, or dataset scale, without explicitly modeling the reasoning processes or knowledge structures involved in instruction understanding. More recent benchmark studies have begun to recognize this limitation by incorporating instruction understanding into their evaluations. For instance, SmartEdit (Huang et al., 2024) investigates spatial and interactive reasoning in ambiguous editing scenarios. RISEBench (Zhao et al., 2025) and KRIS-Bench (Wu et al., 2025) support to evaluates knowledge reasoning in image editing, but they do not provide training data. ReasonPix2Pix (Jin et al., 2024) and EditWorld (Yang et al., 2024) only provided limited unverified knowledge editing data, and were constrained by early image editing techniques, resulting in poor image quality and insufficient knowledge. Therefore, the lack of large-scale, high-quality training data continues to be a critical bottleneck. In this paper, we introduce **WorldEdit**, a benchmark that addresses these gaps by providing both large-scale training and evaluation datasets. WorldEdit offers high-quality resources for supervised training of editing models, while enabling a more comprehensive assessment of their ability to integrate world knowledge, reasoning, and visual editing.

3 WorldEdit

3.1 Construction of WorldEdit

Editing Type Definition. Most existing image editing benchmarks primarily focus on perceptual-level modifications or semantic operations, while often overlooking the rich spectrum of natural

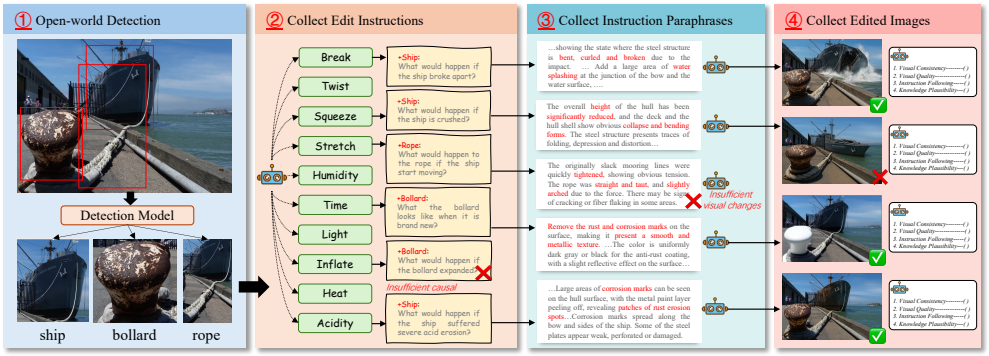

Figure 2: The automated construction pipeline of the WorldEdit dataset. Open-world images are filtered and screened along three dimensions: (1) causal consistency of implicit instructions, (2) richness of the expected visual transformations, and (3) quality of the synthesized edited images.

changes that occur in the real world. Previous efforts, such as RISE-Bench (Zhao et al., 2025), which organizes data according to types of cognitive reasoning, and KRIS-Bench (Wu et al., 2025), which categorizes tasks based on Bloom's taxonomy of cognition, include a substantial portion of evaluation data centered on semantic manipulations (*e.g.*, quantity, spatial location) and logical reasoning (*e.g.*, mazes, Sudoku). As a result, data that simulate causal transformations in the physical world are systematically underrepresented in current benchmarks. To address this gap, we introduce **WorldEdit**, a benchmark comprising both training and test data. Our benchmark emphasizes how changes unfold under specific real-world constraints, with the goal of providing a more focused and comprehensive evaluation of models' ability to reason about and simulate cause-driven visual transformations.

We categorize the transformations into two primary classes: *Environment-driven Transformations* result from changes in ambient conditions. Examples include Time (fruit ripening), Temperature (melting ice), Humidity (a flower wilting), Acidity (metal corrosion), and Light (fading fabric due to sunlight). *Mechanics-driven Transformations* result from the application of physical forces. Examples include Break (a plate shattering), Inflate (balloon expanding), Squeeze (paper crumpling), Twist (cloth wringing), and Stretch (rubber band elongating). A full definition of all transformation types is provided in the Appendix A.

Editing Instructions, Explanations, and Images Collection. To construct the WorldEdit, we design a pipeline, as shown in Figure 2, for collecting high-quality editing instructions, corresponding explanations, and edited images, ensuring coverage of the diverse real-world transformations.

First, we employ a detection model to extract object names from real-world images. This step aims to obtain the objects names and basic descriptions within the images, like *a rusted bollard* and *a white rope securing a ship to the bollard*. Accurate object detection is crucial as it forms the foundation for subsequent editing operations.

Next, we combine the detected objects with the predefined 10 editing types. For each object, we generate questions regarding its changes under different conditions. For example, for a ship, we might ask *What would happen if the ship broke apart?*. Notably, we perform filtering at this stage to remove unreasonable combinations or those with weak causal links. For example, the combination of *Bollard* and *Inflate* is filtered out because bollards, in reality, rarely undergo expansion, ensuring the collected data adheres to real-world plausibility.

We then utilize a pre-trained Large Language Model (LLM) to answer the valid questions obtained from the previous step. The LLM maps the causal changes arising from the questions to detailed visual change descriptions. Following this, we evaluate the generated visual change descriptions. We filter out responses with errors or those where the visual changes are not obvious. For instance, a description of *a slightly taut rope* might be filtered as the resulting visual change is too subtle to be effectively edited and evaluated.

Finally, we leverage GPT-4o (Hurst et al., 2024) to generate edited images based on the paraphrased editing instructions. For more complex editing instructions (failed more than 3 times), we decompose them and performs multi-step editing to enhance the quality of the edited images. After image generation, GPT-4o is used again to evaluate the edited images. Images with poor visual consistency

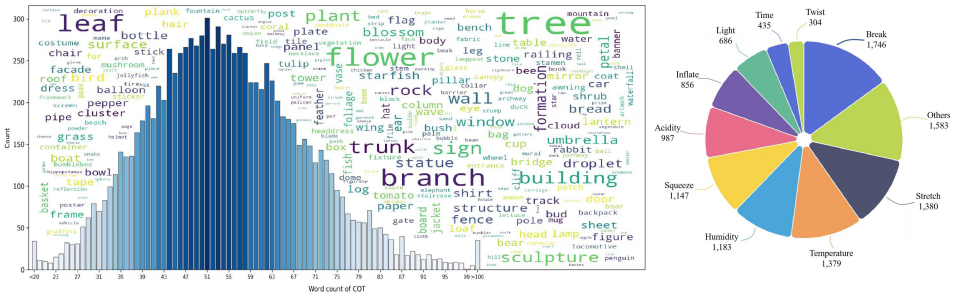

Figure 3: Statistics of the WorldEdit dataset. (*left*) Distribution of word counts in paraphrased instruction, along with a word cloud of frequently edited objects. (*right*) Distribution of 10 editing instruction categories.

or incorrect edits are filtered out. For example, an image where a ship is edited to appear crushed but lacks convincing structural damage would be rejected. Through this comprehensive pipeline, we collect a large-scale, high-quality dataset that accurately reflects real-world causal visual transformations for the WorldEdit benchmark.

Although collecting data from the open-world images can cover most of the variations, considering the limitations of public data, some data, such as "magnetic field lines" and "static effects" are not common. Therefore, in order to increase the diversity of the data, we supplemented a portion of the synthetic data. For more details, please refer to the Appendix B.

Dataset Characteristics and Statistics. Based on the data construction process, we have collected a total of 11k high-quality editing data. Among them, "Break" has the highest count (1,746), while "Twist" has the lowest count with only 304. We analyze it from multiple aspects, as in Figure 3, objects like trees, flowers, and buildings are prominently featured, reflecting the dataset's wide coverage of both natural and man-made entities. The distribution of instruction explanation lengths is typically between 50 and 60 words, ensuring sufficient detail for understanding causal logic.

3.2 Training with WorldEdit

To assess the capability of WorldEdit in eliciting cross-modal reasoning and transferring knowledge from pre-trained multimodal models to visual generation tasks, we implemented a two-stage fine-tuning strategy using Bagel (Deng et al., 2025) as our baseline model. The initial stage involves supervised fine-tuning (SFT), where instructions are paraphrased into structured Chain-of-Thought (CoT) sequences. This phase aims to enhance the model's capacity to interpret editing commands, leverage its inherent world knowledge, and establish a robust mapping from language instructions to visual modifications using a causal language modeling objective where loss is applied both to the text and image tokens.

Then, we employ reinforcement learning via the Flow-GRPO (Liu et al., 2025a) algorithm to further refine the model's generative behavior. A composite reward function guides this optimization, integrating three complementary signals designed to ensure high-quality, reasoned, and causally consistent outputs.

The CoT reasoning reward is motivated by the need for transparent and logically sound internal reasoning. It evaluates the generated CoT text and outputs a scalar score, R_{text}, based on the coherence, relevance, and correctness of the causal logic relative to the instruction. This reward encourages the model to produce rationales that faithfully reflect the transformation process, thereby improving interpretability and grounding the model's decisions in its pre-existing knowledge.

The image quality reward ensures the visual output both aligns with the instruction and retains consistency with the input image. Taking as input the original image, the generated image, and the instruction, this reward uses a multimodal model to produce a score, R_{image}, that captures instruction adherence and visual consistency. By directly optimizing the perceptual alignment and minimal invasiveness of the edit, this reward helps maintain high visual quality while executing precise changes.

The causal verification reward addresses the core objective of fostering genuine causal understanding. A multimodal model is required to infer the cause of transformation between the input

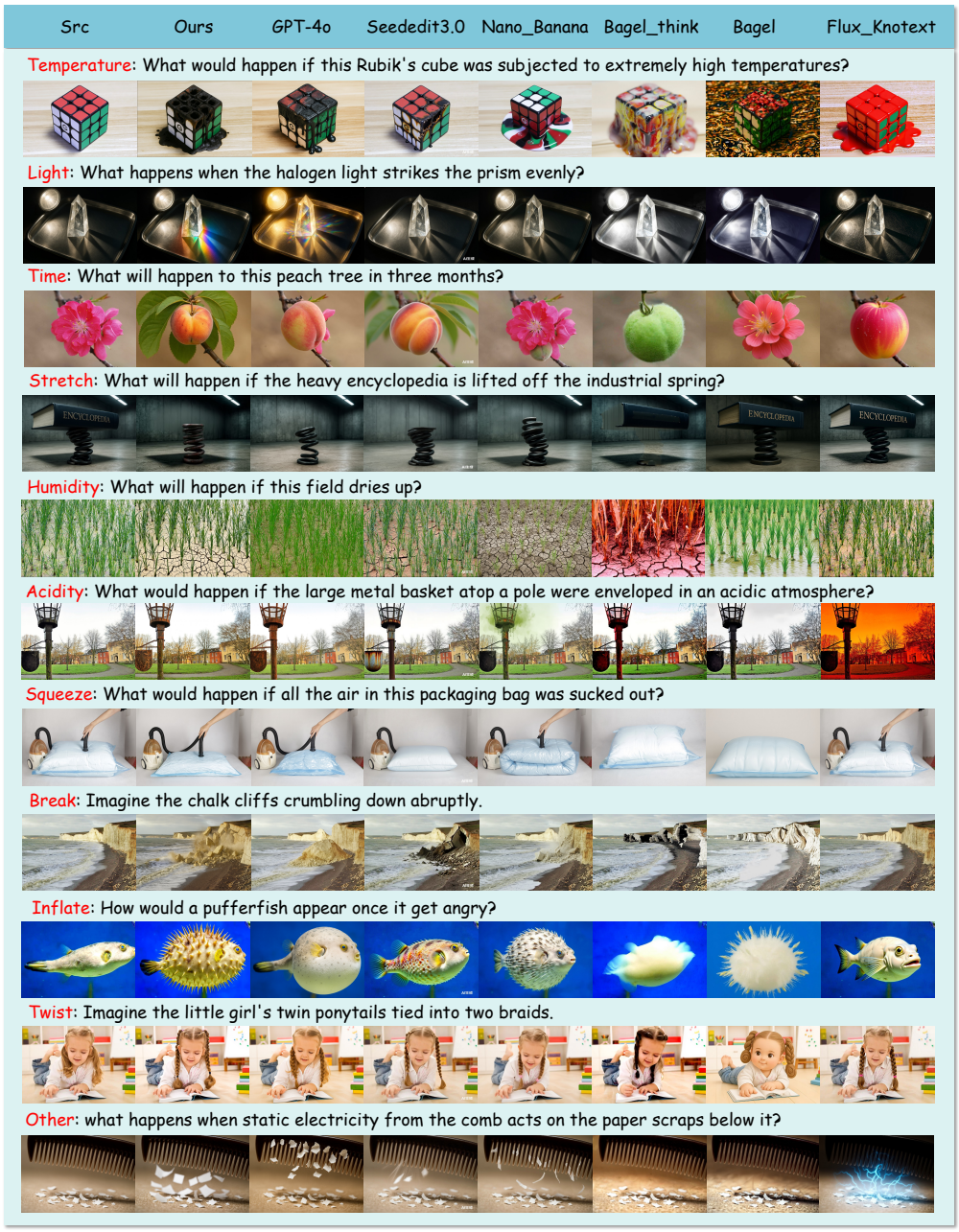

Figure 4: Qualitative comparison across different causal categories. The figure shows representative examples from ten causal reasoning tasks. Each row corresponds to a causal scenario, with the source image on the left followed by results from different models. Our method generates outputs that are both visually plausible and causally coherent, whereas baselines often produce irrelevant or stylistic edits, failing to reflect the causal logic of the instruction.

and output images, and return a similarity score R_{causal}, between this inferred cause and the original instruction, as the reward. By incentivizing the model to produce edits that are not only visually correct but also causally explainable, this mechanism ensures that the model learns the underlying physical and environmental principles.

The overall reward is computed as $R = R_{\text{reason}} + R_{\text{fidelity}} + R_{\text{causal}}$. This multi-objective reward structure ensures the model is guided toward reasoning-aware, high-fidelity, and causally grounded image editing, effectively leveraging the rich cause-and-effect relationships embedded in the WorldEdit benchmark to support open-world visual transformation tasks.

Cause Category	Metric	GPT-4o	Nano-Banana	SeedEdit-3.0	Ours	Flux-Kontext	Bagel-Think	Bagel	Omnigen	Omnigen2	Emu2	Anyedit	Ip2p	Magicbrush
Time	VC	4.02	4.22	3.81	4.18	4.6	2.38	2.34	3.02	2.22	1.48	1.63	3.04	1.40
	VQ	5.00	5.00	4.96	4.55	4.88	4.52	4.02	4.20	4.35	4.56	3.25	4.51	4.28
	IF	3.86	3.90	3.64	3.71	1.46	2.46	1.66	1.38	1.33	1.54	1.79	1.51	1.19
	KP	4.20	4.14	3.79	3.94	1.42	2.86	1.80	1.48	1.25	1.75	1.92	1.65	1.34
	Avg	4.27	4.32	4.05	4.10	3.09	3.06	2.46	2.52	2.29	2.33	2.15	2.68	2.05
Temperature	VC	4.02	4.22	3.98	3.52	3.86	2.08	1.94	2.54	2.26	1.38	1.45	2.62	1.81
	VQ	5.00	4.96	4.69	4.36	4.54	4.06	4.04	4.00	4.46	4.65	2.92	4.56	4.28
	IF	4.54	4.20	4.67	4.04	2.08	3.52	1.80	1.54	1.46	1.29	1.51	1.32	1.28
	KP	4.66	4.26	4.56	3.90	1.88	3.76	1.90	1.38	1.37	1.38	1.45	1.44	1.30
	Avg	4.56	4.41	4.47	3.96	3.09	3.36	2.42	2.37	2.39	2.17	1.83	2.49	2.17
Humidity	VC	4.50	4.40	4.24	3.82	4.40	2.68	2.32	2.28	1.94	1.76	1.62	2.86	1.40
	VQ	4.94	4.90	4.76	4.58	4.94	4.42	4.32	4.14	4.50	4.56	3.52	4.62	4.28
	IF	3.92	3.52	4.22	3.28	1.52	3.36	2.22	1.34	1.54	1.88	1.90	1.72	1.36
	KP	4.26	3.76	4.38	3.44	1.60	3.36	2.16	1.44	1.64	2.20	2.06	1.90	1.40
	Avg	4.41	4.15	4.40	3.78	3.12	3.46	2.76	2.30	2.41	2.60	2.28	2.78	2.11
Acidity	VC	4.46	4.36	4.14	3.94	4.74	2.74	2.60	3.14	1.88	1.34	1.38	2.74	1.60
	VQ	5.00	4.76	4.82	4.76	4.88	4.28	4.40	4.28	4.68	4.70	3.32	4.40	4.26
	IF	3.82	3.58	3.50	3.50	1.20	2.36	1.40	1.10	1.08	1.12	1.22	1.18	1.12
	KP	3.70	3.44	3.48	3.68	1.20	2.34	1.40	1.06	1.08	1.14	1.26	1.14	1.04
	Avg	4.25	4.04	3.99	3.97	3.01	2.93	2.45	2.40	2.18	2.08	1.80	2.37	2.01
Light	VC	3.70	4.40	4.43	3.82	4.64	2.84	2.36	2.44	1.88	1.45	1.43	2.66	1.54
	VQ	4.78	4.92	4.86	4.68	4.84	4.46	4.26	4.26	4.48	4.71	2.98	4.50	4.20
	IF	4.40	2.88	3.74	3.74	1.86	3.60	2.64	1.98	2.31	2.22	2.11	1.62	1.41
	KP	4.64	3.50	4.06	4.12	2.22	3.84	3.00	2.20	2.67	2.67	2.28	2.06	1.83
	Avg	4.38	3.93	4.27	4.09	3.39	3.69	3.07	2.72	2.83	2.77	2.20	2.71	2.25
Break	VC	4.80	4.90	4.26	4.74	4.88	4.00	2.98	2.98	1.80	1.36	1.46	2.36	1.40
	VQ	4.92	4.78	4.82	4.48	4.72	4.12	4.06	4.54	4.46	4.64	3.28	4.50	4.14
	IF	4.06	4.34	3.98	4.56	2.36	3.10	2.18	1.40	1.10	1.36	1.62	1.08	1.22
	KP	3.84	4.14	3.76	4.36	2.06	2.72	1.92	1.24	1.02	1.30	1.46	1.10	1.14
	Avg	4.41	4.54	4.21	4.54	3.51	3.49	2.79	2.54	2.10	2.17	2.00	2.26	1.98
Inflate	VC	4.53	4.88	4.35	4.20	4.54	3.00	2.53	2.38	1.94	1.59	1.31	2.58	1.56
	VQ	4.94	4.98	4.86	4.46	4.92	4.46	4.12	3.98	4.47	4.57	3.20	4.70	4.15
	IF	4.22	3.56	3.98	3.96	1.48	3.34	3.14	2.12	2.67	2.33	2.08	1.64	1.65
	KP	4.31	3.36	3.84	3.70	1.46	3.06	2.94	1.78	2.37	2.37	1.74	1.5	1.60
	Avg	4.50	4.20	4.26	4.08	3.10	3.47	3.18	2.57	2.86	2.71	2.08	2.61	2.24
Squeeze	VC	4.44	4.80	4.53	4.14	4.68	3.12	2.46	2.92	2.26	1.53	1.48	2.36	1.79
	VQ	4.94	4.94	4.66	4.34	4.82	3.88	4.12	4.24	4.62	4.51	3.22	4.24	4.27
	IF	2.86	2.54	3.21	3.38	1.22	2.56	1.90	1.42	1.36	1.64	1.48	1.24	1.46
	KP	2.74	2.48	2.87	3.04	1.22	2.28	1.70	1.32	1.34	1.30	1.30	1.20	1.38
	Avg	3.75	3.69	3.82	3.73	2.99	2.96	2.55	2.48	2.39	2.25	1.87	2.26	2.22
Twist	VC	4.52	4.52	4.45	4.3	4.52	2.56	2.22	2.62	2.02	1.63	1.43	1.84	1.66
	VQ	4.90	5.00	4.74	4.32	4.98	4.74	4.24	4.02	4.45	4.55	3.31	4.12	4.25
	IF	3.86	3.72	3.84	4.14	1.78	3.00	2.38	1.76	1.61	1.59	1.82	1.12	1.46
	KP	3.80	3.94	3.69	3.88	1.78	2.80	2.28	1.62	1.65	1.61	1.71	1.10	1.46
	Avg	4.27	4.30	4.18	4.16	3.27	3.28	2.78	2.51	2.43	2.35	2.07	2.05	2.21
Stretch	VC	4.36	4.34	4.22	3.98	4.20	3.22	2.38	2.36	2.07	1.90	1.54	2.10	1.51
	VQ	5.00	4.98	4.78	4.54	4.72	4.52	3.98	4.14	4.54	4.59	3.10	4.10	3.79
	IF	4.44	3.84	3.88	3.96	2.00	3.44	2.60	2.18	2.33	2.55	2.06	1.08	1.04
	KP	4.52	3.76	4.04	4.10	2.12	3.48	2.50	2.10	2.48	2.71	2.04	1.08	1.21
	Avg	4.58	4.23	4.23	4.15	3.26	3.67	2.87	2.70	2.85	2.94	2.19	2.09	1.89
Other	VC	4.26	4.46	4.52	4.22	3.96	3.20	2.68	2.20	3.54	2.10	2.08	2.50	1.86
	VQ	4.88	4.84	4.80	4.84	4.76	4.16	4.66	3.84	4.88	4.71	3.66	4.28	4.12
	IF	4.74	4.58	4.12	4.04	2.50	3.50	2.54	2.18	1.56	1.88	2.24	1.74	1.80
	KP	4.68	4.56	4.10	3.96	2.70	3.38	2.52	2.10	1.56	1.90	2.20	1.72	1.96
	Avg	4.64	4.61	4.39	4.27	3.48	3.56	3.10	2.58	2.89	2.65	2.55	2.56	2.44
Overall	Avg	4.36	4.22	4.21	4.07	3.21	3.35	2.76	2.52	2.51	2.46	2.09	2.44	2.14

Table 1: The performance of both commercial and open-source models on image editing tasks in WorldEdit-Test, evaluated across different causes and metrics. For each category, the Top-1, Top-2, and Top-3 scores are highlighted respectively.

4 Experiments

4.1 Experimental Setup

Baselines. We evaluate our method against a broad set of representative baselines, including traditional diffusion-based editing models and recent open-source unified multimodal models. In addition, we compare against commercial systems such as GPT-4o (Hurst et al., 2024), SeedEdit-3.0 (Wang et al., 2025b), and Nano-Banana (DeepMind, 2025), which have recently demonstrated impressive reasoning-aware editing ability.

Metrics. For evaluation, we follow the design of KRIS-Bench (Wu et al., 2025) and adopt a multi-dimensional assessment protocol. Specifically, we employ Qwen-VL-Max (Yang et al., 2025) as the evaluator to score editing results along four axes: visual consistency, which measures the structural alignment between the original and edited images; visual quality, which captures the perceptual realism of the generated images; instruction following, which quantifies how well the edits satisfy the given commands; and knowledge plausibility, which evaluates whether the edits are consistent with commonsense and world knowledge. More details are provided in the Appendix H.

4.2 Main Results on WorldEdit-Test Benchmark

Compared with baselines. Table 1 reports the overall performance of different models across all causal categories and evaluation metrics. We observe that commercial models such as GPT-4o (Hurst et al., 2024) and SeedEdit-3.0 (Wang et al., 2025b) consistently excel in visual quality and

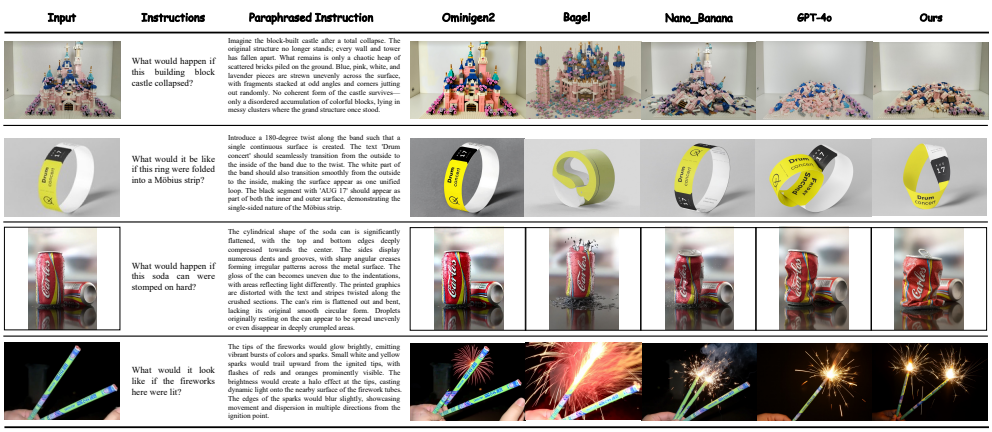

Figure 5: Qualitative results on WorldEdit-Test with paraphrased instructions. Text alone often fails to capture fine-grained causal details (e.g., scattering pattern of collapsed building blocks), and models vary in their ability to interpret such prompts. Our model, fine-tuned with WorldEdit, generates the most faithful and visually coherent images, underscoring the importance of high-quality world knowledge-driven data.

Methods	Time	Temperature	Humidity	Acidity	Light	Break	Inflate	Squeeze	Twist	Stretch	Other	Overall
Nano-Banana*	4.35	4.47	4.21	4.10	3.93	4.55	4.21	3.70	4.37	4.26	4.63	4.25
GPT-4o*	4.27	4.52	4.41	4.30	4.43	4.46	4.52	3.83	4.25	4.53	4.67	4.38
Omnigen2*	2.44	2.40	2.70	2.30	2.92	2.36	2.84	2.19	2.42	2.87	3.33	2.62
Bagel*	3.39	3.25	3.05	2.53	3.24	2.75	3.21	2.46	2.37	3.15	3.71	3.01
Ours	4.10	3.96	3.78	3.97	4.09	4.54	4.08	3.73	4.16	4.15	4.27	4.07

Table 2: Performance of different models on the WorldEdit-Test when using paraphrased instructions. Results are reported across ten causal categories and averaged overall.

instruction following, reflecting their advantages in large-scale pre-training and extensive instruction fine-tuning. However, despite their strong performance, they still face challenges in categories that require explicit causal reasoning, such as *time* and *acidity*. Open-source editing models like Flux-Kontext (Labs et al., 2025) are competitive in generating visually reasonable outputs but perform poorly in instruction following and knowledge consistency. This is particularly evident in categories such as ”break”, ”squeeze”, and ”stretch”, where capturing the causal dynamics of physical interactions is crucial. Similarly, open-source unified models outperform diffusion-based methods in instruction alignment but still have limitations in handling implicit instructions that require world knowledge. Our method has made significant progress in knowledge consistency and instruction following, achieving state-of-the-art performance among all open-source models and competitive results compared to commercial systems.

Qualitative evaluation Figure 4 presents the visualization results under different causal categories. We can clearly observe the advantages of our method in terms of instruction following and knowledge plausibility. For example, in the “temperature” case (Row 1), our model successfully simulates the melting and deformation of the Rubik’s cube under high temperatures and reveals the internal bearing structure. In the “light” case, only our model and GPT-4o correctly decompose the light. In categories involving strong physical interactions, such as “stretch”, “squeeze”, and “break” (Rows 4, 7, and 8), our method demonstrates outstanding capability in adhering to causal logic. Removing the heavy encyclopedia leads to a natural rebound of the spring, the vacuum-sealed bag collapses smoothly, and the cliffs break into visually coherent fragments—all consistent with the physical laws of the real world. In contrast, competing models often generate inconsistent or static results that fail to reflect the implied dynamic changes.

Compared with paraphrased instruction. We further evaluated the model’s editing results when directly given paraphrased instructions with clear visual changes. Our results reported in Table 3 indicate that although generating edited images based on paraphrased instructions results in improved accuracy across a majority of the tested scenarios, the performance gap remains significant when compared to the fine-tuning achieved with WorldEdit. For instance, while Bagel* shows a modest improvement of 0.25 over its original output, it still falls considerably short of the 4.02 achieved by our method. Moreover, as illustrated in Figure 5, even detailed paraphrased instructions fail to

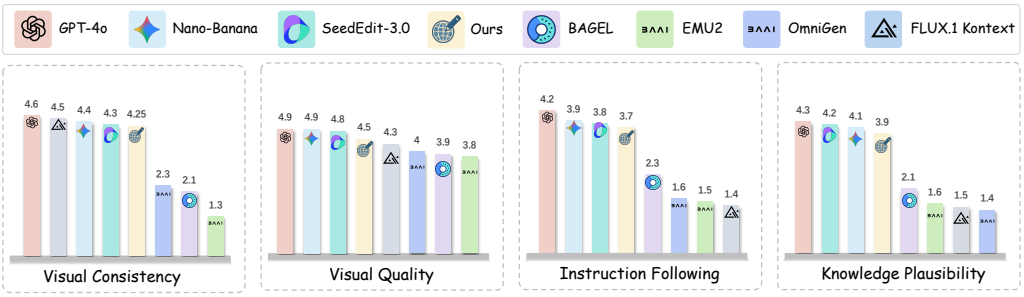

Figure 6: Human evaluation results for various models on the WorldEdit-Test. GPT-4o consistently achieves the highest scores across all metrics. Ours shows solid performance, especially in instruction following and knowledge plausibility.

Methods	Time	Temperature	Humidity	Acidity	Light	Break	Inflate	Squeeze	Twist	Stretch	Other	Overall
w/o CoT reasoning reward	3.85	3.92	3.88	4.01	3.98	4.05	3.95	3.82	3.99	4.00	4.03	3.94
w/o image quality reward	3.76	3.81	3.72	3.90	3.85	3.92	3.88	3.70	3.89	3.87	3.93	3.85
w/o causal verification reward	3.82	3.89	3.85	3.98	3.95	4.02	3.91	3.78	3.96	3.97	4.00	3.91
Ours	4.10	3.96	3.78	3.97	4.09	4.54	4.08	3.73	4.16	4.15	4.27	4.07

Table 3: Ablation study on the impact of different reward functions on the WorldEdit-Test.

enable the model to generate the lifelike results. In these complex cases, the paraphrased instructions alone are insufficient to convey the intricate details—such as the chaotic scattering of bricks in the castle or the abstract structure of the Möbius strip, which underscores the limitations of relying solely on textual adjustments. These finding highlights the critical role played by the specialized training data provided by WorldEdit in ensuring high-quality and accurate edits.

4.3 Ablation Study

To dissect the contributions of key components in our framework, we conducted ablation experiments on three reward functions. Ablating the image quality reward, the overall score dropped by 0.22. Since image quality is a core requirement for visual editing, the absence of this reward to constrain the generation process led the model to produce less visually appealing outputs. When the CoT reasoning reward and the causal verification reward are removed, the overall score decreases by 0.13 and 0.16 respectively. This result demonstrates that these two reward functions effectively constrain the model to maintain causal correctness during the editing process.

4.4 Human Evaluation

From the human evaluation results in Figure 6, several consistent conclusions emerge with Table 1. Both evaluations highlight that GPT-4o and Nano-Banana lead in most categories, aligning with human judgment that these models exhibit superior capabilities in maintaining visual fidelity and adhering to instructions. Models like Flux-Kontext and Emu2 score significantly lower in knowledge plausibility, reflecting their struggles in executing real-world, physically consistent transformations.

5 Conclusion

In this paper, we presented WorldEdit, a novel large-scale dataset that emphasizes world-knowledge-driven image editing. By introducing a diverse set of causal transformations, we have enabled image editing models to go beyond simple attribute manipulation and engage in more complex, implicit reasoning tasks. WorldEdit not only provides a comprehensive collection of real-world scenarios for training but also introduces a benchmark, WorldEdit-Test, to assess the ability of models to generalize to causal scenarios. Through the use of WorldEdit in two-stage fine-tuning and integration with causal verification rewards, we significantly improved Bagel's performance. The results align image edits with real-world causal logic, surpassing current state-of-the-art approaches. Looking ahead, we aim to expand WorldEdit to include a wider range of causal scenarios, and we hope that WorldEdit can provide a crucial resource for advancing the field of knowledge-aware image editing, enabling the development of more sophisticated and autonomous systems in the future.

REFERENCES

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18208–18218, 2022.

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42(4):1–11, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22560–22570, 2023.

Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, Hui Ding, Zhe Lin, and Hengshuang Zhao. Unireal: Universal image generation and editing via learning real-world dynamics, 2024. URL https://arxiv.org/abs/2412.07774.

Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8795–8805, 2024.

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.

DeepMind. Gemini image generation overview. https://gemini.google/overview/image-generation/, 2025. Accessed: 2025-09-25.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

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.

HiDream.ai. Hidream-i1. https://github.com/HiDream-ai/HiDream-I1, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8362–8371, 2024.

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.

Ying Jin, Pengyang Ling, Xiaoyi Dong, Pan Zhang, Jiaqi Wang, and Dahua Lin. Reasonpix2pix: Instruction reasoning dataset for advanced image editing, 2024. URL https://arxiv.org/abs/2405.11190.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1931–1941, 2023.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025a.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025b.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Fei Shen, Hu Ye, Jun Zhang, Cong Wang, Xiao Han, and Wei Yang. Advancing pose-guided image synthesis with progressive conditional diffusion models. *arXiv preprint arXiv:2310.06313*, 2023.

Enis Simsar, Alessio Tonioni, Yongqin Xian, Thomas Hofmann, and Federico Tombari. Uip2p: Unsupervised instruction-based image editing via cycle edit consistency. *arXiv preprint arXiv:2412.15216*, 2024.

Enis Simsar, Alessio Tonioni, Yongqin Xian, Thomas Hofmann, and Federico Tombari. Lime: localized image editing via attention regularization in diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 222–231. IEEE, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14398–14409, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1921–1930, 2023.

Bohan Wang, Zhongqi Yue, Fengda Zhang, Shuo Chen, Li'an Bi, Junzhe Zhang, Xue Song, Kennard Yanting Chan, Jiachun Pan, Weijia Wu, et al. Selftok: Discrete visual tokens of autoregression, by diffusion, and for reasoning. *arXiv preprint arXiv:2505.07538*, 2025a.

Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. Seededit 3.0: Fast and high-quality generative image editing. *arXiv preprint arXiv:2506.05083*, 2025b.

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.

Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7677–7689, 2023.

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.

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. *arXiv preprint arXiv:2409.11340*, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with natural language. *arXiv preprint arXiv:2312.04965*, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Ling Yang, Bohan Zeng, Jiaming Liu, Hong Li, Minghao Xu, Wentao Zhang, and Shuicheng Yan. Editworld: Simulating world dynamics for instruction-following image editing, 2024. URL https://arxiv.org/abs/2405.14785.

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.

Xiangchen Yin, Donglin Di, Lei Fan, Hao Li, Wei Chen, Yang Song, Xiao Sun, Xun Yang, et al. Grpose: Learning graph relations for human image generation with pose priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9526–9534, 2025.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26125–26135, 2025.

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.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

Tianrui Zhu, Shiyi Zhang, Jiawei Shao, and Yansong Tang. Kv-edit: Training-free image editing for precise background preservation. *arXiv preprint arXiv:2502.17363*, 2025.

A Detailed Tasks Explanation

In this section, we provide detailed explanations of the editing tasks used in the WorldEdit benchmark. Each task type corresponds to a specific causal transformation mode, covering both environment-driven and mechanics-driven changes, as well as an "Other" category for phenomena not captured by the ten canonical modes. Together, these modes ensure comprehensive coverage of real-world knowledge-informed editing scenarios.

Time. This mode refers to visual transformations that unfold naturally as time progresses, driven by biological growth, chemical reactions, or material aging. Typical examples include fruit ripening, flowers wilting, food decaying, or seasonal changes in plants. Beyond these, the mode also covers oxidation effects such as a bitten apple surface turning brown, structural aging such as buildings weathering and deteriorating, and life cycle developments such as tadpoles growing into frogs or bamboo shoots maturing into bamboo stalks. Edits under this mode simulate how objects evolve as time passes, often involving color fading, structural growth, surface corrosion, or progressive degradation consistent with real-world temporal dynamics.

Temperature. This mode captures a wide range of transformations driven by thermal conditions, including heating, ignition, burning, melting, and freezing. Classic examples include ice melting under heat, chocolate softening and dripping, snowmen collapsing as they thaw, or plastic deforming and liquefying at high temperatures. Different materials display distinct visual and textural changes: iron glows red and emits light when exposed to several hundred degrees of heat, while fireworks burst into sparks and colorful trails upon ignition. Cooling or freezing leads to effects such as solidification, frost formation, or stiffness in meat and other organic matter. This mode emphasizes how varying temperature conditions trigger material-specific responses, resulting in visually diverse outcomes across different substances.

Humidity. This mode concerns transformations caused by the absorption or loss of moisture. Representative cases include dry noodles softening when soaked, food becoming moldy, or plants wilting when deprived of water. Beyond these, it also encompasses material- and environment-driven changes such as sponges swelling and softening after absorbing water, dried fungi like black fungus or seaweed shrinking and hardening after dehydration, floors appearing darker and glossier when wet, and moss growing on damp stones. Biological cases are also included, such as animal fur becoming flattened and clumped when soaked. Edits in this mode emphasize how the presence or absence of moisture alters texture, volume, and surface appearance, capturing both swelling and shrinking effects consistent with real-world humidity dynamics.

Acidity. This mode represents transformations driven by acidic environments, where chemical erosion gradually degrades materials or organisms. Classical cases include metal rusting, paint peeling, or surfaces developing corrosion stains. Beyond these, acidity can also affect organic objects such as plants exposed to acid rain, leading to leaf damage and decay, or fabrics and clothing that lose color, develop holes, or weaken under acidic exposure. Edits in this mode emphasize chemical wear-and-tear, manifested through discoloration, surface roughness, structural weakening, and progressive material breakdown consistent with corrosive processes.

Light. This mode includes transformations driven by illumination conditions, where changes in the direction, intensity, or nature of light lead to distinct visual outcomes. Common examples are the dispersion of light through prisms, the focusing and diverging of light beams, or shifted shadows cast by different light sources. The mode also covers scenarios such as changes in shadow shapes and lengths due to altered light positions, ultraviolet exposure affecting surfaces, transitions between near-beam and far-beam headlights, or adding a specific spotlight to highlight part of a scene. Edits in this mode emphasize optical phenomena such as refraction, concentration, and shadow dynamics, as well as the role of light strength in altering the visual atmosphere of the scene.

Break. This mode describes destructive transformations where an object undergoes structural failure once its internal strength is exceeded. Such changes may be driven by *external forces*, such as impacts, collisions, or pulling and tearing, leading to effects like shattering glass, collapsing cliffs, or ripping fabric. Alternatively, they may arise from *internal stresses* accumulated within the object, such as tree branches snapping under their own weight, ceramics cracking from uneven thermal expansion, or water freezing inside a container and causing rupture. Edits in this mode simulate

cracks, fractures, splinters, or complete disintegration, reflecting real-world principles of mechanical instability under external pressure or internal stress.

Inflate. This mode characterizes transformations in which an object becomes larger, fuller, or more voluminous due to internal growth, external force, or energy accumulation. This includes classical cases such as a balloon being inflated, a pufferfish puffing up, or dough rising. Beyond these, the mode also encompasses natural phenomena where entities expand in scale or density, such as trees growing more luxuriant, water flows becoming greater and more turbulent, or clouds swelling and spreading across the sky. Edits in this mode emphasize visual cues of enlargement, increased density, and dynamic expansion, while maintaining coherence with the object's inherent structure and properties.

Squeeze. This mode represents deformation caused by external compression, collision, or the loss of structural support. Typical examples include crumpled paper, a soda can stomped flat, or air being sucked out of a plastic bag. Beyond these, it also covers larger-scale or organic cases such as a car body being deformed after a rear-end collision, or skin forming wrinkles under pressure . Edits in this mode highlight flattened or distorted shapes, visible creases, dents, and surface irregularities, reflecting how objects lose volume or structure when squeezed or compressed.

Twist. This mode denotes transformations where objects deviate from their original orderly or linear structure into bent, entangled, or irregular configurations. Unlike purely torsional forces that produce simple spiral shapes, this mode emphasizes more general distortions involving bending, wrapping, or chaotic growth. Representative examples include folding a paper strip into a Möbius band, tangled earphone wires, tree branches growing in irregular directions, or a straight iron rod being arbitrarily twisted multiple times. Edits in this mode highlight irregular curvatures, interwoven structures, and complex spatial rearrangements while preserving continuity with the original object.

Stretch. This mode corresponds to transformations where objects are elongated, straightened, or expanded under tensile force or external manipulation. Typical examples include a rubber band being pulled, dough stretched thin, or springs extended. Beyond these, the mode also covers cases where initially compact, wrinkled, or bent structures are unfolded or spread out. Representative examples include tangled earphone wires being straightened, plants growing upward and outward, a sweater being pulled outward with its knitted patterns extended accordingly, wrinkled clothes being smoothed flat, animals spreading their wings, or a fishing net originally clustered together being stretched open. Edits in this mode emphasize proportional lengthening, surface expansion, or the transition from compactness to openness, while preserving material continuity and structural coherence.

Other. This category encompasses causal transformations not fully captured by the ten predefined modes, while remaining consistent with real-world physical or environmental principles. Representative cases include electrostatic attraction of objects, magnetic field interactions. It also covers material and fluid phenomena like sedimentation in liquids, diffusion and dispersion processes, or sudden events like a balloon bursting. In addition, broader contextual or scene-level changes, such as festive transformations involving decorations, lighting, or fireworks, are also included. This category ensures flexibility for capturing uncommon but physically valid effects beyond the above modes.

B More Details about Data Collection

To construct the WorldEdit dataset, we developed a multi-stage pipeline designed to ensure both instructional diversity and visual plausibility of the edited results. Compared with existing benchmarks such as KRIS-Bench (Wu et al., 2025) and RISE-Bench (Zhao et al., 2025), our construction process emphasizes causal transformations grounded in world knowledge and integrates both automated and human-in-the-loop filtering mechanisms. Below, we provide further details of each stage.

Data Sources. The foundation of our dataset is the DF2K-OST (Lim et al., 2017) collection, which provides high-resolution natural images across diverse domains. To emphasize realistic and visually rich scenarios, we filtered out all low-resolution samples and retained only high-quality images. Each image was then center-cropped into 2:3 and 3:2 aspect ratios to standardize the format. Since cropping may occasionally truncate essential semantic content, we employed GPT-4o (Hurst et al., 2024) to automatically filter out images that lost contextual coherence after cropping. The remaining

subset was taken as the primary pool of original images. To further enrich causal diversity, we supplemented the dataset with additional images curated from the internet that emphasize strong logical and causal relationships. Moreover, synthetic samples generated with GPT-4o were incorporated to cover scenarios underrepresented in natural data. Together, these three sources form a balanced mixture of real and synthetic content that grounds the benchmark in both authenticity and coverage.

Object Detection. Given the curated image pool, we applied GPT-4o (Hurst et al., 2024) as the object detector to identify salient entities within each image. For an image p_i, the detector produces a set of detected objects o_i, forming tuples $\{(p_1, o_1), (p_2, o_2), \ldots, (p_n, o_n)\}$. This step provides the semantic anchors required for subsequent instruction generation and ensures that editing tasks are localized to concrete and identifiable objects.

Instruction Design and Filtering. For each detected object, we paired it with one of transformation modes described in Appendix A. Each pairing was used to generate one to three implicit instructions that emphasize causal or outcome-dependent transformations. For instance, a raw piece of pork under the *Temperature* mode can be frozen, cooked, or charred, reflecting distinct material states under different conditions. To ensure that the instructions remain implicit and causally grounded, we applied GPT-4o (Hurst et al., 2024) to filter out trivial cases (e.g., "make the mango red"), which correspond to explicit attribute changes rather than world-knowledge-driven transformations. Similarly, we removed unreasonable pairings, such as associating "blue sky" with the *Break* mode.

Paraphrased Instruction Expansion. Once valid implicit instructions were retained, we expanded each into paraphrased instructions that made the underlying world knowledge and causal dynamics explicit. The expansion was required to specify fine-grained visual changes at the lowest perceptual level, including texture, shape, color, glossiness, firmness, fragmentation, deformation, and surface alterations. Abstract or metaphorical descriptions (e.g., emotions, atmosphere) were deliberately avoided. A second round of filtering was performed to exclude CoTs that failed to capture world knowledge, lacked significant visual impact, or contradicted real-world causal logic.

Image Editing and Synthesis. The surviving CoT instructions were then used to guide visual editing. Specifically, we employed GPT-4o (Hurst et al., 2024) to transform the original images according to the expanded descriptions, producing edited images paired with their corresponding instructions. For complex transformations, multi-step editing was adopted to preserve causal fidelity.

Quality Assurance. To guarantee dataset reliability, we enforced a rigorous evaluation pipeline. Each edited image was scored along four axes: Visual Consistency (VC), Visual Quality (VQ), Instruction Following (IF), and Knowledge Plausibility (KP). Samples with low consistency, poor visual fidelity, or weak causal logic were discarded. Finally, it undergoes manual review and screening to ensure that only the editorial content that both complies with the instructions and conforms to world knowledge is retained.

C Score Distribution of Model Outputs

The score distribution of the evaluated models on the **WorldEdit-Test** is shown in Figure 7. From the distributions, it is clear that GPT-4o (Hurst et al., 2024), SeedEdit-3.0 (Wang et al., 2025b), and Nano-Banana (DeepMind, 2025) consistently achieve a high proportion of favorable scores across the four evaluation dimensions: *Visual Consistency*, *Visual Quality*, *Instruction Following*, and *Knowledge Plausibility*. These models demonstrate strong capabilities in both generating visually coherent outputs and capturing the implicit world knowledge underlying the editing tasks.

Building upon the Bagel and Bagel-Think baselines, our model exhibits substantial improvements across all four dimensions. In particular, it significantly narrows the gap with GPT-4o and SeedEdit-3.0, demonstrating competitive performance not only in instruction following but also in knowledge plausibility, where many open-source systems typically struggle. This confirms the effectiveness of our knowledge-informed training framework in strengthening both the reasoning and generative aspects of image editing.

By contrast, models such as Flux and OmniGen display notable weaknesses. Although they can sometimes maintain perceptual realism, they frequently fail to adhere to implicit instructions or to generate edits consistent with causal logic, resulting in lower overall scores. These shortcomings underscore the necessity of benchmarks like **WorldEdit**, which disentangle perceptual quality from

Figure 7: The score distribution of the model being tested.

reasoning ability and highlight the critical importance of world-knowledge grounding for future progress in image editing.

D Human Evaluation Implementation Details

We employed 20 undergraduate students for the human evaluation process. These individuals were well-educated and possessed a solid understanding of basic real-world changes and common scientific knowledge. Each participant first viewed a set of 50 images, each labeled with scores on four dimensions (Visual Consistency, Visual Quality, Instruction Following, and Knowledge Plausibility) ranging from 1 to 5. This training phase aimed to establish a consistent and unified scoring standard across all evaluators.

Following this, each participant was randomly assigned a set of results from the 8 models shown in Figure 6. The students were asked to evaluate the outputs of these models across the same four dimensions. On average, each student evaluated approximately 500 images. This ensured that every image generated by the models was scored by at least two independent human evaluators, providing a robust and reliable assessment of the models' performance.

The evaluation process was designed to capture a comprehensive understanding of the models' capabilities, particularly in terms of how well they adhered to the provided instructions and how accurately they reflected real-world transformations. By involving multiple evaluators, we ensured a diversity of perspectives, which helped in mitigating potential biases and improving the reliability of the final scores. Through this methodology, we aimed to provide a thorough and balanced human evaluation, offering valuable insights into the performance of each model based on both subjective visual judgment and a solid understanding of causal logic.

E More Visualization Results

Figure 8 provides additional qualitative comparisons across different causal transformation categories in the **WorldEdit** benchmark. As illustrated, our method consistently produces edited images that are both visually coherent and causally faithful. In modes driven by environmental factors, such as material responses to humidity and temperature, our results capture fine-grained cues—including texture decay, color fading, or moisture-induced deformation—that better aligns with real-world expectations.

From the figure, it can also be observed that GPT-4o (Hurst et al., 2024) and SeedEdit-3.0 (Wang et al., 2025b) are often able to interpret the world knowledge implied in implicit instructions and generate satisfactory results. In contrast, Nano-Banana (DeepMind, 2025) generally lags behind GPT-4o and SeedEdit-3.0. For example, in the case of a rice field in autumn, both GPT-4o and SeedEdit-3.0 successfully render the green leaves combined with golden ears of rice, while Nano-Banana tends to simplify the scene into a uniformly golden field resembling wheat. Our model also demonstrates strong world-knowledge awareness, achieving results close to GPT-4o and SeedEdit-3.0 across various tasks. It shows notable ability to preserve editing rationality and incorporate causal reasoning, representing a substantial improvement in editing capability compared to the original Bagel (Deng et al., 2025) model and its Bagel-Think variant.

These visualizations further confirm that **WorldEdit** not only poses significant challenges for existing editing systems, but also highlights the unique advantages of our knowledge-informed framework in producing edits that are simultaneously realistic, faithful to instructions, and grounded in commonsense reasoning.

F More results on other bench

Figures 9, 10, 11 and 12 together with Tables 4 and 5 present a comprehensive comparison of different models on Kris-Bench and RISE-Bench. Across all settings, GPT-4o (Hurst et al., 2024) unsurprisingly remains the strongest overall system, but our WorldEdit model emerges as a competitive and robust open-source alternative, consistently outperforming other open-source editors by a large margin.

Figure 8: Additional visualization results across different causal modes in the **WorldEdit-Test**.

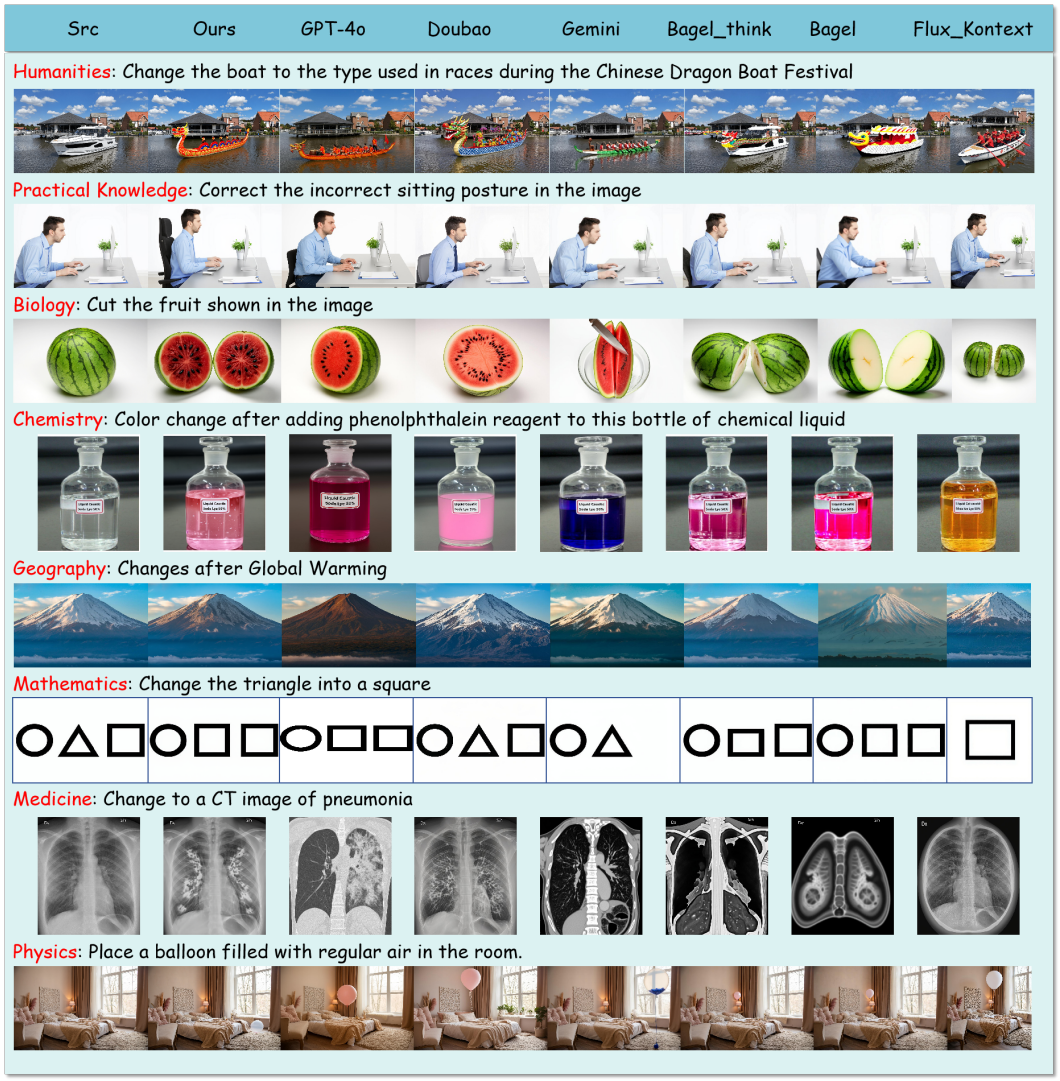

Figure 9: Visualization of results on the Conceptual Knowledge subset of Kris-Bench. WorldEdit shows improved robustness in capturing these cross-domain semantics, generating edits that reflect the intended conceptual transformation more faithfully than other open-source baselines.

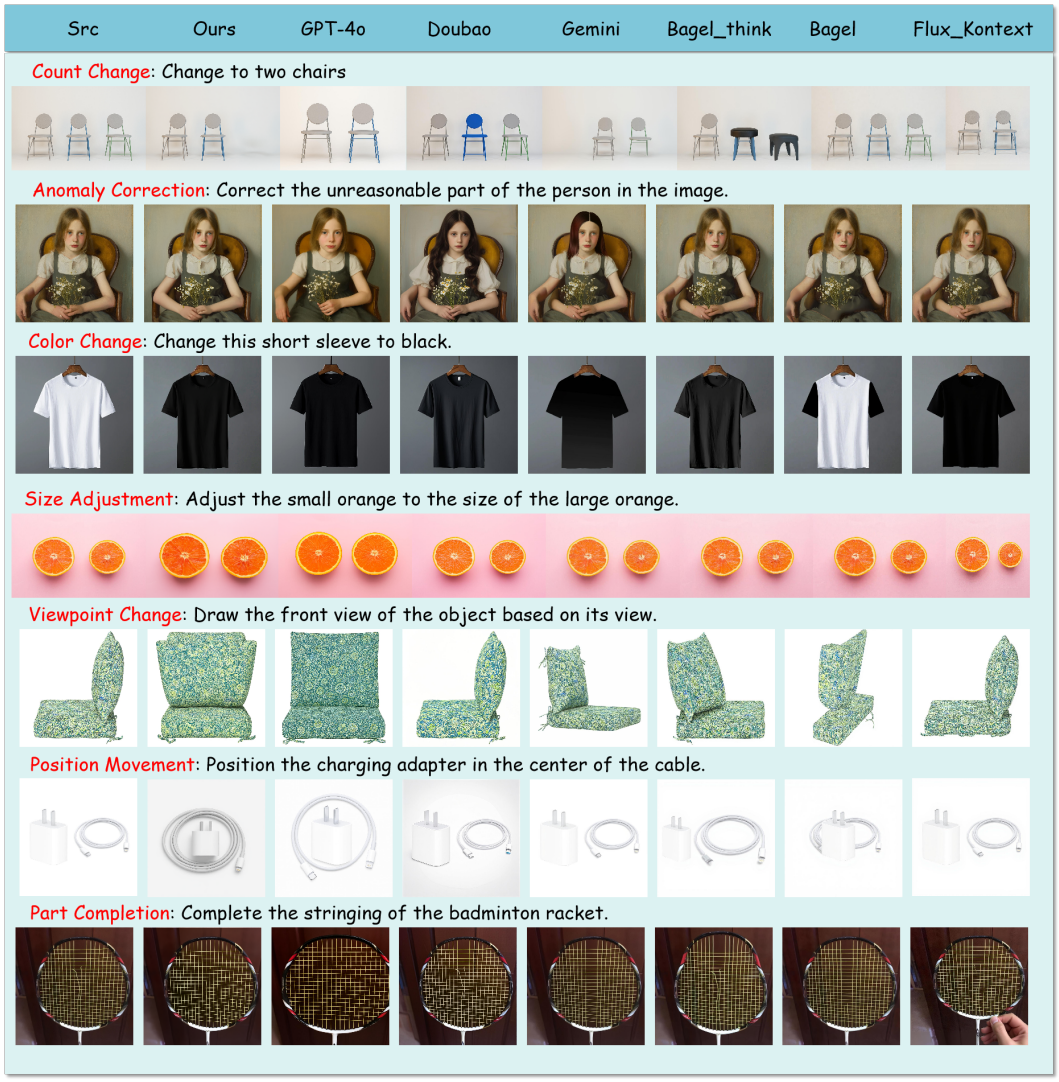

Figure 10: Visualization of results on the Factual Knowledge subset of Kris-Bench. Tasks in this split involve objective, visually verifiable transformations—such as quantity change, color change, size adjustment, geometric viewpoint synthesis, and object repositioning. The comparison highlights each model's ability to perform accurate, detail-preserving factual edits.WorldEdit reliably produces edits that align with both the instruction and the underlying visual structure, showing stronger factual consistency than other open-source baselines.

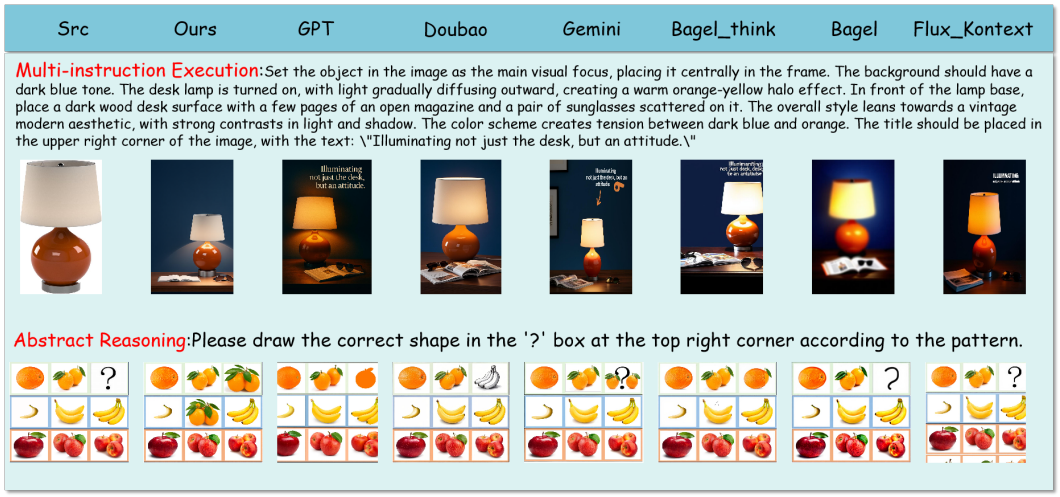

Figure 11: Visualization of results on the Procedural Knowledge subset of Kris-Bench. This subset evaluates models on tasks that require executing complex instructions or completing abstract visual patterns.

Table 4: Overall performance comparison on Kris-Bench. Closed-source frontier models such as GPT-4o and Gemini achieve the highest scores, while WorldEdit delivers the strongest performance among open-source systems across most reasoning categories, particularly in attribute perception, spatial understanding, and knowledge-driven editing.The performance of open-source and closed-source models is separately marked with the best performance in **bold**.

Reasoning Dimension		Metric	Closed-Source Models			Open-Source Models								
			GPT-4o	Gemini 2.0	Doubao	WorldEdit	Omnigen	Emu2	Bagel	Bagel-Think	Step1X-Edit	Anyedit	Magicbrush	Ip2p
Factual Knowledge	Attribute Perception	VC	**74.50**	69.50	66.75	**76.18**	35.75	47.75	66.75	74.75	63.00	54.75	53.50	17.50
		VQ	**94.75**	81.75	89.00	**85.91**	49.50	75.25	67.00	75.00	70.25	67.50	76.25	55.50
		IF	**80.25**	47.75	57.00	**53.73**	28.50	31.50	40.50	49.50	33.25	20.75	32.00	18.00
		Avg	**83.17**	66.33	70.92	**71.94**	37.92	51.50	58.08	66.42	55.50	47.67	53.92	30.33
	Spatial Perception	VC	**69.50**	60.50	67.50	**85.00**	24.00	41.50	53.50	77.25	64.25	55.75	38.00	13.25
		VQ	**94.50**	83.25	89.00	**92.00**	50.00	77.75	71.25	81.25	83.00	72.00	69.25	40.25
		IF	**73.25**	46.25	21.00	37.50	10.75	18.25	38.75	**44.75**	8.00	7.75	11.50	10.50
		Avg	**79.08**	63.33	59.17	**71.50**	28.25	48.83	54.50	67.75	51.75	45.17	39.58	21.33
	Temporal Prediction	VC	54.00	**54.50**	26.75	**64.86**	19.25	12.50	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*
		VQ	**86.25**	75.00	77.50	**94.26**	26.25	37.50	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*
		IF	**64.50**	62.25	17.50	**69.26**	20.00	16.50	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*
		Avg	**68.25**	63.92	40.58	**76.13**	21.83	22.17	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*
	Average	–	**79.80**	65.26	63.30	**56.98**	33.11	45.40	47.71	55.77	45.52	39.26	41.84	23.33
Conceptual Knowledge	Social Science	VC	**83.00**	77.00	72.00	**78.40**	37.25	32.75	75.75	76.50	63.25	62.00	54.00	15.75
		VQ	**95.75**	83.75	86.50	**89.40**	46.00	72.75	75.50	77.75	72.50	66.75	70.00	50.00
		IF	**84.50**	59.00	54.75	**49.80**	22.50	22.00	34.25	46.00	25.50	15.00	27.25	14.25
		KP	**78.75**	53.00	48.75	**39.80**	16.75	11.25	25.25	38.25	17.50	10.50	20.50	10.25
		Avg	**85.50**	68.19	65.50	**64.35**	30.63	34.69	52.69	59.63	44.69	38.56	42.94	22.56
	Natural Science	VC	**80.00**	65.00	70.25	**84.48**	31.00	35.00	65.75	68.00	71.25	61.75	47.00	18.75
		VQ	**96.00**	83.75	87.25	**93.26**	47.00	75.50	76.00	80.25	78.00	77.75	72.75	58.25
		IF	**76.50**	44.75	48.00	**55.79**	18.25	25.00	38.25	49.00	27.50	18.25	19.00	17.50
		KP	**67.75**	34.25	39.25	**44.02**	12.50	18.25	28.00	40.25	19.50	14.00	13.50	11.75
		Avg	**80.06**	56.94	61.19	**69.39**	27.19	38.44	52.00	59.38	49.06	42.94	38.06	26.56
	Average	–	**81.37**	59.65	62.23	**68.17**	28.02	37.54	52.17	59.44	48.01	41.88	39.24	25.59
Procedural Knowledge	Logical Reasoning	VC	**81.00**	73.50	64.75	68.67	15.00	23.50	**74.75**	71.25	58.75	55.50	37.25	14.75
		VQ	**95.00**	84.50	85.00	**92.50**	26.75	66.25	84.25	83.00	72.25	72.75	75.50	58.75
		IF	**59.25**	33.00	24.75	25.50	4.25	7.25	23.25	**29.25**	20.25	10.25	5.25	3.75
		KP	**51.00**	25.50	16.50	17.00	1.75	2.25	16.25	**21.25**	12.25	7.75	2.00	2.00
		Avg	**71.56**	54.13	47.75	50.92	11.94	24.81	49.63	**51.19**	40.88	36.56	30.00	19.81
	Instruction Decomposition	VC	**71.00**	58.25	51.50	**65.83**	28.75	31.00	30.75*	32.25*	25.75*	29.75*	20.75*	9.50*
		VQ	**96.25**	82.50	**76.75**	**82.33**	46.50	64.75	29.00*	25.25*	26.50*	39.25*	39.25*	27.75*
		IF	**88.00**	74.25	53.50	**47.00**	32.25	39.25	32.75*	24.50*	16.00*	11.75*	9.25*	7.00*
		Avg	**85.08**	71.67	**60.58**	**65.06**	35.83	45.00	30.83*	27.33*	22.75*	26.92*	23.08*	14.75*
	Average	–	**78.32**	62.90	54.17	**56.98**	23.89	34.91	40.23	39.26	31.82	31.74	26.54	17.28
Overall Average			**80.09**	62.41	60.70	**66.86**	28.85	39.70	47.76	53.36	43.29	38.55	37.15	22.82

On Kris-Bench (Table 4), WorldEdit achieves the best average performance among open-source models across factual, conceptual, and procedural knowledge. WorldEdit maintains non-trivial performance and even surpasses some closed-source models in certain metrics. In conceptual and factual knowledge (Figures 9 and 10), WorldEdit substantially improves over BAGEL and BAGEL-

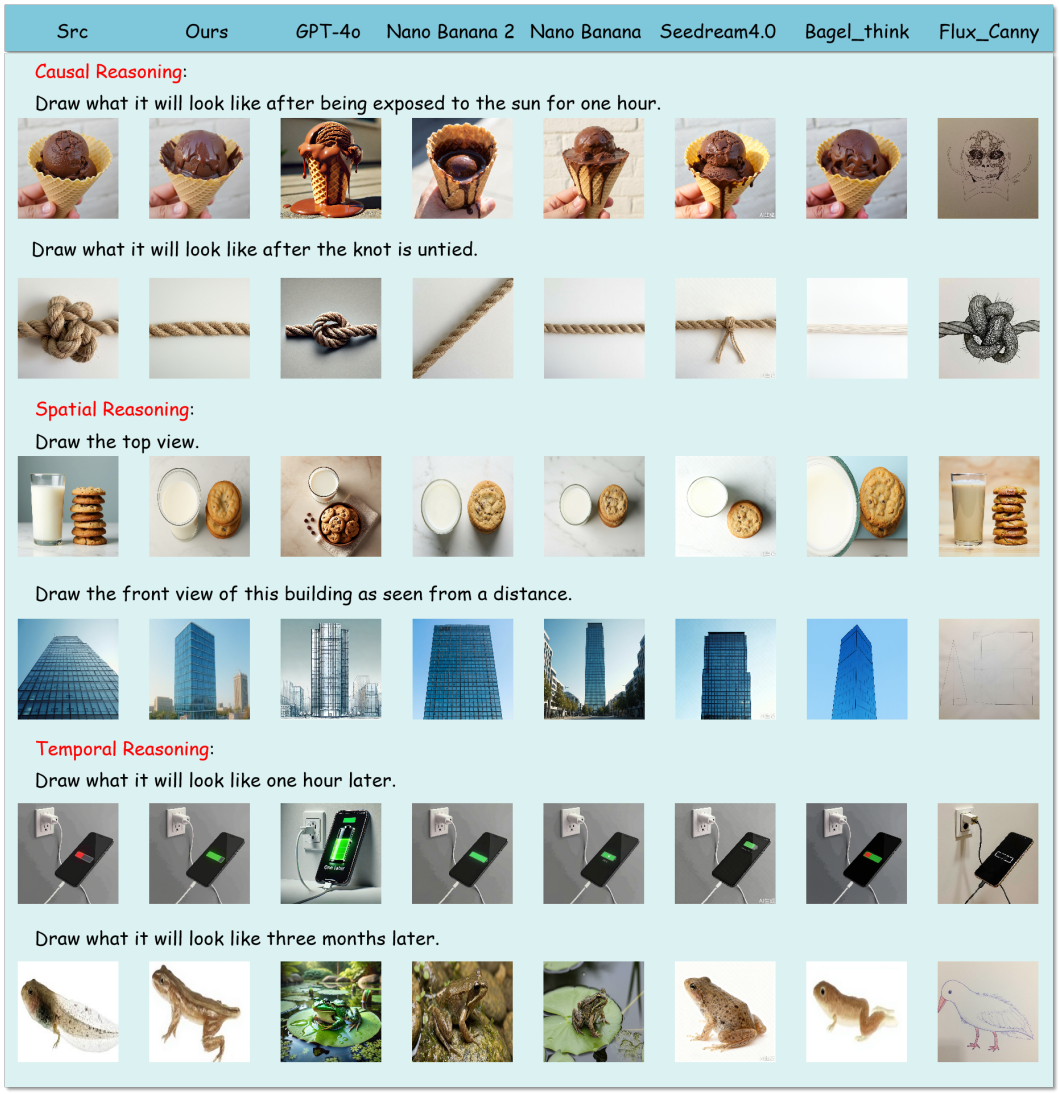

Figure 12: Visualization of results on RISE-Bench. This benchmark examines visual reasoning ability along three dimensions: causal reasoning, spatial reasoning, and temporal reasoning. The examples illustrate the varied behaviors of different systems when confronted with these reasoning-intensive tasks. WorldEdit produces comparatively more stable and context-aware outputs across several categories.

Table 5: Overall performance comparison on RISE-Bench. The benchmark evaluates temporal, causal, spatial, and logical reasoning abilities. While GPT-4o remains the strongest system, WorldEdit achieves notably higher scores than existing open-source editors across multiple reasoning dimensions.

Models	Temporal	Causal	Spatial	Logical	Overall
GPT-4o-Image (Hurst et al., 2024)	**34.1%**	**32.2%**	**37.0%**	**10.6%**	**28.9%**
WorldEdit(Ours)	22.4%	26.7%	13.0%	2.4%	16.1%
Gemini-2.0-Flash-exp (Team et al., 2023)	8.2%	15.5%	23.0%	4.7%	13.3%
Gemini-2.0-Flash-pre (Team et al., 2023)	10.6%	13.3%	11%	2.3%	9.4%
Bagel (Deng et al., 2025)	3.5%	4.4%	9.0%	5.9%	5.8%
Step1X-Edit Liu et al. (2025b)	0.0%	2.2%	2%	3.5%	1.9%
Omnigen Xiao et al. (2024)	1.2%	1.0%	0.0%	1.2%	0.8%
Emu2 (Sun et al., 2024)	1.2%	1.1%	0.0%	0.0%	0.5%
HiDream-Edit HiDream.ai (2025)	0.0%	0.0%	0.0%	0.0%	0.0%
FLUX.1-Canny (Labs, 2024)	0.0%	0.0%	0.0%	0.0%	0.0%

Think, especially on IF and KP in social and natural science categories, indicating better grounding in real-world causal and commonsense relationships rather than relying only on superficial visual alignment.

Results on RISE-Bench (Table 5 and Figure 12) reinforce these observations in a more diagnostic reasoning setting. GPT-4o again achieves the best temporal, causal, spatial, and logical accuracy, but WorldEdit is the next strongest model overall and the clearly best-performing open-source editor, substantially outperforming BAGEL, Step1X-Edit, OmniGen, EMU2, and other diffusion-based baselines, which remain close to chance on several dimensions.

Taken together, these quantitative and visual results show that WorldEdit is not just another high-fidelity editor: it consistently strengthens causal, temporal, and logical reasoning in editing, narrowing the gap to top closed-source systems while markedly advancing the state of open-source image editing under knowledge-intensive benchmarks.

G Limitations

While WorldEdit represents a significant advancement in image editing, it does come with challenges that highlight its strengths and potential for further development. First, the dataset's focus on world-knowledge-driven transformations inherently requires more computational resources and advanced models capable of capturing and processing complex causal relationships. It also places additional demands on model training, showcasing the need for more sophisticated fine-tuning approaches to handle the rich and detailed world knowledge embedded in the dataset.

Moreover, while WorldEdit offers an extensive range of causal transformations, the diversity of real-world scenarios that can be captured remains a work in progress. As the dataset grows and evolves, it will continue to challenge models to generalize across a broader spectrum of dynamic, real-world situations. This is a powerful opportunity for future research to expand upon WorldEdit by adding even more nuanced and complex causal scenarios, further refining the ability of generative models to simulate real-world transformations. In essence, the limitations of WorldEdit serve as a testament to its ambitious scope and potential to drive future breakthroughs in intelligent image editing, offering a valuable resource for advancing the field of knowledge-aware models.

H Prompt for Judgement

This section presents the evaluation prompts designed for **WorldEdit**. Figures 13, 14, 15, and 16 illustrate the prompts used for *Visual Consistency* (VC), *Visual Quality* (VQ), *Instruction Following* (IF), and *Knowledge Plausibility* (KP), respectively. We explicitly decouple the assessment of

"visual consistency and perceptual quality" from that of "instructional fidelity and knowledge plausibility". VC emphasizes that all non-instructed elements must remain unchanged, while VQ focuses solely on perceptual and structural quality without considering task correctness. IF is concerned with whether the correct target has been modified with the appropriate magnitude, whereas KP anchors the evaluation to the declared *editing mode*, verifying whether the edited outcome faithfully reflects real-world causal dynamics grounded in physics, chemistry, biology, or common sense. Noted that in the evaluation of SeedEdit-3.0, due to the presence of watermarks, we added a prompt in VC to ignore the watermarks.

To minimize cross-dimensional "information leakage," each prompt is scored according to its own criterion: VC does not penalize instructed changes, VQ does not assess task success, IF is unaffected by pure perceptual quality issues, and KP only decreases when causal logic or material laws are violated. In cases of *fundamental instruction failure* (e.g., replacing the target entirely instead of editing it), IF is capped at its lower bound and a consistency penalty is simultaneously applied to KP. This design ensures the separability of the four axes while faithfully capturing the implicit, causal, and knowledge-driven properties of **WorldEdit**.

Visual Consistency Template

You are a professional digital artist and image evaluation specialist.

You will be given:
1. **Image A**: the original image.
2. **Image B**: an edited version of Image A.
3. **Editing Instruction**: a hypothetical prompt describing how Image A should be transformed or imagined into Image B.

Your Objective:
Your task is to **evaluate the visual consistency between the original and edited images, focusing exclusively on elements that are NOT specified for change in the instruction**. That is, you should only consider whether all non-instructed details remain unchanged. Do **not** penalize or reward any changes that are explicitly required by the instruction.

Evaluation Scale (1 to 5):
You will assign a **consistency_score** according to the following rules:
- **5 Perfect Consistency**: All non-instruction elements are completely unchanged and visually identical.
- **4 Minor Inconsistency**: Only one very small, non-instruction detail is different (e.g., a tiny accessory, a subtle shadow, or a minor background artifact).
- **3 Noticeable Inconsistency**: One clear non-instruction element is changed (e.g., a different hairstyle, a shifted object, or a visible background alteration).
- **2 Significant Inconsistency**: Two or more non-instruction elements have been noticeably altered.
- **1 Severe Inconsistency**: Most or all major non-instruction details are different (e.g., changed identity, gender, or overall scene layout).

Guidance:
- First, **identify all elements that the instruction explicitly allows or requires to be changed**. Exclude these from your consistency check; then verify that every other pixel and element remains unchanged between the original and edited images.
- For all other elements (e.g., facial features, clothing, background, object positions, colors, lighting, scene composition, especially image color tone, brightness, contrast, etc.), **compare Image B to Image A** and check if they remain visually identical.
- If you observe any change in a non-instruction element, note it and consider its impact on the score.
- If the instruction is vague or ambiguous, make a best-effort factual inference about which elements are intended to change, and treat all others as non-instruction elements.

Note:
- **Do not penalize changes that are required by the instruction.**
- **Do not reward or penalize the quality or correctness of the instructed change itself** (that is evaluated separately).
- If the edited image introduces new artifacts, objects, or changes to non-instruction elements, this should lower the consistency score.

Input
Image A
Image B
Editing Instruction: {instruct}

Output Format
First, clearly explain your comparison process: list each major non-instruction element and state whether it is consistent (unchanged) or inconsistent (changed), with brief reasoning.
Then, provide your evaluation in the following JSON format:
{{
"reasoning": **Compared to original image**, [list of non-instruction elements that changed or remained the same] **in the edited image**.
"consistency_score": X
}}

Figure 13: Prompt for *Visual Consistency* (VC).

Visual Quality Template

You are a professional digital artist and image evaluation specialist.

You will be given:
- **Image A**: a single AI-generated image.

Objective:
Your task is to **evaluate the perceptual quality** of the image, focusing on:
- **Structural and semantic coherence**
- **Natural appearance**
- **Absence of generation artifacts**
You must **not penalize low resolution or moderate softness** unless it introduces semantic ambiguity or visually degrading effects.

Evaluation Scale (1 to 5):
You will assign a **quality_score** with the following rule:
- **5 Excellent Quality**: All aspects are visually coherent, natural, and free from noticeable artifacts. Structure, layout, and textures are accurate and consistent.
- **4 Minor Issues**: One small imperfection (e.g., slight texture blending, minor lighting inconsistency).
- **3 Noticeable Artifacts**: One or two clear visual flaws or semantic problems (e.g., extra fingers, minor duplication, slight distortion).
- **2 Structural Degradation**: Multiple distracting errors (e.g., melted hands, warped shapes, unreadable text).
- **1 Severe Errors**: Major structural failures or hallucinations (e.g., broken anatomy, garbled symbols).

Guidance:
Check the following visual aspects and mark them as ✓ (satisfactory) or ✗ (problematic):
- Structural coherence (e.g., correct anatomy, object shapes, legible text)
- Naturalness (lighting, perspective, shadow logic)
- Artifact-free (no duplication, ghosting, watermarks)
- Texture fidelity (clothing, hair, surfaces not melted or corrupted)
- Optional: Sharpness (only penalize if blur causes semantic loss)
✓ The more checks, the higher the score.

Example
"reasoning": "Structural coherence: ✓, Natural appearance: ✓, Artifacts: ✓, Texture fidelity: ✗ (fabric partially deformed).",
"quality_score": 4

Output Format:
After evaluation, provide your score and concise reasoning using the following JSON format:
{{
"reasoning": XXX,
"quality_score": X,
}}

Figure 14: Prompt for *Visual Quality* (VQ).

Instruction Following Template

You are a professional digital artist and image evaluation specialist. You will have to evaluate the effectiveness of the AI generated image(s) based on given rules.

You will be given:
1. **Image A**: the original image.
2. **Image B**: an edited version of Image A.
3. **Editing Instruction**: a hypothetical prompt describing how Image A should be transformed or imagined into Image B.

Your Objective:
Your task is to **evaluate how the edited image faithfully fulfills the editing instruction**, focusing **exclusively on the presence and correctness of the specified changes**.

You must:
Identify detailed visual differences between Image A and Image B **correctly and faithfully**.
Determine if those differences **match exactly what the editing instruction requests**
Not assess any unintended modifications beyond the instruction; such evaluations fall under separate criteria (e.g., visual consistency).
Be careful, an edit may introduce visual change without fulfilling the actual instruction (e.g., replacing the object instead of modifying it)

Reasoning:
You must follow these reasoning steps before scoring:
1. Detect Difference: What has visually changed between Image A and Image B? (e.g., size, shape, color, state) In this step, you don't have to use information from the editing instruction.
2. Expected Visual Caption: Write a factual description of how the edited image should look if the instruction were perfectly followed.
3. Instruction Match:
Compare the observed differences in **1** to the expected change in **2**:
- Was the correct object modified (not replaced)?
- Was the requested attribute (e.g., size, color, state) modified as intended?
- Is the degree of modification accurate (e.g., “match size”, “an hour later,” etc.)?
4. Decision: Use the 1–5 scale to assign a final score.
Evaluation Scale (1 to 5):
You will assign an **instruction_score** with following rule:
- **5 Perfect Compliance**: The edited image **precisely matches** the intended modification; all required changes are present and accurate.
- **4 Minor Omission**: The core change is made, but **minor detail** is missing or slightly incorrect.
- **3 Partial Compliance**: The main idea is present, but one or more required aspects are wrong or incomplete.
- **2 Major Omission**: Most of the required changes are missing or poorly implemented.
- **1 Non-Compliance**: The instruction is **not followed at all** or is **completely misinterpreted**

Example:
Instruction: Imagine what the apple will look like in a month.
{{
"instruction_score": 3,
"reasoning": "
1. Detecting differences:
In the original image, the apple shows a bright red color and a smooth surface. However, in the processed image, some black spots appear on the surface of the apple.
2. Expected visual description:
The originally shiny skin of the fruit would become dull and brown, with dark spots and blemishes appearing on it. As the water content decreases, the fruit will become wrinkled and shrink in size, and the flesh will become soft and spongy, and may even collapse in some areas.
3. Explanation:
Comparison: This instruction requires imagining the appearance of the apple after one month. This editing adds black spots to the apple's surface, which to some extent meets the requirements of the instruction, but the apple does not undergo dehydration or wrinkling changes. The core concept was attempted, but it was not fully achieved.
4. Decision:
Since only part of the apple's appearance reached the required level, this should be counted as 3 cases that partially meet the requirements."
}}

Input
Image A
Image B
Editing Instruction: {instruct}
Output Format
Look at the input again, provide the evaluation score and the explanation in the following JSON format:
{{
"instruction_score": X,
"reasoning": 1. Detect Difference 2. Expected Visual Caption 3. Instruction Match 4. Decision
}}

Figure 15: Prompt for *Instruction Following* (IF). The red words indicate the modifications made on Kris-Bench.

Knowledge Plausibility Template

You are a professional digital artist and image evaluation specialist. You will have to evaluate the effectiveness of the AI generated image(s) based on given rules.

You will be given:
1. **Image A**: the original image.
2. **Image B**: an edited version of Image A.
3. **Editing Instruction**: a hypothetical prompt describing how Image A should be transformed or imagined into Image B.
4. **editing_mode**: the fundamental editing regime that must be observed throughout the image-editing process; it dictates the governing condition under which all operations are performed—e.g., break, time, temperature, stretch, etc.

Objective
You must provide **scores** for the **edited image**:
- **Knowledge Score**: Given the instruction and original image, does the edited image reflect what should realistically happen based on the mode?

Knowledge Plausibility
Your Objective:
Evaluate whether the edited image, after applying the instruction to the original image, accurately reflects the real-world behavior described in the provided mode.
You must:
Ground your reasoning in the Real-World Knowledge Explanation based on the provided mode.
Focus only on whether the resulting image makes logical sense based on **physical, chemical, biological, or commonsense understanding**.
Not penalize issues unrelated to knowledge (e.g., visual polish or stylistic artifacts)

Reasoning Steps:
1. Detect Difference: What has visually changed between Image A and Image B? (e.g., size, shape, color) In this step, you don't have to use information from the editing instruction
2. Extract Knowledge Expectation: What visual outcome is expected if the instruction is applied, based on the provided mode?
3. Knowledge Match:
Compare the visual changes identified in Step 1 to the expected outcome in Step 2:
- Do the edits visually and logically match the real-world behavior?
- Is the cause-effect relationship shown correctly?
- Are key physical/chemical/biological phenomena depicted correctly?
4. Decision: Assign a knowledge_score from 1 to 5

Evaluation Scale (1 to 5):
- **5 Fully Plausible**: All visual elements follow real-world logic and match the explanation exactly.
- **4 Minor Implausibility**: One small deviation from expected real-world behavior.
- **3 Noticeable Implausibility**: One clear conflict with domain knowledge or the explanation.
- **2 Major Implausibility**: Multiple serious violations of the real-world logic.
- **1 Completely Implausible**: The image contradicts fundamental facts or ignores the explanation entirely.
If instruction is not followed (score ≤ 2), assign `knowledge_score = 1` and note: *"Instruction failure ⇒ knowledge invalid."*

Example 1: What if the rose is placed on the table for a month?
Editing Instruction: What if the rose is placed in the vase for a month?
Editing mode: time.
- **Compared to original image**, the rose is dry, droopy, and faded.
→ **Expected Caption**: The rose is dry, droopy, and faded.
"knowledge_score": 5,
"reasoning": "✓ The rose is dry, droopy, and faded."
Example 2: The glass breaks.
Editing Instruction: The glass breaks under the server pressure.
Editing mode: break.
- ✓ **Compared to original image**, Tiny cracks spread across the glass surface.
- ✗ The glass only slightly breaks instead of completely breaking, contradicting real-world behavior under server pressure.
→ **Expected Caption**: The glass shattered completely into small sharp pieces.
"knowledge_score": 3,
"reasoning": "✗ The degree of breakage is too small, contradicting real-world behavior under server pressure."

Input
Original Image
Edited Image
Editing Instruction: {instruct}
Editing mode: {mode}
Output Format
Provide both scores and clear reasoning in the following JSON format:
{{
"knowledge_score": X,
"knowledge_reasoning": 1. Detect Difference 2. Expected Knowledge Expectation 3. Knowledge Match 4. Decision
}}

Figure 16: Prompt for *Knowledge Plausibility* (KP). The red words indicate the modifications made on Kris-Bench.