VGEdit: Unlocking Video Generation Priors for Reasoning-Informed Image Editing

Haiquan Lu, Gongfan Fang, Xinyin Ma, and Xinchao Wang*

National University of Singapore
haiquanlu@u.nus.edu, xinchao@nus.edu.sg

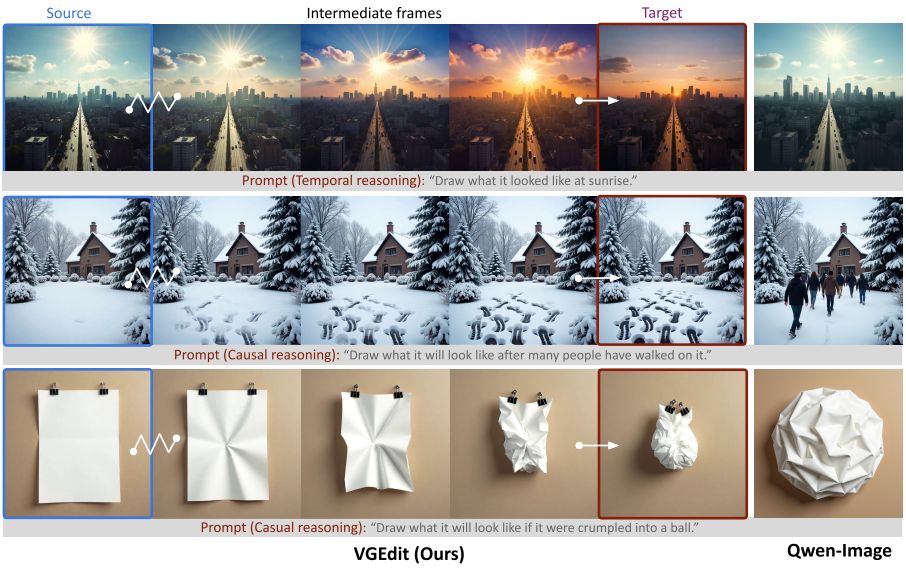

Fig. 1: By reframing reasoning-informed image editing as a video generation process, our method produces temporally coherent intermediate frames as an implicit reasoning trajectory, enabling the final frame to produce the desired edited result.

Abstract. Image editing has become an important capability of modern visual generative models. Most existing approaches generate the target image directly given an input image and an editing instruction. While effective for simple appearance modifications, this paradigm often fails when edits require multi-step inference over spatial, temporal, causal, or commonsense relationships, which entails reasoning about how the scene should evolve through a series of plausible intermediate states from the source image to the desired result. Notably, modern video generation models naturally capture such transformation processes: by modeling how scenes evolve over time, they encode strong temporal coherence and

* Corresponding author.

implicit physical priors. However, directly applying video models to image editing remains challenging due to the mismatch between editing instructions and video generation objectives. To bridge this gap, we propose VGEdit, a reinforcement learning framework that empowers video generation models for image editing. Our method consists of two components: (1) *Instruction transfer*: an MLLM converts the editing instruction into a video description, reframing the edit as a plausible transformation process. (2) *Reinforcement learning*: we adapt the video model with reward-based optimization, treating intermediate frames as implicit reasoning chains and the last frame as the target edited image. This formulation eliminates the need for ground-truth editing videos and enables direct optimization from rewards that capture instruction faithfulness, contextual consistency, and user preference alignment. Experiments show that VGEdit significantly outperforms prior methods on reasoning-informed image editing benchmarks, demonstrating that connecting image editing and video generation in a process-centric manner provides a scalable path toward more physically grounded and reasoning-capable visual editing.

Keywords: Image Editing · Video Generation Models

1 Introduction

Recent large-scale generative models [7, 18, 25, 45, 50, 55] have deeply transformed the paradigm of image editing, enabling text-driven modifications across a wide range of applications, including social media, e-commerce, education, creative design, and simulation. These models exhibit strong instruction-following ability and controllability. However, despite their impressive visual fidelity, the physical consistency and reasoning required for complex edits remain largely underexplored. This limitation is particularly evident in reasoning-informed image editing [48, 59], where the task extends beyond simple semantic modification and requires explicit reasoning to produce physically grounded edits [48, 59]. In particular, the model needs to perform multi-step inference over spatial, temporal, causal, and commonsense relationships while preserving physical coherence throughout the editing process [59].

Meanwhile, large-scale video generative models have exhibited emerging capabilities in visual understanding, reasoning, and physical-world simulation [13, 19, 31, 40, 43, 49]. These abilities enable models to synthesize temporally coherent and logically consistent video content conditioned on the contextual information present in the current frame. A growing body of recent work highlights this potential across diverse domains, including visual perception [43], control [8], manipulation [20], puzzle solving [24], and mathematical reasoning [37, 43]. Such visual ability has naturally inspired recent efforts to repurpose video generation for image editing by reformulating editing as a temporal transformation process from the source image to the desired result [32, 47]. However, a substantial gap remains between video generation and reasoning-informed image editing. Zero-shot prompting does not naturally align complex editing instructions with the objectives of video generation, while directly fine-tuning video models on

source–target editing pairs underutilizes their latent world knowledge and provides no explicit supervision for the intermediate transformation process. This raises a central question: **how can we more effectively leverage video foundation model priors for reasoning-informed image editing?**

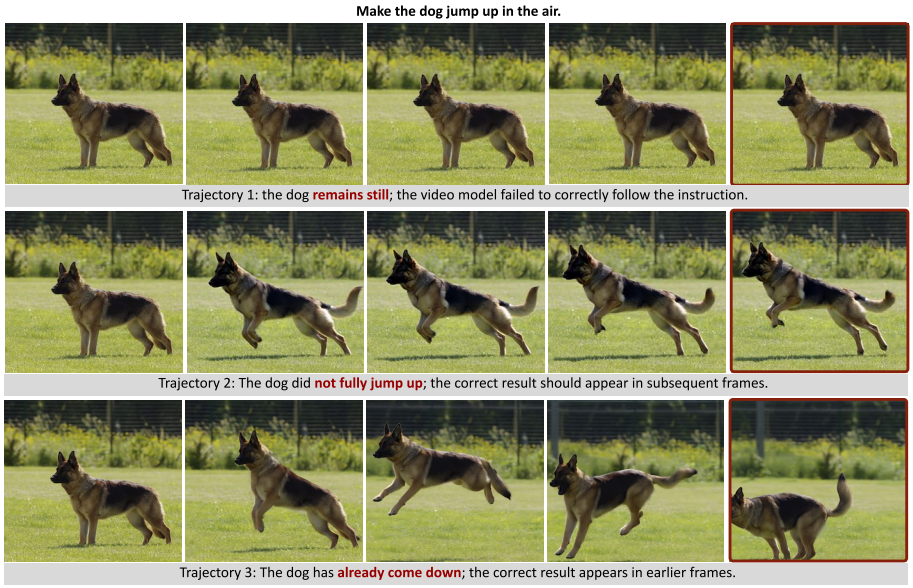

Fig. 2: Zero-shot video generation can produce diverse plausible trajectories with substantially different temporal realizations. This highlights the mismatch between video generation and image editing, which our method addresses by optimizing the full trajectory with rewards over intermediate and final frames.

Ideally, the model should use its world knowledge, physical priors, and temporally coherent intermediate frames as an implicit reasoning trajectory toward the final edited image, as illustrated in Fig. 1. However, this goal is non-trivial in practice: as shown in Fig. 2, zero-shot sampling often produces diverse yet plausible trajectories with substantially different temporal realizations, whereas image editing requires the target state to be realized at the terminal frame. This temporal under-specification leads to high variance and unreliable edits, suggesting that effective adaptation should optimize the entire trajectory to favor coherent transitions and consistent terminal satisfaction of the instruction.

Building on these motivations, we present VGEdit, a reinforcement learning framework that unlocks video foundation models' priors for reasoning-informed image editing. As illustrated in Fig. 3, VGEdit connects image editing and video generation through a two-stage framework. (1) **Instruction transfer**. To reduce the mismatch between complex editing instructions and video generation objectives, we use a multimodal large language model (MLLM) to convert the

source image and editing instruction into a compact, video-oriented prompt. Given a small set of in-context exemplars, the MLLM reframes the edit as a physically plausible transformation process from the source image to the target image, encouraging the model to generate coherent intermediate frames while steering the final frame toward the desired edit. (2) **Reinforcement learning adaptation**. We further adapt the video model using reinforcement learning, without requiring ground-truth image-editing videos. We treat each generated video as an edit trajectory, where the intermediate frames serve as a latent reasoning process and the final frame serves as the edited result. The model is optimized using reward feedback over both intermediate and final frames, encouraging instruction faithfulness, contextual consistency, and plausible transition dynamics. To obtain reliable rewards, we use a frozen pretrained MLLM to score a collage of uniformly sampled frames, which mitigates last-frame bias and provides a more stable preference signal. This stage progressively steers the model toward task-aligned generations and improves its ability to translate world knowledge and physical priors into reasoning-informed edits.

Extensive experiments demonstrate that our approach significantly outperforms previous methods on reasoning-centric image editing tasks, producing edits that are more instruction-faithful, visually coherent, and semantically grounded, while also achieving strong performance on general image editing benchmarks. These results suggest that recasting image editing as a video-based transformation process offers a promising direction for more controllable and reasoning-capable image editing. In summary, our contributions are as follows:

1. We propose an instruction transfer mechanism that uses an MLLM to convert editing instructions into video-oriented descriptions, narrowing the gap between image-editing objectives and video generation.
2. We introduce an RL strategy that optimizes the video generation model with MLLM-based rewards over both intermediate and final frames, enabling it to better exploit its latent world knowledge for reasoning-informed edits.
3. We demonstrate that our framework, VGEdit, turns video generation into an effective paradigm for reasoning-informed image editing, offering a scalable path toward physically plausible and reasoning-capable image editing.

2 Related Work

2.1 Reasoning-Informed Image Editing.

Reasoning-informed image editing aims to connect high-level semantic understanding and instruction-level reasoning with precise visual manipulation [48, 59]. Early editing approaches operate on pretrained diffusion models without additional training, enabling controllable edits through techniques such as partial denoising [26], attention manipulation [11, 42], mask-guided blending [1, 41], and latent inversion [16]. Despite their flexibility, these methods often struggle with complex semantics and physically grounded, reasoning-intensive edits.

Recent advances in image editing have been largely driven by large-scale foundation models. Instruction-based methods such as InstructPix2Pix [4] learn mappings from paired images and edit instructions, while larger models, such as UniReal [6] and FLUX.1 Kontext [18], further improve alignment and fidelity through scaling. Unified multimodal frameworks [21,22,45,46,51] further integrate visual understanding and generation, enabling open-domain and interactive editing, as exemplified by OmniGen [50], Qwen-Image-Edit [45], and Bagel [7].

More recently, several works have begun to repurpose video generation for image editing by reformulating editing as a temporal transformation process from the source image to the desired result [32,47]. This perspective is appealing because it introduces intermediate visual states that may capture the progression of an edit, rather than relying solely on direct one-step image manipulation. However, a substantial gap remains between general video generation and reasoning-informed image editing. In particular, zero-shot prompting does not reliably align complex editing instructions with the objectives of video generation, while directly fine-tuning video models on source-target editing pairs underutilizes their latent world knowledge and provides no explicit supervision over the intermediate transformation process.

Despite this progress, current approaches still struggle with edits that require deeper logical reasoning and physically consistent transformations, especially when multi-step spatial, temporal, causal, and commonsense inference is involved. Our work builds on the emerging video-based perspective, but goes beyond prior methods by explicitly leveraging the process priors of video generation models and adapting them through trajectory-level learning, enabling reasoning-informed image editing via implicit visual reasoning rather than direct one-step manipulation.

2.2 Video Generation Models.

Large-scale video generative models [5, 29, 33–36, 39, 44] have exhibited emerging capabilities in visual understanding, temporal reasoning, and physical-world simulation, enabling applications in perception, physical modeling, manipulation, and reasoning through video generation [9, 15, 37, 43]. A key aspect of these models is their ability to interpret the current frame and generate subsequent frames that evolve according to coherent physical and semantic dynamics. This work builds on these world-simulation priors and investigates how they can be more effectively adapted to reasoning-informed image editing.

2.3 Reinforcement learning for Visual Generation.

Reinforcement learning (RL) [10,28] has become a standard paradigm for aligning foundation models with human preferences and improving reasoning ability. Recent work extends it to visual generation. Prior approaches adapt policy optimization or preference learning to diffusion, including DDPO [3], ReFL [52], DiffusionDPO [38] and GRPO-style methods [23, 53] that couple optimization with stochastic samplers. However, RL for visual generation is challenging due

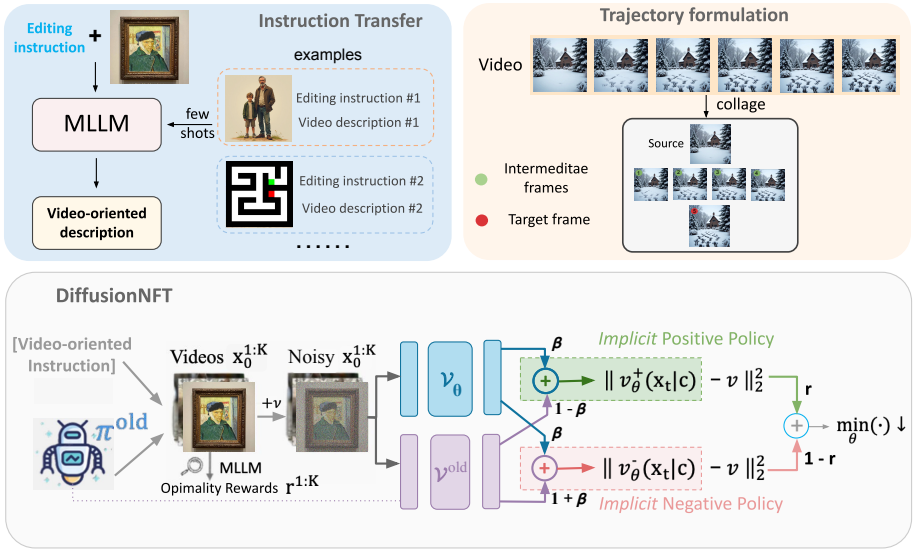

Fig. 3: Overview of VGEdit. The framework consists of two parts: (1) Instruction transfer, an MLLM converts the input edit instruction into a video-oriented temporal editing description; and (2) RL optimization, we formulate each sampled video into a frame collage for MLLM-based reward scoring, and update the latent video diffusion model using DiffusionNFT.

to the mismatch between iterative sampling and standard policy optimization, which can introduce biased signals or reduce sample diversity. DiffusionNFT [60] addresses this with an online RL formulation that directly optimizes the diffusion forward process via a negative-aware objective, avoiding likelihood estimation and better preserving generation quality.

3 Methodology

3.1 Overview

We propose **VGEdit**, a reinforcement learning framework that adapts video foundation models for reasoning-informed image editing. Instead of directly predicting an edited image, VGEdit reformulates image editing as a short video generation problem, where temporally coherent intermediate frames describe a plausible transformation process from the source image to the target image. The final frame corresponds to the edited output, while intermediate frames serve as an implicit reasoning trajectory.

Our framework consists of two components:

1. **Instruction transfer**: converting a source image and editing instruction into a video-oriented temporal editing description compatible with video generation models.

2. **Reinforcement learning adaptation**: optimizing the video model with reward feedback so that generated trajectories satisfy editing objectives and human preferences, without requiring ground-truth editing videos.

Given a source image x and an editing instruction e, our goal is to generate a video

$$V = \{v_0, v_1, \ldots, v_T\}, \tag{1}$$

where $v_0 = x$ and the final frame v_T is the edited image. The intermediate frames $\{v_1, \ldots, v_{T-1}\}$ form a physically plausible transformation process.

3.2 Instruction Transfer

A central challenge in applying video generation models to image editing is the mismatch between editing instructions and video generation objectives. Image editing instructions typically specify only the desired final appearance, whereas video models require a temporal description of how the scene evolves; using the instruction directly as a prompt often yields unstable or unrealistic results. As illustrated in Fig. 2, even with the same prompt and video generation model, zero-shot sampling can yield diverse trajectories with substantially different temporal realizations, whereas image editing requires a specific target state to be achieved at a specific time.

We address this by reformulating editing as a temporal transformation process. As shown in Fig. 3, we use an MLLM to convert the source image and editing instruction into a temporal editing description that narrates a physically plausible sequence of intermediate states, while keeping irrelevant factors (*e.g.*, camera viewpoint) fixed unless necessary. Specifically, this temporal description specifies (i) the initial state anchored to the source image, (ii) a short sequence of physically grounded intermediate transitions, and (iii) the final state corresponding to the intended edit.

To improve reliability, we employ in-context learning with several exemplar triples (source image, instruction, video-oriented description). Conditioned on this description, the video model generates coherent intermediate frames and a final frame that satisfies the desired edit, effectively turning image editing into process generation and better leveraging the model's temporal, causal, and physical priors. Concrete examples of instruction-to-description transfer are provided in the supplementary material.

3.3 Reinforcement Learning Adaptation

Although video foundation models encode rich world-simulation priors and exhibit strong physical consistency, they do not reliably produce task-aligned edits and often show high generation variance as shown in Fig. 2. Moreover, collecting ground-truth videos that satisfy reasoning-informed editing for supervised adaptation is impractical. Naturally, reinforcement learning offers a framework to address these challenges, as it enables the model to learn from reward feedback that evaluates both intermediate and final frames, encouraging instruction faithfulness, contextual consistency, and preference-aligned transformation dynamics.

Trajectory formulation. To enable trajectory-level reward signals while keeping evaluation efficient and less biased toward the final frame, we treat the generated video $V = \{v_0, \ldots, v_T\}$ as an edit transformation process, where intermediate frames act as latent reasoning steps and the last frame v_T is the edited result. Inspired by [17,32], we uniformly sample frames from the trajectory and construct a compact representation (a collage) denoted as $C(V)$ as shown in Fig. 3. This collage can then be fed to the frozen MLLM judge for reward computation, substantially reducing the cost of evaluating long video sequences while mitigating evaluation bias from relying on a single frame.

Reward feedback. We leverage a pretrained MLLM as a training-free reward model to evaluate editing results. For each editing instance, the video generator produces an edit trajectory $V = \{v_0, \ldots, v_T\}$, from which we build a compact collage representation $C(V)$ (as mentioned above). The MLLM is then prompted with the source image, the editing instruction, and the collage, and returns a scalar reward that reflects how well the generated trajectory satisfies the editing requirement:

$$R(V) = f_{\text{MLLM}}(x, e, C(V)). \tag{2}$$

Following [21], we compute rewards from the MLLM's logit distribution over score tokens rather than sampling a single token. Let $X = (x, e, C(V))$ denote the MLLM input and let $R_{<n}$ be the preceding output tokens. The probability of the next token is

$$p(r_n \mid X, R_{<n}) = \text{softmax}(D(X, R_{<n}))\,, \tag{3}$$

where $D(\cdot)$ outputs the logits.

We restrict scoring to a predefined set of rating tokens $\mathcal{M}$. Each token $r \in \mathcal{M}$ is mapped to a scalar score $w(r)$ representing its rating level. We then compute a continuous reward as the expectation over the token distribution:

$$R(V) = s_{\text{con}}(X) = \sum_{r \in \mathcal{M}} w(r) \cdot p(r_n = r \mid X, R_{<n}). \tag{4}$$

Compared to sampling a single token or using discrete scores, this logit-based expectation provides a smoother and more stable reward signal and captures the MLLM's internal uncertainty, thereby reducing hallucinations and reward variance during training.

DiffusionNFT. We adopt Diffusion Negative-aware Fine-Tuning (DiffusionNFT) [60] as our RL training method. Unlike prior online diffusion RL methods [3,23,38], DiffusionNFT performs policy optimization on the forward diffusion process via a flow-matching objective, rather than optimizing over the reverse denoising trajectory. By contrasting reward-partitioned positive and negative samples, it integrates reinforcement signals directly into a standard diffusion training objective. As a native off-policy formulation, it decouples policy training

from data sampling, supports arbitrary black-box solvers during sampling, and eliminates the need for likelihood estimation. This is particularly beneficial in our setting, as it enables sample-efficient off-policy optimization while supporting arbitrary samplers.

In our setting, as shown in Fig. 3, a latent video diffusion model parameterized by the denoiser $\epsilon_\theta(z_t, t, c)$ generates an edit trajectory V conditioned on the transferred prompt c. We compute a reward $R(V)$ using the frozen MLLM judge (Eq. (2)) and convert it into an optimality probability $r(V, c) \in [0, 1]$:

$$r(V,c) = \frac{1}{2} + \frac{1}{2}\,\mathrm{clip}\left(\frac{R(V) - \mathbb{E}_{V\sim p_{\theta^{\mathrm{old}}}(\cdot|c)}[R(V)]}{Z_c}, -1, 1\right), \tag{5}$$

where θ^{old} is a frozen reference model and $Z_c > 0$ is a normalizing factor.

DiffusionNFT defines implicit positive/negative denoisers by mixing the current and reference models:

$$\epsilon_\theta^+(z_t, t, c) = (1-\beta)\,\epsilon_{\theta^{\mathrm{old}}}(z_t, t, c) + \beta\,\epsilon_\theta(z_t, t, c), \tag{6}$$

$$\epsilon_\theta^-(z_t, t, c) = (1+\beta)\,\epsilon_{\theta^{\mathrm{old}}}(z_t, t, c) - \beta\,\epsilon_\theta(z_t, t, c), \tag{7}$$

and updates θ by minimizing the reward-weighted contrastive regression:

$$\begin{aligned}\mathcal{L}_{\mathrm{NFT}}(\theta) = \mathbb{E}_{V\sim p_{\theta^{\mathrm{old}}}(\cdot|c),\, z_0, \epsilon, t}\Big[& r(V,c)\,\|\epsilon_\theta^+(z_t,t,c) - \epsilon\|_2^2 \\ & + (1 - r(V,c))\,\|\epsilon_\theta^-(z_t,t,c) - \epsilon\|_2^2\Big],\end{aligned} \tag{8}$$

where $z_t = \sqrt{\bar\alpha_t}\,z_0 + \sqrt{1-\bar\alpha_t}\,\epsilon$ is the noised latent. We periodically refresh the reference parameters $\theta^{\mathrm{old}} \leftarrow \theta$ for stable online adaptation. The general pipeline of our method is illustrated in the supplementary material.

4 Experiments

4.1 Experimental Setups

Evaluation. For quantitative evaluation, we use two reasoning-centric image editing benchmarks, RISE [59] and KRIS [48], which assess instruction-following edits that require multi-step reasoning. RISE emphasizes reasoning-informed editing across temporal, causal, spatial, and logical dimensions, while KRIS provides a diagnostic taxonomy that groups edits by factual, conceptual, and procedural knowledge. In addition, we evaluate on a general-purpose editing benchmark, ImgEdit [54], to verify that video-generation-based methods remain effective for standard visual editing tasks. For all three benchmarks, we use GPT-4o or GPT-4.1 [27] as the automatic evaluator, following the official benchmark implementations. We additionally include a human user study in the supplementary material.

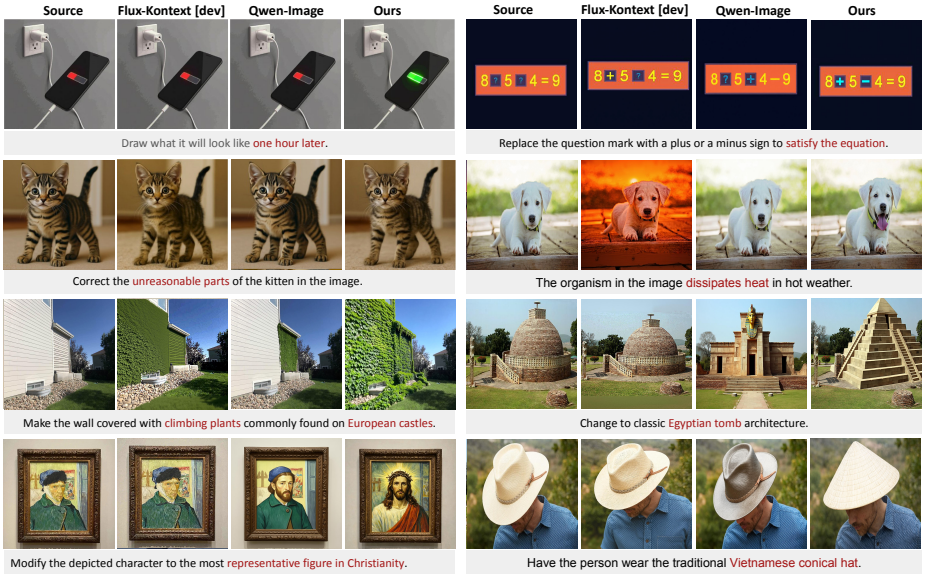

Fig. 4: Qualitative results on reasoning-informed image editing. VGEdit produces edits that are more instruction-faithful and physically consistent, while maintaining subject identity and scene coherence.

Implementation Details. We use Qwen3-VL-30B-A3B-Instruct [2] as the MLLM for both instruction transfer and reward evaluation, providing rewards that reflect reasoning correctness, contextual/identity consistency, and overall visual quality. For video generation, we adopt HunyuanVideo-1.5 [44] and generate trajectories with 31 frames. Training uses a group size of 24 and a batch size of 6, and is conducted on 2 nodes, each equipped with 8 H200 GPUs. We fine-tune the model using LoRA [12] with an adapter strength of 64 and a rank of 32. We optimize the model with the Muon optimizer [14], which accelerates convergence and improves training stability. Additional implementation details and concrete prompt templates are provided in the supplementary material.

Compared Methods. We mainly compare VGEdit with representative baselines from two categories. The first category includes large-scale image generation model-based methods. Specifically, we consider Step1X-Edit [25] and FLUX.1-Kontext-Dev [18]. The second category includes unified multimodal frameworks, which jointly model visual understanding and generation and have shown strong potential for open-domain and reasoning-related editing. Specifically, we compare with Bagel [7], OmniGen2 [46], UniCoT [30] and Qwen-Image [45]. Together, these methods provide a diverse set of baselines, covering both foundation editing models and general multimodal models.

Table 1: Quantitative results on RISE-Bench. We report category-wise accuracy (%) evaluated by GPT-4.1 on Temporal, Causal, Spatial, Logical, and the Overall average.

Models	Temporal (%)	Causal (%)	Spatial (%)	Logical (%)	Overall (%)$_{\uparrow}$
Step1X-Edit [25]	0.0	2.2	2.0	3.5	1.9
Bagel [7]	2.4	5.6	14.0	1.2	6.1
Bagel-Think [7]	5.9	17.8	21.0	1.2	11.9
OmniGen2 [46]	1.2	1.0	0.0	1.2	0.8
UniCoT [30]	8.2	18.9	20.0	1.2	12.5
Flux.1-Kontext-Dev [18]	2.3	5.5	13.0	1.2	5.8
Qwen-Image [45]	4.7	10.0	17.0	2.4	8.9
HunyuanVideo-1.5 [44]	3.6	2.5	15.8	2.8	6.2
Ours	**13.7**	**34.5**	**25.0**	**16.1**	**22.3**

4.2 Main Results

Reasoning-informed image editing. Table 1 and Table 2 summarize results on the reasoning-centric benchmarks RISE-Bench [59] and KRIS-Bench [48]. VGEdit consistently outperforms prior image editing baselines (including multi-model image editing methods [7, 30, 45]) on both benchmarks. Compared to image-only editors, which largely rely on single-step appearance manipulation and lack explicit process modeling or physical priors, VGEdit benefits from temporally coherent intermediate frames that enforce a plausible transformation trajectory. This process-centric generation helps avoid physically inconsistent shortcuts by enforcing coherent transitions and preserving contextual consistency.

The improvements are especially pronounced on categories that require multi-step inference (*e.g.*, temporal/causal reasoning in RISE and procedural knowledge in KRIS), where standard instruction-following editors often produce visually plausible but physically inconsistent or logically incorrect edits. As illustrated in Fig. 4, image-only editors may satisfy surface-level appearance cues while violating commonsense dynamics (*e.g.*, incorrect cause-effect outcomes, physically or commonsense-violating results, or broken object interactions). In contrast, VGEdit leverages the video model's implicit world-simulation and physical grounding priors to maintain temporal continuity and contextual consistency throughout the edit trajectory, resulting in edits that are more coherent, physically plausible, and logically aligned with the instruction.

General editing. We further evaluate VGEdit on the general-purpose ImgEdit benchmark [54] as shown in Tab. 3. Despite being designed for reasoning-intensive edits, VGEdit remains competitive and often improves over strong baselines on standard editing tasks, indicating that video generation based methods are promising path towards visual editing tasks. Qualitatively, VGEdit tends to preserve subject identity and contextual details more reliably while producing smoother, more physically grounded transformations (as shown in Fig. 5). These results suggest that trajectory-level alignment improves not only complex reasoning

Table 2: Quantitative comparisons on KRIS-Bench. We report the composite score for each category and the average Instruction Following (IF), Visual Consistency (VC), Visual Quality (VQ).

Models	Attribute Percep.	Spatial Percep.	Social Science	Natural Science	Logical Reason.	Instruction Decomp.	Overall
Step1X-Edit [25]	55.50	51.75	44.69	49.06	40.88	22.75	43.29
Bagel [7]	61.39	62.08	50.21	46.26	30.21	48.44	48.69
Bagel-Think [7]	60.39	61.19	49.06	47.44	29.44	48.36	48.71
OmniGen2 [46]	65.41	53.36	50.46	45.30	32.19	56.36	49.24
UniCoT [30]	67.94	73.72	59.45	53.19	40.97	54.67	56.76
Flux.1-Kontext-Dev [18]	70.78	69.20	51.27	52.05	45.82	73.67	57.35
Qwen-Image [45]	72.57	**79.92**	61.45	56.38	48.57	78.44	62.77
HunyuanVideo-1.5 [44]	65.67	61.92	53.87	51.40	52.78	63.39	57.71
Ours	**75.88**	77.92	**69.04**	**75.20**	**62.29**	**86.56**	**72.48**

edits but also general instruction-following behavior, and that repurposing video generation as a process solver is broadly beneficial for image editing.

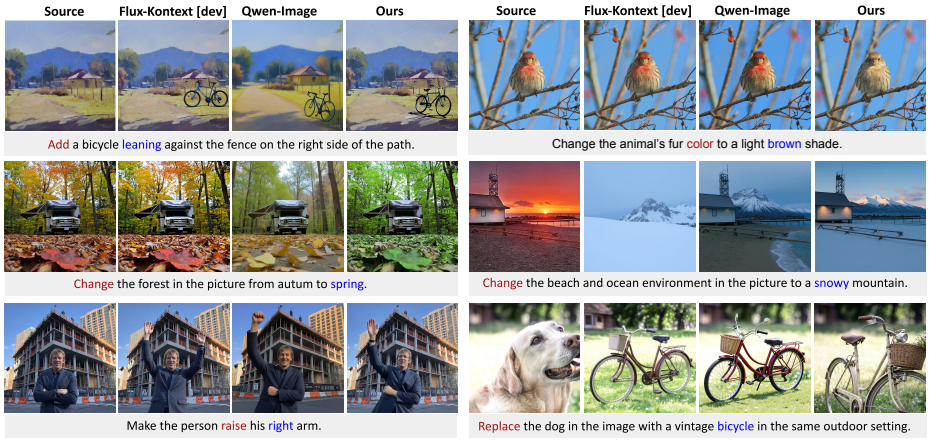

Fig. 5: Qualitative results on general image editing (ImgEdit). VGEdit remains competitive on standard editing tasks while preserving visual quality and content consistency, demonstrating good performance beyond reasoning-centric settings.

4.3 Analysis and Ablations

Effect of instruction transfer and RL adaptation. We first study the contribution of the two core components of VGEdit: instruction transfer and RL adaptation. Table 4 compares three settings: (i) the original HunyuanVideo-1.5

Table 3: Quantitative performance on general editing benchmark (ImgEdit). We report category-wise and overall scores on various editing tasks. We use GPT-4.1 for evaluation following the original paper.

Models	Add	Adjust	Extract	Replace	Remove	BackG	Style	Hybrid	Action	Overall↑
MagicBrush [56]	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.83
Instruct-Pix2Pix [4]	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.20	1.46	1.88
AnyEdit [55]	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
UltraEdit [58]	3.44	2.81	2.13	2.96	1.45	2.83	3.76	1.91	2.98	2.70
ICEdit [57]	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
Step1X-Edit [25]	**3.88**	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
Bagel [7]	3.81	3.59	1.58	3.85	3.16	3.39	4.51	2.67	4.25	3.42
UniWorld-V1 [22]	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
OmniGen2 [46]	3.57	3.06	1.77	3.74	3.20	3.57	4.81	2.52	4.68	3.44
Flux.1-Kontext-Dev [18]	3.76	3.45	2.15	3.98	2.94	3.78	4.38	2.96	4.26	3.52
Qwen-Image [45]	4.38	4.16	3.43	**4.66**	4.14	4.38	4.81	3.82	4.69	4.27
GPT-Image [27]	4.61	4.33	2.90	4.35	3.66	4.57	**4.93**	**3.96**	4.89	4.20
HunyuanVideo-1.5 [44]	2.65	1.28	3.01	3.52	2.13	2.74	3.33	4.33	2.82	2.87
Ours	3.82	**4.56**	**4.14**	4.56	**4.40**	**4.48**	4.40	3.69	**4.95**	**4.33**

Table 4: Ablation study on core components of VGEdit (instruction transfer and RL adaptation) on RISE-Bench, KRIS-Bench and ImgEdit.

Models	Reasoning-informed		General	Average
	RISE-Bench	KRIS-Bench	ImgEdit	
HunyuanVideo-1.5	6.2	57.71	2.87	22.26
w/ Instruction Transfer	15.2	67.96	3.56	28.91
w/ Instruction Transfer and RL	22.3	72.48	4.33	33.04

in a zero-shot manner, (ii) HunyuanVideo-1.5 with instruction transfer only, and (iii) the full VGEdit with both instruction transfer and RL adaptation.

The zero-shot video model struggles with complex editing instructions: the generated transformations are frequently unrelated to the requested edits and rarely reach the desired final state, reflecting the mismatch between image-editing instructions and the native objective of video generation. Adding instruction transfer substantially improves instruction alignment by converting the original edit into a video-oriented transformation description, making the model more likely to generate meaningful intermediate transitions and produce the correct edited result. However, the outputs remain inconsistent across prompts, indicating that task reformulation alone is insufficient to reliably induce task-consistent edit trajectories. With RL adaptation, the full VGEdit achieves the strongest and most consistent performance, as reward feedback over intermediate and final frames steers the model toward trajectories that better align with the editing instruction and visual context, improving both the final edited image and the coherence of the transformation process.

Overall, the two components play complementary roles: instruction transfer provides a task formulation that is compatible with video generation, while RL adaptation aligns the generative dynamics with the editing objective. Their combination is therefore essential for reliable reasoning-informed image editing.

Table 5: Ablation study of the reward formulation on RISE-Bench, KRISE-Bench, and ImgEdit.

Reward Design	Reasoning-informed		General	Average
	RISE-Bench	KRIS-Bench	ImgEdit	
Final-frame + Sampling	15.7	69.34	3.52	29.52
Final-frame + Logit	20.4	70.43	4.03	31.62
Trajectory-level + Sampling	17.5	68.94	4.01	30.15
Trajectory-level + Logit (Ours)	22.3	72.48	4.33	33.04

Ablation of reward formulation. We next analyze the design of the reward used for RL adaptation. In particular, we study two key choices: (i) final frame vs. full trajectory, and (ii) logit-based expectation or sampling-based discrete token selection. The quantitative comparison is reported in Tab. 5.

When the reward is computed using only the final frame, optimization becomes noticeably less stable and even downstream performance drops. This is because final-frame evaluation ignores the intermediate transformation process and tends to favor outputs that appear visually plausible at the endpoint, even when the overall edit trajectory is logically inconsistent with the instruction. In contrast, trajectory-level reward, implemented by scoring a collage of uniformly sampled frames, provides a richer supervision signal by exposing the evaluator to both the evolution of the scene and the final outcome. This leads to more reliable preference signals and consistently better editing results.

We further compare two reward computation strategies. Sampling-based reward estimation introduces additional variance due to stochastic token sampling from the evaluator, making training less stable. By contrast, our logit-based expectation directly aggregates the evaluator's token probabilities, producing a smoother and less noisy reward signal. As shown in Tab. 5, combining trajectory-level evaluation with logit-based scoring yields the best overall performance, confirming that both design choices are important for effective RL adaptation.

5 Conclusion

We presented **VGEdit**, a reinforcement learning framework that repurposes video foundation models for reasoning-informed image editing by formulating editing as a temporal transformation process. Through instruction transfer, editing instructions are converted into temporally grounded prompts, and a reward-driven adaptation stage aligns the model using MLLM-based feedback without requiring ground-truth editing videos. Experiments show that VGEdit significantly improves reasoning-centric editing performance while remaining competitive on general editing tasks, demonstrating that the world knowledge, physical priors, and temporal consistency of video models can be effectively harnessed for complex visual editing, and offering a scalable path toward more physically plausible and reasoning-capable visual editing.

Acknowledgements

This project is supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2026, WBS: A-8004345-00-00), the National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity R&D Programme and CyberSG R&D Cyber Research Programme Office (Award: CRPO-GC1-NTU-002).

References

1. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18208–18218 (2022)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)
4. 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)
5. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024)
6. 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: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12501–12511 (2025)
7. 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)
8. Fang, G., Ma, X., Wang, X.: In-video instructions: Visual signals as generative control. arXiv preprint arXiv:2511.19401 (2025)
9. Fei, Z., Qiu, D., Li, D., Yu, C., Fan, M.: Video diffusion transformers are in-context learners. arXiv preprint arXiv:2412.10783 (2024)
10. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
11. 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)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
13. Huang, S., Wu, J., Zhou, Q., Miao, S., Long, M.: Vid2world: Crafting video diffusion models to interactive world models. arXiv preprint arXiv:2505.14357 (2025)
14. Jordan, K., Jin, Y., Boza, V., Jiacheng, Y., Cesista, F., Newhouse, L., Bernstein, J.: Muon: An optimizer for hidden layers in neural networks (2024), `https://kellerjordan.github.io/posts/muon/`

15. Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., Feng, J.: How far is video generation from world model: A physical law perspective. arXiv preprint arXiv:2411.02385 (2024)
16. 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)
17. Kim, W., Choi, C., Lee, W., Rhee, W.: An image grid can be worth a video: Zero-shot video question answering using a vlm. IEEE Access **12**, 193057–193075 (2024)
18. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742
19. Li, D., Fang, Y., Chen, Y., Yang, S., Cao, S., Wong, J., Luo, M., Wang, X., Yin, H., Gonzalez, J.E., et al.: Worldmodelbench: Judging video generation models as world models. arXiv preprint arXiv:2502.20694 (2025)
20. Li, H., Sun, L., Hu, Y., Ta, D., Barry, J., Konidaris, G., Fu, J.: Novaflow: Zero-shot manipulation via actionable flow from generated videos. arXiv preprint arXiv:2510.08568 (2025)
21. 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)
22. 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)
23. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. arXiv preprint arXiv:2505.05470 (2025)
24. Liu, J., Teshome, W., Ghimire, S., Sznaier, M., Camps, O.: Solving masked jigsaw puzzles with diffusion vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23009–23018 (2024)
25. 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)
26. 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)
27. Open, A.: Introducing gpt-4.1 in the api (2025)
28. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
29. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024)
30. 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)

31. Ren, Z., Wei, Y., Guo, X., Zhao, Y., Kang, B., Feng, J., Jin, X.: Videoworld: Exploring knowledge learning from unlabeled videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29029–29039 (2025)
32. 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)
33. Seawead, T., Yang, C., Lin, Z., Zhao, Y., Lin, S., Ma, Z., Guo, H., Chen, H., Qi, L., Wang, S., et al.: Seaweed-7b: Cost-effective training of video generation foundation model. arXiv preprint arXiv:2504.08685 (2025)
34. Team, G.: Mochi 1. https://github.com/genmoai/models (2024)
35. Team, V.: Veo: A text-to-video generation system. Tech. rep., Google Deepmind (2025), https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf
36. Technology, K.: Kling ai: Text-to-video and image-to-video generation model. https://klingai.com/global/ (2024), accessed: 13 November 2025
37. Tong, J., Mou, Y., Li, H., Li, M., Yang, Y., Zhang, M., Chen, Q., Liang, T., Hu, X., Zheng, Y., et al.: Thinking with video: Video generation as a promising multimodal reasoning paradigm. arXiv preprint arXiv:2511.04570 (2025)
38. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8228–8238 (2024)
39. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
40. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. Advances in Neural Information Processing Systems **36**, 7594–7611 (2023)
41. Wang, Y., Zhang, W., Xu, H., Jin, C.: Dreamtext: High fidelity scene text synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28555–28563 (2025)
42. Wang, Y., Zhou, C., Xu, H.: Enhancing object coherence in layout-to-image synthesis. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2025)
43. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv preprint arXiv:2509.20328 (2025)
44. Wu, B., Zou, C., Li, C., Huang, D., Yang, F., Tan, H., Peng, J., Wu, J., Xiong, J., Jiang, J., et al.: Hunyuanvideo 1.5 technical report. arXiv preprint arXiv:2511.18870 (2025)
45. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324
46. 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)

47. 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)
48. 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)
49. Xiang, J., Liu, G., Gu, Y., Gao, Q., Ning, Y., Zha, Y., Feng, Z., Tao, T., Hao, S., Shi, Y., et al.: Pandora: Towards general world model with natural language actions and video states. arXiv preprint arXiv:2406.09455 (2024)
50. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13294–13304 (2025)
51. Xin, Y., Qin, Q., Luo, S., Zhu, K., Yan, J., Tai, Y., Lei, J., Cao, Y., Wang, K., Wang, Y., et al.: Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. arXiv preprint arXiv:2510.06308 (2025)
52. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 15903–15935 (2023)
53. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)
54. 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)
55. 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: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26125–26135 (2025)
56. 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)
57. Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: Enabling instructional image editing with in-context generation in large scale diffusion transformer. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
58. Zhao, H., Ma, X.S., Chen, L., Si, S., Wu, R., An, K., Yu, P., Zhang, M., Li, Q., Chang, B.: Ultraedit: Instruction-based fine-grained image editing at scale. Advances in Neural Information Processing Systems **37**, 3058–3093 (2024)
59. 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)
60. Zheng, K., Chen, H., Ye, H., Wang, H., Zhang, Q., Jiang, K., Su, H., Ermon, S., Zhu, J., Liu, M.Y.: Diffusionnft: Online diffusion reinforcement with forward process. arXiv preprint arXiv:2509.16117 (2025)