PhyEditBench: A Real-World Multi-Stage Benchmark for Physics-Aware Image Editing

Shengbin Guo*,[1], Shaokang He*,[2], Chaoyue Meng[3], Shengpeng Xiao[4], Xunzhi Xiang[1], Shaofeng Zhang[5], and Qi Fan[1,✉]

[1] Nanjing University [2] SDU [3] HRBEU [4] SDUTCM [5] USTC
shengbinguo2022@gmail.com, sk_he@mail.sdu.edu.cn, fanqi@nju.edu.cn

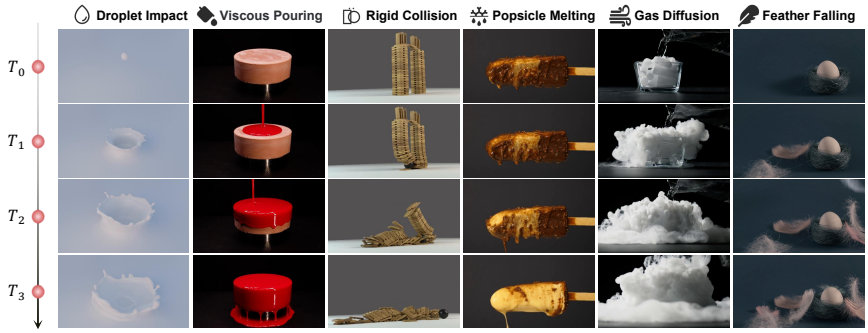


Fig. 1: High-quality, high-resolution, real-world examples from PhyEditBench, which encompasses a diverse range of complex physical processes.

Abstract. While instruction-based image editing, enabled by multimodal generative models, has advanced significantly, existing benchmarks lack comprehensive evaluation of physics-based reasoning—a critical capability for handling real-world scenarios. To address this, we introduce PhyEditBench, a benchmark designed to assess the physical understanding of editing models. Guided by a hierarchical taxonomy, we establish 4 primary classes and 12 subclasses. It comprises 238 **high-quality, high-resolution, real-world** instances—meticulously extracted from videos to capture authentic physical dynamics, alongside 35 synthetic Anti-Physics instances. Our empirical analysis of current SOTA editing methods exposes substantial limitations in their physics-based reasoning. We further propose a training-free baseline named **PhyWorld** that uses test-time scaling and a latent reduction strategy. PhyWorld outperforms comparable models and suggests that the video generation process can effectively serve as a reasoning mechanism for image editing. The project page is available at `https://github.com/Previsior/PhyEditBench`.

Keywords: Physics-aware image editing · Video-based benchmark · Evaluation metrics · Visual Reasoning · World Models

* Equal contribution, ✉ Corresponding author.

1 Introduction

The rapid evolution of multi-modal generative models has driven unprecedented breakthroughs in both image and video generation [4, 6, 41, 42, 51]. Building upon these powerful foundational models, instruction-based image editing has emerged as a crucial and highly practical downstream task. It enables users to manipulate specific visual content using natural language prompts [5, 10, 11, 20, 22, 27, 31, 34, 43, 44, 54, 55, 63]. Unlike generating content from scratch, the editing task requires a delicate equilibrium: models must accurately execute the given instructions while meticulously preserving the structural integrity and task-irrelevant semantics of the source image [2, 24, 37].

Early instruction-based editing models primarily focused on low-level visual transformations, such as global style transfer, color adjustment, or simple object replacement [5,20,36]. However, as user demands grow more sophisticated, there is an increasing need for editing models capable of handling complex instructions that require deep cognitive reasoning. To systematically assess and improve these capabilities, recent research has pivoted towards reasoning-centric image editing. A variety of reasoning benchmarks [17, 38, 57, 66] and editing frameworks have been introduced [19, 23, 28–30, 40, 49, 52, 60, 65], pushing the boundaries of image editing from pixel manipulation to semantic-level deduction.

Despite these advancements, existing reasoning benchmarks exhibit a critical limitation: they predominantly focus on spatial layout, logic puzzles, or attribute binding [14,17,38,56–58,66]. In essence, they often reduce the evaluation of "reasoning" to advanced instruction following and visual perception, but critically lack scenarios involving real-world physical dynamics. Consequently, these benchmarks fall short in accurately evaluating a model's intrinsic understanding of the real world. A robust editing model designed for real-world applications must not only understand *what* to change semantically but also *how* physical laws govern those state transitions.

To bridge this gap, we introduce **PhyEditBench**, a comprehensive and challenging benchmark specifically designed to assess the physics-based reasoning capabilities of image editing models. As shown in Fig. 1, PhyEditBench is meticulously constructed from high-quality, high-resolution real-world videos to capture authentic physical dynamics. We conduct an extensive empirical analysis of current state-of-the-art (SOTA) open-source and closed-source editing models on our benchmark. The evaluation reveals substantial limitations: traditional editing methods perform sub-optimally, frequently generating physically implausible artifacts or pasting objects without logical state transitions. This indicates that current image editing paradigms, which heavily rely on static statistical priors, struggle to reason about the dynamic evolution of the physical world.

Recently, the remarkable progress in world models and video generation has offered a promising new perspective [6,12,26,51]. Since video models are trained to predict subsequent frames, they inherently learn to simulate physical laws and maintain temporal causality. Inspired by this, we present a training-free framework named **PhyWorld** that leverages pretrained video generation models for physics-aware image editing. Following recent paradigms [43, 56], we formulate

the editing task as a temporal transformation process and interpret the intermediate generated frames as an implicit reasoning process. Furthermore, drawing inspiration from scaling laws at inference time [18], we construct our method upon an evolutionary Test-Time Scaling (TTS) algorithm and a Video Reward Model [33] to iteratively optimize and substantially enhance the output quality. By integrating a latent reduction strategy, our framework ensures strict fidelity to the source image while achieving physically plausible editing results.

In summary, our main contributions are three-fold:

- We introduce **PhyEditBench**, a novel, high-quality, real-world benchmark dedicated to evaluating physical reasoning in multi-modal image editing.
- We provide a comprehensive empirical analysis of current SOTA editing models, exposing their critical limitations in understanding and executing real-world physical dynamics.
- We propose a training-free editing baseline named **PhyWorld** that harnesses video generation models as reasoning engines. Augmented with Test-Time Scaling and a latent reduction strategy, our method outperforms comparable models in physically grounded editing tasks.

2 Related Works

Instruction-based Image Editing and Benchmarks. The rapid development of diffusion models [39, 42] and Multi-modal Large Language Models (MLLMs) [1,32,46] has catalyzed a paradigm shift in image editing. Early instruction-based editing methods, such as InstructPix2Pix [5] and MagicBrush [63], primarily focused on aligning text prompts with visual representations to perform low-level manipulations, including global style transfer, color adjustment, and simple object replacement. However, as user instructions become increasingly complex, these models often struggle to comprehend the underlying logic, relying instead on superficial semantic matching.

To address this, recent research has pivoted toward reasoning-centric image editing, empowering models with deep cognitive abilities to interpret complex, multi-step, or counterfactual instructions [16, 21, 29]. Consequently, a variety of benchmarks have been proposed to systematically evaluate these advanced capabilities. For instance, KRIS-Bench [57] and RISEBench [66] evaluate models across diverse cognitive dimensions, including spatial reasoning, conceptual knowledge, and logic puzzles. Similarly, UniREditBench [17] and WiseEdit [38] introduce a unified evaluation framework for reasoning-based editing. Despite their comprehensive coverage of spatial and semantic logic, these benchmarks exhibit a critical blind spot: they inherently lack the evaluation of *real-world physical dynamics*. They predominantly test whether a model knows *what* to change (e.g., modifying attributes or layouts) rather than *how* an object's state evolves under physical laws (e.g., gravity, fluid dynamics, or deformation). Our proposed **PhyEditBench** directly addresses this gap by introducing high-resolution, real-world instances dedicated exclusively to physics-based reasoning.

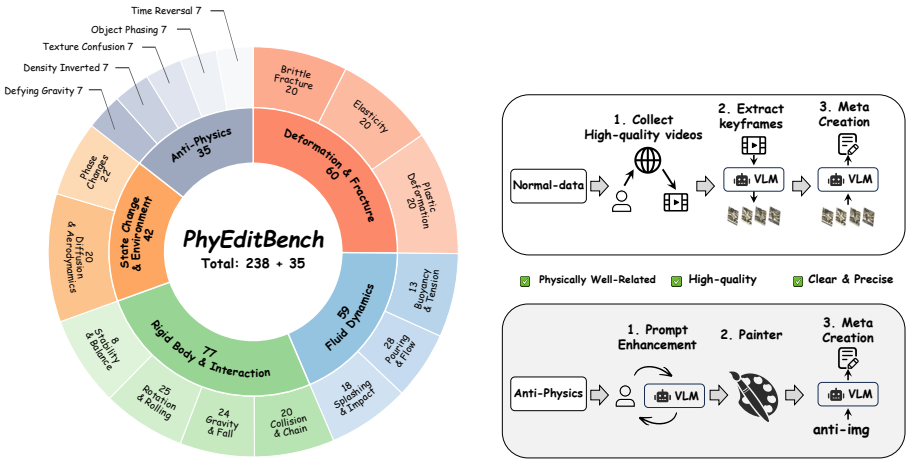
Table 1: Comparison between our PhyEditBench and previous related datasets. Existing image editing benchmarks lack real-world physical dynamics and multi-state granularity, while physical video datasets are not tailored for instruction-based image editing. PhyEditBench bridges this gap. (✓: Yes, ∼: Partial, ✗: No)

Category	Dataset	Editing Task	Real-world	High-Res.	Physics-aware	Multi-State
Image Editing Benchmarks	RISEBench [66]	✓	✗	✓	✗	✗
	KRIS-Bench [57]	✓	∼	✓	∼	✗
	UniREditBench [17]	✓	∼	✓	✗	✗
	WiseEdit [38]	✓	∼	✓	✗	✗
Video & Physics Datasets	CLEVRER [59]	✗	✗	✗	✓	✓
	Physion [3]	✗	✗	∼	✓	✓
	NewtonGen [61]	✗	✗	✓	✓	✓
	Action100M [8]	✗	✓	∼	✗	✓
	Sth-Sth-V2 [15]	✗	✓	✗	✓	✓
Ours	**PhyEditBench**	✓	✓	✓	✓	✓

Physical Reasoning in Vision Models. Understanding intuitive physics is a fundamental hallmark of machine intelligence. Early explorations in physical reasoning primarily focused on Visual Question Answering (VQA) and video prediction tasks. Datasets such as CLEVRER [59] and Physion [3] evaluate a model's ability to predict collisions, stability, and dynamic events. While foundational, these datasets are overwhelmingly constructed using 3D rendering engines (e.g., Blender, MuJoCo), resulting in simplified, synthetic environments that fail to capture the complexity, textures, and unpredictable nature of the real world.

Recently, the intersection of physical reasoning and generative AI has garnered significant attention. Researchers have observed that despite generating visually stunning images and videos, modern diffusion models frequently violate basic physical principles, producing hallucinations such as reversed gravity or unnatural fluid dynamics [62, 64]. To mitigate this, pioneering works like NewtonGen [61] and PhyGDPO [7] have attempted to explicitly inject Newtonian dynamics or physics-guided reward models into the generation process. However, these efforts are largely confined to text-to-video generation or rely heavily on external physical simulators. To date, there is a conspicuous absence of a benchmark designed to evaluate how well *general image editing models* understand and manipulate real-world physical states. As compared in Tab. 1, existing resources fall into two disjoint extremes: current editing benchmarks lack physical and multi-state granularity, while physical video datasets are ill-suited for instruction-guided editing tasks. By curating a taxonomy of real-world physical interactions, our benchmark serves as a crucial testbed to bridge this gap.

World Models and Video Generation. The recent emergence of large-scale video generation models, often referred to as "world models" (e.g., Sora [6], Stable Video Diffusion [4, 51]), has demonstrated unprecedented capabilities in simulating the physical world. By training on massive amounts of sequential data to predict subsequent frames, these models implicitly internalize physical laws, temporal causality, and object persistence [26, 41, 51].



(a) Benchmark Overview **(b)** Data Construction

Fig. 2: **(a)** shows our benchmark taxonomy and data volume. **(b)** illustrates the data construction pipeline.

This profound temporal and physical understanding presents a novel pathway for solving complex image editing tasks. Instead of treating editing as a static, single-step pixel transformation, recent state-of-the-art approaches have begun to formulate image editing as a temporal generation process. Methods such as CoF [50], Frame2Frame [43], and ChronoEdit [56] leverage pretrained video diffusion models to generate intermediate transition frames, effectively utilizing the video generation process as an implicit reasoning mechanism. Inspired by this paradigm, we leverage world models to address physics-based editing. Specifically, our training-free framework employs a test-time optimization approach to enhance generation quality. By transforming static physics editing instructions into a video-guided reasoning trajectory, our method harnesses the physically plausible reasoning capabilities inherent in pretrained video generation models. Consequently, it outperforms most traditional static editing models on physically demanding tasks, despite maintaining a compact 5B parameter size.

3 PhyEditBench

3.1 Overview

PhyEditBench is a benchmark for evaluating physics-based reasoning in instruction-guided image editing. Unlike existing benchmarks that primarily emphasize semantic correctness or local appearance edits, PhyEditBench targets *physical process understanding*: models must produce visually faithful edits that follow plausible physical dynamics and maintain scene invariants.

Benchmark composition. Guided by a hierarchical taxonomy, PhyEditBench contains 4 primary classes and 12 subclasses spanning common real-world physical phenomena. The benchmark includes 238 instances extracted from real-world videos to capture authentic physical dynamics, and an additional 35 synthetic *Anti-Physics* instances that deliberately violate physical laws. Fig. 2 (a) provides an overview of the taxonomy and data distribution.

Instance format. Fig. 3 (a) shows the composition of normal data points. Each instance in the physics-process subset is represented as a *four-state trajectory*: `input`, `intermediate 1`, `intermediate 2`, and `output`. These four frames are sampled from a single real video to depict a temporally coherent physical transition from an initial stable state to a final state. To evaluate both coarse and fine-grained physical understanding, each instance is annotated with: (i) one **global instruction** describing the overall edit goal from `input` to `output`; (ii) three **step instructions** describing the intended transition for each consecutive pair of states; (iii) concise **explanations** of the underlying physical process; and (iv) **invariants** (e.g., viewpoint and background) that should remain unchanged. This design supports two complementary evaluation settings: (i) *step-wise editing*, which tests whether a model can follow physically meaningful intermediate checkpoints; and (ii) *global editing*, which tests whether the model can infer plausible intermediate dynamics from a high-level instruction.

Why intermediate states? Physical processes often unfold gradually and contain latent constraints that are not captured by a single end-state comparison [4, 6, 9, 53]. Intermediate checkpoints in PhyEditBench enable fine-grained diagnosis: a model may reach the final state while violating the physical trajectory (e.g., implausible motion direction or inconsistent material evolution), which would be exposed by mismatches at `intermediate 1`/`intermediate 2`. Consequently, this multi-stage structure provides a significantly stronger and more rigorous probe for physics-grounded reasoning compared to standard one-shot edits.

3.2 Taxonomy of Physical Types

To systematically cover different real-world physics while maintaining the interpretability of the benchmark, we designed a hierarchical taxonomy. Inspired by previous works [3, 45, 59], each subclass is defined by a set of physical principles and representative scene patterns, enabling both category-level analysis and fine-grained diagnosis.

Deformation & Fracture. This primary class captures shape change and material failure under external forces, emphasizing how object geometry and integrity evolve with impact, compression, and elastic recovery. It includes three subclasses: *Brittle Fracture*, *Plastic Deformation*, and *Elasticity*.

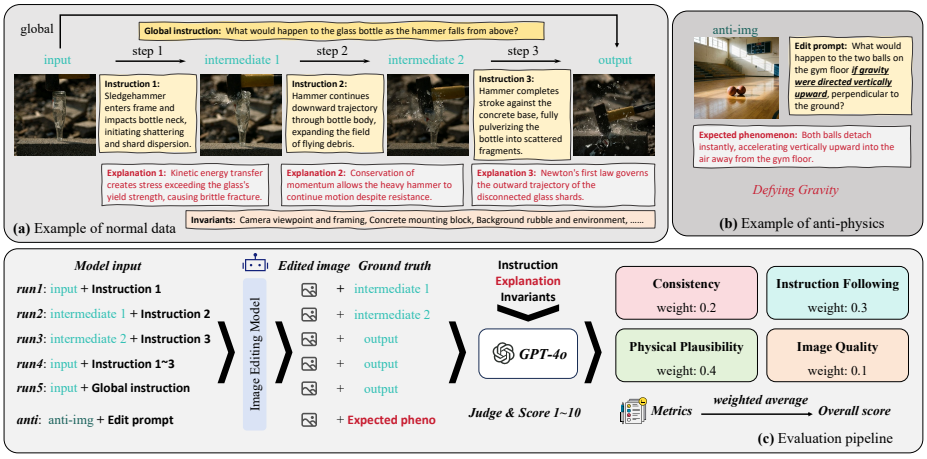


Fig. 3: **(a)** shows the form of normal data points, including pictures, editing instructions, explanations, and invariants. **(b)** depicts the data point form of anti-physics, including original images, editing instructions that violate physics, and expected phenomena. **(c)** illustrates our benchmark scoring pipeline.

Fluid Dynamics. This class focuses on complex liquid motion and multi-phase phenomena, where realistic edits require coherent free-surface behavior, plausible flow patterns, and physically consistent interactions between fluids and interacting objects. To systematically cover these dynamics, it includes three distinct subclasses: *Splashing & Impact*, *Pouring & Flow*, and *Buoyancy & Tension*.

Rigid Body & Interaction. This class covers rigid-body motion and contact-driven interactions governed by gravity, momentum transfer, friction, stability, and rotation. It includes four subclasses: *Gravity & Fall*, *Collision & Chain*, *Stability & Balance*, and *Rotation & Rolling*.

State Change & Environment. This class describes transformations induced by environmental factors such as heat transfer and air/gas dynamics, which often manifest as gradual changes with characteristic visual signatures. It includes two subclasses: *Phase Changes* and *Diffusion & Aerodynamics*.

Anti-Physics Data point. As illustrated in Fig. 3 (b), an Anti-Physics data point is meticulously designed to evaluate a model's ability to process counterfactual physical conditions. Each instance comprises three components: a source image, an edit prompt, and an expected phenomenon. The edit prompt explicitly injects a physical rule that contradicts common real-world experience. The inclusion of Anti-Physics instances is essential to decouple genuine physical reasoning from memorized statistical priors. Conventional models often rely on pre-training biases (e.g., assuming "knives always cut apples") and ignore explicit textual constraints [13, 35, 48, 59]. By introducing counterfactual scenarios, we force models

to suppress these inherent visual habits and execute strict deductive reasoning based solely on the prompt. Success in these instances thus demonstrates true dynamic physical deduction rather than mere pattern matching.

3.3 Data Construction

The overall pipeline is shown in Fig. 2 (b). For real-world instances, we sourced royalty-free videos from public stock platforms. We utilized a Vision-Language Model for coarse keyframe proposal and metadata drafting, followed by rigorous human verification and correction of temporal orders, keyframe alignment, and physical invariants to ensure high quality. For Anti-Physics data, the source images were synthesized using modern generative models, with the VLM formulating the counterfactual prompts and expected phenomena, followed by human auditing. Details are provided in the Appendix.

3.4 Evaluation Pipeline

Following previous work [17, 38, 57, 66], PhyEditBench evaluates editing models by generating edited images under standardized inputs and then scoring the results by four dimensions with a unified VLM-based judge (GPT-4o). Fig. 3 (c) illustrates the overall evaluation pipeline.

Model inputs and run types. For the physics-process subset, each instance contains four states (`input`, `intermediate 1`, `intermediate 2`, `output`), and we define five run types to probe both fine-grained and holistic understanding. Runs **TypeA–TypeC** perform step-wise editing on consecutive state pairs: `input` $\rightarrow$ `intermediate 1`, `intermediate 1` $\rightarrow$ `intermediate 2`, and `intermediate 2` $\rightarrow$ `output`, each taking the corresponding input image and step instruction to produce one edited image. **TypeD** applies the three-step instructions jointly starting from `input` and evaluates only the final state. **TypeE** performs global editing from `input` to `output` using the high-level instruction. For the Anti-Physics subset, the editing model takes a single input image and a counterfactual edit prompt, and outputs one edited image.

VLM-based scoring. Given the model output, GPT-4o scores each run type using the provided instruction or edit prompt, optional physical explanation, and invariants, together with the relevant reference images (ground-truth targets when available). Specifically, GPT-4o assigns a score in $\{1,\ldots,10\}$ with a brief rationale for four complementary dimensions: **1. Consistency**, preservation of invariants and non-target content. **2. Instruction Following**, faithfulness to the instruction, and alignment with the intended target state. **3. Physical Plausibility**, whether the edit reflects physically coherent dynamics consistent with the provided explanation or expected phenomenon. **4. Image Quality**, visual realism, and absence of artifacts. We compute the final score as a weighted average of these four dimensions, using fixed weights across all runs and both subsets. Details are provided in the Appendix.

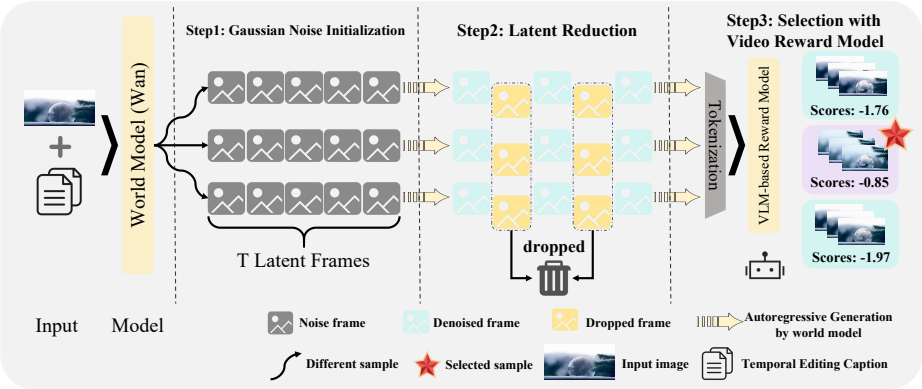


Fig. 4: Overview of the proposed PhyWorld pipeline. The editing process begins by initializing multiple Gaussian noise samples. During generation, a latent reduction strategy dynamically drops intermediate frames to compress the sequence and improve efficiency. Finally, a Video Reward Model evaluates all generated candidates, selecting the optimal sequence whose final frame is then extracted as the editing result.

4 Method

4.1 Overview

This work introduces **PhyWorld**, a strong **training-free** baseline that leverages pretrained video generation models for image editing. Aligning with [43, 56], we formulate editing as a temporal transformation, interpreting intermediate frames as a reasoning process similar to ChronoEdit [56]. Inspired by [18], we build PhyWorld upon an evolutionary Test-Time Scaling (TTS) algorithm and a Video Reward Model [33] to elevate output quality, while incorporating a latent reduction strategy to ensure efficiency. Our method achieves superior performance on our benchmark, outperforming same-category methods with comparable parameter sizes [43] as well as most existing open-source models. The overall architecture is depicted in Fig. 4.

4.2 Editing Prompt Enhancement

Conventionally, text-based image editing methods operate on an image-text pair (I_s, c), where the text c serves as guidance for the specific modifications to be applied to I_s. Similar to Frame2Frame [43], our framework transforms the static editing prompt into a format suitable for video generation. Specifically, we adapt the original instruction into a *Temporal Editing Caption* [43], designed to describe the transition process between the input and output images. Leveraging recent advances in vision-language models (VLMs), we employ Qwen-3.5 Max [47], a state-of-the-art VLM, to perform this enhancement. In particular, the model reasons about the actions described in the editing prompt, analyzes

their physical procedure, and extends them into a detailed description of the underlying physical process. The prompt is shown in the Appendix.

4.3 Test-Time Scaling

To enhance video generation quality, EvoSearch [18] proposes an evolutionary search algorithm for Text-to-Video tasks that identifies the optimal output. The search process is conducted via latent selection at specific denoising timesteps using a video reward model [33], formulated as:

$$R(\boldsymbol{x}_{t_i}) = \mathbb{E}_{\boldsymbol{x}_0 \sim p_0(\boldsymbol{x}_0|\boldsymbol{x}_{t_i})} \left[r(\boldsymbol{x}_0) | \boldsymbol{x}_{t_i} \right], \tag{1}$$

The method first randomly initializes a population of latents $\{\boldsymbol{x}_T^i\}_{i=1}^{k_{\text{start}}}$ at timestep T, where k_{start} denotes the initial population size in the population size schedule $k = \{k_{\text{start}}, k_{t_1}, \ldots, k_{t_j}, \ldots\}$. At each evolution timestep t_j in evolution schedule $\mathcal{T} = \{T, \ldots, t_j, \ldots, t_n\}$, EvoSearch performs evolutionary optimization on the current latents by denoising the latents into videos and scoring them according to the formulation above Eq. (1). The method then stores the latents at each evolution timestep along with their corresponding video scores. Through Top-K selection, tournament selection, and mutation operations, the method selects a subset of latents to continue the subsequent denoising process. Despite its improvements in video generation quality, this approach incurs substantial computational overhead, resulting in low efficiency. Furthermore, its application is restricted to Text-to-Video generation. In this work, we adapt this method to the Image-to-Video (I2V) architecture and strike a balance between the efficacy of EvoSearch and computational cost by performing a single search step at the end of the generation process. This strategy does not significantly compromise performance while substantially improving efficiency.

4.4 Video Generation

We leverage the pretrained Wan2.2 video generation model [51] as our backbone. Specifically, we employ its TI2V-5B variant, which utilizes the Wan2.2-VAE for efficient latent compression, achieving a spatial-temporal compression ratio of $4\times 4\times 16$. Formally, let C denote the input instruction and I denote the input image. The generation process initiates with 5 Gaussian noise samples, which undergo denoising conditioned on I. We utilize 121 frames as reasoning tokens and 30 sampling timesteps during generation. To optimize computational efficiency, we follow the approach in [56] by employing a method termed the latent reduction strategy, as depicted in Fig. 4. Specifically, this strategy is applied at sampling timesteps 10 and 20. While the initial 121 frames are encoded into 31 VAE latent tokens, the sequence length is reduced to 21 and 11 tokens at these respective timesteps by pruning intermediate latent states. After the generation process, the best output is determined by the video reward model [33] and the final frame of the output is selected as the edited image conditioned on I and C.

Table 2: Main experimental results of editing models.

Dimension	Model	closed-source			open-source								video-based		
		GPT-Image-1.5	Gemini-2.5	Seedream4.0	OmniGen2	InstructPix2Pix	FLUX.1-Kontext-dev	Step1X-Edit	Qwen-Image-Edit	UniWorld-V2	BAGEL	BAGEL-Think	Frame2Frame	PhyWorld	ChronoEdit-14B
normal data															
Avg.	Overall	8.23	6.51	8.47	5.36	5.61	5.27	6.79	6.71	7.08	6.09	6.43	5.38	6.43	**8.51**
Avg. by Metrics	Cons.	8.84	7.21	9.69	6.35	7.08	7.03	9.24	8.41	8.53	7.82	8.53	7.96	7.58	**9.67**
	Inst.	8.21	6.13	8.22	4.23	4.16	4.56	5.93	6.35	6.78	5.47	5.72	4.08	5.89	**7.98**
	Phys.	8.30	6.64	8.22	5.54	5.58	4.81	6.19	6.37	6.73	5.69	5.96	4.73	6.30	**8.42**
	Imag.	6.78	5.72	7.75	5.96	7.12	5.71	6.87	5.79	6.45	6.04	6.26	6.65	6.26	**8.13**
Avg. by Classes	Defor.	8.36	7.36	7.43	3.71	4.24	5.40	3.16	6.41	6.60	5.55	5.77	5.25	6.29	**8.38**
	Fluid.	8.30	6.79	9.04	6.32	6.41	5.37	7.20	7.38	7.43	6.43	6.82	5.43	6.55	**8.74**
	Rigid.	7.91	5.87	8.65	5.51	5.49	4.99	6.77	6.10	6.99	6.21	6.55	5.17	6.10	**8.32**
	State.	8.51	6.07	8.84	6.05	6.67	5.44	7.13	7.33	7.44	6.14	6.59	5.86	6.94	**8.79**
Avg. by Types	TypeA	8.16	6.55	8.60	5.44	5.51	5.53	7.16	7.01	7.43	6.01	6.53	5.62	6.59	**8.64**
	TypeB	8.29	6.52	8.60	5.40	6.23	5.51	7.44	7.09	7.41	7.06	7.08	6.04	6.59	**8.68**
	TypeC	8.35	6.43	8.74	5.30	5.97	5.36	7.16	6.83	7.46	6.87	7.01	5.70	6.34	**8.51**
	TypeD	8.32	6.68	8.27	5.54	4.98	5.02	6.02	6.36	6.46	5.49	5.83	4.35	6.27	**8.49**
	TypeE	8.01	6.37	8.15	5.08	5.36	4.94	6.15	6.27	6.64	4.99	5.72	5.17	6.36	**8.23**
anti-physics															
Avg.	Overall	7.04	7.07	6.95	4.74	4.45	6.07	5.34	6.16	5.79	5.11	6.16	3.81	**6.39**	4.99
Avg. by Metrics	Cons.	9.03	8.71	8.20	8.17	7.60	**8.60**	8.23	7.74	8.17	7.77	7.89	6.80	7.91	6.89
	Inst.	7.20	7.03	7.29	4.09	3.69	5.71	4.83	6.29	5.77	4.86	**6.34**	2.51	**6.34**	4.66
	Phys.	5.51	5.91	5.78	2.89	2.54	4.43	3.51	4.83	4.06	3.29	5.00	2.51	**5.29**	3.63
	Imag.	8.71	8.54	8.11	7.23	8.09	**8.63**	8.40	7.91	8.06	7.83	6.83	6.89	7.86	7.69

5 Experiments

5.1 Evaluation Models and Settings

To evaluate representative instruction-based image editing approaches, we benchmark a diverse set of models covering both closed-source and open-source systems, as well as different generation paradigms. For traditional image editing, we include three closed-source models: **GPT-Image-1.5** [22], **Gemini-2.5-flash-image** [10], and **Seedream4.0** [44], together with open-source models: **OmniGen2** [55], **InstructPix2Pix** [5], **FLUX.1-Kontext-dev** [27], **Step1X-Edit** [34], **Qwen-Image-Edit** [54], **UniWorld-V2** [31], and **BAGEL** [11]. Beyond conventional editors, we additionally evaluate video-generation-based editing methods: **Frame2Frame** [43], **ChronoEdit** [56], and **PhyWorld**, which perform image editing by leveraging the video generation process as an implicit reasoning mechanism. The detail information of each model can be referred to the Appendix.

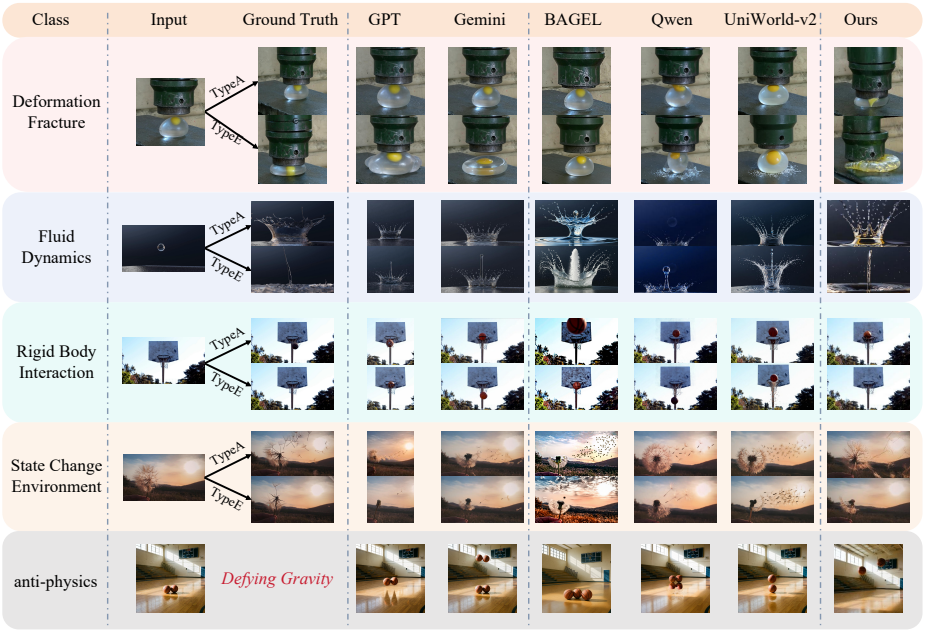


Fig. 5: Qualitative comparison on physical reasoning tasks across five distinct categories. Each row presents two editing variations (Type A/E). PhyWorld demonstrates superior physical plausibility and closer alignment with ground truth, notably excelling in the anti-physics scenario.

5.2 Main Results

Overall Performance. The empirical findings summarized in Tab. 2 encompass performance across both normal and anti-physical data subsets, offering a comprehensive view of model capabilities. Within the conventional data setting, ChronoEdit-14B secures the top position, with Seedream4.0 trailing by a narrow margin. Turning to other open-source alternatives, UniWorld-V2 emerges as a strong performer, whereas our approach attains a highly competitive ranking—remarkably, with the most compact architecture (5B parameters vs. 14B in ChronoEdit-14B and the similarly performing BAGEL-Think, and 19B in Step1X-Edit-V1P1). It is also worth highlighting that BAGEL-Think surpasses its predecessor BAGEL, implying that the tasks curated in our benchmark place substantial demands on physical reasoning proficiency.

Shifting focus to the anti-physical subset, a different picture emerges: neither closed-source nor open-source models demonstrate robust competence in tackling physically counterfactual scenarios. In this challenging regime, Gemini-2.5 maintains its leadership among proprietary systems, while PhyWorld distinguishes itself as the strongest open-source contender, closing the performance disparity relative to closed-source solutions. These findings validate that our framework effectively leverages intermediate video generation frames as reasoning tokens,

while the Test-Time Scaling (TTS) strategy unlocks the pre-trained model's inherent physical reasoning capabilities. Overall, the suboptimal performance exhibited by both open-source and closed-source models on our benchmark underscores a critical limitation: current image editing models require significant advancements in reasoning about physical processes.

Analysis by Metrics. Our protocol evaluates Consistency, Instruction Following, Physical Plausibility, and Image Quality. Physical Plausibility serves as the core metric for assessing real-world reasoning. In the conventional subset, GPT-Image-1.5 and ChronoEdit-14B lead the closed-source and open-source models, respectively. Notably, our training-free method secures a highly competitive place among open-source solutions. This performance is particularly encouraging considering that our framework operates without additional fine-tuning and relies entirely on a compact 5B Image-to-Video backbone architecture.

Analysis by Classes. Performance varies significantly across physical categories. As shown in Tab. 2, *Deformation & Fracture* and *Fluid Dynamics* are particularly challenging for most open-source static models due to complex topological changes. Conversely, closed-source models (e.g., Seedream4.0) and video-based models like ChronoEdit-14B exhibit remarkable proficiency. PhyWorld demonstrates robust, balanced performance across all categories, highlighting the versatility of video priors.

Analysis by Types. Evaluating across run types reveals the compounding difficulty of long-horizon reasoning tasks. Most models achieve peak performance in short-term, single-step transitions (TypeA/B) but degrade significantly in global (TypeE) or joint multi-step (TypeD) settings. This pattern exposes the vulnerability of traditional static models to error accumulation during extended physical processes. In contrast, video-based frameworks effectively mitigate this degradation by strictly adhering to temporal causality, underscoring the inherent advantage of continuous temporal modeling for complex state transitions.

5.3 Assessment of Evaluation Protocol

To ensure the reliability and fairness of our automated Vision-Language Model (VLM) evaluator, we conducted a human validation study. We randomly sampled instances across different temporal states of physical evolution (TypeA, TypeD, TypeE) as well as counterfactual scenarios (Anti-Physics). Human raters were provided with the source images, edit prompts, and the generated outputs from four representative models. They were instructed to rank the models across four dimensions: Visual Consistency (VC), Instruction Following (IF), Physical Plausibility (PP), and Image Quality (IQ). We then calculated the Kendall τ rank correlation coefficient to measure the alignment between human judgments and the VLM's automated rankings [25]. Details can be found in the Appendix.

As illustrated in Fig. 6, the empirical results validate our evaluation protocol by demonstrating strong human-model alignment across critical reasoning dimensions. Specifically, the Kendall tau rank correlation for instruction following

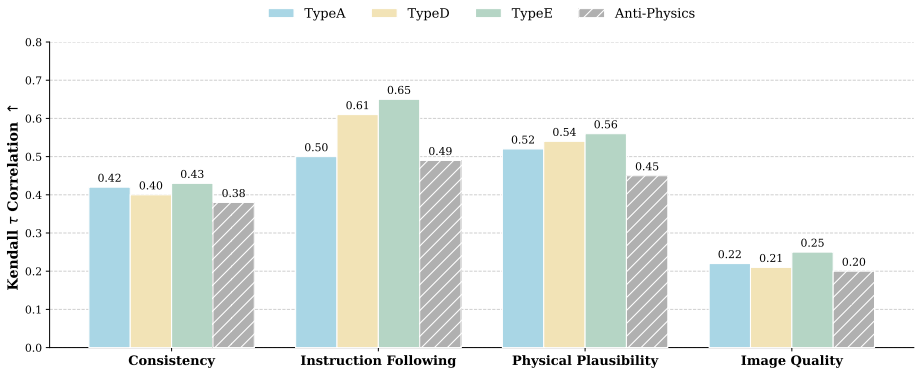


Fig. 6: Correlation between human and VLM evaluations across different physical stages and metrics. We report the Kendall τ rank correlation coefficient on four metrics.

and physical plausibility reaches up to 0.65 and 0.56, respectively, during the final output stage, logically increasing as the physical transformations become more visually pronounced. Furthermore, the evaluator maintains robust agreement when assessing counterfactual anti-physics scenarios, proving its competence in interpreting unnatural dynamics despite the inherent subjectivity of such tasks. Conversely, the correlation for low-level image quality remains notably lower. This divergence is a well-documented characteristic of modern multi-modal models, which excel at high-level semantic logic but inherently lack human sensitivity to subtle pixel artifacts. Ultimately, the substantial correlation in physical deduction and instruction execution definitively establishes our automated pipeline as a reliable, scalable, and physically grounded metric for PhyEditBench.

6 Conclusion

In this paper, we introduced **PhyEditBench**, a pioneering high-resolution benchmark designed to rigorously evaluate the physics-based reasoning capabilities of instruction-guided image editing models. Unlike previous datasets that focus on static semantic modifications, our benchmark encompasses diverse real-world physical dynamics across fine-grained temporal stages, alongside challenging counterfactual anti-physics scenarios. Our comprehensive evaluation of current SOTA models exposed their critical limitations in understanding intuitive physics and dynamic state transitions. To bridge this gap, we proposed PhyWorld, a training-free framework that harnesses the temporal causality of pre-trained video generation models as an implicit reasoning engine. By integrating test-time scaling with a latent reduction strategy, our method achieves comparable physical plausibility and visual consistency. Ultimately, we hope PhyEditBench will inspire future research toward equipping multi-modal generative models with a robust and dynamic understanding of the physical world.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China, under Grant Nos. 62192783, 62276128, and 62406140; the Young Elite Scientists Sponsorship Program by China Association for Science and Technology, under Grant No. 2023QNRC001; the Key Research and Development Program of Jiangsu Province, under Grant No. BE2023019; and the Jiangsu Natural Science Foundation, under Grant Nos. BK20221441 and BK20241200. The authors would like to thank the support of Huawei Ascend Cloud Ecological Development Project.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bar-Tal, O., Ofri-Amar, D., Fridman, R., Kasten, Y., Dekel, T.: Text2live: Text-driven layered image and video editing. In: European conference on computer vision. pp. 707–723. Springer (2022)
3. Bear, D.M., Wang, E., Mrowca, D., Binder, F.J., Tung, H.Y.F., Pramod, R., Holdaway, C., Tao, S., Smith, K., Sun, F.Y., et al.: Physion: Evaluating physical prediction from vision in humans and machines. arXiv preprint arXiv:2106.08261 (2021)
4. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
5. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
6. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Ramesh, A., Klar, M., Sohl-Dickstein, J., et al.: Video generation models as world simulators. OpenAI Technical Report (2024), `https://openai.com/research/video-generation-models-as-world-simulators`
7. Cai, Y., Li, K., Jia, M., Wang, J., Sun, J., Liang, F., Chen, W., Juefei-Xu, F., Wang, C., Thabet, A., et al.: Phygdpo: Physics-aware groupwise direct preference optimization for physically consistent text-to-video generation. arXiv preprint arXiv:2512.24551 (2025)
8. Chen, D., Kasarla, T., Bang, Y., Shukor, M., Chung, W., Yu, J., Bolourchi, A., Moutakanni, T., Fung, P.: Action100m: A large-scale video action dataset. arXiv preprint arXiv:2601.10592 (2026)
9. Chen, Z., Zhou, Q., Shen, Y., Hong, Y., Sun, Z., Gutfreund, D., Gan, C.: Visual chain-of-thought prompting for knowledge-based visual reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1254–1262 (2024)
10. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)

11. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
12. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7346–7356 (2023)
13. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
14. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
15. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE international conference on computer vision. pp. 5842–5850 (2017)
16. Han, F., Jiao, Y., Chen, S., Xu, J., Chen, J., Jiang, Y.G.: Controlthinker: Unveiling latent semantics for controllable image generation through visual reasoning. arXiv preprint arXiv:2506.03596 (2025)
17. Han, F., Wang, Y., Li, C., Liang, Z., Wang, D., Jiao, Y., Wei, Z., Gong, C., Jin, C., Chen, J., et al.: Unireditbench: A unified reasoning-based image editing benchmark. arXiv preprint arXiv:2511.01295 (2025)
18. He, H., Liang, J., Wang, X., Wan, P., Zhang, D., Gai, K., Pan, L.: Scaling image and video generation via test-time evolutionary search. arXiv preprint arXiv:2505.17618 (2025)
19. He, Z., Qu, X., Li, Y., Zhu, T., Huang, S., Cheng, Y.: Diffthinker: Towards generative multimodal reasoning with diffusion models. arXiv preprint arXiv:2512.24165 (2025)
20. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
21. Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., Zhou, J., Dong, C., Huang, R., Zhang, R., et al.: Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8362–8371 (2024)
22. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
23. Jiao, S., Lin, Y., Zhong, Y., She, Q., Zhou, W., Lan, X., Huang, Z., Yu, F., Yu, Y., Zhao, Y., et al.: Thinkgen: Generalized thinking for visual generation. arXiv preprint arXiv:2512.23568 (2025)
24. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6007–6017 (2023)
25. Kendall, M.G.: The treatment of ties in ranking problems. Biometrika **33**(3), 239–251 (1945)
26. Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Schindler, G., Hornung, R., Birodkar, V., Yan, J., Chiu, M.C., et al.: Videopoet: A large language model for zero-shot video generation. arXiv preprint arXiv:2312.14125 (2023)

27. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
28. Li, H., Jiang, L., Yan, Q., Song, Y., Kang, H., Liu, Z., Lu, X., Wu, B., Cai, D.: Thinkrl-edit: Thinking in reinforcement learning for reasoning-centric image editing. arXiv preprint arXiv:2601.03467 (2026)
29. Li, H., Zhang, M., Zheng, D., Guo, Z., Jia, Y., Feng, K., Yu, H., Liu, Y., Feng, Y., Pei, P., et al.: Editthinker: Unlocking iterative reasoning for any image editor. arXiv preprint arXiv:2512.05965 (2025)
30. Li, Y., Liu, L., Zhang, X., Xue, W., Luo, W., Guo, Y., Tian, Q.: Cogniedit: Dense gradient flow optimization for fine-grained image editing. arXiv preprint arXiv:2512.13276 (2025)
31. Li, Z., Liu, Z., Zhang, Q., Lin, B., Wu, F., Yuan, S., Yan, Z., Ye, Y., Yu, W., Niu, Y., et al.: Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. arXiv preprint arXiv:2510.16888 (2025)
32. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
33. Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Qin, W., Xia, M., et al.: Improving video generation with human feedback. arXiv preprint arXiv:2501.13918 (2025)
34. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
35. Marcus, G.: The next decade in ai: Four steps towards robust artificial intelligence. arxiv. arXiv preprint arXiv:2002.06177 (2020)
36. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
37. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6038–6047 (2023)
38. Pan, K., Chen, W., Qiu, H., Yu, Q., Bu, W., Wang, Z., Zhu, Y., Li, J., Tang, S.: Wiseedit: Benchmarking cognition-and creativity-informed image editing. arXiv preprint arXiv:2512.00387 (2025)
39. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
40. Qin, L., Gong, J., Sun, Y., Li, T., Yang, M., Yang, X., Qu, C., Tan, Z., Li, H.: Uni-cot: Towards unified chain-of-thought reasoning across text and vision. arXiv preprint arXiv:2508.05606 (2025)
41. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
43. Rotstein, N., Yona, G., Silver, D., Velich, R., Bensaïd, D., Kimmel, R.: Pathways on the image manifold: Image editing via video generation. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7857–7866 (2025)
44. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025)
45. Spelke, E.S.: Principles of object perception. Cognitive science **14**(1), 29–56 (1990)
46. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
47. Team, Q.: Qwen3.5: Accelerating productivity with native multimodal agents (February 2026), `https://qwen.ai/blog?id=qwen3.5`
48. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic compositionality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248 (2022)
49. Tian, Y., Yang, L., Yang, J., Wang, A., Tian, Y., Zheng, J., Wang, H., Teng, Z., Wang, Z., Wang, Y., et al.: Mmada-parallel: Multimodal large diffusion language models for thinking-aware editing and generation. arXiv preprint arXiv:2511.09611 (2025)
50. Tong, C., Chang, M., Zhang, S., Wang, Y., Liang, C., Zhao, Z., An, R., Zeng, B., Shi, Y., Dai, Y., et al.: Cof-t2i: Video models as pure visual reasoners for text-to-image generation. arXiv preprint arXiv:2601.10061 (2026)
51. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
52. Wang, K., Chen, R., Zheng, T., Huang, H.: Imagent: A unified multimodal agent framework for test-time scalable image generation. arXiv preprint arXiv:2511.11483 (2025)
53. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
54. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
55. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
56. Wu, J.Z., Ren, X., Shen, T., Cao, T., He, K., Lu, Y., Gao, R., Xie, E., Lan, S., Alvarez, J.M., et al.: Chronoedit: Towards temporal reasoning for image editing and world simulation. arXiv preprint arXiv:2510.04290 (2025)
57. Wu, Y., Li, Z., Hu, X., Ye, X., Zeng, X., Yu, G., Zhu, W., Schiele, B., Yang, M.H., Yang, X.: Kris-bench: Benchmarking next-level intelligent image editing models. arXiv preprint arXiv:2505.16707 (2025)
58. Ye, Y., He, X., Li, Z., Lin, B., Yuan, S., Yan, Z., Hou, B., Yuan, L.: Imgedit: A unified image editing dataset and benchmark. arXiv preprint arXiv:2505.20275 (2025)

59. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: Clevrer: Collision events for video representation and reasoning. arXiv preprint arXiv:1910.01442 (2019)
60. Yin, F., Liu, S., Han, Y., Wang, Z., Xing, P., Wang, R., Cheng, W., Wang, Y., Li, A., Yin, Z., et al.: Reasonedit: Towards reasoning-enhanced image editing models. arXiv preprint arXiv:2511.22625 (2025)
61. Yuan, Y., Wang, X., Wickremasinghe, T., Nadir, Z., Ma, B., Chan, S.H.: Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. arXiv preprint arXiv:2509.21309 (2025)
62. Zhan, Y.W., Wang, X., Chen, H., Feng, T., Feng, W., Wang, R., Li, G., Li, Q., Zhu, W.: Phyvllm: Physics-guided video language model with motion-appearance disentanglement. arXiv preprint arXiv:2512.04532 (2025)
63. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems **36**, 31428–31449 (2023)
64. Zhang, T., Yu, H.X., Wu, R., Feng, B.Y., Zheng, C., Snavely, N., Wu, J., Freeman, W.T.: Physdreamer: Physics-based interaction with 3d objects via video generation. In: European Conference on Computer Vision. pp. 388–406. Springer (2024)
65. Zhang, Z., Chen, Z., Yang, Z., Yang, Y.: Are image-to-video models good zero-shot image editors? arXiv preprint arXiv:2511.19435 (2025)
66. Zhao, X., Zhang, P., Tang, K., Zhu, X., Li, H., Chai, W., Zhang, Z., Xia, R., Zhai, G., Yan, J., et al.: Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. arXiv preprint arXiv:2504.02826 (2025)