PICABENCH: HOW FAR ARE WE FROM PHYSICALLY REALISTIC IMAGE EDITING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Image editing has achieved remarkable progress recently. Modern editing models could already follow complex instructions to manipulate the original content. However, beyond completing the editing instructions, the accompanying physical effects are the key to the generation realism. For example, removing an object should also remove its shadow, reflections, and interactions with nearby objects. Unfortunately, existing models and benchmarks mainly focus on instruction completion but overlook these physical effects. So, at this moment, *how far are we from physically realistic image editing?* To answer this, we introduce **PICABench**, which systematically evaluates physical realism across eight sub-dimension (spanning optics, mechanics, and state transitions) for most of the common editing operations (add, remove, attribute change, *etc*). We further propose the **PICAEval**, a reliable evaluation protocol that uses VLM-as-a-judge with per-case, region-level human annotations and questions. Beyond benchmarking, we also explore effective solutions by learning physics from videos and construct a training dataset **PICA-100K**. After evaluating most of the mainstream models, we observe that physical realism remains a challenging problem with large rooms to explore. We hope that our benchmark and proposed solutions can serve as a foundation for future work moving from naive content editing toward physically consistent realism.

1 INTRODUCTION

Recent advances in instruction-based image editing have brought remarkable progress (Wu et al., 2025a; Batifol et al., 2025; OpenAI, 2025; Google, 2025; ByteDance, 2025; Liu et al., 2025; Cai et al., 2025). In particular, with the emergence of unified multi-modal models (Deng et al., 2025; Lin et al., 2025; Wu et al., 2025b), they can seamlessly follow natural language instructions and produce visually compelling, semantically coherent edits. These systems have demonstrated strong generalization capabilities across diverse domains, establishing a new standard for controllable and high-quality image manipulation.

However, the realism of image editing depends not only on semantic accuracy but also on the correct rendering of physical effects. Even simple operations like object addition or removal often trigger complex interactions with lighting, shadows, and object support in the scene. Existing benchmarks overlook this limitation by solely emphasizing semantic fidelity and visual consistency. Although some recent benchmarks (Wu et al., 2025c; Li et al., 2025) attempt to probe scientific-plausible editing capabilities, their test cases diverge from common user-edit scenarios but focus on scientific domains with specific physical or chemistry knowledge. Consequently, we lack a clear understanding of *how far we are from physically realistic image editing.*

To address this gap, we introduce **PICA** (**P**hys**IC**s-**A**ware) Bench—a diagnostic benchmark designed to evaluate physical realism in image editing beyond semantic fidelity. Drawing on common requirements in real-world editing applications Taesiri et al. (2025), we categorize physical consistency into three intuitive dimensions that are often overlooked in typical editing tasks: *Optics*, *Mechanics*, and *State Transition*. These dimensions were selected to reflect common but under-penalized error types, such as unrealistic lighting effects, impossible object deformations, or implausible state changes. Together, they span eight sub-dimensions, each defined by concrete, checkable criteria: *Optics* includes light propagation, reflection, refraction, and light-source effects; *Mechanics* captures deformation and causality; and *State Transition* addresses both global and local state changes. This

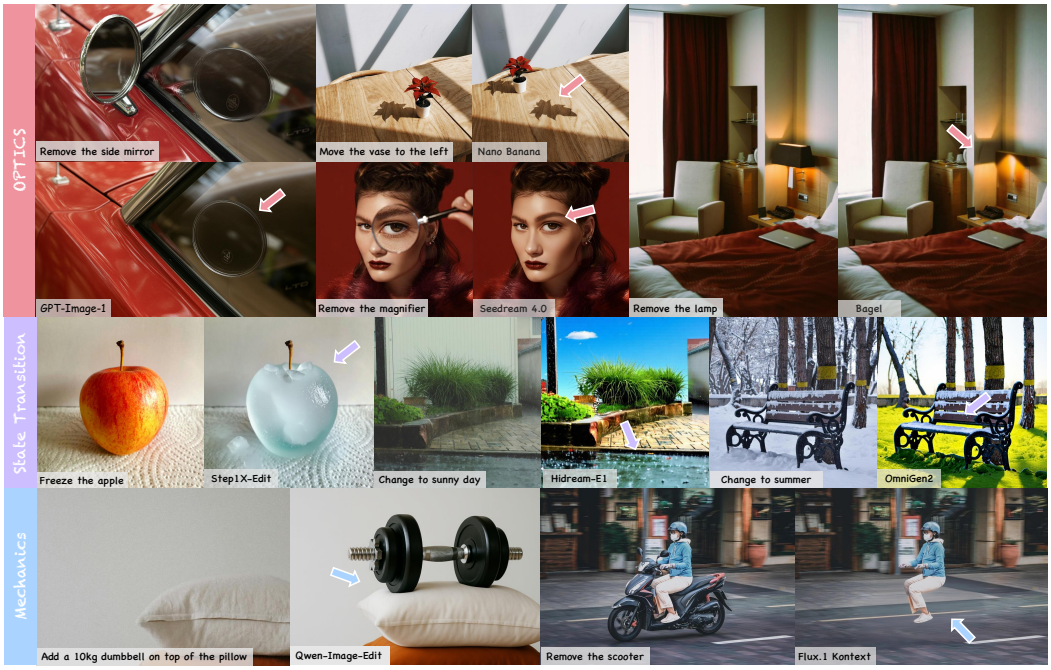


Figure 1: **Challenging cases from PICABench.** Despite providing instruction-aligned outputs, current SoTA models still struggle with generating physically realistic edits, resulting in unharmonized lighting, deformation, or state transitions with *common editing operations*.

fine-grained taxonomy facilitates systematic assessment of whether edited images adhere to principles such as lighting consistency, structural plausibility, and realistic state transitions. Together, it enables comprehensive evaluation and targeted diagnosis of physics violations in image editing models.

With the carefully curated test cases, evaluating the physical correctness remains challenging. We introduce **PICAEval**, a reliable and interpretable protocol tailored for physics-aware assessment. While existing VLM-as-Judge setups (Wu et al., 2025c; Niu et al., 2025; Sun et al., 2025; Zhao et al., 2025) offer a convenient way to automate evaluation, they typically rely on general prompts without grounding in physical principles. As a result, these setups often lack sensitivity to nuanced physical violations and may produce hallucinated judgments when faced with subtle or localized cues. Facing this challenge, PICAEval adopts targeted, per-example **Q&A** aligned with specific physical sub-dimensions, substantially improving diagnostic accuracy. To further reduce hallucination, we incorporate **grounded human-annotated key regions** (e.g., reflection surfaces, contact interfaces), directing the model's attention to physically relevant evidence. This protocol yields high agreement with human assessments, offering a reliable measurement for physical correctness.

Beyond evaluation, we provide a strong baseline by learning physics from videos. Specifically, we present **PICA-100K**, a synthetic dataset of 100k editing examples constructed from videos. Prior work (Yu et al., 2025b; Chen et al., 2025; Chang et al., 2025; Cao et al., 2025a) has shown that editing pairs derived from videos can enhance the quality and robustness of editing models. Motivated by recent advances in video generation approaching world-simulator (Wan et al., 2025), we design an automatic pipeline that integrates a text-to-image model as a scene renderer and an image-to-video model as a state-transition simulator. From the generated videos, we extract temporally coherent editing pairs and further recalibrate multi-level editing instructions using GPT-5. Our experiments shows that finetuning on PICA-100K significantly improves the baseline model's capability to generate physically realistic editing results without sacrificing semantic quality.

We benchmark 11 open- and closed-source image editing models across diverse architectures and scales. PICABench comprehensively distinguishes models based on their level of physical awareness, while PICA-100K effectively improves model performance. As shown in Fig. 1, modeling physical realistic transformations is still challenging for current SoTA models, which underlines the significance of advancing from semantic editing toward physically grounded image manipulation in the future. Our main contributions could be summarized as follows.

- We introduce **PICABench**, a comprehensive and fine-grained benchmark for physics-aware image editing. It covers diversified physical effects (eight sub-dimensions) and includes the great majority of commonly required editing operations in practical applications.
- We propose **PICAEval**, a region-aware, VQA-based evaluation protocol that incorporates human-annotated key regions to provide interpretable and reliable assessments for physical correctness, improving robustness to subtle errors compared to general scoring prompts.
- We construct **PICA-100K**, a large-scale dataset derived from synthetic videos, and show that fine-tuning existing models (*e.g.* FLUX.1 Kontext) on this dataset effectively enhances their physical consistency while preserving semantic fidelity.

2 Related Work

2.1 Instruction-based Image Editing Models

Recent advances in instruction-based image editing have led to substantial progress in controllable and diverse visual manipulation (Ye et al., 2025a; Yu et al., 2025a; Zeng et al., 2025; Jin et al., 2024; Huang et al., 2024). Prior approaches implement image editing in a training-free manner (Yang et al., 2023; Pan et al., 2023; Couairon et al., 2022). Recent training-based methods such as HiDream-E1.1 (Cai et al., 2025), Step1X-Edit (Liu et al., 2025), FLUX.1 Kontext (Batifol et al., 2025), and Qwen-Image-Edit (Wu et al., 2025a) improve edit quality, responsiveness, and instruction alignment, while unified frameworks (e.g., Bagel (Deng et al., 2025), OmniGen2 (Wu et al., 2025b), UniWorld-V1 (Lin et al., 2025)) integrate instruction-following, visual reasoning, and multi-task learning to support diverse tasks like free-form manipulation, future-frame prediction, multiview synthesis, segmentation, and composition. Closed-source systems (e.g., GPT-Image-1 OpenAI (2025), Seedream 4.0 (ByteDance, 2025), Nano-Banana Google) further demonstrate strong user-intent alignment and high visual fidelity across text-to-image and image-to-image workflows. However, despite these gains, most approaches prioritize semantic and perceptual quality and often neglect physical constraints, leading to artifacts such as unrealistic shadows, refractions, and deformations, underscoring the need for physics-aware editing.

2.2 Instruction-Based Image Editing Benchmarks

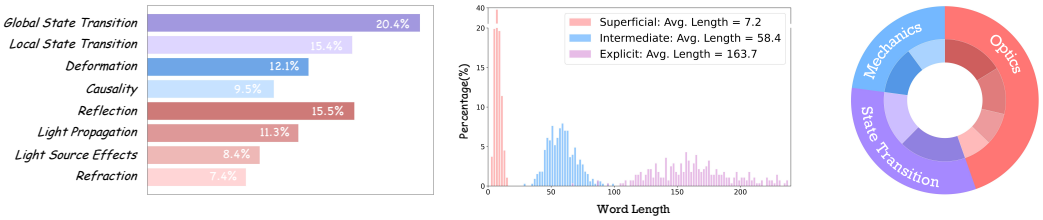
Instruction-based image editing benchmarks have evolved from early reliance on semantic (DINO, CLIP (Zhang et al., 2023; Wang et al., 2023; Ma et al., 2024)) and pixel-level (PSNR, SSIM) metrics, which capture similarity but miss fine-grained semantic alignment, to modern "VLM-as-a-Judge" evaluations (Wu et al., 2025c; Niu et al., 2025; Zhao et al., 2025; Sun et al., 2025; Ye et al., 2025b; Liu et al., 2025; Cao et al., 2025b) that use vision-language models to rate instruction adherence, perceptual quality, and realism across diverse, complex prompts. While these LLM-based approaches enable general multi-dimensional scoring, they are prone to overlooking physically implausible edits (e.g., unrealistic lighting, deformations, or object interactions) and can hallucinate, allowing visually appealing yet inconsistent outputs to score well. To close this gap, we introduce a physics-aware benchmark and the PICAEval—a region-grounded, QA-based metric that evaluates physical consistency through localized, interpretable assessments anchored to specific regions of interest.

3 Method

In this section, we first give an overall introduction of PICABench, a benchmark structured to evaluate physical realism in image editing. We then dive into the construction steps, begin with the data curation pipeline, which pairs diverse images with multi-level editing instructions. Next, we present PICAEval, a region-grounded evaluation protocol for reliable assessment. Finally, we propose PICA-100K, a synthetic dataset built from videos, and show how fine-tuning on it provides a strong baseline for improving physics-aware editing.

3.1 PICABench

We introduce the task coverage and overall statistics of PICABench. Our benchmark focuses on three core dimensions of physical realism: *Optics*, *Mechanics*, and *State Transition*, which reflect common



(a) Distribution of qa pairs (b) Word length distribution of edit instruction (c) Distribution of PICABench

Figure 2: **Statistics Analysis of PICABench.** PICABench is a comprehensive benchmark designed to evaluate physical realism of image editing models across eight sub-dimentinons. Fig. 2(a) shows distribution of QA pairs. Fig. 2(b) presents words length distribution of editing instruction across three levels of prompt. Fig. 2(c) provides a perspective on overall composition of PICABench.

yet overlooked failure modes such as unrealistic lighting, implausible deformations, and invalid state changes. As shown in Fig. 2(c), the benchmark includes 984 editing samples spanning these three dimensions, further divided into eight sub-dimensions with concrete and checkable criteria—ranging from optical effects, to mechanical plausibility, and to realistic state transitions.

Optics. This category evaluates whether edited images follow the basic physical rules of light, including how it casts shadows, reflects from surfaces, bends through transparent materials, and interacts with light sources. Edits should produce shadows, reflections, refractions, and light-source effects that align with the scene's geometry and lighting—matching shadow direction and occlusion, enabling view- and shape-dependent reflections, ensuring smooth background distortion through transparent media, and maintaining consistent color, softness, and falloff for added light sources. These effects, while often subtle, are key to making edits appear natural and physically believable.

Mechanics. This category evaluates whether edited objects remain mechanically and causally consistent with the scene. Deformation should follow material properties—rigid objects must retain shape, while elastic ones deform smoothly with consistent texture and geometry. Causality covers a broader range of physically plausible effects, including structural responses to force redistribution, agent reactions to added or removed stimuli, and environmental changes that alter object behavior, all of which must follow consistent physical or behavioral laws.

State transition. This category evaluates whether environmental and material changes unfold in a physically coherent manner, either across the entire scene or within localized regions. **Global state transitions**, such as changes in time of day, season, or weather, must update all relevant visual cues consistently—ranging from lighting and shadows to vegetation, surface conditions, and atmospheric effects. These changes require coordinated, scene-wide modifications that follow natural temporal or environmental progression. **Local state transitions**, on the other hand, involve targeted physical changes confined to specific objects or regions. These include phenomena such as wetting, drying, melting, burning, freezing, wrinkling, splashing, or fracturing. Edits must integrate smoothly with surrounding context, preserve material boundaries, and maintain plausible causal triggers.

3.2 Data Curation

To enable reliable, fine-grained evaluation of physics-aware image editing (PAIE), we curate benchmark entries that pair natural images with editing instructions explicitly designed to test physical consistency. Our data curation pipeline is aligned with the taxonomy in Sec. 3.1 and structured into two stages: *Data Collection* and *Edit Instruction Construction*. A visual overview is shown in Fig. 4.

Data collection. We begin by defining a structured vocabulary mapped to the eight sub-dimensions. To broaden the coverage, we use GPT-5 to expand this vocabulary into a rich keyword set encompassing materials, lighting contexts, and long-tail phenomena. We then use these keywords to retrieve candidate images from licensed and public sources. We prioritize visually diverse scenes that exhibit salient physical cues, such as directional lighting, transparent or reflective media, deformable objects, or phase-changeable substances. Human annotators filter duplicates and artifacts and tag applicable sub-dimensions for each image to support subsequent annotation.

Figure 3: **Statistics Analysis of PICABench.** We present illustrative examples from eight sub-dimensions. Key regions are annotated to help reduce hallucination for VLMs.

Instruction construction. Each retained image is paired with a human-written natural language instruction that induces a physics-relevant edit, grounded in the scene's physical affordances and designed to implicitly target a specific sub-dimension. To assess not only whether models can follow surface-level commands but also whether they can internalize and apply physical knowledge under varying prompt conditions, we construct three levels of instruction complexity: *superficial* prompts that issue plain edit commands without explanations, which probe models' intrinsic physical priors and align with realistic usage scenarios; *intermediate* prompts that include a brief rationale grounded in physical rules, serving as reasoning cues to activate physical knowledge; and *explicit* prompts that further describe the expected results of the edit, minimizing ambiguity to strictly assess visua capabilities. We use GPT-5 to expand each human-authored instruction into these three forms, followed by manual review to ensure clarity, factual correctness, and alignment with the visual context. For each sample, the benchmark retains a canonical version of the instruction.

3.3 PICAEVAL

Evaluating physics-aware image editing (PAIE) remains challenging. Unlike semantic fidelity or perceptual quality, physical realism is inherently contextual: it depends not only on the edited content but also on its alignment with the physical constraints implied by the original scene and instruction. Moreover, there is no reference image to serve as ground truth, and general prompting strategies such as "Is this edit correct?" often yield vague or hallucinated responses from VLMs.

To address this, we introduce **PICAEval**, a region-grounded, question-answering based metric designed to assess physical consistency in a modular, interpretable manner. Inspired by recent metrics like VDCscore (Chai et al., 2024), PICAEval decomposes each evaluation instance into multiple region-specific verification questions that can be reliably judged by a VLM. Each benchmark entry is paired with a curated set of spatially grounded yes/no questions designed to probe whether the edited output image preserves physical plausibility within key regions. These questions are tied to visually observable physical phenomena—such as shadows, reflections, object contact, or material deformation—and are anchored to human-annotated regions of interest (ROIs). This design

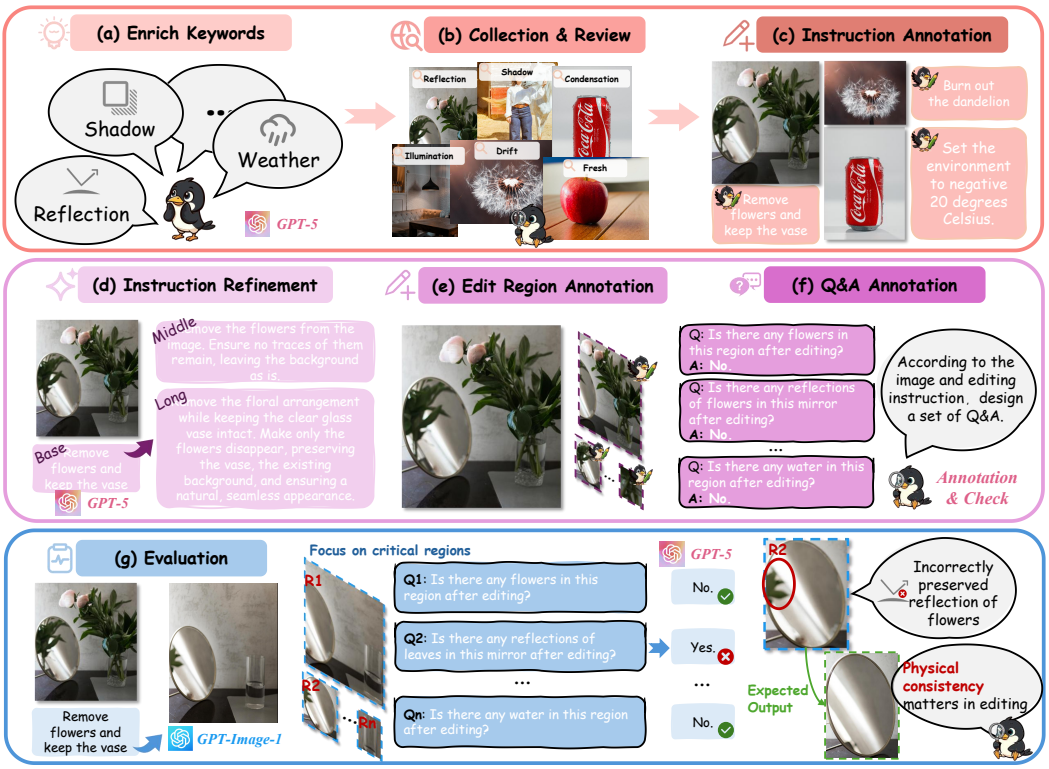


Figure 4: **Overall pipeline for benchmarks construction and evaluation**. (a–b) We enrich a physics-specific keyword set and retrieve diverse candidate images. (c–d) Human-written editing instructions are expanded into three levels of complexity using GPT-5. (e) Annotators mark physics-critical regions. (f) Spatially grounded yes/no questions are generated to evaluate physical plausibility. (g) During evaluation, VLMs answer each question with reference to the edited region.

encourages localized, evidence-based reasoning and reduces the influence of irrelevant image content on the VLM's judgment.

Evaluation pipeline. As illustrated in Fig. 4(e–f), the evaluation proceeds as follows: (1) Annotators mark key regions in the input image where physics-critical evidence is expected to appear post-editing (e.g., reflective surfaces, deformation zones, cast shadows); (2) Using the edit instruction and region, GPT-5 generates a set of 4–5 binary QA pairs per entry, which are then manually reviewed for clarity and coverage; (3) At test time, a VLM (e.g., GPT-5) is prompted with the edited image, instruction, region, and question, and produces an answer constrained to the visible content within the region.

PICAEval is computed as the proportion of questions for which the VLM answer exactly matches the reference label. Compared to direct prompting, this QA-based protocol offers three key advantages: (i) spatial grounding reduces hallucination, (ii) decomposition increases interpretability and robustness, and (iii) the format better mirrors how humans evaluate physical plausibility—through concrete, localized checks. We report quantitative comparisons and per-subdimension breakdowns to enable diagnostic analysis of physics-aware image editing capabilities in Sec. 4.

3.4 Strong Baseline: Learning Physical Realism from Videos

To address the limitations identified in Sec. 3.1, we introduce **PICA-100K**, a purely synthetic dataset designed to improve physics-aware image editing. Our decision to use fully generated data is driven by three primary motivations. **First**, prior work (Yu et al., 2025b; Chen et al., 2025; Cao et al., 2025a; Chang et al., 2025) has demonstrated that constructing image-editing data from video is an effective strategy for enhancing model performance, particularly for capturing real world dynamics. **Second**, building large-scale, real-world datasets tailored to physics-aware editing is prohibitively expensive and labor-intensive. **Third**, the rapid progress in generative modeling has unlocked new

Figure 5: **PICA-100K construction pipeline**. We first curate structured prompts for scene and subject composition, refined by GPT-5 and rendered using FLUX.1-Krea-dev for text-to-image generation. Motion-based edit instructions are created via GPT-5 and applied using Wan2.2-14B to synthesize short videos depicting physical transformations. The first and last frames, along with the edit instruction, form the image pairs for training.

possibilities: state-of-the-art text-to-image models (Labs, 2024) can now generate highly realistic and diverse images, while powerful image-to-video (I2V) models such as Wan2.2-14B (Wan et al., 2025) simulate complex dynamic processes with remarkable physical fidelity. Together, these generative priors enable the creation of training data with precise and controllable supervision signals, which are essential for training models to perform fine-grained, physically realistic edits. We find that fine-tuning the baseline on PICA-100K enhances the model's performance in real-world evaluation.

PICA-100K dataset. As shown in Fig. 5, we begin by constructing two structured prompt dictionaries: a Subject Dictionary and a Scene Dictionary, which include a wide array of subjects and environments (e.g., "a tea pot," "a black kitchen table"). These entries are paired using handcrafted text-to-image (T2I) templates and further refined using GPT-5, resulting in high-quality natural language instructions. The refined instructions are passed to the FLUX.1-Krea-dev (Lee et al., 2025) to generate static source images that are both visually realistic and semantically diverse.

Next, we generate motion-oriented instructions to simulate physical edits. This is accomplished by designing a series of I2V instruction templates, describing plausible motion-based changes such as rotations, movements, or tilts. These templates are expanded using GPT-5 to improve clarity and behavioral precision. The motion instructions (e.g., "remove the tea pot," "tilt the vase until it tips over," or "swing the lantern gently in the wind") are then applied to the corresponding images using Wan2.2-14B-I2V, which synthesizes short video clips depicting the intended physical transformations.

For each video, we extract the first and last frames to construct a (source, edited) image pair. These pairs, along with the corresponding instruction, are used to form supervision signals. GPT-5 is employed to annotate each pair automatically, labeling the final frame as the preferred output. This pipeline eliminates the need for manual labeling while maintaining high annotation consistency.

Our final dataset contains 100,000 instruction-based editing samples distributed across eight physics categories. The experimental results in Sec. 4 demonstrate that this pipeline can effectively generate high-quality data, significantly enhancing model performance on physics-aware image editing tasks.

Comparison with related works. PICA-100K is closely related to recent efforts Chang et al. (2025); Rotstein et al. (2025) that utilize video priors on image editing tasks. It differs from them in both motivation and methodology. ByteMorph (Chang et al., 2025) is primarily designed for non-rigid image editing, emphasizing visually salient motions such as articulation, deformation, and large pose or viewpoint changes. However, focus on large motions may hurt models' ability to keep non-edited region unchanged. Rotstein et al. (2025) proposes a training-free method, which focuses on zero-shot feasibility. It directly leverages a video generation model to simulate the editing process. Our work instead targets physical realism, which represents implicit physics principle of real world. Also, our data pipeline allows for more controllable generation where the non-edited regions remain stable.

Training paradigm. To demonstrate the effectiveness of PICA-100K, we fine-tune FLUX.1-Kontext-dev (Batifol et al., 2025), a 12B flow-based diffusion transformer for image editing. We employ LoRA (Hu et al., 2022) with a rank of 256 for fine-tuning. The model is trained using a batch size of 64 and optimized using the AdamW optimizer with a learning rate of 10^{-5}. The entire fine-tuning procedure is conducted over 10,000 optimization steps on 8 NVIDIA H200 GPUs.

Table 1: **Quantitative comparison on PICABench** evaluated by GPT-5 for instruction-based editing models, where Acc ↑, Con ↑, LP, LSE, GST, LST denote Accuracy (%) and Consistency (db), Light propagation, Light Source Effects, Global State Transition, Local State Transition respectively. and indicates the best and second best score in a category, respectively.

Model	LP		LSE		Reflection		Refraction		Deformation		Causality		GST		LST		Overall	
	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑
GPT-Image-1	55.65	18.81	66.75	20.13	64.05	19.07	43.36	18.64	62.42	20.37	50.00	20.08	73.87	36.20	58.72	22.71	61.46	22.95
Nano Banana	50.00	29.72	54.63	31.39	63.79	26.90	35.50	27.74	57.45	27.42	52.11	28.44	64.38	40.63	56.25	32.81	56.46	31.22
Seedream 4.0	54.77	25.49	65.80	28.27	68.69	23.82	38.75	27.00	59.11	27.27	51.05	26.71	71.14	36.76	58.98	33.20	60.84	29.05
Bagel	43.82	28.57	51.54	32.08	56.96	28.78	29.00	24.31	44.87	28.08	40.51	31.20	55.09	35.15	43.75	32.05	47.52	30.48
Bagel-Think	43.46	31.33	55.34	29.80	55.67	33.01	35.77	27.66	48.51	29.79	42.62	34.11	53.62	36.91	48.05	34.19	49.10	32.70
DiMOO	36.75	27.70	33.97	33.26	30.28	24.00	26.56	23.93	33.77	30.65	32.49	27.39	21.23	49.42	24.35	36.09	28.92	32.73
OmniGen2	46.29	20.46	49.41	28.85	58.76	25.10	27.37	23.22	44.21	25.51	40.93	28.05	49.71	38.52	34.90	27.80	45.28	27.84
Uniworld-V1	37.99	18.89	42.99	20.48	48.71	19.16	26.29	19.06	42.05	19.16	33.76	18.10	31.80	17.54	33.20	19.56	37.30	18.90
Hidream-E1	43.46	20.46	52.26	25.38	59.15	20.18	32.52	21.17	47.19	22.37	39.45	22.15	61.06	34.75	45.18	24.17	49.76	24.39
Step1X-Edit	42.05	29.46	53.68	31.26	58.89	29.50	30.89	31.58	48.34	31.51	49.79	32.09	58.02	35.43	47.53	30.19	50.42	31.47
Qwen-Image-Edit	52.12	22.03	59.14	26.28	64.82	23.80	35.50	26.54	50.50	26.42	48.95	24.94	63.60	36.72	54.17	28.44	55.62	27.42
Flux.1 Kontext	48.23	29.21	57.48	29.61	62.24	27.83	28.46	28.22	51.32	31.50	51.05	31.44	53.82	39.03	45.31	33.95	51.06	31.90
Flux.1 Kontext+*SFT*	49.12	30.42	59.38	30.69	64.95	28.37	30.89	28.40	50.17	31.74	46.62	31.87	51.17	40.82	44.79	34.41	51.88	32.71

Table 2: **Performance across different prompt specificity levels.** Model performance improves with prompt specificity.

Model	LP		LSE		Reflection		Refraction		Deformation		Causality		GST		LST		Overall	
	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑
Bagel-superficial	46.42	36.17	42.43	39.84	48.11	35.18	46.85	30.10	44.56	36.43	44.76	37.64	43.54	35.66	41.85	36.47	44.62	36.28
Bagel-intermediate	53.96	26.28	62.15	22.93	54.95	30.73	48.95	22.92	54.01	26.49	42.89	30.14	53.79	34.00	40.84	28.55	51.21	28.97
Bagel-explicit	58.03	17.63	66.90	18.44	57.90	20.95	51.40	18.68	58.29	18.81	53.37	23.71	66.12	33.53	57.00	21.41	59.56	23.06
Flux.1 Kontext-superficial	51.06	29.70	59.33	29.96	52.71	28.85	41.61	28.49	51.52	32.11	40.19	33.89	47.18	37.84	38.53	31.92	47.47	32.39
Flux.1 Kontext-intermediate	50.68	28.15	62.68	27.77	54.83	28.46	34.97	28.71	52.23	30.09	49.33	31.07	48.07	38.13	43.87	31.69	50.19	31.28
Flux.1 Kontext-explicit	57.64	27.77	66.90	25.25	59.67	27.35	40.21	26.96	56.51	28.67	57.74	27.99	65.68	36.11	52.24	29.57	59.11	29.40
Qwen-Image-Edit-superficial	58.99	22.61	64.08	27.43	62.62	23.86	60.84	25.06	54.37	24.65	48.18	26.79	62.85	35.42	52.53	25.75	57.99	27.24
Qwen-Image-Edit-intermediate	60.93	23.62	67.78	24.82	60.50	24.82	46.50	27.22	57.58	27.49	51.30	26.66	60.55	36.20	50.51	28.24	57.56	28.06
Qwen-Image-Edit-explicit	57.06	20.18	69.72	22.79	63.09	22.63	52.10	26.10	58.11	24.95	59.29	23.99	66.79	35.03	60.46	25.08	62.09	25.78
Nano Banana-superficial	55.71	29.32	58.80	32.21	56.37	27.67	48.25	27.31	59.54	28.31	54.17	30.17	59.51	39.00	52.24	30.18	56.32	31.29
Nano Banana-intermediate	56.48	28.89	62.85	30.70	58.96	28.06	46.85	28.12	61.68	27.96	60.12	28.12	61.59	38.56	53.68	30.54	58.96	30.74
Nano Banana-explicit	59.00	29.01	62.15	30.04	61.67	27.43	51.40	27.97	62.57	28.52	63.45	29.14	61.14	38.66	56.57	31.02	60.62	30.90

4 Experiment

4.1 Evaluation Details

We evaluate 11 closed- and open-source models, covering most recent image-editing and unified vision-language systems. Closed-source systems include GPT-Image-1 (OpenAI, 2025), Nano Banana (Google, 2025), and Seedream 4.0 (ByteDance, 2025). Open-source baselines include FLUX.1 Kontext (Batifol et al., 2025), Step1X-Edit (Liu et al., 2025), Bagel (Deng et al., 2025), Bagel-Think (Deng et al., 2025), HiDream-E1.1 (Cai et al., 2025), UniWorld-V1 (Lin et al., 2025), OmniGen2 (Wu et al., 2025b), Qwen-Image-Edit (Wu et al., 2025a), and DiMOO (Team, 2025). All input images are resized proportionally to a maximum resolution of 1024 on the longer side prior to evaluation. To ensure fairness and reproducibility, we run all models using their default settings from official repositories or web APIs on H200 GPUs.

For PICAEval, we first use the provided annotation masks to crop the edited region from the image. The cropped region is then resized proportionally to 1024 on the longer side before being passed to the VQA-based evaluator. This ensures standardized input size while preserving relevant physical cues within the editing region. We report results using both the current state-of-the-art closed-source model (GPT-5) and the leading open-source alternative (Qwen2.5-VL–72B) as VLM evaluator. For consistency evaluation, we compute PSNR over the non-edited regions by masking out the predicted edit area, thereby measuring how well models preserve the original content outside the editing scope.

4.2 Benchmark Results

We are still far from physically realistic image editing. Tab. 1 presents a comprehensive evaluation of existing methods. All open-source models score below 60 on the benchmark, and only the closed-source models—GPT-Image-1 and Seedream 4.0—slightly exceed this threshold. These results

Table 3: **Ablation Results.** We construct a real-video-based dataset (Mira400K). The model trained on Mira400K underperforms, highlighting the effectiveness of our targeted synthetic data pipeline.

Model	LP		LSE		Reflection		Refraction		Deformation		Causality		GST		LST		Overall	
	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑	Acc ↑	Con ↑
Flux.1 Kontext	66.06	29.46	70.28	30.76	71.40	28.31	43.46	29.59	58.66	31.48	57.29	32.30	63.41	36.39	56.16	32.01	61.98	31.66
+MIRA400K	63.11	27.53	70.08	29.75	70.90	26.87	44.65	28.64	60.61	29.77	53.06	30.59	61.97	38.23	54.37	30.69	60.60	30.71
+PICA100K	69.45	29.92	73.98	31.19	74.60	27.99	47.34	28.92	64.87	31.54	57.95	32.34	66.37	37.94	60.15	32.42	65.19	31.99

underscore a persistent gap in the ability of current image editing models to generate physics-aware and physically realistic outputs.

The gap between understanding and physical realism. Among open-source models, unified architectures consistently underperform compared to dedicated image editing models. Although unified MLLMs attempt to integrate visual understanding and generation within a single framework, the presumed advantage of enhanced world understanding does not translate into improved physical realism. This suggests that stronger understanding alone is insufficient, and effectively coupling understanding with generation remains an open challenge. Tab. 2 presents performance across different prompt specificity levels. As shown in Tab. 2, model accuracy improves as prompts become more detailed. The decrease of consistency can be attributed to the trade-off between improving physical realism and preserving non-edited image regions. However, the gain from intermediate prompts is much smaller than that from explicit prompts. We speculate this is due to the lack of internalized physics principles, which prevents models from leveraging the additional information. Interestingly, the Bagel model outperforms Flux Kontext under explicit prompts, likely because its unified architecture enhances long-text comprehension. Notably, even with explicit prompts that explicitly specify the editing regions, the overall performance still remains below 60.

Video data helps physics learning. Fine-tuning FLUX.1-KONTEXT on our PICA-100K dataset yields consistent improvements across multiple dimensions of physical realism. As shown in Appendix A.5, our model consistently produces more physically plausible results, while other models often exhibit unrealistic lighting effects, implausible object deformations, or invalid state changes. Quantitative results in Tab. 1 further support this: our fine-tuned model achieves a +2.01% improvement in optics accuracy and a +0.27% gain in mechanics accuracy over the base model. In addition, it demonstrates better overall physical consistency, improving from 31.90% to 32.71%. These findings suggest that synthetic supervision signals derived from videos can effectively enhance a model's capacity for physics-aware image editing. They also validate the effectiveness of our video-to-image data generation pipeline in capturing diverse and complex physical phenomena. However, we observe a slight drop in State Transition Accuracy, possibly due to limitations in directly using first and last frames of a video to represent meaningful state changes. We plan to explore more fine-grained strategies to extract temporal context and leverage intermediate frames.

We also experimented with using real video data to construct an image editing dataset. Following the data pipeline of Unireal (Chen et al., 2025), we employed Miradata (Ju et al., 2024) to generate 400K edited images (Mira400K) and trained the model under the same settings. However, as shown in Tab. 3, the model trained on Mira400K performed even worse in overall accuracy. This further demonstrates the efficiency and effectiveness of our proposed data generation pipeline.

4.3 VALIDITY OF PICAEVAL

We conduct a human study using Elo ranking to further validate the effectiveness of PICAEval. As shown in Fig. 6, PICAEval achieves higher correlation with human judgments than the baseline. This result demonstrates that our per-case, region-level human annotations and carefully designed questions effectively mitigate VLM hallucinations, leading to outcomes that better reflect human preferences. Additional details of the human study are provided in Appendix A.4.

5 LIMITATIONS AND FUTURE DIRECTIONS

While our approach demonstrates clear benefits in physics-aware image editing, it has several limitations. First, the PICA-100K dataset, though effective, is built using a relatively simple generation pipeline and remains limited in scale. Second, our model is trained purely via supervised finetuning

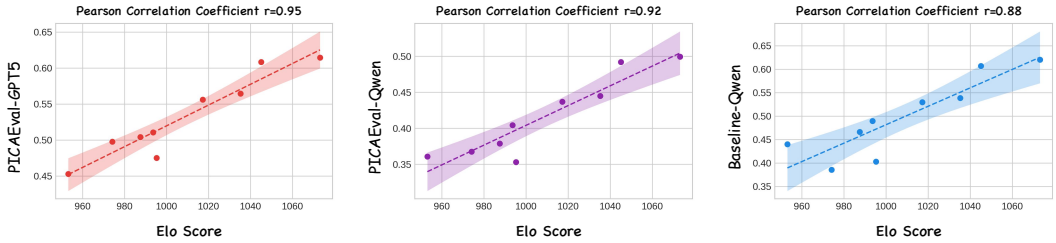


Figure 6: **Alignment between evaluation results and human preference.** We make Pearson correlation analysis between Elo scores from human study and different settings. PICAEval-GPT5, PICAEval-Qwen use GPT-5 and Qwen2.5-VL-72B as the evaluator respectively. Baseline-Qwen adopts Qwen2.5-VL-72B but without edit region annotations. Results show that incorporating stronger VLMs and region-level infomation yields higher alignment with human preference.

(SFT), which brings modest gains but may underexploit the full potential of data. Third, the current framework only supports single-image inputs, lacking the ability to incorporate multi-image or multi-condition contexts. In future work, we aim to enhance the data pipeline, explore RL-based post-training, and extend the model to support more expressive conditioning formats.

6 Conclusion

We present PICABench, a new benchmark for evaluating physical realism in image editing, along with PICAEval, a region-grounded, QA-based metric for fine-grained assessment. Our results show that current models, still far from producing physically realistic edits. To improve this, we introduce PICA-100K, a synthetic dataset derived from videos. Fine-tuning on this dataset significantly boosts physical consistency, demonstrating the promise of video-based supervision. We hope our benchmark, metric, and dataset can drive progress toward physics-aware image editing.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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

ByteDance. Seedream 4.0. 2025. URL https://seed.bytedance.com/en/seedream4_0/.

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.

Mingdeng Cao, Xuaner Zhang, Yinqiang Zheng, and Zhihao Xia. Instruction-based image manipulation by watching how things move. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2704–2713, 2025a.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, Bo Qu, Wenhai Wang, Yu Qiao, Dajuin Yao, and Yihao Liu. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding, 2025b. URL https://arxiv.org/abs/2507.14533.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024.

Di Chang, Mingdeng Cao, Yichun Shi, Bo Liu, Shengqu Cai, Shijie Zhou, Weilin Huang, Gordon Wetzstein, Mohammad Soleymani, and Peng Wang. Bytemorph: Benchmarking instruction-guided image editing with non-rigid motions. *arXiv preprint arXiv:2506.03107*, 2025.

Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12501–12511, 2025.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

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.

Google. Nano banana. 2025. URL https://gemini.google/overview/image-generation/.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

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.

Ying Jin, Pengyang Ling, Xiaoyi Dong, Pan Zhang, Jiaqi Wang, and Dahua Lin. Reasonpix2pix: instruction reasoning dataset for advanced image editing. *arXiv preprint arXiv:2405.11190*, 2024.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024.

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

Sangwu Lee, Titus Ebbecke, Erwann Millon, Will Beddow, Le Zhuo, Iker García-Ferrero, Liam Esparraguera, Mihai Petrescu, Gian Saß, Gabriel Menezes, and Victor Perez. Flux.1 krea [dev]. https://github.com/krea-ai/flux-krea, 2025.

Jialuo Li, Wenhao Chai, Xingyu Fu, Haiyang Xu, and Saining Xie. Science-t2i: Addressing scientific illusions in image synthesis, 2025. URL https://arxiv.org/abs/2504.13129.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

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.

Yiwei Ma, Jiayi Ji, Ke Ye, Weihuang Lin, Zhibin Wang, Yonghan Zheng, Qiang Zhou, Xiaoshuai Sun, and Rongrong Ji. I2ebench: A comprehensive benchmark for instruction-based image editing. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:41494–41516, 2024.

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

OpenAI. Gpt-image-1. 2025. URL https://openai.com/index/introducing-4o-image-generation/.

Zhihong Pan, Riccardo Gherardi, Xiufeng Xie, and Stephen Huang. Effective real image editing with accelerated iterative diffusion inversion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 15912–15921, 2023.

Noam Rotstein, Gal Yona, Daniel Silver, Roy Velich, David Bensaid, and Ron Kimmel. Pathways on the image manifold: Image editing via video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7857–7866, 2025.

Kaiyue Sun, Rongyao Fang, Chengqi Duan, Xian Liu, and Xihui Liu. T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation. *arXiv preprint arXiv:2508.17472*, 2025.

Mohammad Reza Taesiri, Brandon Collins, Logan Bolton, Viet Dac Lai, Franck Dernoncourt, Trung Bui, and Anh Totti Nguyen. Understanding generative ai capabilities in everyday image editing tasks. *arXiv preprint arXiv:2505.16181*, 2025.

Alpha VLLM Team. Lumina-dimoo: A unified masked diffusion model for multi-modal generation and understanding, 2025. URL https://github.com/Alpha-VLLM/Lumina-DiMOO.

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.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18359–18369, 2023.

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

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*, 2025b.

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*, 2025c.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18381–18391, 2023.

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025a.

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*, 2025b.

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, 2025a.

Xin Yu, Tianyu Wang, Soo Ye Kim, Paul Guerrero, Xi Chen, Qing Liu, Zhe Lin, and Xiaojuan Qi. Objectmover: Generative object movement with video prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17682–17691, 2025b.

Bohan Zeng, Ling Yang, Jiaming Liu, Minghao Xu, Yuanxing Zhang, Pengfei Wan, Wentao Zhang, and Shuicheng Yan. Editworld: Simulating world dynamics for instruction-following image editing. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12674–12681, 2025.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*, 2023.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Hao Li, Zicheng Zhang, Guangtao Zhai, Junchi Yan, Hua Yang, Xue Yang, and Haodong Duan. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *arXiv preprint arXiv:2504.02826*, 2025.

A MORE DETAILS OF PICABENCH

A.1 TASK DEFINITION

A.1.1 OPTICS

Light propagation requires shadows that are geometrically consistent with the dominant light source, including direction, length, softness, and occlusion. Typical failure modes include misaligned or missing cast shadows and flat shading that ignores occluders.

Reflection consistency demands view-dependent behavior for specular highlights and mirror reflections. Mirror images must preserve pose and depth; highlight positions should vary with surface curvature and viewpoint. Failures include "floating" reflections or highlights that remain fixed despite evident shape or view changes.

Refraction requires continuous, coherent background distortion through transparent or translucent media. When edited objects involve glass or water, background edges should bend and scale according to interface geometry, with preserved edge continuity. Discontinuous refractive boundaries or inverted distortions indicate violations.

Light-source effects evaluate whether new light-introducing edits (like "add a lamp") are consistent with the global illumination context—color casts, shadow penumbra, and brightness falloff should integrate naturally with the scene. Common issues include mismatched color temperatures, overly hard shadows, or inconsistent falloff relative to distance.

A.1.2 MECHANICS

Deformation assesses whether shape changes respect expected material properties. Rigid objects should not bend plastically; elastic deformations should be smooth and bounded. Texture and patterning should warp consistently with geometry rather than tear or duplicate. For instance, changing a chair's height should not collapse its frame or produce rubber-like bending.

Causality requires physically plausible contacts and supports under gravity. Edited objects should not float, interpenetrate, or rest in unstable equilibria (e.g., a heavy object balanced on a non-supporting point). Support relations must imply load transfer and stability. Violations include hovering objects, impossible stacking, and intersecting geometries that break solidity.

A.1.3 STATE TRANSITION

Global transitions affect the entire scene (e.g., day-to-night, dry-to-wet, solid-to-molten). Changes must propagate consistently: illumination color and intensity should update across surfaces; wetness should alter reflectance and darkening on all relevant materials; phase changes should be coherent and, when implied, justified by scene-level cues (e.g., a pervasive heat source). Inconsistencies include night skies with daylight shadows or partial melting without corresponding global evidence.

Local transitions involve spatially confined edits (e.g., adding steam, charring an edge, or melting a corner). These effects must integrate with nearby context and causal cues. Steam implies heat and moisture and may induce local condensation; flames produce light spill and secondary reflections; partial melting should respect material boundaries and continuity. When localized changes ignore surrounding context or violate material behavior, the edit becomes physically implausible.

A.2 MORE SCORE RESULTS

Tab. 4 lists the performance of models on PICABench, evaluated by Qwen2.5-VL-72B (Bai et al., 2025). It can be seen that the general rule and conclusion are similar to those suggested by Tab. 1: Most models have very low scores (below 60), indicating a fatal gap in the ability to generate physics-aware images.

A.3 DETAILS OF BENCHMARK METRICS

Table 4: **Quantitative comparison on PICABench** evaluated by Qwen2.5VL-72B for instruction-based editing models, where Acc, Con, LP, LSE, GST, LST denote Accuracy (%) and Consistency (db), Light propagation, Light Source Effects, Global State Transition, Local State Transition respectively. ■ and ■

Model	LP		LSE		Reflection		Refraction		Deformation		Causality		GST		LST		Overall	
	Acc	Con	Acc	Con	Acc	Con	Acc	Con	Acc	Con	Acc	Con	Acc	Con	Acc	Con	Acc	Con
GPT-Image-1	54.59	18.81	51.07	20.13	44.85	19.07	45.26	18.64	52.48	20.37	41.14	20.08	56.75	36.20	47.66	22.71	49.94	22.95
Nano Banana	45.05	29.72	38.72	31.39	44.07	26.90	36.86	27.74	48.68	27.42	41.77	28.44	48.53	40.63	44.40	32.81	44.50	31.22
Seedream 4.0	52.30	25.49	53.92	28.27	50.64	23.82	41.46	27.00	45.70	27.27	42.41	26.71	53.62	36.76	47.79	33.20	49.22	29.05
Bagel	36.22	28.57	33.02	32.08	36.98	28.78	34.42	24.31	35.26	28.08	37.34	31.20	36.50	35.15	31.64	32.05	35.28	30.48
Bagel-Think	42.40	31.33	37.53	29.80	36.60	33.01	37.40	27.66	37.58	29.79	36.71	34.11	40.22	36.91	35.68	34.19	38.12	32.70
DiMOO	31.27	27.70	19.24	33.26	22.42	24.00	26.29	23.93	29.47	30.65	30.17	27.39	15.66	49.42	26.04	36.09	24.20	32.73
OmniGen2	42.58	20.46	35.63	28.85	41.11	25.10	33.60	23.22	35.93	25.51	34.39	28.05	36.20	38.52	28.65	27.80	36.08	27.84
Uniworld-V1	33.04	18.89	24.47	20.48	30.67	19.16	21.14	19.06	31.95	19.16	28.06	18.10	18.40	17.54	27.86	19.56	26.68	18.90
Hidream-E1.1	37.81	20.46	32.54	25.38	38.40	20.18	30.62	21.17	35.10	22.37	32.49	22.15	43.44	34.75	34.51	24.17	36.74	24.39
Step1X-Edit	38.52	29.46	34.44	31.26	39.18	29.50	27.64	31.58	39.57	31.51	39.24	32.09	43.74	35.43	32.94	30.19	37.88	31.47
Qwen-Edit	47.88	22.03	44.66	26.28	43.94	23.80	40.11	26.54	41.72	26.42	38.82	24.94	47.16	36.72	41.54	28.44	43.70	27.42
Flux.1 Kontext	41.34	29.21	39.43	29.61	44.72	27.83	28.46	28.22	39.07	31.50	50.63	31.44	40.51	39.03	36.46	33.95	40.44	31.90
Flux.1 Kontext+*SFT*	45.76	30.42	42.28	30.69	46.13	28.37	32.25	28.40	41.89	31.74	47.05	31.87	40.90	40.82	37.76	34.41	41.96	32.71

We provide detailed definition of accuracy and consistency as follows. Let N be the total number of annotated QA pairs, $\hat{a}_i$ be the VLM-predicted answer for the i-th question, a_i be the reference answer, and $\mathbb{I}(\cdot)$ the indicator function. The accuracy is defined as:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{a}_i = a_i) \tag{1}$$

We use psnr of non-edited region as consistency. For each image pair, we compute the PSNR over non-edited pixels, using a binary mask $M_i(p)$ where $M_i(p) = 1$ denotes an edited pixel and $M_i(p) = 0$ otherwise. Define the non-edited region as $\Omega_i = \{p \mid M_i(p) = 0\}$, where p indexes pixels. Let I_i^{src} be the source image and I_i^{edit} the edited image.

The MSE (mean squared error) over the non-edited region is:

$$\text{MSE}_i = \frac{1}{|\Omega_i|}\sum_{p\in\Omega_i}\left\|I_i^{\text{edit}}(p) - I_i^{\text{src}}(p)\right\|_2^2 \tag{2}$$

Then, the per-sample consistency score (PSNR) is:

$$\text{Con}_i = 10\cdot\log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}_i}\right) \tag{3}$$

Finally, the dataset-level consistency is computed as the average across all N samples:

$$\text{Con} = \frac{1}{N}\sum_{i=1}^{N}\text{Con}_i \tag{4}$$

Here, MAX is the maximum pixel value (e.g., 255 for 8-bit images).

A.4 Detailed Human Evaluation Protocol

Study setup. We use the Rapidata[1] platform to conduct pairwise human preference comparisons for evaluating image editing quality. Each trial presents a reference image and two model outputs (A/B) under a fixed unified instruction:

> Select the image that more closely matches the editing instruction.

The A/B order is randomized per trial. Annotators are compensated at or above local minimum wage.

Datasets and models. We evaluate 10 models over the PICABench dataset at three difficulty levels (*superficial*, *intermediate*, *explicit*), forming 45 unordered model pairs per item. For each difficulty,

[1] https://www.rapidata.ai/

we sample 50 items via stratified sampling over the `physics_law` taxonomy. Each item yields 45 comparisons, each judged by 3 annotators, resulting in 20,250 votes per split.

Elo computation. To aggregate preferences, we use a robust Elo rating system. For a match between model A and B with current ratings (R_A, R_B), the expected win probability of A is:

$$E_A = \frac{1}{1 + 10^{\frac{R_B - R_A}{S}}}, \tag{5}$$

where $S = 400$ is the scaling factor.

Given the vote ratio $s_A \in [0, 1]$ for model A, with $s_B = 1 - s_A$, the ratings are updated as:

$$\begin{aligned} R'_A &= \max(R_{\min},\ R_A + K_{\text{eff}}(s_A - E_A)), \\ R'_B &= \max(R_{\min},\ R_B + K_{\text{eff}}(s_B - E_B)), \end{aligned} \tag{6}$$

where $K_{\text{eff}} = K \cdot \frac{v}{5}$ adjusts for vote count $v = v_A + v_B$, and $K = 24$ is the base step size.

Robust aggregation. To reduce order effects and improve stability, we shuffle the comparison stream and re-run Elo updates for $T = 50$ rounds. The final Elo score for model m is computed as:

$$\bar{R}_m = \frac{1}{T}\sum_{t=1}^{T} R_m^{(t)}, \quad \sigma_m = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(R_m^{(t)} - \bar{R}_m\right)^2}. \tag{7}$$

Parameter setting. Table 5 summarizes the Elo configuration used in all human evaluations.

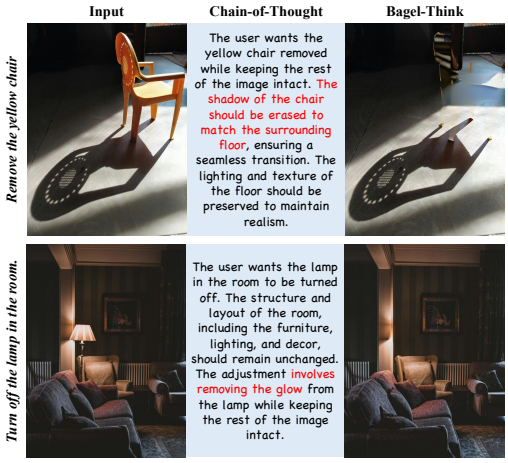


Figure 7: Examples of Bagel's reasoning trace.

Table 5: Elo parameter setting.

Parameter	Value
Initial Elo rating	1,000
Elo scaling factor S	400
Base K-factor	24
Minimum Elo rating $R_{\min}$	700
Number of rounds T	50
Votes per match	3
Model pairs per item	45
Items per difficulty	50
Benchmark splits	3
Total comparisons per split	6,750
Total votes per split	20,250

A.5 More visualization

Fig. 8-11 presents generated images of various models prompted by samples in our PICABench. The prompts cover all eight physics laws and three complexity levels. They demonstrate that the performance of these models varies considerably in complying with physical laws.

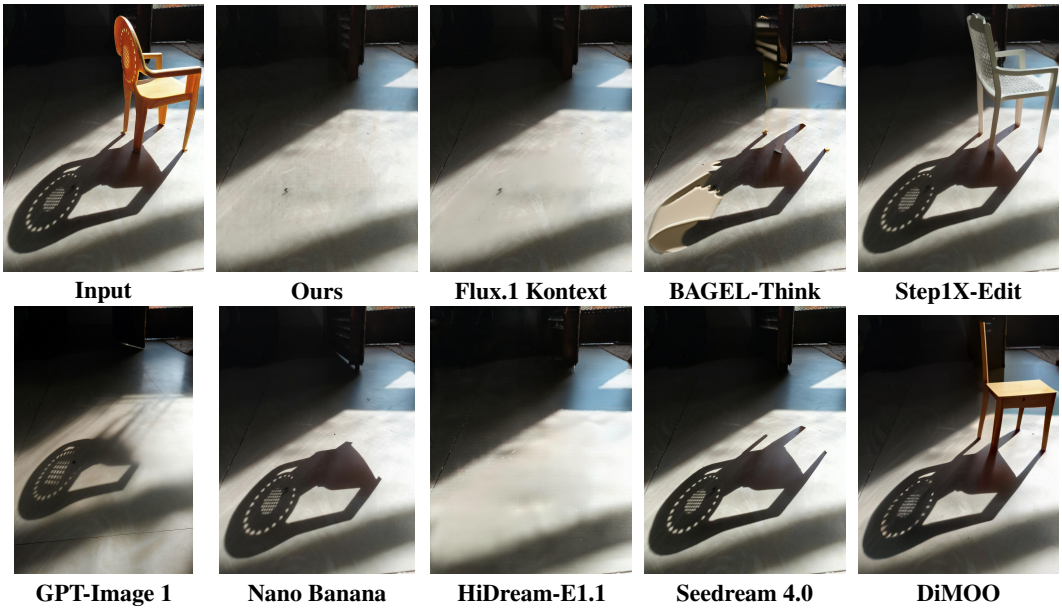
Most models either just perform superficial edits and ignore the physics law, or completely fail to understand the instruction. Only a few models, including ours, can yield physically plausible images in most cases. Therefore, the ability to follow physical laws is crucial but lacking in most models, and by PICABench we hope to draw the community's attention to this critical problem.

Moreover, we show Bagel's think process in Fig. 7. As shown in Fig. 7, model successfully reasons the correct results in its chain-of-thought, yet fails to execute them in the generated image.

A.6 System Prompt for QA Generation

To generate QA pairs, we design a system prompt as follows.

Superficial Prompt: Remove the yellow chair



Superficial Prompt: Move the potted plant to left side of the table.

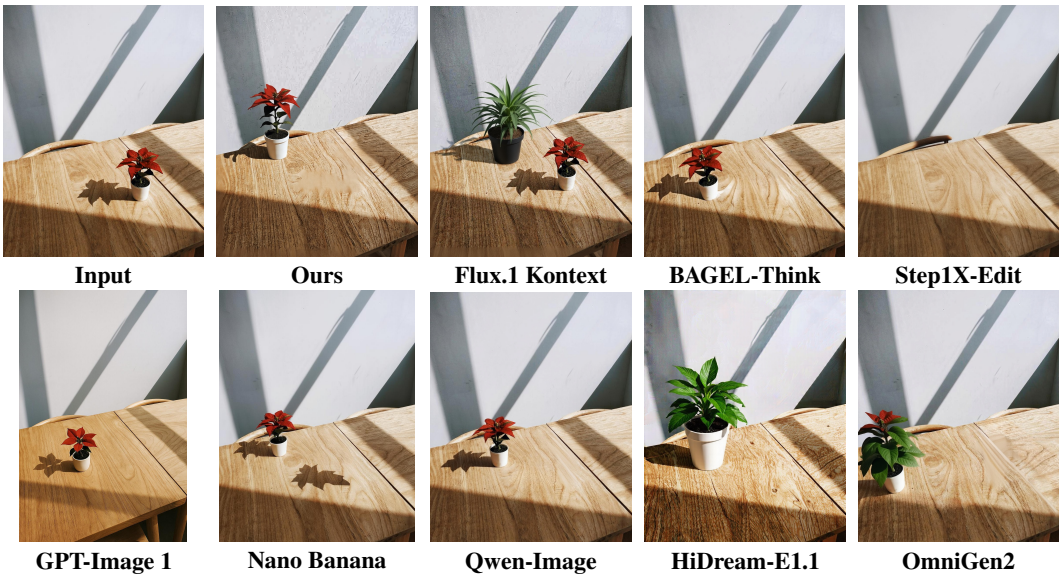
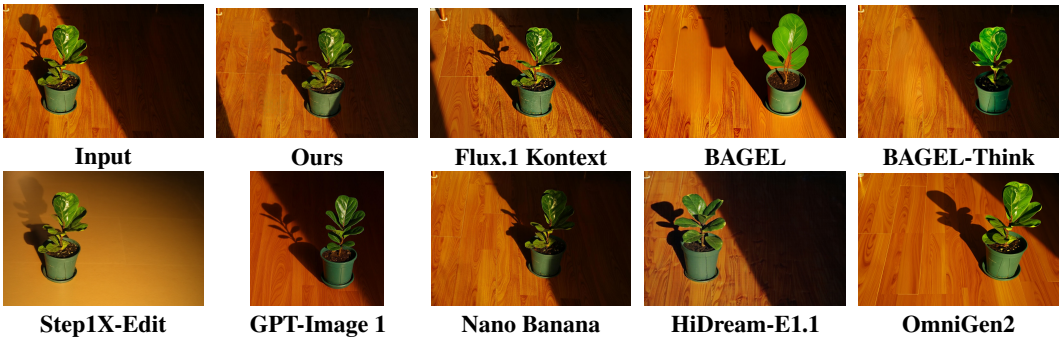


Figure 8: Examples of how models follow the law of light propagation in optics (superficial propmts).

System Prompt for QA Generation

You are an expert in image editing evaluation. Your task is to generate specific, targeted QA pairs to assess the success of this image editing task.
EDITING TASK CONTEXT:
- Edit Instruction: {`edit_instruction`}
- Physics Law: {`physics_law`}
- Operation Type: {`operation`}

Intermediate Prompt: Move the potted plant as a whole to the right side of the image, keeping it upright on the flat ...



Explicit Prompt: Reposition the dark ceramic mug together with its round cork coaster from its current spot to the right half of ...

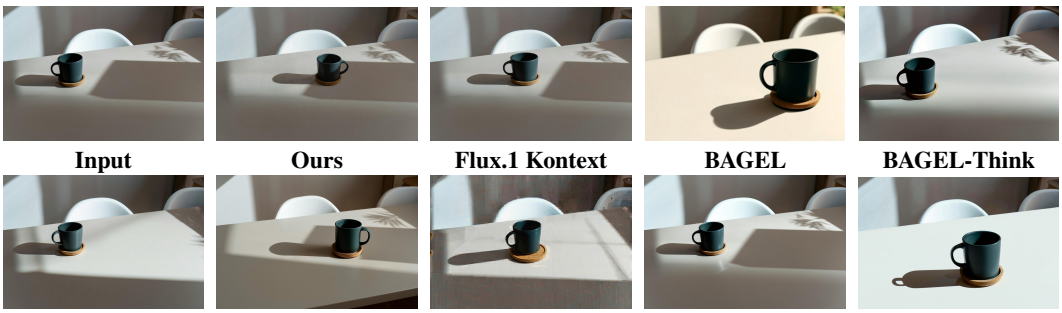


Figure 9: Examples of how models follow the law of light propagation in optics (intermediate & explicit propmts).

CRITICAL CONSTRAINT:
The evaluator will ONLY see the final edited image and the edit instruction. They **CANNOT** see the original image. Therefore, all questions must be answerable based solely on the final image.
GENERATE QUESTIONS FOR TWO CATEGORIES:
1. EDITING COMPLETION ASSESSMENT
Your goal: verify that the specific changes requested in the instruction are visible in the final image.
- Always **explicitly localize** the target object using a locator phrase *within* the noun phrase.
- Locator phrases may use: *position* (left/right/top...), *relative position*, *ordinal* (leftmost...), *attributes* (color/size...), *relationships* (attached to...).
- Focus on directly observable characteristics in the result.
2. PHYSICS CONSISTENCY ASSESSMENT
Your goal: evaluate whether the final image respects the laws of {`physics_law`}.
- Check for physically impossible or unrealistic arrangements.
- Assess object states, positions, contacts, shadows, reflections, etc.
- Evaluate *current* physical state only, not the editing process.
MANDATORY SINGLE-CRITERION RULE
- Each question must test **exactly one** observable predicate.
- **Do not** use “and”, “or”, “while”, “both”, etc.
- Connectors may be used inside locator phrases only.
- Valid predicates: present/absent, is color X, located at Y, touching, casting shadow, number equals N.
QUESTION FORM GUIDELINES
- **Removal**: Ask for absence, e.g., “Is the [locator] [object] absent?”
- **Addition**: Ask for presence.
- **Move**: Ask for new position relative to anchor.
- **Attribute**: Ask for color/texture/text on the object.

Superficial Prompt: Turn on the lamp on the bedside table.

Superficial Prompt: Turn off the lamp in the room.

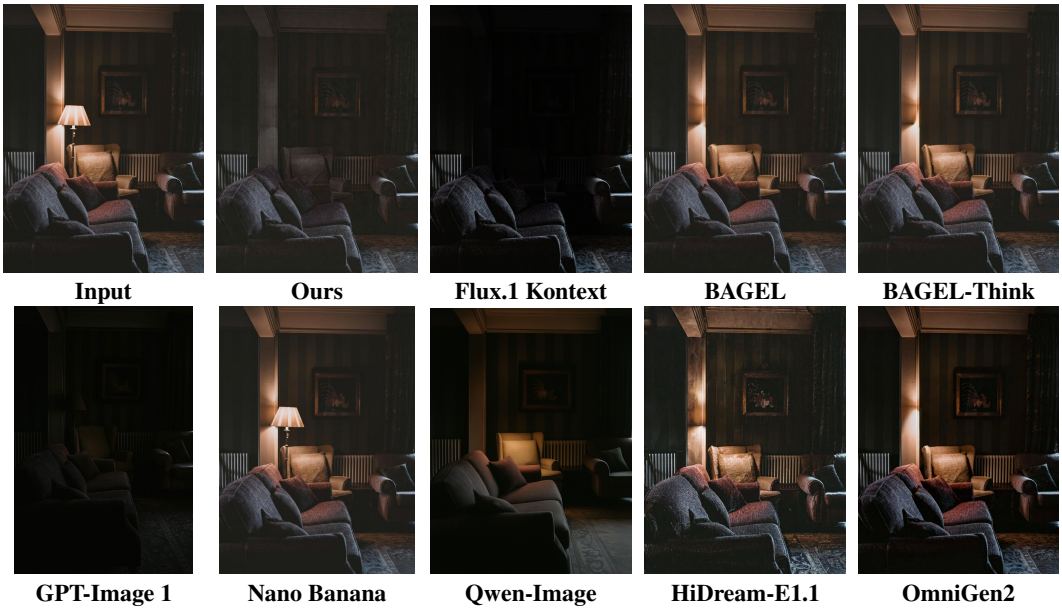


Figure 10: Examples of how models follow the law of light source effects in optics (superficial propmts).

- **Count**: Ask about exact number of localized targets.
- Use clear and concrete language. Avoid vague terms like “some”, “appears to”, “looks like”.

REQUIREMENTS

- Keep questions concise and clear.
- Use simple language; split complex checks.

- Avoid ambiguity and ensure single interpretation.
- Use locator phrases when categories appear multiple times.
- Frame questions positively.
- Cover all key aspects with multiple atomic questions (each addressing a different predicate).

CRITICAL: Every answer must be `"Yes"` or `"No"` — no other values are acceptable.

OUTPUT FORMAT:

```
{
  "Editing Completion QA": [
    {"question": "...", "answer": "Yes"},
    {"question": "...", "answer": "No"}
  ],
  "Physics Consistency QA": [
    {"question": "...", "answer": "Yes"},
    {"question": "...", "answer": "No"}
  ]
}
```

BAD EXAMPLES (DO NOT OUTPUT)
- "Is there a table?" (ambiguous)
- "Is the central table removed and is the floor clean?" (two predicates)

GOOD EXAMPLES
- "Is there a round wooden table in the center foreground?"
- "Is there a small blue cup on the right edge of the desk?"
- "Is there a traffic cone placed on the left side of the crosswalk?"
- "Is the leftmost of the two vases red?"
- "Is the shadow of the lamp cast toward the lower-right, consistent with a top-left light source?"

Final Reminder:
Questions must evaluate the *final image state*, not the editing history.
Avoid rewording the same question multiple times — each question must test a **different** aspect.

B More details about PICA-100K

B.1 System Prompt for Image-to-Video Captioning

To generate physics-aware captions for image-to-video generation, we design a system prompt that instructs the model to describe one physically plausible, visually salient content change observable over 3–5 seconds. The model is not allowed to reference the source image, prompt, or editing. The full system prompt used is shown below.

System Prompt for I2V Caption Generation

You are an expert writer of image-to-video captions (3–5 s).
You will receive ONE input image. DO NOT ask questions. DO NOT mention "image/photo/edit/prompt".

Goal
- Produce ONE concise motion caption that creates a VISUALLY OBVIOUS content change consistent with the physical law.
- "Content change" means change the state of light source (add/remove/move/change color or intensity), add/remove/move/replace an object, or alter a local/global material state.
- The camera is secondary: keep camera static unless a tiny move is necessary for visibility.

Thinking Steps (internal only)
1) Parse the scene: pick 1–2 salient objects; identify light, surfaces, supports, deformables, reflective/refractive media.
2) Choose ONE change that the specified physical law can plausibly cause. Prefer highly visible ones.
3) Make it measurable: include motion/visual cues (e.g., "slides right by half its width", "shadow doubles").
4) Keep identities and layout stable unless global changes are implied.
5) Default: camera static. If necessary, use one simple motion (e.g., slow push-in).

Law Playbook (pick one)
- Light_Source_Effects: turn on/off lamp; move lamp to affect all lit areas.
- Light_Propagation: move object to change shadow shape/position under key light.

- Reflection: place object near mirror; reflection should match highlights.
- Refraction: move object behind glass/water to create distortion.
- Deformation: add weight to soft object (e.g., pillow) to create indentation.
- Causality: remove support or add offset weight to cause collapse/tilt/slide.
- Local: change wet/dry/frozen/burnt/fractured/etc. with local visual cues.
- Global: simulate time/season/weather shift with coherent lighting/material change.

Constraints

- Duration: 3–5 seconds, single continuous shot.
- Primary change must be content-based.
- No object added/removed unless required by the change.
- Use specific, visible nouns (e.g., "mirror", "glass of water", "pillow").
- Use physics cue words (e.g., shadow, reflection, warping, indentation).
- Avoid stories or naming the law.
- End the sentence with camera state: e.g., "camera static".

Output Format

Return ONLY valid JSON:

```
{"i2v_prompt": "..." }
```

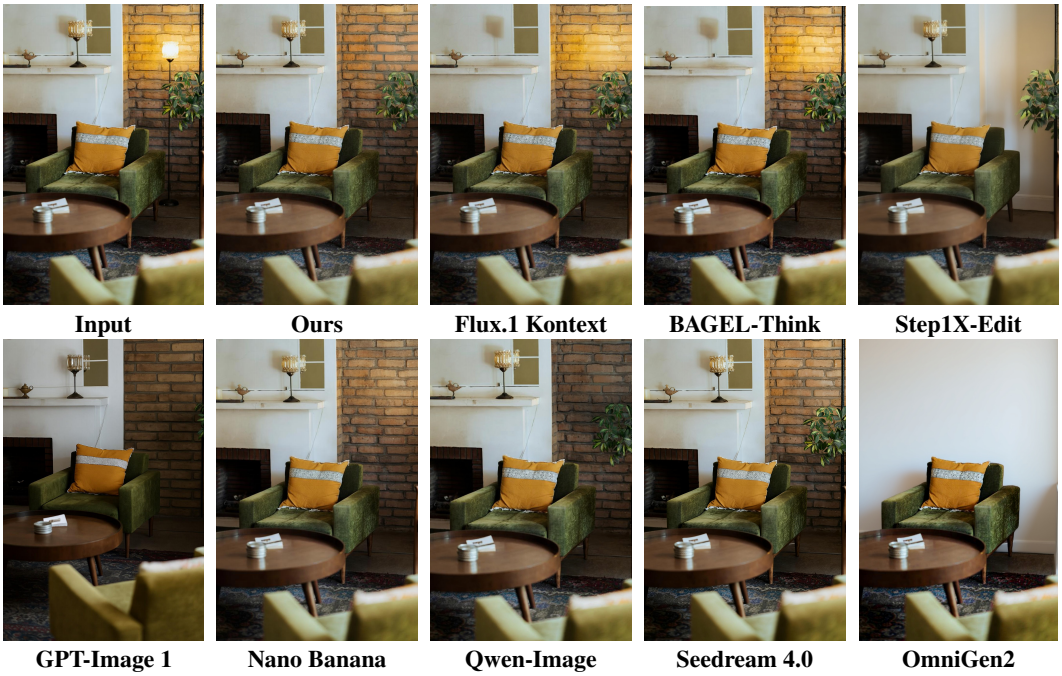
Examples (for reference only)

- "The desk lamp turns off and all previously lit areas fall into dimness, camera static."
- "The ceramic mug slides right by half its width and its shadow shortens under left key light, camera static."
- "The spoon moves behind the glass and warps due to refraction, camera static."
- "A dumbbell compresses the pillow, forming a deep indentation and partial rebound, camera static."
- "Dense droplets form on the fabric and darken the surface, camera static."
- "Light shifts to sunset; shadows grow longer and warmer, camera static."

B.2 Example of PICA-100K

Fig. 12 shows some examples in PICA-100K dataset. For each pair, we focus on the manifestation of physical laws.

Intermediate Prompt: Remove the tall floor lamp next to the plant on the right and also remove all illumination it produced. Update ...



Explicit Prompt: In the snowy dusk forest scene, a wooden post at center-left carries two lanterns, with the upper-left lantern currently glowing ...

Figure 11: Examples of how models follow the law of light source effects in optics (intermediate & explicit propmts).

Figure 12: Examples of PICA-100K dataset.