PhysEdit: Physically Consistent Image Editing via Causal Enforcement

Siqi Wan[1*], Jingwen Chen[2], Yehao Li[2], Yingwei Pan[2], Ting Yao[2], and Tao Mei[2]

[1] University of Science and Technology of China
[2] HiDream.ai Inc. China
wansiqi4789@mail.ustc.edu.cn
{chenjingwen, liyehao, pandy, tiyao}@hidream.ai, tmei@live.com

Abstract. While instruction-guided image editing has seen significant strides, most existing models fail to preserve physical laws and causality. This limitation stems from the sparsity of physical supervision signals, coupled with the static, non-causal architecture of prevailing frameworks, which together hinder a deep comprehension of the physical world. To bridge this gap, we present PhysEdit, a novel framework that enforces physical consistency in image editing through causal generation and physics-aware reinforcement learning. Specifically, by harnessing the intrinsic physical priors of video generative models, we curate a high-quality dataset, dubbed PhysEdit-50K, wherein every sample encapsulates the complex physical causalities and dynamics involved in the transition between input and edited images. Building upon this, a specialized Causal Image Editor is devised to factorize the image editing process along the temporal evolution of visual transformations, thereby internalizing the underlying physical laws and boosting causal coherence. Simultaneously, an additional regularization term is incorporated to maintain visual consistency in unedited regions beyond the simple flow matching objective, mitigating the impact of undesired viewpoint shifts or camera motions in the dataset synthesized by video models. Furthermore, we integrate physics-aware reinforcement learning with a tailored fine-grained reward to steer the editing process toward better adherence to physical laws. Extensive experiments on PICABench and ImgEdit-Bench demonstrate that our PhysEdit significantly outperforms state-of-the-art baselines, yielding physically consistent and plausible edits. Code is publicly available at: https://github.com/HiDream-ai/PhysEdit/.

Keywords: Image Editing · Diffusion Transformer

1 Introduction

The field of visual content generation has advanced rapidly in recent years, driven primarily by innovations in diffusion model architectures [25, 26] and their training paradigms [11, 16, 32]. These foundational breakthroughs have subsequently

[*] This work was performed at HiDream.ai.

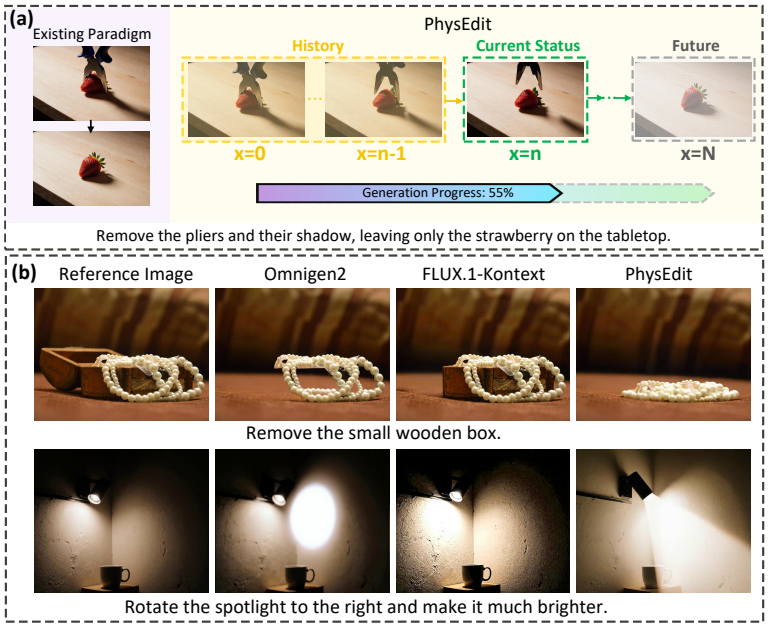

Fig. 1: (a) In contrast to existing discrete single-step editing paradigms, our proposed PhysEdit casts image editing as a continuous causal process. Through progressive synthesis of intermediate visual states, it captures underlying physical dynamics and ensures consistency across the entire editing trajectory. (b) Examples generated by existing state-of-the-art editing models and PhysEdit.

spurred a wide array of complex downstream applications, including instruction-guided image editing [3,6,18,43], subject-driven customization [21,33,41,46], and 3D asset generation [15, 17, 27, 40]. In this paper, we focus on a critical yet underexplored task: physically consistent image editing. This requires generative models to not only faithfully follow the semantic intent of a user instruction, but also holistically model the cascading physical effects in the real world.

Existing state-of-the-art instruction-based image editing frameworks [3, 14, 18,35,45,47] have demonstrated exceptional generation and generalization capabilities, enabling sophisticated visual manipulations driven by open-vocabulary prompts. Nevertheless, a closer inspection reveals a fundamental limitation: existing methodologies primarily optimize for semantic alignment and visual consistency, frequently bypassing the fundamental physical laws that govern natural scenes. This oversight frequently results in edits that are physically implausible, undermining the realism and practical utility of generated outputs. Consider a stacked-object scenario (illustrated in the first row of Fig. 1 (b)) where removing the bottom object should, under gravity, cause the top object to fall and occupy the vacated space. However, prevailing models (e.g., OmniGen2 [37]) trivially erase the target object, leaving the upper item suspended in mid-air in violation of physical realism. We attribute this profound deficiency in following physical laws to two primary bottlenecks inherent in current image editing paradigms.

First, standard image editing benchmarks [3,31,42] are predominantly composed of static input-output pairs, where the discontinuous and sparse nature of such supervision fundamentally hinders models from learning continuous real-world physical processes. Second, current architectures are fundamentally static and non-causal. By performing direct single-step mapping from input to output (left side in Fig. 1 (a)), this kind of model circumvents the continuous chain of physical causality that should underpin realistic transformation.

To address these challenges, we propose to explicitly enforce physical causality in image editing from both data and architectural perspectives. On the data front, we construct PhysEdit-50K, a high-quality dataset featuring dense physical and causal dynamics. Specifically, we synthesize continuous editing trajectories by distilling physical priors and world knowledge from large-scale video generative models such as Wan2.2 [34]. A fine-grained, instance-adaptive filtering mechanism is then applied to rigorously guarantee the physical plausibility of the generated visual sequences. Building upon this curated dataset, we present PhysEdit, a novel framework that reformulates conventional one-step image editing as a progressive, causal generation process (right side in Fig. 1). At its core, we devise a specialized Causal Diffusion Transformer (cDiT) that progressively synthesizes intermediate visual states over multiple steps, with each state conditioned strictly on its preceding historical context. This design is paired with a self-forcing training mechanism that mitigates the train-test discrepancy, thereby strengthening temporal causality during inference. Furthermore, to counter unintended viewpoint shifts and camera jitters inherent in visual trajectories synthesized from video priors, we introduce a consistency regularization term (CL) that anchors unedited background regions via distillation from a frozen single-step teacher model, which preserves the spatial coherence essential for common editing scenarios. Finally, physics-aware reinforcement learning (PARL) is further incorporated to optimize the model over complex causal interactions with a fine-grained reward, boosting physical consistency and generalization capability.

In summary, our key contributions are threefold: (1) We curate PhysEdit-50K, a high-quality dataset that provides dense physical supervision, moving beyond conventional static input-output pairs. (2) We present a novel framework, PhysEdit, which reformulates static image editing as a continuous, causal generation process, a paradigm shift that effectively addresses the physical implausibility inherent in single-step discrete generation. (3) We further incorporate a new regularization term to preserve spatial consistency of unedited regions and physics-aware reinforcement learning to enhance physical causality. These components collectively yield substantial performance gains over existing baselines on PICABench [28] and ImgEdit-Bench [42].

2 Related Work

Instruction-guided Image Editing. Instruction-guided image editing has garnered significant attention and emerged as a pivotal research direction within computer vision and generative AI. InstructPix2Pix [3] pioneers early efforts us-

ing synthetic editing triplets, while MagicBrush [45] and AnyEdit [43] improve quality with human-annotated data. To handle complex user intent, SmartEdit [12] integrates multimodal large language models (MLLMs) to comprehend input instructions. By scaling up datasets and model architectures, state-of-the-art editing models [2, 4, 36] have reached a new frontier in instruction adherence, consistency preservation, and generation fidelity. For instance, Qwen-Image-Edit [36] integrates MLLMs and DiT, leveraging multimodal understanding for precise editing. Furthermore, an emerging trend consolidates generative and editing tasks within a unified framework. OmniGen2 [37] provides a versatile foundation that leverages dual decoding pathways for generation, editing, and multimodal understanding. Similarly, Unified-IO [20] employs a unified autoregressive transformer to handle this broad spectrum of tasks. More recently, HiDream-O1-Image [4, 5] introduces a natively unified image generative foundation model capitalizing on a pixel-space Diffusion Transformer, enabling image synthesis, editing and personalized generation within a shared architecture.

Physically Consistent Image Editing. Despite the formidable editing capabilities demonstrated by the aforementioned methods, they frequently struggle to maintain physical consistency during the manipulation process. Recognizing this limitation, recent works have begun integrating physical priors into image editing. EditWorld [44] pioneers this direction by incorporating world models to explicitly simulate physical dynamics, enabling grounded scene transformations that adhere to environmental constraints. BAGEL [8] leverages the world knowledge encoded in multimodal large language models to reason about phsycially plausible editing outcomes. Later, PICABench [28] and UniReal [7] borrow physical priors from powerful video generative models to construct training data with sampled two frames as static input-output image pairs, which are then used to train editing models. Meanwhile, F2F [29] and ChronoEdit [38] treat image editing as video generation, leveraging video models' inherent temporal consistency to capture complex motion and ensure high-fidelity physical adherence.

3 Method

In this section, we present our proposed method designed to enable physically consistent image editing. To clearly articulate our approach, we organize this section into three main parts. To begin with, we describe the data construction pipeline for PhysEdit-50K dataset with dense physical supervision in Sec. 3.1. Then, the architecture of our proposed PhysEdit and the new consistency loss are comprehensively explained in Sec. 3.2. Ultimately, Sec. 3.3 demonstrates our physics-aware reinforcement learning strategy, which utilizes a fine-grained reward function to further steer PhysEdit toward enhanced compliance with complex physical laws and causality.

3.1 PhysEdit-50K

Although several datasets [9, 10, 28] have been released for physics-aware image editing, they are confined to static input-output image pairs without intermedi-

ate state sequences. We argue that such sparse supervision hinders the ability of image editing model to internalize the underlying physical laws and causal relationships, resulting in edits that lack physical coherence. To address this challenge, we present a data synthesis pipeline to collect high-quality visual editing trajectories that fully capture the continuous evolution of visual transformations.

Specifically, taking the source image I and instruction c from each sample in PICA-100K [28] as initial conditions, we leverage the physical priors embedded in state-of-the-art video generative models $\mathcal{F}_{vid}$, such as Wan2.2 [34], to generate a continuous, dynamic evolutionary video $V = \mathcal{F}_{vid}(I, c)$. Particularly, the first frame in the generated V is the source image I, the last frame is treated as the edited image I^e, and the frames in between capture the rich physical transformation dynamics.

To further ensure the quality and physical consistency of the synthesized data, we introduce an adaptive fine-grained verification mechanism. Given the diversity of editing tasks across different samples, evaluating them with a fixed and global criterion often fails to account for the sample-specific multi-dimensional artifacts (e.g., identity degradation, instruction misalignment, or physical inaccuracies). Consequently, we decompose each editing instruction c into a sequence of more granular fine-grained criteria $\{p^1, p^2, \ldots, p^K\}$, each p^k designed to intervene on a specific aspect of the involved visual transformation. We then employ a Multimodal Large Language Model (MLLM) to perform binary assessments on whether the last frame I^e in the generated video V satisfies each fine-grained criterion p^k. After filtering out videos that fail these multifaceted checks, N frames are uniformly sampled from each qualified video to compose a dynamic editing trajectory of intermediate images depicting the transition from the source image to the edited image. Finally, we curate a high-quality dataset, i.e., PhysEdit-50K, that furnishes dense, physically-grounded supervision beyond existing datasets.

3.2 PhysEdit

Conventional diffusion-based image editing models are typically confined to producing a single static output based on the source image and input instruction. To capture rich physical dynamics and complex causal relationships, one potential strategy is to task the model with generating a full sequence of intermediate images. However, under such a paradigm, the bidirectional influence between "future" and "past" states compromises temporal causality. While autoregressive models [13,22,39] naturally enforce causality, their reliance on discrete tokenization inevitably sacrifices visual fidelity.

To address these challenges, we propose a novel framework, dubbed PhysEdit, that novelly reformulates image editing as a progressive, causal process following physical laws. PhysEdit unfolds autoregressively across a temporal trajectory of N images, while performing diffusion-based denoising within each image to preserve spatial fidelity. The overview of our PhysEdit is illustrated in Fig. 2.

Specifically, let c, $\mathbf{x}_0$ and $\mathbf{x}_{1:N}$ be the editing instruction, the clean latent code of source image and the visual trajectory with N images, respectively. The

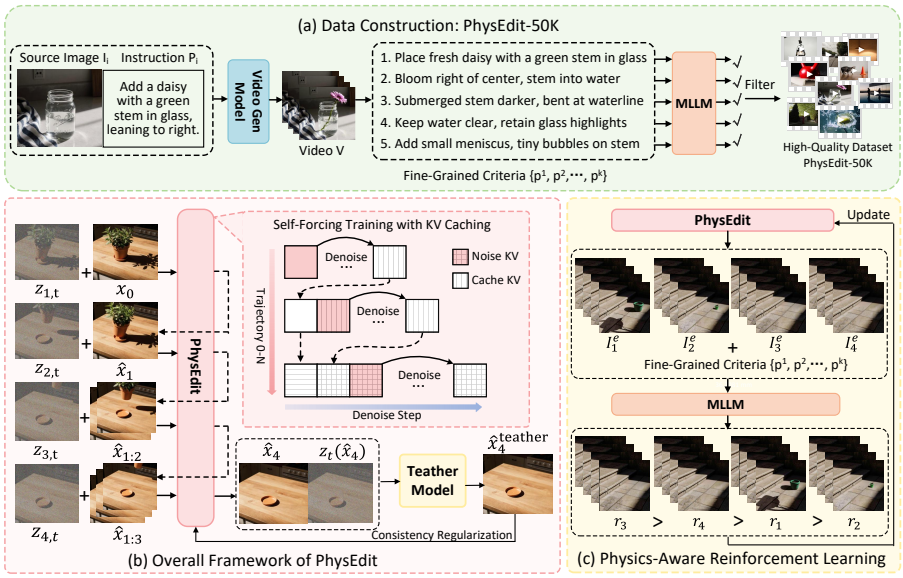

Fig. 2: (a) Data construction pipeline for our PhysEdit-50K. We synthesize PhysEdit-50K by harnessing the physical priors of video generative models to capture continuous visual transformations, followed by a sample-adaptive verification mechanism to ensure physical correctness. (b) The overall framework our PhysEdit. PhysEdit reformulates single-step image editing as a multi-step, causality-aware process. Rather than directly "jumping" from source to edited image, it progressively generates intermediate visual transition states, with each step conditioned on self-generated historical context. KV caching is adopted to facilitate self-forcing training. An additional consistency regularization term is devised to preserve spatial fidelity in unedited regions via guidance from a single-step teacher model. (c) Physics-Aware Reinforcement Learning. An MLLM serves as a physics expert to perform fine-grained assessments of the generated outputs, providing informative reward signals that further steer the model toward better adherence to physical laws and causality.

autoregressive generation process in our PhysEdit is defined by the following factorization of conditional probabilities:

$$p_\theta(\mathbf{x}_{1:N}|\mathbf{x}_0, c) = \prod_{i=1}^{N} p_\theta(\mathbf{x}_i|\mathbf{x}_{<i}, \mathbf{x}_0, c), \tag{1}$$

where θ denote the parameters of our PhysEdit. For the spatial generation of each intermediate image $\mathbf{x}_i$, PhysEdit follows the general rectified flow framework [16, 19]. The forward process is defined as a linear interpolation between the clean latent $\mathbf{x}_i$ and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. At a random timestep $t \in [0, 1]$, the perturbed state $\mathbf{z}_{i,t}$ is given by

$$\mathbf{z}_{i,t} = t\mathbf{x}_i + (1-t)\boldsymbol{\epsilon}. \tag{2}$$

To reverse this process, we adopt a Diffusion Transformer (DiT) to parameterize the time-dependent velocity field $\mathbf{v}_t^\theta$, which governs the deterministic flow from the standard Gaussian prior to the data manifold.

Self-Forcing Training. To mitigate exposure bias in autoregressive training with standard teacher-forcing paradigm, we implement a self-forcing strategy that conditions the current flow on the model's own past predictions, thereby eliminating the train-inference distribution mismatch. Formally, we denote the history context for $\mathbf{z}_{i,t}$ as $\hat{\mathbf{x}}^\theta_{<i}$, which is generated on-the-fly during training via autoregressive rollout. To ensure computational feasibility, we employ a few-step Ordinary Differential Equation solver for fast generation of each preceding frame and further leverage a Key-Value Caching mechanism to incrementally retain intermediate activations of the history context. This design eliminates redundant computation for efficient training without sacrificing trajectory fidelity.

Consistency Regularization. A challenge in training on video-synthesized image editing data arises from potential viewpoint drifts or camera motion inherent in video generative models. We observe that naively optimizing the our PhysEdit with the standard flow matching objective may inadvertently compromises the spatial consistency of unedited regions. Therefore, we introduce an additional consistency loss that enables the editing model to simultaneously capture dense physical state transitions and intricate causality, while preserving unedited regions as effectively as single-step editing models.

Specifically, the pre-trained single-step editing model is adopted as the teacher model ϕ_{teacher}. Conditioned on the previous history of $N-1$ generated intermediate images $\{\hat{\mathbf{x}}_i\}_{i=1}^{N-1}$, our PhysEdit predicts the denoised image $\mathbf{z}^\theta_{N,t}$ at timestep t and approximates its clean latent $\hat{\mathbf{x}}^\theta_N$. Then, $\hat{\mathbf{x}}^\theta_N$ is corrupted to a randomly sampled timestep t by injecting Gaussian noise ϵ according to Eq. 2, yielding the noisy latent $\mathbf{z}_t(\hat{\mathbf{x}}^\theta_N)$. Taking the editing instruction c, the source image $\mathbf{x}_0$ as inputs, the teacher model ϕ_{teacher} predicts the reference velocity field

$$\mathbf{v}_{\text{teacher}} = \phi_{teacher}(\mathbf{z}_t(\hat{\mathbf{x}}^\theta_N), \mathbf{x}_0, c, t). \tag{3}$$

$\mathbf{v}_{\text{teacher}}$ is further utilized to derive a refined version of the editing

$$\hat{\mathbf{x}}_N^{\phi_{teacher}} = \mathbf{z}_t(\hat{\mathbf{x}}^\theta_N) + (1-t)\mathbf{v}_{\text{teacher}}. \tag{4}$$

Therefore, our consistency loss is defined as:

$$\mathcal{L}_{consist}(\theta) = \mathbb{E}_{t\sim[0,1],\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[sg(\hat{\mathbf{x}}_N^{\phi_{teacher}}) - \hat{\mathbf{x}}^\theta_N], \tag{5}$$

where $sg(\cdot)$ is short for "stop gradient". Intuitively, this loss treats the teacher's output as a fixed target, pushing the generator's initial prediction $\hat{\mathbf{x}}^\theta_N$ to align with the high-fidelity manifold defined by the tearch model.

3.3 Physics-Aware Reinforcement Learning

To further generalize the model's capacity for physically-consistent image editing, we introduce a second-stage reinforcement learning phase to post-train

our PhysEdit. Specifically, we leverage a Multimodal Large Language Model (MLLM), denoted as $\mathcal{M}$, to serve as a physics expert with massive open-world knowledge. Rather than relying on a holistic score, we adopt a scoring mechanism consistent with our dataset curation pipeline (Sec. 3.1). This approach ensures that the reward signal is both dense, interpretable and physically grounded by employing a fine-grained adaptive evaluation rubric for each sample. This verification explicitly captures a set of K physical constraints and causal relationships inferred by the instruction c. $\mathcal{M}$ evaluates the edited image I^e against these K constraints sequentially within a global context, and the final reward R is calculated as the satisfaction rate of these constraints:

$$R(I^e, c) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}\left(\mathcal{M}(I^e, p^k) = \text{True}\right), \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function. We optimize our PhysEdit using Group Relative Policy Optimization (GRPO) [30] with a KL divergence penalty as a regularization term. For each instruction c, we sample a group of G edited images $\{I_i^e\}_{i=1}^G$ from the old policy $\pi_{\theta_{old}}$ (i.e., the optimized PhysEdit) to compute relative advantages, optimizing the objective:

$$\mathcal{J}(\theta) = \mathbb{E}_{c,\{I_i^e\}\sim\pi_{\theta_{old}}} \left[\frac{1}{G} \sum_{i=1}^{G} \left(\min\left(\rho_i \hat{A}_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)\hat{A}_i\right) - \beta \mathbb{D}_{KL}(\pi_\theta \| \pi_{ref}) \right) \right], \tag{7}$$

where $\rho_i = \frac{\pi_\theta(I_i^e|c)}{\pi_{\theta_{old}}(I_i^e|c)}$ is the probability ratio, and π_{ref} is the frozen reference initialization from the pre-trained PhysEdit. The advantage $\hat{A}_i$ is computed using the physics-aware rewards $r_i = R(I_i^e, c)$ within the sampled group:

$$\hat{A}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}. \tag{8}$$

By maximizing this fine-grained physics reward, PhysEdit effectively transcends pure appearance modification, achieving continuous, causal edits that are strictly governed by real-world physical knowledge.

4 Experiments

4.1 Experimental Setups

Datasets. To enable physically grounded image editing, we construct PhysEdit-50K, a new dataset featuring dense supervision of physical dynamics. Each sample comprises a source image, an editing instruction, and a paired video synthesized by Wan2.2 [34] that depicts the progressive causal evolution specified by the instruction. After rigorous filtering for physical validity and semantic fidelity using Qwen2.5-VL-72B [1], the resulting PhysEdit-50K comprises 51,460 high-quality image-video pairs spanning 8 distinct physical interaction categories. For model training, we uniformly sample N frames from each video to form the continuous editing trajectories.

Evaluation Benchmarks. To comprehensively evaluate our proposed PhysEdit, we conduct evaluation on two benchmarks: one targeting physical consistency and the other assessing general editing performance. First, we adopt PICABench [28] to assess the physical plausibility of generated edits. This benchmark comprises 900 real-world samples spanning 8 sub-dimensions: light propagation, light source effects, reflection, refraction, deformation, causality, and global/local state transitions. In line with real-world editing scenarios, we use superficial prompts (i.e., plain commands probing intrinsic physical priors) as the editing instructions. Two metrics are computed per sub-dimension: Accuracy and Consistency. Accuracy is scored by GPT-5.1 [24] following [28], evaluating clarity, factual correctness, and visual context alignment. Consistency is measured via Peak Signal-to-Noise Ratio (PSNR) calculated over unedited regions, with the predicted edited area masked out. Second, we benchmark general editing capabilities on ImgEdit-Bench, a test set comprising 734 high-quality samples across 9 representative tasks: add, remove, alter, replace, style transfer, background change, motion change, hybrid edit, and action. Following prior work, we employ GPT-4.1 [23] to evaluate instruction adherence, editing quality, and detail preservation, reporting the average score across all dimensions.

Implementation Details. The training process consists of two stages. (1) Supervised Fine-Tuning. We fine-tune PhysEdit on our curated dataset (PhysEdit-50K) for 2 epochs with a batch size of 8 and a constant learning rate of 1e-5. We apply a condition dropout rate of 0.1 to text prompts and source images to enable classifier-free guidance (CFG) during inference. (2) Physics-Aware Reinforcement Learning. We further optimize our model for one epoch using Group Relative Policy Optimization (GRPO) [30] to enhance physical consistency. Training uses a group size of 4. The frozen Qwen2.5-VL-7B [1] serves as the reward model for fine-grained physics-aware reward assessment.

4.2 Results

Quantitative Results on PICABench. Table 1 compares our PhysEdit with existing state-of-the-art image editing models on the PICABench-Superficial benchmark. We initialize our PhysEdit from the pre-trained OmniGen2 [37] and Qwen-Image-Edit [36], yielding two variants: PhysEdit and PhysEdit$_{\text{Qwen}}$. As shown, our proposed method consistently achieves the highest accuracy across all evaluated dimensions. Remarkably, PhysEdit surpasses the strongest baseline, Flux.1 Kontext [2], by a substantial absolute margin of 8.6% in Overall Accuracy. We attribute this gain to the reformulation of single-step editing as a progressive, causality-aware process within our PhysEdit, combined with its architectural innovations. These design choices enable the model to learn rich physical dynamics from dense visual transformation trajectories, leading to improved physical consistency. In addition to strong physical realism, our model achieves record-breaking performance in Overall Consistency, which stems from our consistency regularization term. It is also worth noting that PhysEdit exhibits pronounced superiority in modeling complex light-matter interactions that

Table 1: Quantitative comparison results on PICABench-Superficial [28]. Acc ↑ and Con ↑ denote Accuracy (%) and Consistency (dB), respectively. LP, LSE, GST, and LST stand for Light Propagation, Light Source Effects, Global State Transition, and Local State Transition, respectively.

Model	LP	LSE	Reflection	Refraction	Deformation	Causality	GST	LST	Overall
Accuracy↑									
MagicBrush [45]	39.43	36.33	44.46	30.16	41.94	36.52	41.42	33.08	37.91
Instruct-Pix2Pix [3]	37.11	38.14	43.58	29.33	39.72	37.68	40.33	30.79	37.09
AnyEdit [43]	42.76	44.07	43.71	33.89	40.66	39.33	42.71	36.49	40.49
Step1X-Edit [18]	45.92	47.44	53.46	34.21	45.94	42.42	48.79	38.57	44.59
Bagel [8]	46.50	39.39	49.27	42.74	44.25	39.29	46.80	42.75	43.87
Uniworld-V1 [14]	42.37	34.50	46.04	46.05	40.10	39.52	22.85	39.50	37.68
OmniGen2 [37]	50.90	48.52	56.01	42.22	43.23	40.10	44.33	37.64	45.36
Flux.1 Kontext [2]	54.90	57.36	57.37	38.91	47.88	38.26	48.79	41.21	48.08
Qwen-Image-Edit [36]	62.33	61.52	62.31	55.70	49.56	48.61	64.23	53.47	57.21
Chronoedit [38]	63.35	60.15	64.87	62.05	53.67	50.92	62.55	53.10	58.83
PhysEdit (Ours)	59.57	64.81	63.33	62.50	52.30	49.54	48.92	43.93	56.72
PhysEdit$_{\text{Qwen}}$ (Ours)	**63.87**	**65.91**	**65.47**	**65.41**	**54.00**	**51.38**	**65.23**	**54.51**	**60.70**
Consistency↑									
MagicBrush [45]	24.45	23.94	24.75	22.29	24.33	22.21	7.79	17.48	20.91
Instruct-Pix2Pix [3]	22.76	22.36	24.51	22.12	24.52	20.00	8.52	16.43	20.15
AnyEdit [43]	23.63	24.00	25.49	23.44	25.38	24.16	11.79	18.16	22.01
Step1X-Edit [18]	30.38	27.53	29.32	27.37	29.71	26.92	8.75	20.92	25.11
Bagel [8]	29.12	29.23	28.11	28.36	33.12	27.51	10.48	24.53	26.30
Uniworld-V1 [14]	18.50	19.96	18.59	17.48	18.82	18.11	17.62	19.41	18.48
OmniGen2 [37]	25.69	28.34	27.78	24.84	29.28	26.93	12.18	25.89	24.12
Flux.1 Kontext [2]	27.58	25.97	26.92	26.76	28.86	29.69	12.52	25.70	24.57
Qwen-Image-Edit [36]	30.97	29.52	28.23	28.21	32.29	29.42	13.84	27.52	27.50
Chronoedit [38]	30.88	30.86	28.90	28.30	33.51	29.83	13.21	27.82	27.91
PhysEdit (Ours)	30.01	30.21	28.18	27.99	33.18	27.59	13.89	26.76	27.23
PhysEdit$_{\text{Qwen}}$ (Ours)	**31.41**	**30.85**	**29.76**	**29.06**	**33.71**	**29.82**	**13.90**	**28.32**	**28.35**

conventional image editing models typically struggle with. For instance, in Refraction, PhysEdit achieves an accuracy of 62.50% , substantially outperforming the next best model, Uniworld-V1 [14] at 46.05%, corresponding to a relative improvement of 35.7%. This dominant trend extends to Light Source Effects (LSE, 64.81%) and Reflection (63.33%). Furthermode, while the baseline Qwen-Image-Edit already exhibits strong physical consistency (Overall Accuracy 57.21%) owing to its large capacity, our PhysEdit$_{\text{Qwen}}$ further improves performance by a substantial margin of 3.49%. This yields a new state-of-the-art, with an Overall Accuracy of 60.70% and an Overall Consistency of 28.35 dB. These gains are particularly pronounced on demanding physical scenarios, delivering absolute improvements of 4.44% on Deformation and 9.71% on Refraction. Overall, these observations confirm the merits of our framework, showcasing its ability to render physically plausible and complex image edits.

Quantitative Results on ImgEdit-Bench. We report the quantitative results on the ImgEdit-Bench in Table 2. As shown, the OmniGen2-based PhysEdit achieves a competitive Overall score of 3.75 among baselines with comparable model capacity. Notably, PhysEdit surpasses the formidable baseline, FLUX.1 Kontext [2], by a relative margin of 6.5%. When adopting Qwen-Image-Edit as

Table 2: Quantitative comparison results on ImgEdit-Bench [42]. "Overall" is calculated by averaging all scores across tasks.

Model	Add↑	Adjust↑	Extract↑	Replace↑	Remove↑	Background↑	Style↑	Hybrid↑	Action↑	Overall ↑
MagicBrush [45]	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.90
Instruct-Pix2Pix [3]	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.20	1.46	1.88
AnyEdit [43]	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
Step1X-Edit [18]	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
BAGEL [8]	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
UniWorld-V1 [14]	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
OmniGen2 [37]	3.57	3.06	1.77	3.74	3.20	3.57	4.81	2.52	4.68	3.44
FLUX.1 Kontext [2]	3.76	3.45	2.15	3.98	2.94	3.78	4.38	2.96	4.26	3.52
Qwen-Image-Edit [36]	4.21	4.09	3.24	4.35	4.01	4.24	4.93	3.67	4.76	4.16
Chronoedit [38]	4.31	4.02	3.25	4.03	**4.32**	4.30	4.68	3.73	4.83	4.16
PhysEdit (Ours)	3.87	3.22	2.88	3.93	3.44	3.94	4.85	2.98	4.73	3.75
PhysEdit$_{\text{Qwen}}$ (Ours)	**4.38**	**4.14**	**3.35**	**4.50**	4.27	**4.46**	**4.96**	**3.78**	**4.85**	**4.29**

Table 3: Quantitative comparison results on RISEBench [48]. Accuracy (%) for each dimension is reported. "Overall" is calculated by averaging across tasks.

Model	Temporal	Causal	Spatial	Logical	Overall
Step1X-Edit [18]	0.0	2.2	2.0	3.5	1.9
BAGEL [8]	2.4	5.6	14.0	1.2	6.1
UniWorld-V1 [14]	1.9	6.3	15.2	1.1	6.2
OmniGen2 [37]	1.2	1.0	0.0	1.2	0.8
FLUX.1 Kontext [2]	2.3	5.5	13.0	1.2	5.8
Qwen-Image-Edit [36]	4.7	10.0	17.0	2.4	8.9
PhysEdit$_{\text{Qwen}}$ (Ours)	**5.2**	**10.4**	**18.1**	**2.6**	**9.1**

the backbone, PhysEdit$_{\text{Qwen}}$ further improves the Overall score to 4.29, substantially outperforming the strong backbone (4.16). This improvement underscores the broad applicability of our approach beyond physics-focused scenarios. We observe particularly pronounced improvements in the "Extract" and "Remove" tasks, where PhysEdit outperforms the runner-up by 0.61 and 0.2, respectively. These two editing operations are challenging because they necessitate not only the accurate manipulation of the target object but also the cohesive adjustment of associated physical phenomena, such as coupled illumination and cast shadows. The substantial gains achieved on these difficult cases underscore our PhysEdit's exceptional capacity to enforce rigorous physical consistency. Furthermore, there is a risk that fine-tuning PhysEdit for physical realism inadvertently degrades spatial consistency preservation, as training data synthesized by video generative models often contains viewpoint shifts or camera motions. By leveraging guidance from a single-step teacher editing model, our approach maintains spatial fidelity in tasks that demand precise spatial alignment between the input and edited image, such as "Add" and "Style Transfer". This robustness directly validates the effectiveness of our proposed consistency loss.

Quantitative Results on RISEBench. To further comprehensively validate the adherence to physical laws and complex causality, we benchmark our approach on RISEBench [48]. As shown in Table 3, our PhysEdit$_{\text{Qwen}}$ consistently achieves the best performance across all evaluated dimensions. Notably, PhysEdit$_{\text{Qwen}}$ yields an absolute improvement of 0.2% over Qwen-Image-Edit [36] in the Overall score, reaching an accuracy of 9.1%. We attribute this gain

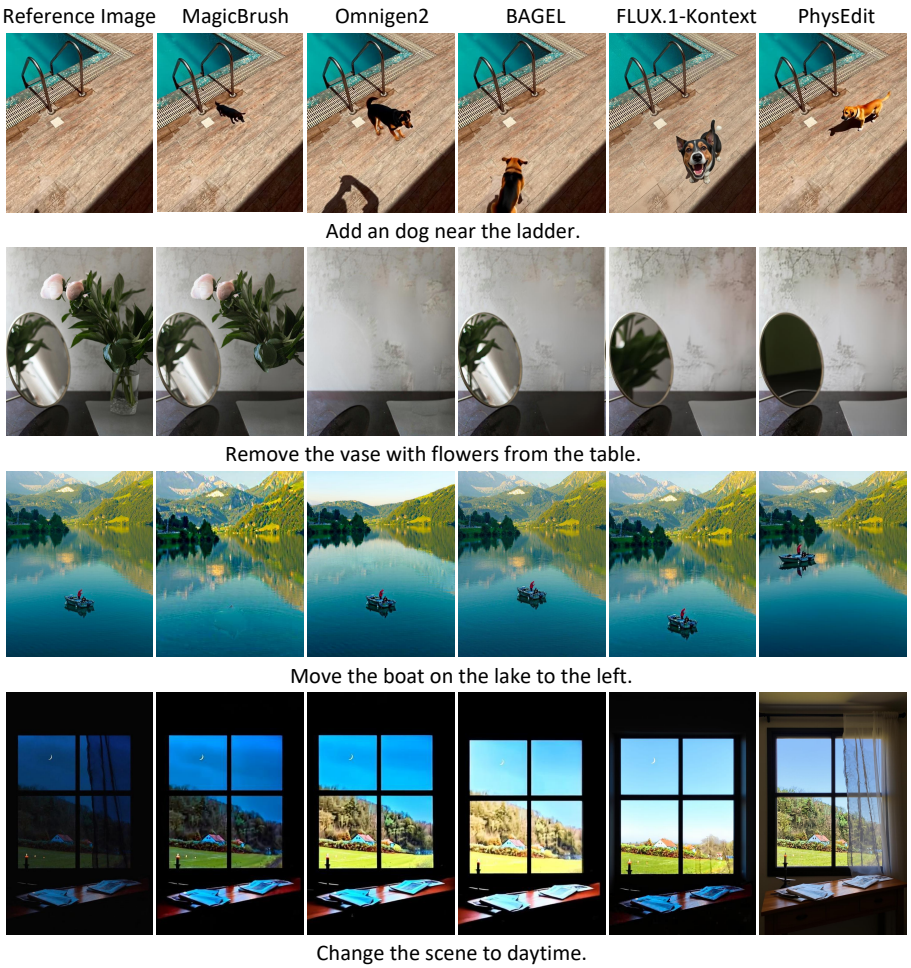

Fig. 3: Qualitative comparisons on PICABench-Superficial.

to the integration of our specialized Causal Image Editor and physics-aware reinforcement learning. Compared to conventional static editing models, our PhysEdit$_{\text{Qwen}}$ exhibits pronounced superiority in complex visual editing tasks. For instance, on the "Causal" dimension, PhysEdit$_{\text{Qwen}}$ attains an accuracy of 10.4%, substantially outperforming the next best model (10.0%) by a clear margin. This advantage extends to "Spatial" (18.1%) and "Temporal" (5.2%) dimensions. Overall, these results confirm the merits of our proposed method.

Qualitative Results. Fig. 3 presents a qualitative comparison between our proposed PhysEdit and conventional instruction-based editing models. While existing baselines demonstrate basic instruction-following capabilities, they frequently struggle to maintain physical realism. For example, in the top row, FLUX.1-Kontext [2] generates a dog with distorted proportions and completely omits the required cast shadows. In the second row, most baselines remove the

Fig. 4: Examples of generated results by our PhysEdit on PICABench-Superficial.

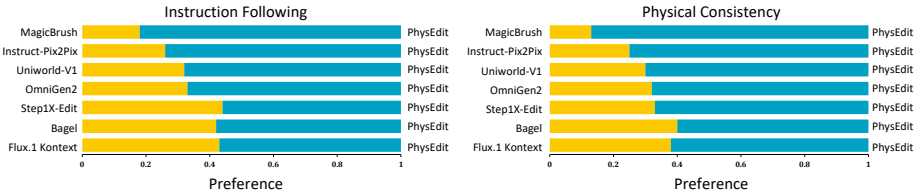

Fig. 5: User study on 100 image-instruction pairs randomly sampled from PICABench and ImgEdit-Bench.

vase but fail to eliminate its corresponding mirror reflection. Similarly, when tasked with transitioning a scene to daytime (bottom row), existing image editing models leave the interior unnaturally dark. In contrast, our PhysEdit consistently executes these edits with high physical fidelity, yielding results that align with real-world physical constraints. Furthermore, Fig. 4 visualizes several editing trajectories generated by PhysEdit. Through progressive causal modeling of the continuous visual transformation, our model overcomes the abrupt visual discontinuities of single-step models and ensures physically grounded transitions during the editing process.

User Study. A comprehensive user study is conducted to examine whether the edited images align with human perception and physical realism. We randomly sample 100 image-instruction pairs (50 from PICABench and 50 from ImgEdit-Bench) for evaluation. 15 participants from diverse backgrounds (i.e., computer vision (5), digital art (4), physics (3), and general users (3)) are invited and asked to conduct pairwise comparisons between our PhysEdit and seven competing approaches (Flux.1 Kontext [2], OmniGen2 [37], Uniworld-V1 [14], Step1X-Edit [18], Bagel [8], MagicBrush [45], and Instruct-Pix2Pix [3]). For each pair of generated images, users are asked to choose the winner based on two criteria: 1) instruction following (i.e., whether the edit accurately reflects the textual instruction while preserving the unedited background), and 2) physical consistency (i.e., faithful adherence to physical laws in rendering associated physical effects such as realistic shadows and reflections). Fig. 5 summarizes the averaged preference rates by all participants across the 100 test samples. As shown, our

Table 4: Ablation study on the effect of pivotal components in our PhysEdit. **cDiT**, **CL**, and **PARL** denote Causal DiT, Consistency Loss, and Physics-Aware Reinforcement Learning, 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 ↑
Base	50.90	25.69	48.52	28.34	56.01	27.78	42.22	24.84	43.23	29.28	40.10	26.93	44.33	12.18	37.64	25.89	45.36	24.12
Base + cDiT	58.13	29.12	63.96	29.97	58.98	28.05	57.52	25.35	48.87	32.01	45.12	27.47	45.77	13.40	39.44	26.26	52.23	26.45
Base + cDiT + CL	58.87	29.52	64.33	30.03	59.16	28.17	62.00	26.60	50.12	32.27	46.12	27.08	46.69	13.78	39.89	26.69	53.40	26.76
Base + cDiT + CL + PARL	59.57	30.01	64.81	30.21	63.33	28.18	62.50	27.99	52.30	33.18	49.54	27.59	48.92	13.89	43.93	26.76	56.72	27.23

Table 5: Ablation study on the effect of editing trajectory length N.

N	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 ↑
1	52.81	27.03	56.36	29.20	58.39	28.01	52.50	24.99	46.87	30.17	42.65	27.01	44.02	12.89	37.95	26.04	48.94	25.67
2	53.61	27.98	59.67	29.87	58.90	28.02	57.36	25.18	47.12	30.27	44.25	27.05	45.04	13.08	38.47	26.07	50.55	25.94
4	58.13	29.12	63.96	29.97	58.98	28.05	57.52	25.35	48.87	32.01	45.12	27.47	45.77	13.40	39.44	26.26	52.23	26.45
6	58.21	29.45	63.60	29.90	59.38	28.10	57.67	25.80	49.36	32.00	45.73	27.43	45.59	13.48	39.76	26.52	52.41	26.58

PhysEdit significantly outperforms the rival models by a large margin regarding both instruction following and physical consistency.

4.3 Discussions

Effect of Pivotal Components in PhysEdit. To rigorously isolate and quantify the contribution of each pivotal component in our PhysEdit, we conduct a progressive ablation study on PICABench-Superficial, as detailed in Table 4. **cDiT**, **CL**, and **PARL** denote Causal DiT, Consistency Loss, and Physics-Aware Reinforcement Learning, respectively. The transition from the vanilla **Base** model to our **cDiT** architecture drives a profound performance leap, boosting Overall Accuracy from 45.36% to 52.23%. This improvement is particularly striking in the "Light Source Effects (LSE)" category, where accuracy surges by an absolute 15.44% (from 48.52% to 63.96%). These results underscores the value of our approach: transforming editing from a discrete one-step task into a progressive, continuous process and modeling its rich dynamics through a causal generative architecture to effectively integrate physical knowledge. Furthermore, the introduction of **CL** proves indispensable for both instruction following and spatial detail preservation. Specifically, the **Base + cDiT + CL** configuration achieves a remarkable absolute improvement of 1.17% and 0.31 dB in Overall Accuracy and Consistency, respectively, compared to the **Base + cDiT** setup. This validates that our consistency regularization strategy successfully alleviates unintended modifications and faithfully preserving the structural integrity of unedited regions as instructed. A final boost of +3.32% in Overall Accuracy is achieved through **PARL**. Notably, performance on the "Causality" subset improves from 46.12% to 49.54%. This final piece of evidence confirms that fine-grained assessments of physical validity in PARL effectively steers PhysEdit toward improved physical consistency.

Effect of Editing Trajectory Length N. To systematically investigate the influence of the editing trajectory length N, we conduct an ablation study on PICABench-Superficial using our **Base + cDiT**. Table 4 summarizes the quantitative results across configurations with $N \in \{1, 2, 4, 6\}$. As expected, increas-

Table 6: Ablation study on the impact of our synthesized PhysEdit-50K. **SFT** denotes Supervised Fine-Tuning on PhysEdit-50K.

Frame	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 ↑
Base	50.90	25.69	48.52	28.34	56.01	27.78	42.22	24.84	43.23	29.28	40.10	26.93	44.33	12.18	37.64	25.89	45.36	24.12
Base + SFT	52.81	27.03	56.36	29.20	58.39	28.01	52.50	24.99	46.87	30.17	42.65	27.01	44.02	12.89	37.95	26.04	48.94	25.67
PhysEdit	59.57	30.01	64.81	30.21	63.33	28.18	62.50	27.99	52.30	33.18	49.54	27.59	48.92	13.89	43.93	26.76	56.72	27.23

ing N consistently improves both Accuracy (Acc) and Consistency (Con) across most metrics. Specifically, extending the trajectory length from $N = 1$ to $N = 4$ yields substantial absolute gains of 3.29% and 0.78 dB in Acc and Con, respectively. We attribute this improvement to the fact that a larger N enables PhysEdit to model denser, more continuous physical dynamics the target editing process, ensuring stronger adherence to underlying physical laws and causality. However, as N is further increased from 4 to 6, performance gains saturate with only marginal improvements. This suggests that, within the limited temporal window of the generated short videos, $N = 4$ is sufficient to cover the key intermediate states of physical evolution. It enables the model to capture essential state transitions while maintaining a balanced computational and memory cost.

Impact of Synthesized PhysEdit-50K. To better understand the contribution of our curated PhysEdit-50K, we analyze the performance under different data utilization strategies, with results reported in Table 6. Fine-tuning the single-step **Base** model on static input-output pairs sampled from PhysEdit-50K provides only limited benefits. Conversely, training PhysEdit on dense visual transformation sequences yields massive enhancements, elevating Acc to 56.72% and Con to 27.23 dB. This stark contrast validates our core designs (i.e., **cDiT, CL, PARL**), highlighting that learning from continuous temporal dynamics, rather than static endpoints, promotes physically consistent edits.

5 Conclusion

In this paper, we present PhysEdit, a novel framework designed to overcome the pervasive issue of physical implausibility in instruction-guided image editing. Diverging from conventional methods that treat editing as a static, single-step mapping, our approach fundamentally reformulates the task as a continuous, causal generation process. Specifically, we first leverage the physical priors inherent in advanced video generation models to curate a new dataset, PhysEdit-50K, which captures dense visual transformation dynamics rather than mere isolated input-output pairs. Building upon this data, a specialized Causal Diffusion Transformer is devised to model continuous editing trajectories and the underlying physical laws. Furthermore, we integrate a novel consistency regularization term and physics-aware reinforcement learning strategy to boost the integrity of unaltered regions and physical plausibility. Extensive evaluations on PICABench and ImgEdit-Bench demonstrate that PhysEdit outperforms existing baselines in both instruction adherence and physical realism.

Acknowledgement. This work was supported by the Key Science & Technology Project of Anhui Province No. 202523o09050002.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
3. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023)
4. Cai, Q., Chen, J., Chen, Y., Li, Y., Long, F., Pan, Y., Qiu, Z., Zhang, Y., Gao, F., Xu, P., et al.: Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. arXiv preprint arXiv:2505.22705 (2025)
5. Cai, Q., Chen, J., Gao, C., Gong, Z., Li, Y., Mei, T., Pan, Y., Peng, Y., Qiu, Z., Yao, T., Yu, K., Zhang, Y., et al.: Hidream-o1-image: A natively unified image generative foundation model with pixel-level unified transformer. arXiv preprint arXiv:2605.11061 (2026)
6. Chen, J., Pan, Y., Yao, T., Mei, T.: Controlstyle: Text-driven stylized image generation using diffusion priors. In: ACMMM (2023)
7. Chen, X., Zhang, Z., Zhang, H., Zhou, Y., Kim, S.Y., Liu, Q., Li, Y., Zhang, J., Zhao, N., Wang, Y., et al.: Unireal: Universal image generation and editing via learning real-world dynamics. In: CVPR (2025)
8. 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)
9. Fang, R., Duan, C., Wang, K., Huang, L., Li, H., Yan, S., Tian, H., Zeng, X., Zhao, R., Dai, J., et al.: Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. arXiv preprint arXiv:2503.10639 (2025)
10. 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)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
12. Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., Zhou, J., Dong, C., Huang, R., Zhang, R., et al.: Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In: CVPR (2024)
13. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. In: NeurIPS (2024)
14. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147 (2025)
15. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: CVPR (2023)
16. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
17. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: ICCV (2023)

18. 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)
19. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
20. Lu, J., Clark, C., Zellers, R., Mottaghi, R., Kembhavi, A.: Unified-io: A unified model for vision, language, and multi-modal tasks. arXiv preprint arXiv:2206.08916 (2022)
21. Mou, C., Wu, Y., Wu, W., Guo, Z., Zhang, P., Cheng, Y., Luo, Y., Ding, F., Zhang, S., Li, X., et al.: Dreamo: A unified framework for image customization. In: SIGGRAPH (2025)
22. Mu, J., Vasconcelos, N., Wang, X.: Editar: Unified conditional generation with autoregressive models. In: CVPR (2025)
23. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
24. OpenAI: Gpt-5 technical report (2025), `https://openai.com/`
25. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
26. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
27. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
28. Pu, Y., Zhuo, L., Han, S., Xing, J., Zhu, K., Cao, S., Fu, B., Liu, S., Li, H., Qiao, Y., et al.: Picabench: How far are we from physically realistic image editing? arXiv preprint arXiv:2510.17681 (2025)
29. Rotstein, N., Yona, G., Silver, D., Velich, R., Bensaïd, D., Kimmel, R.: Pathways on the image manifold: Image editing via video generation. In: CVPR (2025)
30. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
31. Sheynin, S., Polyak, A., Singer, U., Kirstain, Y., Zohar, A., Ashual, O., Parikh, D., Taigman, Y.: Emu edit: Precise image editing via recognition and generation tasks. In: CVPR (2024)
32. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. arXiv preprint arXiv:2303.01469 (2023)
33. Wan, S., Chen, J., Cai, Q., Pan, Y., Yao, T., Mei, T.: VTON-VLLM: Aligning virtual try-on models with human preferences. In: NeurIPS (2025)
34. Wan, T.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
35. Wang, Q., Zhang, B., Birsak, M., Wonka, P.: Instructedit: Improving automatic masks for diffusion-based image editing with user instructions. arXiv preprint arXiv:2305.18047 (2023)
36. 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)
37. 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)
38. Wu, J.Z., Ren, X., Shen, T., Cao, T., He, K., Lu, Y., Gao, R., Xie, E., Lan, S., Alvarez, J.M., Gao, J., Fidler, S., Wang, Z., Ling, H.: Chronoedit: Towards temporal reasoning for in-context image editing and world simulation. In: ICLR (2026)

39. Xiong, J., Liu, G., Huang, L., Wu, C., Wu, T., Mu, Y., Yao, Y., Shen, H., Wan, Z., Huang, J., et al.: Autoregressive models in vision: A survey. arXiv preprint arXiv:2411.05902 (2024)
40. Yang, H., Chen, Y., Pan, Y., Yao, T., Chen, Z., Ngo, C.W., Mei, T.: Hi3d: Pursuing high-resolution image-to-3d generation with video diffusion models. In: ACMMM (2024)
41. Ye, J., Jiang, D., Wang, Z., Zhu, L., Hu, Z., Huang, Z., He, J., Yan, Z., Yu, J., Li, H., et al.: Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. arXiv preprint arXiv:2508.09987 (2025)
42. 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)
43. Yu, Q., Chow, W., Yue, Z., Pan, K., Wu, Y., Wan, X., Li, J., Tang, S., Zhang, H., Zhuang, Y.: Anyedit: Mastering unified high-quality image editing for any idea. In: CVPR (2025)
44. Zeng, B., Yang, L., Liu, J., Xu, M., Zhang, Y., Wan, P., Zhang, W., Yan, S.: Editworld: Simulating world dynamics for instruction-following image editing. In: ACM MM (2025)
45. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. In: NeurIPS (2023)
46. Zhang, Y., Qiu, Z., Liu, Z., Pan, Y., Yao, T., Mei, T.: Evoid: Reinforced evolution for identity-preserving video generation. In: CVPR (2026)
47. Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: Enabling instructional image editing with in-context generation in large scale diffusion transformer. In: NeurIPS (2025)
48. Zhao, X., Zhang, P., Tang, K., Zhu, X., Li, H., Chai, W., Zhang, Z., Xia, R., Zhai, G., Yan, J., Yang, H., Yang, X., Duan, H.: Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. In: NeurIPS (2025)