PRECISE EDITING AND FLEXIBLE REFERENCING FOR INTERACTABLE WORLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

We present EditWorld, a video world model for precise editing and flexible referencing in interactable worlds. Existing video world models primarily focus on navigation, letting users explore generated worlds but offering limited control over how existing world content is modified. EditWorld extends world modeling from exploration to precise modification by streaming editing instructions and reference images during autoregressive generation. To support these capabilities, EditWorld introduces Gated Causal Attention for temporally varying editing conditions and reference images, together with a Sparse Context mechanism that maintains a bounded historical context for long-horizon inference. We further adopt joint autoregressive and bidirectional training with annealed self-resampling, and construct a dedicated data synthesis and annotation pipeline that provides supervision for world editing. We also present WBench-Editing to systematically evaluate streaming world editing capabilities. EditWorld achieves the best overall performance on WBench-Editing with an overall score of 73.8 and an editing score of 80.0, substantially outperforming existing methods on editing-related metrics.

1 INTRODUCTION

Video world models, driven by recent advances in autoregressive video generation, have emerged as a promising substrate for world exploration (Robbyant et al., 2026; Gao et al., 2026; Sun et al., 2025), game generation (Li et al., 2025; Tang et al., 2025), and embodied simulation (Kairos et al., 2026). These models autoregressively generate videos in response to streaming user inputs, such as actions and prompts. Existing video world models, however, have primarily emphasized navigation, focusing on faithful control of camera trajectories and user actions. More recent efforts have begun to extend this capability toward text-driven event generation. YUME 1.5 (Mao et al., 2026) supports text-controlled world events, HY-WorldPlay 1.5 (Sun et al., 2025) enables promptable events across diverse scenes, and LingBot-World and LingBot-World 2.0 (Robbyant et al., 2026; Gao et al., 2026) further expand the range of text-driven events and interactive actions. DreamX-World (DreamX et al., 2026) additionally introduces composable event control through event instruction tuning. Despite this progress, existing approaches mainly focus on triggering or generating new events, rather than precisely modifying specified content already present in the world, such as addition, removal, replacement, and stylization. Moreover, flexible incorporation of content from reference images remains insufficiently explored. XGEN-JING (XGEN-JING, 2026) and ABot-World (Jiang et al., 2026) support identity conditioning from initial references, but do not enable users to interactively and flexibly inject content from different reference images into the generated world over time. As a result, although existing world models increasingly support navigation, text-driven events, and reference-based conditioning, they still lack precise control over editing existing world content and flexible integration of reference images throughout interaction.

To this end, we present **EditWorld**, a video world model for precise editing and flexible referencing in interactable worlds. EditWorld enables users to continuously modify world content through streaming editing instructions and to flexibly incorporate information from reference images during generation. Specifically, our model is built upon an autoregressive video generation framework in which images or videos are used as the world prior, while camera poses, textual prompts, and reference images are incorporated as conditioning signals for navigation and modification. Starting from LingBot-World-Base (Robbyant et al., 2026), a bidirectional video world model, Gated Causal Attention is introduced to support streaming editing instructions and reference images while preserv-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ing causal video generation, and the model is further adapted to autoregressive generation through teacher-forcing (Williams & Zipser, 1989) training. Joint autoregressive and bidirectional objectives (Gao et al., 2026) are adopted to improve condition-following capability. To prevent the video context from growing unboundedly with video length, a Sparse Context mechanism is designed to constrain the historical context to a fixed budget during both training and inference. Furthermore, self-resampling (Guo et al., 2025) is adopted to mitigate error accumulation during autoregressive rollouts and improve fidelity and stability over long-horizon generation. On the data side, a dedicated pipeline is developed. Based on public navigation and editing datasets (Li et al., 2026; Wang et al., 2026a; Zhou et al., 2025; He et al., 2025a; Bai et al., 2026), multiple off-the-shelf methods are leveraged to construct editable world data with detailed editing annotations and reference images, providing supervision for fine-grained content modification and reference-guided generation.

Since existing world model benchmarks primarily evaluate navigation capabilities (Wu et al., 2026a; Lu et al., 2026; Ding et al., 2026; Xu et al., 2026b), we introduce WBench-Editing, a sub-benchmark of WBench (Ying et al., 2026) designed to systematically evaluate and compare the streaming editing capabilities of existing world models. WBench-Editing consists of approximately 150 cases, each spanning 240–480 frames and involving one to three streaming editing instructions, with a subset additionally incorporating reference images. Our model achieves the best overall performance on WBench-Editing, with an overall score of 73.8 and an editing score of 80.0, substantially outperforming existing methods in editing capability. To further demonstrate that our approach preserves strong general world modeling capabilities, we also report results on the original WBench, where our model achieves performance comparable to several commercial world models.

In summary, this paper makes the following contributions:

- A video world model for **precise editing and flexible referencing** is developed, enabling users to continuously modify world content through streaming edit instructions and incorporate content from reference images.
- WBench-Editing, a sub-benchmark of WBench, is introduced to systematically evaluate streaming world editing capabilities.
- Our model achieves the best overall performance on WBench-Editing, with a substantial advantage in editing capability, while maintaining competitive performance on WBench.

2 Related Work

Video World Models. Video world models aim to create infinite worlds with versatile interactions. YUME (Mao et al., 2025), ASTRA (Zhu et al., 2026c), and Matrix-Game (Zhang et al., 2025) pioneer interactive video world modeling by generating videos from input images and enabling world exploration through action or camera-trajectory control. Subsequent efforts have focused on improving viewpoint-control accuracy. HY-World 1.5 (Sun et al., 2025) utilizes PRoPE, which explicitly incorporates camera poses as positional priors for video tokens. LingBot-World (Robbyant et al., 2026) encodes camera poses as Plücker features to inject token-wise spatial information. The Hunyuan-GameCraft series (Li et al., 2025; Tang et al., 2025) maps keyboard and mouse inputs into a shared camera representation space, while the Matrix-Game series (Zhang et al., 2025; He et al., 2025b; Wang et al., 2026b) enables frame-level keyboard and mouse conditioning. DreamX-World (DreamX et al., 2026) introduces E-PRoPE with relative frustum-based encoding, and Wonder (Xu et al., 2026a) further improves camera-pose control through a dense coordinate field. Some works improve long-horizon world consistency by introducing specialized memory modules to mitigate appearance drift when revisiting previously explored locations. AlayaWorld (AlayaWorld et al., 2026) combines an explicit 3D reprojection cache with compressed representations of recent frames. WorldKV (Yi et al., 2026) retrieves evicted KV-cache chunks according to the current camera viewpoint, while StableWorld (Yang et al., 2026) employs Dynamic Frame Eviction to discard frames that have accumulated visual drift. Generation efficiency is another key factor in the practicality of video world models. SANA-WM (Zhu et al., 2026a) leverages linear attention for real-time generation over minute-long horizons, while minWM (Zhao et al., 2026a) and SolarWM (Huang et al., 2026a) provide a fully open-source end-to-end framework for efficient world modeling. ABot-World-0 (Jiang et al., 2026) improves generation speed through a lightweight VAE decoder and efficient attention mechanisms, while MoWorld (Moxin et al., 2026) demonstrates real-time interactive world modeling on NPUs. Beyond navigation and action control, recent works

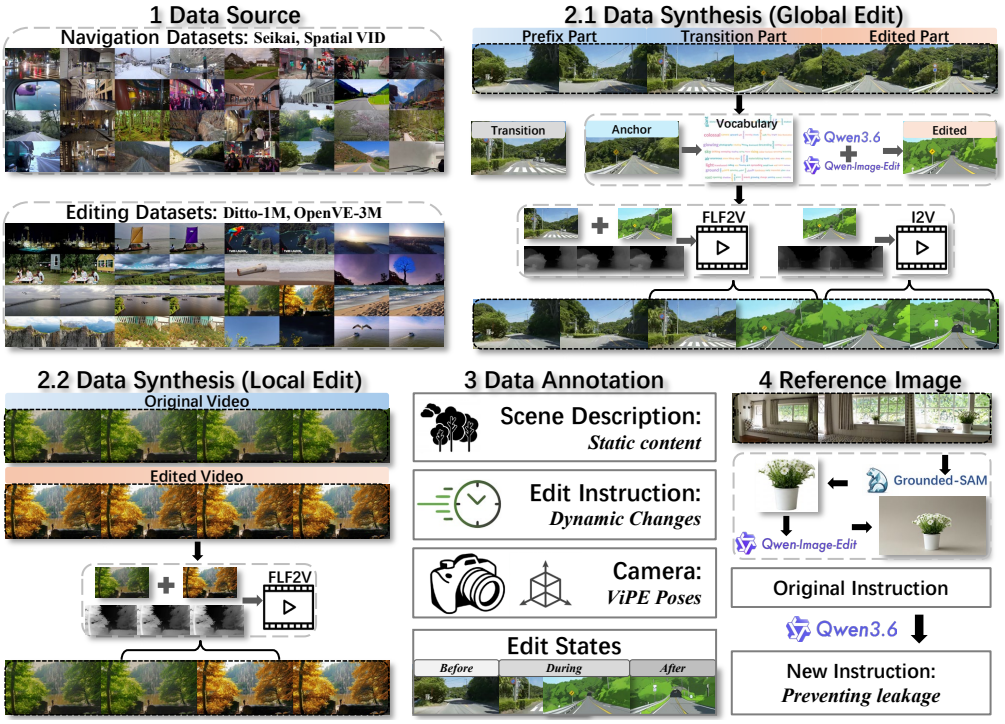


Figure 1: Data pipeline overview. (1) Our raw data is collected from two main sources: open-source navigation datasets and video editing datasets. (2) Sections 2.1 and 2.2 illustrate the synthesis pipelines for constructing global and local editable world data from the source videos. (3) Each synthesized training sample is annotated with four components: scene description, editing instruction, editing state, and camera poses. (4) Reference images are further extracted from the synthesized videos, together with rewritten editing instructions to prevent semantic leakage.

have also explored text-conditioned event generation and initial reference-based identity conditioning. YUME 1.5 (Mao et al., 2026) improves the accuracy of text-controlled event generation, while LingBot-World 2.0 (Gao et al., 2026) supports multiple text-driven events. XGEN-JING (XGEN-JING, 2026) and ABot-World-0 (Jiang et al., 2026) further introduce reference-based identity conditioning, enabling identity information to be preserved during world generation.

Autoregressive Video Generation. Autoregressive video generation serves as a key technology for interactable video world models. Diffusion Forcing (Chen et al., 2024) assigns an independent noise level to each frame, unifying next-frame prediction with full-sequence diffusion. Self-Forcing (Huang et al., 2026b) addresses exposure bias by performing autoregressive rollouts with KV caching during training, such that each frame is conditioned on previously generated outputs. Building on this paradigm, Self-Forcing++ (Cui et al., 2026) samples training clips from self-generated long videos and leverages knowledge from a teacher model, while Self Gradient Forcing (Zhuang et al., 2026) enables gradients from future predictions to propagate through historical KV states. Context Forcing (Chen et al., 2026) further replaces short-context teachers with long-context supervision. Self-Resampling (Guo et al., 2025) performs end-to-end training from scratch while explicitly simulating inference-time errors during training, whereas Reward-Forcing (Zhang et al., 2026a) replaces teacher supervision with reward signals. Causal Forcing (Zhu et al., 2026b) identifies a theoretical inconsistency in distilling autoregressive students from bidirectional teachers, arising from violations of frame-level injectivity. Causal Forcing++ (Zhao et al., 2026b) extends this framework to one/two-step sampling per frame and identifies initialization as a critical bottleneck.

3 Methodology

3.1 Data Pipeline

Existing world model datasets typically consist of navigation videos collected in large-scale environments, providing limited supervision for modifying world content. In contrast, public video editing

datasets usually contain only source–edited video pairs and do not explicitly model the smooth temporal transition from the original state to the edited state. To train a video world model for precise editing and flexible referencing in interactable worlds, we develop a data pipeline for synthesizing long-horizon navigation videos with temporally grounded editing events and reference-image conditioning. As shown in Figure 1, the source data are drawn from two categories: Sekai (Li et al., 2026), OmniWorld (Zhou et al., 2025), and SpatialVID (Wang et al., 2026a) are used as navigation data, while Ditto-1M (Bai et al., 2026) and OpenVE-3M (He et al., 2025a) are used as editing data. We broadly categorize editing operations into global editing and local editing. Global editing refers to holistic changes in the appearance of the world, such as changes in weather, season, time, illumination, color tone, and artistic style, whereas local editing refers to localized modifications to individual elements, including addition, removal, and modification. Separate synthesis pipelines are designed for these two editing categories. Each synthesized video is further annotated with scene descriptions, editing instructions, editing states, camera trajectories, and optional reference images.

Global Editing Data. The goal of global editing data is to provide supervision for holistic world transitions. Given a raw navigation video, a global editing instruction is first constructed, and an anchor frame is selected, after which the edit is propagated across the video. We build an editing vocabulary containing approximately 200 primary global editing attributes and 100 auxiliary attributes. For each video, Qwen3.6-27B (Qwen, 2026) is used to sample one primary attribute and one compatible auxiliary attribute according to the video content. The selected attribute combination is then instantiated into a concrete global editing prompt that is consistent with the current scene. The anchor frame and the editing instruction are subsequently provided to Qwen-Image-Edit (Wu et al., 2025) to generate the edited anchor frame.

Given a high-quality edited anchor frame, the single-frame edit is extended to the full video through a smooth temporal transition. Let the anchor-frame position in the original video be denoted by F_a. We determine the starting point of the transition segment, denoted by F_t, and preserve the original video before F_t as the unedited prefix. For the transition segment, the original frame at F_t is used as the first-frame condition, while the edited anchor frame at F_a is used as the last-frame condition. A VLM is then prompted to describe the smooth state transition between the two boundary frames. The transition prompt and the two boundary frames are fed into a depth-controlled first-last-frame-to-video model, Wan2.2-FLF2V-A14B-Control (Videox-fun, 2026), to synthesize the transition segment. To preserve the spatial structure and camera motion of the original video, depth maps from the corresponding temporal interval are extracted and used as structural control signals, reducing undesired drift in camera motion and scene geometry. For the segment after F_a, we employ the depth-controlled image-to-video model Wan2.2-I2V-A14B-Control (Videox-fun, 2026), using the edited anchor frame as the initial-frame condition and the depth sequence extracted from the original video after F_a as the control signal. Finally, the unedited prefix, generated transition segment, and edited continuation are concatenated to form the complete video.

Local Editing Data. Local editing data is designed to teach the model element-level modification capabilities, including adding, removing, or replacing specific elements in the world. Beyond object-level operations, local editing also covers localized attribute changes, such as modifications to color, material, shape, and state. This portion of the dataset is primarily constructed from publicly available video editing datasets that provide source videos, edited videos, and corresponding editing instructions. To improve data quality, we first apply a two-stage VLM-based filtering pipeline to the collected video pairs. In the first stage, the VLM determines whether the difference between the source and edited videos corresponds to a local editing operation. In the second stage, semantic consistency is evaluated together with the overall video quality. For each video pair that passes filtering, a source frame F_s is selected from the original video and a target frame F_t from the edited video. The source frame is required to clearly present the target element before editing, while the target frame should fully capture the desired post-edit state. The temporal interval between these two endpoint frames is treated as the transition segment. We then provide F_s, F_t, and the original editing instruction to a VLM to generate a transition prompt describing the required visual change. The transition segment is synthesized using the same depth-controlled video generation model as in the global editing pipeline. Finally, the source-video segment before F_s, the generated transition segment, and the edited-video segment after F_t are concatenated to form the final video.

Data Annotation. Our annotations consist of four components: scene description, editing instruction, editing state, and camera poses. The scene description captures the static and invariant content of the scene. To generate this annotation, the synthesized video together with its editing instruction

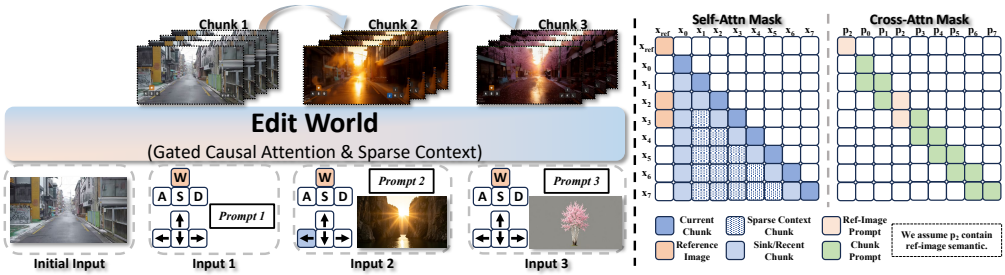


Figure 2: EditWorld. The left panel illustrates how our model uses an initial image as the world prior and autoregressively generates and edits the world in response to user-provided actions, textual prompts, and reference images. The right panel visualizes the attention patterns of our Gated Causal Attention and Sparse Context mechanisms. For simplicity, we illustrate the case with one sink chunk, one recent chunk, one preceding prompt, and one reference image.

is provided to a VLM, which is prompted to describe the scene while excluding elements affected by the editing operation. Although editing instructions are already obtained during the preceding synthesis process, they may be inaccurate or incomplete. We therefore provide the original editing instruction, the synthesized video, and the generated scene description to the VLM, and prompt it to produce a more accurate and detailed editing instruction while avoiding redundancy or conflicts with the scene description. This process decouples the textual conditioning of each video into two complementary components: the scene description, which represents static content, and the editing instruction, which specifies dynamic changes. For editing-state annotations, each video is divided into three temporal states: *before*, *during*, and *after*. This design provides fine-grained temporal supervision for chunk-level editing control. The video is partitioned into chunks, and a VLM is used to assign an editing state to each chunk. Finally, ViPE (Huang et al., 2025) is used to re-estimate the camera intrinsics and extrinsics for all training samples, providing consistent camera annotations.

Reference Images. A subset of the training data is further augmented with reference images to teach the model how to incorporate content from external references into the generated world. For local editing data, Grounded-SAM (Ren et al., 2024) is used to segment the target object from the selected reference frame. The resulting segmentation is then provided to Qwen-Image-Edit (Wu et al., 2025) to repair incomplete or imperfect regions and produce the final reference image. For global editing data, the previously generated edited anchor frame is used as the basis for reference construction. The anchor frame is provided to the image editing model, while a VLM generates an instruction that alters the scene, layout, and environment while preserving the target editing attributes, such as weather, style, or time. This process produces a reference image that retains the desired editing attributes while differing from the original world in scene-level content. We further rewrite the corresponding editing instructions for reference-conditioned samples. Specifically, descriptions of content already conveyed by the reference image are removed from the textual instruction to prevent semantic leakage. As a result, the model cannot rely solely on text to recover the target content and is instead encouraged to extract and incorporate the relevant information from the reference image.

3.2 EditWorld

As shown in the left part of Figure 2, our world model takes an initial frame as the world prior and autoregressively generates an interactable world in response to a stream of user inputs, including textual prompts, actions, and reference images. To support precise editing and flexible referencing during interaction, world generation is formulated as a causal video generation process, where each video chunk is conditioned on preceding observations and the user inputs available up to the current time step. Let $\mathcal{V} = \{x_1, x_2, \ldots, x_T\}$ denote a sequence of video chunks, where $x_t \in \mathbb{R}^{L\times H\times W\times C}$ represents a chunk of L frames at time index t, and let $\mathcal{C} = \{c_1, c_2, \ldots, c_T\}$ denote the corresponding sequence of user inputs. Under the causal formulation, the generation process is factorized as

$$p_\theta(x_{1:T} \mid c_{1:T}) = \prod_{t=1}^{T} p_\theta(x_t \mid x_{<t}, c_{\leq t}). \tag{1}$$

Here, θ denotes the model parameters. Causality is enforced through both the model architecture and the training curriculum, as detailed in the following sections.

3.2.1 CAUSAL VIDEO MODEL

We train a causal video generation model for multi-condition controllable world generation. Our model is built upon LingBot-World-Base (Robbyant et al., 2026), a bidirectional video world model. As illustrated in the right part of Figure 2, Gated Causal Attention is introduced to support streaming editing instructions and reference images during autoregressive generation while preserving temporal causality and continuity. In parallel, we design a Sparse Context mechanism that constrains the video latent context to a fixed budget during inference.

Gated Causal Attention. As described in the data pipeline section, each video is annotated with a scene description, an editing instruction, and chunk-wise editing states. These annotations are used to construct the gated cross-attention mechanism. For each video chunk x_t, the corresponding prompt tokens p_t always contain the scene description prompt p_{scene}, while the editing instruction prompt p_{edit} is activated according to the associated editing state. Specifically, p_{edit} is included in p_t only when x_t is labeled as *during*. To maintain temporal continuity when editing prompts change across chunks, each chunk x_t attends not only to its current prompt p_t, but also to the prompts of the two preceding chunks, p_{t-1} and p_{t-2}. This short-term textual context helps reduce abrupt visual transitions caused by prompt switching.

Following the chunk-by-chunk autoregressive generation paradigm, a chunk-causal self-attention pattern is adopted to preserve temporal causality. To enable causal generation without sacrificing training parallelism, the clean latent chunks are concatenated with their noisy counterparts. Let x_t^{clean} and x_t^{noisy} denote the clean and noisy latent chunks at time step t, respectively. The causal self-attention pattern is defined as

$$\mathcal{S}\left(x_t^{\text{clean}}\right) = \left\{x_j^{\text{clean}} : j \le t\right\}, \qquad \mathcal{S}\left(x_t^{\text{noisy}}\right) = \left\{x_j^{\text{clean}} : j < t\right\} \cup \left\{x_t^{\text{noisy}}\right\}, \tag{2}$$

where $\mathcal{S}(\cdot)$ denotes the set of latent blocks accessible to the corresponding attention query.

Reference images are encoded by the VAE and concatenated with the video latents along the temporal dimension. To incorporate reference images while preserving causality, we introduce a unidirectional gated self-attention mechanism between reference image tokens and video latent tokens. Tokens from each reference image are restricted to attending only to tokens within the same reference image, preventing their representations from being influenced by video tokens or other references. Conversely, the visibility of reference image tokens to a video chunk x_t is gated by its textual context, including p_t, p_{t-1}, and p_{t-2}. Specifically, x_t is allowed to attend to a reference image only when its visible prompts contain semantics referring to that reference. This design explicitly aligns reference conditioning with the textual context of each video chunk, enabling flexible incorporation of reference content at the appropriate generation stage. We further assign reference image tokens a negative temporal RoPE margin m to mitigate direct copy-and-paste behavior. For the i-th reference image, all its tokens are assigned a negative temporal RoPE coordinate $m(i+1)$. This negative temporal offset separates reference tokens from the video tokens, encouraging the model to treat reference images as conditioning signals rather than directly copying their spatial content.

Sparse Context. To prevent the visual context from growing unboundedly during autoregressive inference, we introduce a Sparse Context mechanism together with sparse attention training. When generating the target chunk x_t, the sink chunk x_0 and the two most recent chunks, x_{t-1} and x_{t-2}, are always retained. The sink chunk preserves the initial world state, while the recent chunks provide short-term temporal context for maintaining generation continuity. In addition, we retrieve the k most relevant chunks from the remaining history and include their KV caches in the active context. As a result, the context budget remains fixed throughout autoregressive generation.

To efficiently retrieve relevant historical chunks, we construct a compact query for the current chunk and compact keys for historical chunks by average pooling their pre-RoPE query and key representations. Given the compact query $\bar{Q}_t$ of the current chunk and the compact key $\bar{K}_j$ of a historical chunk, their relevance is measured using cosine similarity:

$$s_{t,j} = \frac{\langle \bar{Q}_t, \bar{K}_j \rangle}{\|\bar{Q}_t\| \, \|\bar{K}_j\|}. \tag{3}$$

The k historical chunks with the highest similarity scores are selected, and the active sparse context is constructed as

$$\mathcal{A}_t = \mathcal{S}_t \cup \mathrm{TopK}_{j \in \mathcal{H}_t}(s_{t,j}, k) \cup \mathcal{R}_t, \tag{4}$$

where $\mathcal{S}_t$, $\mathcal{R}_t$, and $\mathcal{H}_t$ denote the sink chunk, recent chunks, and the remaining historical chunks, respectively. During training, a corresponding sparse context strategy is adopted: the sink and recent chunks are always retained, while a random number of additional chunks are sampled from the remaining history. This exposes the model to diverse sparse historical contexts and improves its robustness to the fixed-budget context used during inference.

3.2.2 Training Curriculum

We adopt teacher-forcing (Williams & Zipser, 1989) to adapt the bidirectional model for autoregressive video generation. We empirically observe that optimizing only a causal generation objective tends to bias the model toward video continuation, while weakening its responsiveness to diverse input conditions. To mitigate this issue, bidirectional and autoregressive objectives are jointly optimized, allowing the model to retain strong condition-following capability while acquiring causal generation behavior. In addition, an annealed self-resampling strategy (Guo et al., 2025) is adopted to improve robustness to error accumulation by progressively exposing the model to its own generated history during training. Finally, we distill the autoregressive model into a few-step model using pCM (Wang et al., 2024), and further incorporate Self Gradient Forcing (Zhuang et al., 2026) to improve the generation quality of the few-step autoregressive model.

Joint Autoregressive and Bidirectional Objectives. Optimizing only the causal video generation objective results in relatively weak responsiveness to user inputs. In contrast, the bidirectional model learns to follow editing instructions and reference images within substantially fewer training steps. Following LingBot-World 2.0 (Gao et al., 2026), we therefore adopt joint autoregressive and bidirectional training. The two branches share the same model parameters and training data, differing only in their attention patterns. For each video chunk x_i, a flow timestep $\tau \sim \mathcal{U}(0,1)$ and Gaussian noise $\epsilon_i \sim \mathcal{N}(0, I)$ are sampled to construct the noisy latent

$$x_i^\tau = (1-\tau)x_i + \tau\epsilon_i. \tag{5}$$

Under the autoregressive attention pattern, the flow velocity of each chunk is predicted using only its causal video context and the conditions available up to the current step:

$$\mathcal{L}_{\mathrm{AR}} = \mathbb{E}_{x,i,\tau,\epsilon}\left[\left\|v_\theta\left(x_i^\tau, \tau \mid x_{<i}, c_{\le i}\right) - (\epsilon_i - x_i)\right\|_2^2\right]. \tag{6}$$

Under the bidirectional attention pattern, all video chunks are jointly denoised with full temporal attention:

$$\mathcal{L}_{\mathrm{BI}} = \mathbb{E}_{x,\tau,\epsilon}\left[\left\|v_\theta\left(x^\tau, \tau \mid c\right) - (\epsilon - x)\right\|_2^2\right]. \tag{7}$$

The final training objective combines the two branches:

$$\mathcal{L} = \mathcal{L}_{\mathrm{AR}} + \lambda_{\mathrm{BI}}\mathcal{L}_{\mathrm{BI}}, \tag{8}$$

where λ_{BI} controls the relative contribution of the bidirectional objective.

4 Experiments

4.1 Comparison on WBench-Editing

Existing world model benchmarks do not comprehensively evaluate editing capabilities and instead focus primarily on navigation performance and video generation quality. To address this gap, we develop WBench-Editing based on the evaluation framework of WBench (Ying et al., 2026). Specifically, approximately 150 cases are redesigned, each spanning 15–30 seconds (240–480 frames), with a subset additionally incorporating reference images. Each case contains one to three editing instructions, and some editing instructions are causally dependent on preceding ones. To evaluate world editing more systematically, we introduce dedicated metrics while retaining most WBench metrics related to navigation and video fidelity. The overall evaluation is organized into six categories: *Editing, Navigation, Quality, Setting, Consistency, and Physical*. Editing capability is assessed from three perspectives: (1) whether the intended modification is correctly executed, (2) the quality of the world after editing, and (3) whether content unrelated to the edit is properly preserved. These editing-related metrics are evaluated through VLM-based question answering.

Table 1: Comparison results on WBench-Editing. All evaluated methods use their autoregressive model variants, and all test cases are evaluated using text-only multi-turn interactive editing instructions. The top four results in each column are highlighted with progressively darker colors.

Model	Overall	Editing	Navigation	Quality	Setting	Consistency	Physical
MatrixGame3 (Wang et al., 2026b)	54.0	18.1	65.7	74.0	42.8	82.5	49.7
JoyAI-Echo-1.5 (Zhang et al., 2026b)	53.8	21.4	48.0	71.9	51.4	80.0	62.0
minWM(Wan 2.1) (Zhao et al., 2026a)	53.2	36.2	48.1	62.8	48.6	66.7	61.5
Astra (Zhu et al., 2026c)	56.1	27.1	65.5	68.5	44.3	87.4	53.4
Alaya-EVOKE (Yin et al., 2026)	57.7	47.2	63.6	71.2	40.2	59.8	56.6
HY-World 1.5 (Sun et al., 2025)	56.7	21.8	62.0	64.8	57.2	89.0	61.7
HY-GameCraft (Li et al., 2025)	58.7	20.2	73.8	75.1	48.4	88.5	55.7
SANA-WM (Zhu et al., 2026a)	59.2	18.5	77.7	72.0	55.0	84.4	58.0
ABot-World (Jiang et al., 2026)	59.3	28.8	70.6	72.2	44.1	80.0	63.0
Zing-0.5 (Seedleap, 2026)	63.5	33.8	64.5	73.0	67.7	89.6	67.5
AlayaWorld (AlayaWorld et al., 2026)	64.7	31.6	82.8	63.8	63.2	93.6	66.7
LingBot-World (Robbyant et al., 2026)	66.0	54.8	71.3	75.8	56.3	73.4	63.3
DreamX-World (DreamX et al., 2026)	65.9	53.9	70.4	73.3	55.7	82.0	63.0
SolarWM(H3) (Huang et al., 2026a)	65.4	20.2	86.7	75.0	67.1	88.3	67.7
LingBot-World 2.0 (Gao et al., 2026)	66.1	47.2	70.7	74.4	61.3	82.2	66.2
YUME 1.5 (Mao et al., 2026)	67.0	54.8	72.9	76.1	53.5	84.3	62.0
EditWorld (Ours)	73.8	80.0	74.3	74.4	60.9	83.8	67.9

Since most existing video world models do not support reference images as interactive conditioning inputs, all models are evaluated using text-only streaming editing instructions for a fair comparison. Table 1 reports the performance of our model and existing world models on WBench-Editing. Because the benchmark requires multiple edits during autoregressive generation, all evaluated methods use their autoregressive variants. Our model achieves the best overall performance with an *Overall* score of 73.8, outperforming the second-best model, YUME 1.5 (Mao et al., 2026), by 6.8 points. The advantage is particularly pronounced on the core *Editing* metric, where our model reaches 80.0, exceeding the second-best score of 54.8 by 25.2 points. This substantial margin demonstrates a stronger ability to accurately execute streaming editing instructions while preserving the surrounding world state. Beyond editing accuracy, our model also achieves the best *Physical* score of 67.9, indicating that the edited worlds maintain physical plausibility after content modification. Meanwhile, the model maintains solid performance across *Navigation*, *Quality*, and *Setting*, with scores of 74.3, 74.4, and 60.9, respectively, together with a *Consistency* score of 83.8. These results show that the substantial improvement in editing capability is achieved while preserving the broader world-modeling capabilities required for coherent long-horizon generation. The upper part of Figure 3 presents multi-turn editing results from our model and competing methods. Our model responds more accurately to instructions that modify the generated world and better preserves world consistency across successive edits. In contrast, other models tend to interpret editing requests as text-driven events; when new instructions are introduced, they often fail to preserve the existing world state and instead generate substantially different scene content.

4.2 Comparison with Reference-Conditioned Video World Models

Several existing models Jiang et al. (2026); XGEN-JING (2026); Huang et al. (2026a) support reference images as initial conditioning inputs. We therefore select the WBench-Editing cases that include reference images and compare our model against these methods on this subset. As shown in Table 2, our model achieves the best *Overall* score of 74.0 and the highest *Editing* score of 74.6, demonstrating strong reference-conditioned editing capability. More importantly, our model supports flexible reference conditioning throughout autoregressive generation: reference images can be introduced at different interaction stages to modify an already generated world, rather than being restricted to a fixed initial condition. As illustrated in the lower part of Figure 3, XGEN-JING can incorporate the referenced *cart*, but because the reference is provided only at initialization, the timing of the reference-guided modification cannot be controlled precisely. As a result, the *cart* and the yellow balloons appear together instead of following the intended multi-turn sequence in which the cart is introduced first and the balloons are added only in the subsequent edit. SolarWM also supports reference-image conditioning, but exhibits weaker responsiveness to such temporally controlled reference-based edits. Additional qualitative comparisons are provided in the appendix.

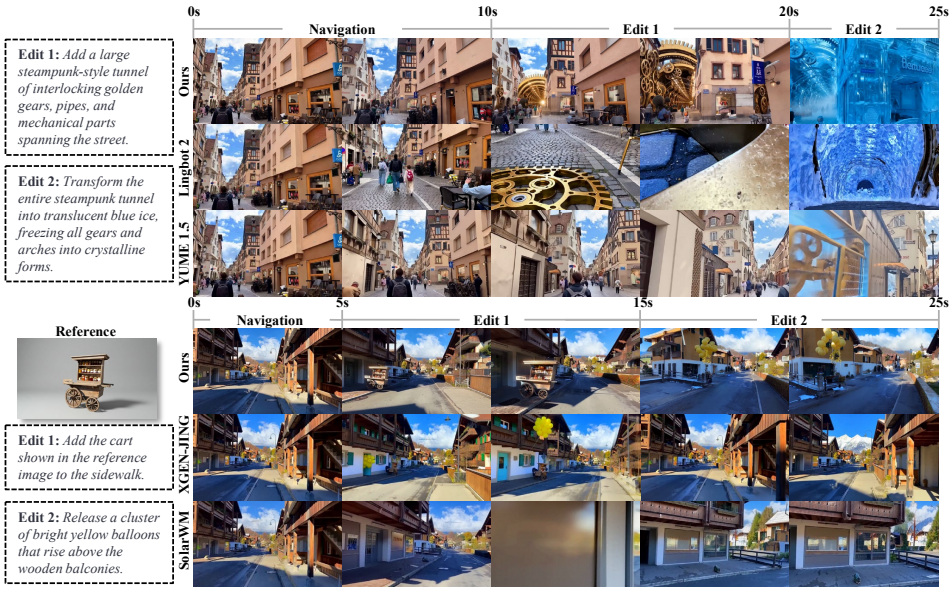


Figure 3: Comparison results with video world models on WBench-Editing. The upper panel presents multi-turn world editing with text-only instructions, while the lower panel shows multi-turn editing with reference-image conditioning. Our model follows the editing instructions more accurately while better preserving the consistency of the overall world. Zoom in for the best view.

Table 2: Comparison results on WBench-Editing (Reference only).

Model	Overall	Editing	Navigation	Quality	Setting	Consistency	Physical
ABot-World (Jiang et al., 2026)	57.3	25.0	66.9	72.0	45.4	80.5	60.9
SolarWM(H3) (Huang et al., 2026a)	64.6	19.8	85.9	74.8	67.6	87.5	64.9
XGEN-JING(BI) (XGEN-JING, 2026)	68.7	63.6	70.4	72.4	70.0	79.8	62.3
EditWorld (Ours)	74.0	74.6	80.5	73.3	65.4	83.1	67.3

4.3 Ablation Study

Individual Training Objectives. We train our model with joint autoregressive and bidirectional objectives to balance causal video continuation with responsiveness to input conditions. To validate this design, we compare it with an ablated variant trained only with the autoregressive objective. As shown in Table 3, joint AR+BI training improves *Editing Overall* from 69.7 to 80.0 and consistently enhances *Editing Presence*, *Semantic Alignment*, and *Editing Completion*. The largest gain is observed in *Detail Accuracy*, which increases from 34.9 to 53.2, suggesting that the bidirectional objective is particularly important for capturing fine-grained editing requirements. Overall, these results indicate that bidirectional supervision substantially strengthens the model's ability to follow input conditions while preserving autoregressive generation capability.

Table 3: Ablation of the training objective on WBench-Editing. Detailed editing-related metrics are reported for models trained with joint objectives and with the autoregressive objective only.

Model	Editing Overall	Editing Presence	Semantic Alignment	Editing Completion	Detail Accuracy	Edit Cleanliness
AR Only	69.7	82.6	79.0	73.5	34.9	78.3
AR + BI	80.0	93.0	89.2	86.2	53.2	78.1

5 Conclusion

This work presents EditWorld, a video world model that moves beyond navigation toward precise world editing and flexible reference-based control. EditWorld enables users to modify generated worlds through streaming instructions, actions, and reference images while maintaining stable long-horizon generation. To support this capability, we develop dedicated model components, training strategies, and a data pipeline tailored to world modification, and introduce WBench-Editing for systematic evaluation of streaming editing. Experiments show that EditWorld substantially improves editing performance while preserving strong general world-modeling capability.

AI USE STATEMENT

We used generative AI tools to assist with creating and editing auxiliary software code, including data processing, evaluation, visualization, debugging, and scripting, and to edit the manuscript for improved clarity, grammar, and readability. We have reviewed all AI-assisted work. All AI-assisted code was reviewed and tested by the authors, and all AI-assisted text was manually reviewed and revised to ensure technical accuracy and consistency with the authors' intended meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Team AlayaWorld, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Mingliang Zhai, et al. Alayaworld: Interactive long-horizon world modeling–full technical report. *arXiv preprint arXiv:2607.18367*, 2026.

Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, et al. Scaling instruction-based video editing with a high-quality synthetic dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37971–37981, 2026.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026.

Jiaxing Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. In *International Conference on Learning Representations*, volume 2026, pp. 85802–85822, 2026.

Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. In *International Conference on Learning Representations*, volume 2026, pp. 103603–103622, 2026.

Team DeepMind. Genie 3: A new frontier for world models, August 2025. URL `https://deepmind.google/models/genie/`.

Kaixin Ding, Xi Chen, Minghong Cai, Zhiyuan Xu, Yiyang Wang, Yuxiang Lu, Junyi Li, Shuyang Chen, Yuan Gao, Xin Tao, et al. Playworld: Benchmarking world models with agent players over long-horizon objectives. *arXiv preprint arXiv:2608.13552*, 2026.

Team DreamX, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, et al. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

Yuwei Guo, Ceyuan Yang, Hao He, Yang Zhao, Meng Wei, Zhenheng Yang, Weilin Huang, and Dahua Lin. End-to-end training for autoregressive video diffusion via self-resampling. *arXiv preprint arXiv:2512.15702*, 2025.

Team Happy Oyster. Happy oyster - real-time world model for interactive creation, April 2026. URL `https://www.happyoyster.com/`.

Haoyang He, Jie Wang, Jiangning Zhang, Zhucun Xue, Xingyuan Bu, Qiangpeng Yang, Shilei Wen, and Lei Xie. Openve-3m: A large-scale high-quality dataset for instruction-guided video editing. *arXiv preprint arXiv:2512.07826*, 2025a.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025b.

Team HiDream. Hidream-o1-world, August 2026. URL https://hidream.ai/.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

Junchao Huang, Guian Fang, Shengju Qian, Xianghao Kong, Zhuoran Zhao, Wei Huang, Yihua Du, Zixin Zhang, Justin Cui, Yuchao Gu, et al. Solarwm: Open data and scalable training for long-horizon video world models. *arXiv preprint arXiv:2609.02886*, 2026a.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026b.

Team InSpatio, Donghui Shen, Guofeng Zhang, Haomin Liu, Haoyu Ji, Hujun Bao, Hongjia Zhai, Jialin Liu, Jing Guo, Nan Wang, et al. Inspatio-world: A real-time 4d world simulator via spatiotemporal autoregressive modeling. *arXiv preprint arXiv:2604.07209*, 2026.

Fan Jiang, Zhaoxu Sun, Mengchao Wang, Ziyu Zhu, Chiyu Wang, Yunpeng Zhang, Wenlin Liu, Yun Wang, Xue Zheng, Rui Sun, et al. Abot-world-0: Infinite interactive world rollout on a single desktop gpu. *arXiv preprint arXiv:2607.19191*, 2026.

Team Kairos, Fei Wang, Shan You, Qiming Zhang, Tao Huang, Zuoyi Fu, Zhisheng Zheng, Yunlong Xi, Feng Lv, Xiaoming Wu, et al. Kairos: A regret-aware native world-action model stack for physical ai, 2026. *URL https://arxiv. org/abs/2606.16533*, 2026.

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2(3):6, 2025.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *Advances in Neural Information Processing Systems*, 38, 2026.

Jinpeng Lu, Dexu Zhu, Haoyuan Shi, Linghan Cai, Guo Tang, Yinda Chen, Jie Cao, Duyu Tang, Yi Zhang, Yong Dai, et al. Current world models lack a persistent state core. *arXiv preprint arXiv:2606.20545*, 2026.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, and Kaipeng Zhang. Yume1. 5: A text-controlled interactive world generation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7752–7761, 2026.

Team Moxin, Deyi Ji, Tianrun Chen, Xin Zhang, Jiale Yang, Qi Zhu, An Zhao, Zihao Xie, Han Wang, Xuanyi Liu, et al. Moworld: A flash world model. *arXiv preprint arXiv:2607.06216*, 2026.

Team Qwen. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Team Robbyant, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Team Seedleap. Zing-0.5: An efficient real-time interactive world model, August 2026. URL https://github.com/seedleap/zing-world-model.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, et al. Lyra 2.0: Explorable generative 3d worlds. *arXiv preprint arXiv:2604.13036*, 2026.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, Linfeng Zhang, et al. Hunyuan-gamecraft-2: Instruction-following interactive game world model. *arXiv preprint arXiv:2511.23429*, 2025.

Team Videox-fun. Videox-fun: A video generation pipeline for diffusion transformer, 2026. URL https://github.com/aigc-apps/VideoX-Fun.

Fu-Yun Wang, Zhaoyang Huang, Alexander W Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased consistency models. *Advances in neural information processing systems*, 37:83951–84009, 2024.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, et al. Spatialvid: A large-scale video dataset with spatial annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42592–42603, 2026a.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026b.

Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

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

Meiqi Wu, Zhixin Cai, Fufangchen Zhao, Xiaokun Feng, Rujing Dang, Bingze Song, Ruitian Tian, Jiashu Zhu, Jiachen Lei, Hao Dou, et al. Omni-worldbench: Towards a comprehensive interaction-centric evaluation for world models. *arXiv preprint arXiv:2603.22212*, 2026a.

Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, et al. Infinite-world: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. *arXiv preprint arXiv:2602.02393*, 2026b.

Team XGEN-JING. Xgen-jing: An egocentric interactive experience model, 2026. URL https://github.com/XGEN-Labs/XGEN-JING/.

Jiacong Xu, Hanwen Jiang, Zhixin Shu, Kalyan Sunkavalli, Vishal M Patel, and Yiqun Mei. Wonder: Video world model done better. *arXiv preprint arXiv:2607.26037*, 2026a.

Xiaojie Xu, Zhengyuan Lin, Kang He, Yukang Feng, Xiaofeng Mao, Yuanyang Yin, Kaipeng Zhang, and Yongtao Ge. Worldmark: A unified benchmark suite for interactive video world models. *arXiv preprint arXiv:2604.21686*, 2026b.

Ying Yang, Zhengyao Lv, Tianlin Pan, Haofan Wang, Binxin Yang, Hubery Yin, Chen Li, Ziwei Liu, and Chenyang Si. Stableworld: Towards stable and consistent long interactive video generation. *arXiv preprint arXiv:2601.15281*, 2026.

Jung Yi, Minjae Kim, Paul Hyunbin Cho, Wooseok Jang, Sangdoo Yun, and Seungryong Kim. Worldkv: Efficient world memory with world retrieval and compression. *arXiv preprint arXiv:2605.22718*, 2026.

Yuanyang Yin, Gongxuan Wang, Yifan Zhan, Chuanhao Li, Kaipeng Zhang, and Feng Zhao. Alaya-evoke: From linear-scaling supervision to endless world. *arXiv preprint arXiv:2608.13546*, 2026.

Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. Wbench: A comprehensive multi-turn benchmark for interactive video world model evaluation. *arXiv preprint arXiv:2605.25874*, 2026.

Jingran Zhang, Ning Li, Yuanhao Ban, Andrew Bai, and Justin Cui. Reward-forcing: Autoregressive video generation with reward feedback. *arXiv preprint arXiv:2601.16933*, 2026a.

Songchun Zhang, Yaowei Li, Junhao Zhuang, Weiyang Jin, Haoyu Wang, Xin Lu, Yilang Sun, Shiyi Zhang, Haoran Li, Xiaoxiao Ma, et al. Echowm: Open and enterable omnimodal world models. *arXiv preprint arXiv:2608.23189*, 2026b.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025.

Min Zhao, Hongzhou Zhu, Bokai Yan, Zihan Zhou, Yimin Chen, Wenqiang Sun, Kaiwen Zheng, Guande He, Xiao Yang, Chongxuan Li, et al. minwm: A full-stack open-source framework for real-time interactive video world models. *arXiv preprint arXiv:2605.30263*, 2026a.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026b.

Xin Zhou and Cong Miao. Astronex-world 1.0: Real-time interactive world model foundation. *arXiv preprint arXiv:2609.20034*, 2026.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, et al. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. *arXiv preprint arXiv:2509.12201*, 2025.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. Sana-wm: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026a.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026b.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jiwen Lu, and Jie Zhou. Astra: General interactive world model with autoregressive denoising. In *International Conference on Learning Representations*, volume 2026, pp. 79167–79184, 2026c.

Junhao Zhuang, Shiyi Zhang, Yuxuan Bian, Yaowei Li, Yawen Luo, Yijun Liu, Weiyang Jin, Songchun Zhang, Xianglong He, Xuying Zhang, et al. Self gradient forcing: Native long video extrapolation. *arXiv preprint arXiv:2607.20368*, 2026.

A Annealed Self-Resampling

At the early stage of training, we adopt standard teacher forcing, where the historical video context consists entirely of clean ground-truth chunks. As training progresses, self-resampling (Guo et al., 2025) is introduced by replacing an increasing proportion of ground-truth history with model-generated chunks. An annealing schedule is used to progressively increase the probability of conditioning on self-generated history. By exposing the model to imperfect contexts produced by its own autoregressive rollouts, the model becomes more robust to error accumulation and better maintains stable, high-quality generation over long horizons. The gradual transition from clean ground-truth context to self-generated context also helps stabilize training throughout the curriculum.

Table 4: Comparison results on WBench. All baseline results are collected from the official WBench Leaderboard, accessed on September 21, 2026. The top four results in each column are highlighted with progressively darker colors.

Model	Average	Quality	Setting	Interaction	Consistency	Physical
Astra (Zhu et al., 2026c)	63.7	67.1	59.6	66.9	73.3	51.4
HY-GameCraft (Li et al., 2025)	68.2	73.0	66.6	66.3	72.6	62.4
MatrixGame2 (He et al., 2025b)	68.7	73.8	67.1	80.3	65.1	57.2
Kairos 3.0 (Kairos et al., 2026)	70.3	74.0	70.3	64.1	82.6	60.4
MatrixGame3 (Wang et al., 2026b)	71.3	75.5	63.6	83.6	74.5	59.3
Infinite-World (Wu et al., 2026b)	72.8	77.0	69.3	75.4	80.0	62.1
YUME 1.5 (Mao et al., 2026)	73.3	77.6	72.4	71.4	80.1	65.2
Astronex-World (Zhou & Miao, 2026)	73.5	78.2	73.5	63.4	83.6	68.6
Fantasy-World (Dai et al., 2026)	73.8	72.4	71.3	71.9	86.4	66.8
InSpatio-World (InSpatio et al., 2026)	73.9	71.5	71.4	73.2	88.4	65.2
Genie 3 (DeepMind, 2025)	73.9	75.2	72.5	73.4	82.6	65.7
ABot-World (Jiang et al., 2026)	74.7	76.8	71.4	84.0	79.5	61.7
DreamX-World (5B AR) (DreamX et al., 2026)	75.0	77.5	80.8	78.6	74.9	63.3
SANA-WM (4-step AR) (Zhu et al., 2026a)	76.0	79.3	76.1	82.2	80.7	61.9
AlayaWorld (AlayaWorld et al., 2026)	76.3	79.3	69.7	80.0	89.5	63.1
Lyra 2.0 (4-step AR) (Shen et al., 2026)	76.4	77.1	73.2	85.6	79.3	66.7
Happy Oyster (Happy Oyster, 2026)	76.8	77.3	74.2	84.9	84.3	63.5
LingBot-World (fast) (Robbyant et al., 2026)	77.4	79.4	77.9	79.2	84.9	65.7
EditWorld (AR, SFT)	77.7	74.8	77.7	80.3	87.8	67.7
HY-World 1.5 (AR distilled) (Sun et al., 2025)	78.1	78.1	72.2	86.8	86.9	66.3
LingBot-World (base-camera) (Robbyant et al., 2026)	78.5	78.9	72.6	80.1	89.9	71.2
LingBot-World v2 (Gao et al., 2026)	79.4	81.8	76.8	82.8	86.5	69.1
EditWorld (AR, 4-step)	79.4	75.2	85.9	78.7	89.2	68.2
Alaya-EVOKE (Yin et al., 2026)	80.8	82.8	83.8	78.6	86.9	72.1
HiDream-O1-World (HiDream, 2026)	80.9	81.0	82.2	80.0	88.0	73.3
Zing-0.5 (Seedleap, 2026)	81.0	80.6	77.8	84.2	88.5	73.8
JoyAI-Echo-1.5 (WM, 4-step) (Zhang et al., 2026b)	81.0	81.1	77.5	87.9	88.3	70.1
XGEN-Jing (4-step AR) (XGEN-JING, 2026)	81.0	81.3	91.6	78.7	83.7	69.8
EditWorld (BI, SFT)	81.2	80.3	84.6	82.9	88.8	69.5
JoyAI-Echo-1.5 (WM, BI) (Zhang et al., 2026b)	81.6	81.5	79.4	86.6	89.8	70.6
XGEN-Jing (BI) (XGEN-JING, 2026)	81.9	82.4	90.3	75.6	88.1	73.3
Alaya-EVOKE-Turbo (Yin et al., 2026)	82.0	81.9	82.1	83.9	88.1	74.0

B Few-Step Distillation

Our real-time autoregressive model is trained using a two-stage distillation strategy, consisting of trajectory distillation followed by distribution distillation. In the first stage, we adopt the Phased Consistency Model (PCM) (Wang et al., 2024) as a warm-up to establish few-step generation capability. PCM partitions the teacher denoising trajectory into multiple phases and enforces consistency within each phase, allowing the student to approximate the full denoising process with only a few updates. We use four function evaluations (NFE) for each video chunk. Starting from the PCM-initialized model, we further perform distribution distillation with Self Gradient Forcing (SGF) (Zhuang et al., 2026) under the same 4-NFE sampling budget. SGF refines the student's generation distribution using autoregressive rollouts conditioned on its own generated history. Specifically, rollout states are first collected without gradient tracking, followed by a parallel reconstruction pass in which the generated historical latents are treated as detached inputs while their key–value representations are recomputed with gradients enabled. This allows losses from subsequent chunks to supervise both their denoising predictions and the encoding of historical context into causal memory, without backpropagating through the entire sequential rollout. Together, the two stages first establish efficient few-step denoising and then improve generation quality and temporal consistency under self-generated contexts, enabling stable long-horizon autoregressive generation.

C Comparison on WBench

To demonstrate that our model maintains strong general world-modeling capability, we report comprehensive results on WBench (Ying et al., 2026). As shown in Table 4, we evaluate our SFT model under both bidirectional and autoregressive inference patterns. EditWorld (BI, SFT) achieves an *Average* score of 81.2, ranking fourth among all evaluated models and improving upon LingBot-World (base-camera) from 78.5 to 81.2. Notably, substantial gains are observed in *Quality*, *Setting*, and *Interaction*, with *Setting* increasing from 72.6 to 84.6. Under causal autoregressive inference, Edit-



Edit: *Change the text on the dark wooden sign hanging above the doorway on the left side of the street to match the sign shown in the reference image.*



Edit: *Add a glowing light trail that rises from the lower left and spirals into the large 3D symbol from the reference image, rotating above the mountain peak.*

(a) Ablation of unidirectional self-attention.



Edit 1: *Add the dog from the reference image near the stone bridge, walking away along the promenade before sitting down to face the water.*



Edit 2: *Cover the stone bridge with dense purple wisteria and hanging flowering vines, with their reflections in the river*

(b) Ablation of short-term textual context.

Figure 4: (a) Unidirectional reference-video attention better preserves fine-grained reference details than bidirectional patterns. (b) Short-term textual context improves temporal continuity across prompt transitions; without preceding prompts, switching editing instructions causes abrupt scene changes and weaker consistency with the preceding video content.

World (AR, SFT) achieves an *Average* score of 77.7, remaining competitive with strong open-source and commercial world models. Furthermore, the distilled 4-step autoregressive model improves the *Average* score to 79.4, with particularly strong performance in *Setting* and *Consistency*, reaching 85.9 and 89.2, respectively. Compared with the autoregressive SFT model, this corresponds to a 1.7-point improvement in *Average* and an 8.2-point gain in *Setting*. These results show that EditWorld retains strong general world-modeling performance across different inference patterns, while supporting precise editing and flexible referencing capabilities.

D Ablation on Unidirectional Self-Attention

To prevent video tokens from influencing the representations of reference-image tokens, we adopt unidirectional self-attention between reference and video tokens. To evaluate this design, we construct an ablated variant in which reference and video tokens are mutually visible through bidirectional self-attention. As shown in Figure 4(a), the proposed unidirectional attention better preserves fine-grained details from the reference image. In contrast, bidirectional attention allows video tokens

to alter the reference representation, resulting in the loss of reference-specific details and weaker preservation of object identity.

E Ablation on Short-term Textual Context

To avoid abrupt visual changes when editing prompts switch across chunks, our cross-attention design allows each video chunk to attend to its current prompt as well as the prompts of the two preceding chunks. To evaluate this short-term textual context, we construct an ablated variant in which each chunk attends only to its current prompt. As shown in Figure 4(b), when the editing prompt changes between frames 160 and 240, the ablated model exhibits a pronounced scene shift and an abrupt temporal transition, with the generated environment becoming inconsistent with the preceding video content. In contrast, incorporating prompts from preceding chunks helps preserve the existing world state and enables smoother transitions across editing instructions.

F Visualization

Figures 5 and 6 present qualitative comparisons between our model and existing video world models on WBench-Editing, including LingBot-World 2.0 (Gao et al., 2026), LingBot-World (Robbyant et al., 2026), YUME 1.5 (Mao et al., 2026), and DreamX-World (DreamX et al., 2026). As shown, our model follows the specified editing instructions more accurately while preserving smooth camera motion and stronger scene consistency across multiple editing rounds. In contrast, competing models often exhibit incomplete edits, unintended changes to unrelated scene content, or substantial deviations from the preceding world state. These qualitative results further demonstrate the advantage of our model in accurately executing sequential world modifications while maintaining consistency over time.

Figures 7 and 8 present qualitative comparisons on WBench-Editing cases with reference-image conditioning. Our model accurately incorporates content from the provided references, including appearance, style, object identity, and fine-grained visual attributes, while preserving the surrounding world context across successive edits. More importantly, we can introduce reference images flexibly at different interaction turns rather than restricting them to a fixed initial condition. This allows users to inject new reference content into an already generated world and combine reference-guided modifications with subsequent editing instructions. In contrast, existing reference-conditioned models often rely on the reference image only at initialization or show weaker control over when and how the referenced content is introduced. These results highlight the advantage of our model in flexible reference conditioning for multi-turn interaction world editing.

Figures 9 and 10 present generation results conditioned on reference images, showing that our model can accurately incorporate referenced objects into the generated world and flexibly modify scene appearance according to the provided visual guidance. Beyond object-level conditioning, the model also supports more abstract reference-based control, such as transferring the visual style of a reference image to the generated world. Figures 11 and 12 further demonstrate multi-turn interactive editing. Across successive editing rounds, our model accurately follows the requested modifications while preserving scene elements unrelated to the current edit.

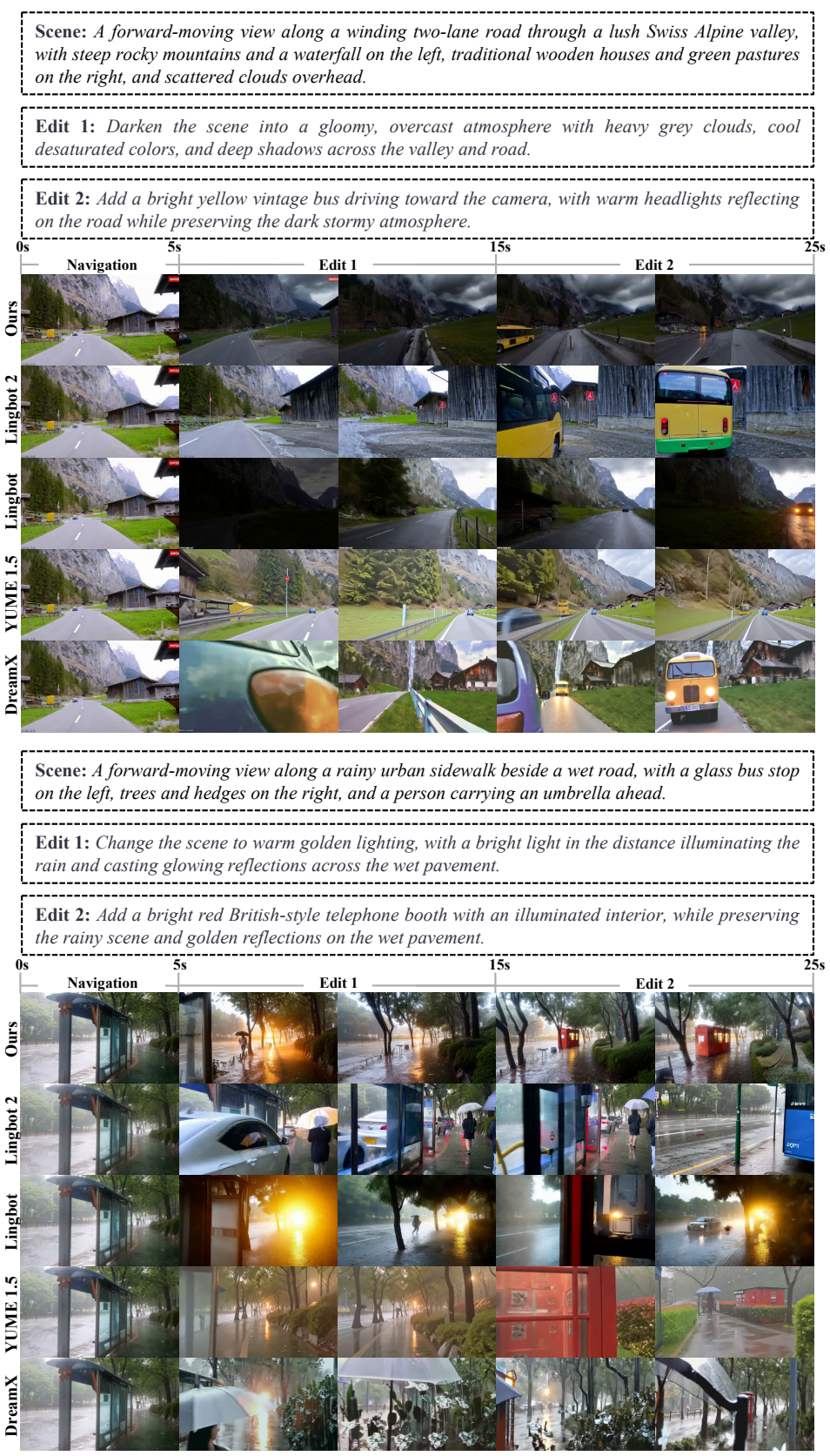


Figure 5: Comparison results with existing video world models on WBench-Edit (text only).



Figure 6: Comparison results with existing video world models on WBench-Edit (text only).

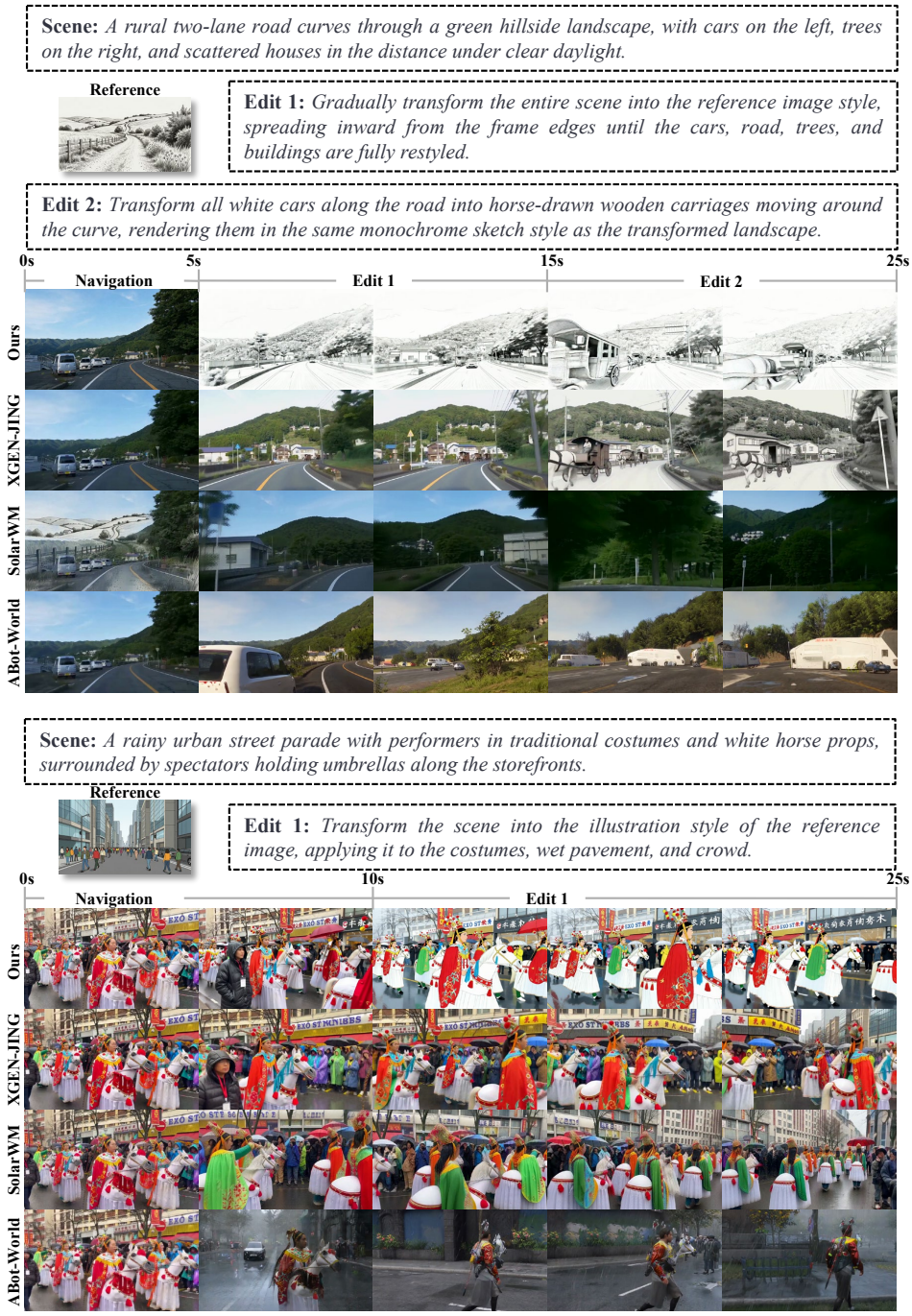


Figure 7: Comparison results with video world models on WBench-Edit (with reference image).

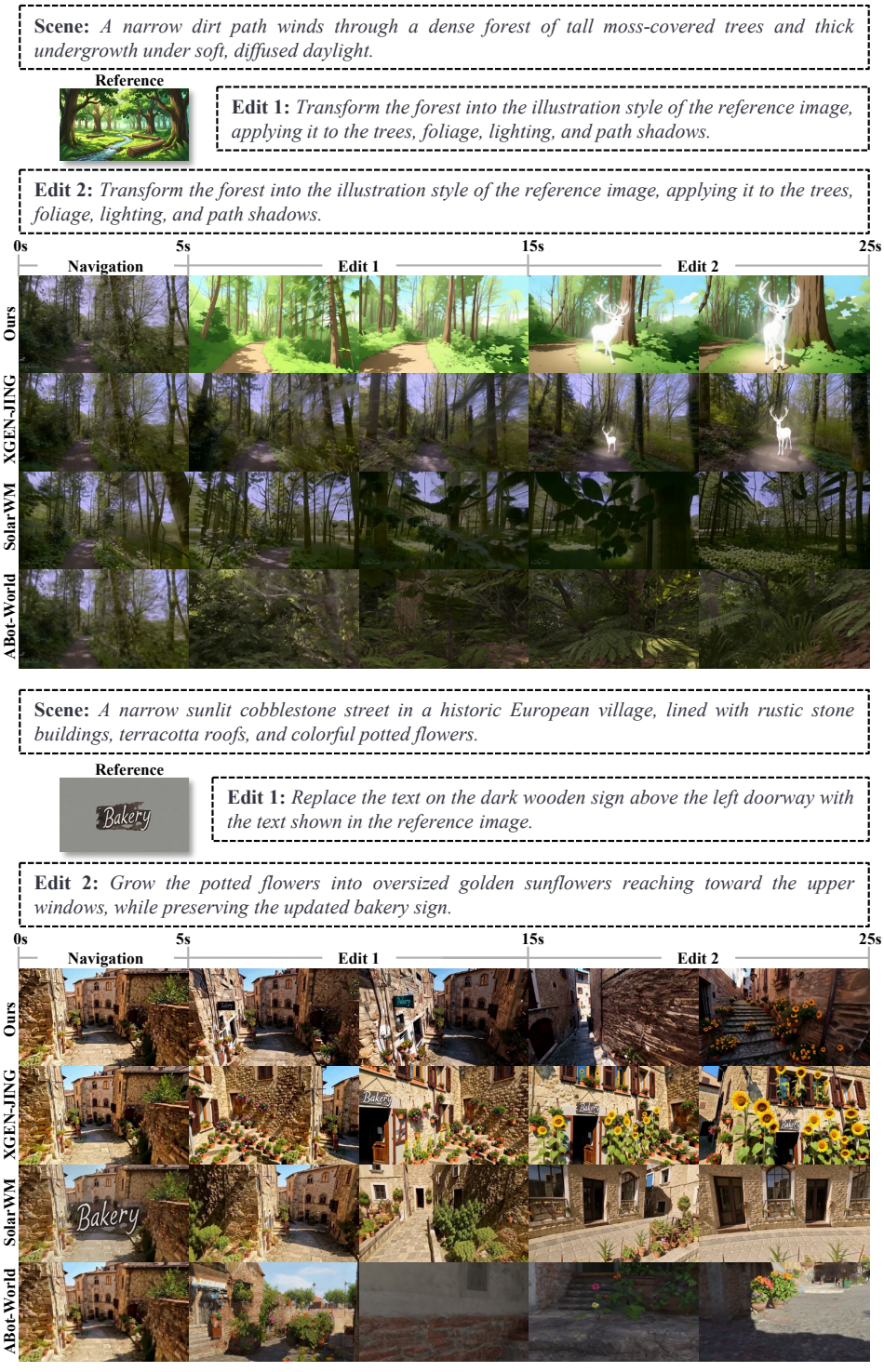


Figure 8: Comparison results with video world models on WBench-Edit (with reference image).

Scene: *An aerial view of a tropical island with massive granite boulders, dense green vegetation, a white sand beach, and clear turquoise water, as the camera slowly moves forward over the bay.*

Edit: *Add a burst of energy near the pier that forms the creature from the reference image, hovering above the water and walking in place while remaining suspended.*



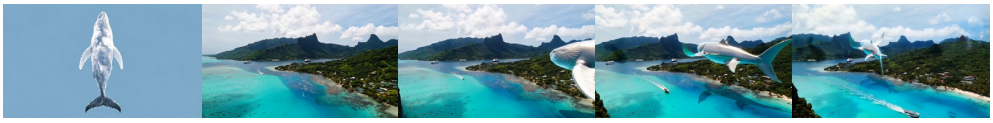
Scene: *A rainy town street curves gently past shops and pedestrians with umbrellas, with the Sunstar Hotel visible through the mist as the camera slowly pans right.*

Edit: *Add the clothesline and drying laundry from the reference image to the right foreground, stretching across the frame and gently swaying in the breeze.*



Scene: *An aerial view of a tropical bay surrounded by lush green mountains, with turquoise water, a small coastal village, and several boats as the camera slowly pans right.*

Edit: *Add the massive creature from the reference image hovering above the bay, then have it swim smoothly through the air with gentle tail and body motion.*



Scene: *A forward-moving view along a gently curving two-lane road in a rural area, with cars on the left, trees on the right, and a green hillside with scattered houses in the distance.*

Edit: *Gradually transform the entire scene into the style of the reference image, spreading inward from the frame edges until the cars, road, trees, and buildings are fully restyled.*



Scene: *A backward-moving view along a quiet, rain-soaked street in a Japanese town at night, with warm lantern-lit storefronts on the left, wet reflective pavement, a parked car, and a lone figure in the distance.*

Edit: *Add the bench from the reference image to the left sidewalk near the lantern-lit storefront as the camera moves backward.*

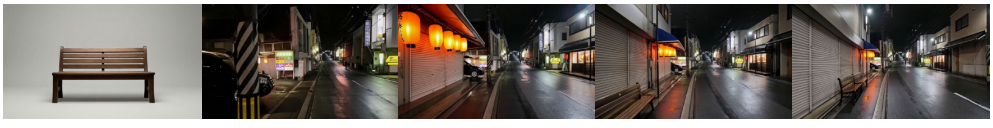


Figure 9: Visualization results of our model. The first image on the left is the reference input.

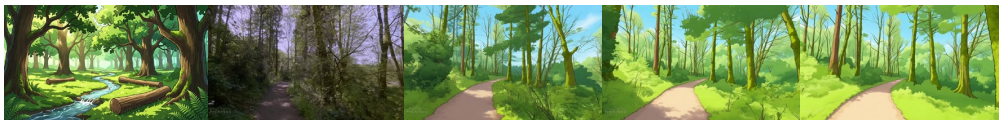
Scene: *A bright, cozy window seat with patterned cushions and curtains in a white-paneled room, overlooking green foliage as the camera slowly zooms in.*

Edit: *Add the potted plant shown in the reference image to the window sill.*



Scene: *A forward-moving view along a narrow dirt path through a dense forest, surrounded by tall moss-covered trees and thick undergrowth under soft, diffused daylight.*

Edit: *Transform the forest into the illustration style of the reference image, applying it to the trees, foliage, lighting, and path shadows.*



Scene: *A forward-moving view along a narrow, rain-soaked urban street at night, with streetlights reflecting on the wet road, buildings on both sides, and red traffic lights visible in the distance.*

Edit: *Transform the entire scene into the illustration style of the reference image, applying it to the buildings, road, foliage, and lighting.*



Scene: *A forward-moving view through a sunny waterfront park, with trees and benches in the foreground, a large ferry docked nearby, and the Statue of Liberty visible across the water.*

Edit: *Transform the scene into the visual style of the reference image, applying it to the trees, water, and boat.*



Scene: *A high-angle aerial view of a turquoise urban river, with historic buildings on the left, a marina filled with small boats on the right, and a stone bridge ahead as the camera moves forward and tilts downward.*

Edit: *Add the dog from the reference image near the stone bridge, walking away along the promenade before sitting down to face the water.*



Figure 10: Visualization results of our model. The first image on the left is the reference input.



Figure 11: Visualization results of multi-turn interactive editing of our model.

Scene: *A forward-moving view along an Amsterdam canal on a sunny day, with traditional canal houses on the right, modern buildings and bicycles on the left, and a houseboat in the foreground.*

Edit 1: *Transform the scene into a warm golden sunset, with an orange-yellow sky, low directional light, long shadows, and deep golden reflections across the canal.*

Edit 2: *Shift the warm golden sunset into bright, cooler daylight, reducing the amber tones and golden reflections.*

Edit 3: *Then transition the bright daylight into a star-filled night, adding green fireflies around the floating garden and shimmering reflections across the dark canal.*



Scene: *A forward-moving view along an urban sidewalk lined with historic stone buildings, storefronts, parked cars, and pedestrians, with a domed building visible in the distance under an overcast sky.*

Edit 1: *Transform the entire scene into the illustration style of the reference image, applying it to the buildings, vehicles, reflections, and pedestrians.*

Edit 2: *Transform the white sedan into a bright red London-style taxi with a black roof and illuminated sign, matching the scene's illustration style.*

Edit 3: *Add heavy snowfall that covers the sidewalk, vehicles, garlands, and stone ledges in snow, while keeping the red taxi prominent.*



Scene: *A forward-moving view along a narrow, wet urban street in winter, with light snow falling, buildings on both sides, and parked scooters visible in the distance under an overcast sky..*

Edit 1: *Change the scene to intense golden lighting from the end of the street, creating a warm glowing haze with strong orange-yellow reflections on the wet pavement and buildings.*

Edit 2: *Cover the street trees and climbing plants with dense pink cherry blossoms, with bright petals drifting through the golden light above the wet street.*

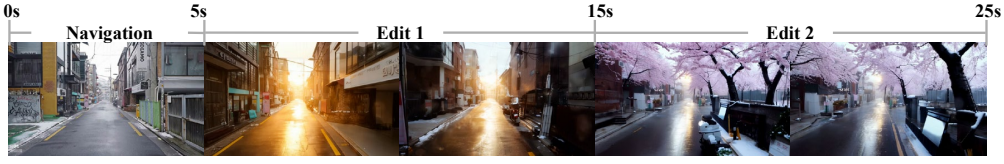


Figure 12: Visualization results of multi-turn interactive editing of our model.