MIX FORCING: CAUSAL PRE-TRAINING VIA MIXTURE ATTENTION FOR INTERACTIVE WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

While interactive video world models require autoregressive generation, directly adapting pretrained bidirectional video diffusion models into autoregressive ones with teacher forcing can disrupt their visual priors and degrade autoregressive generation quality. In this paper, we introduce MIX FORCING, a causal pre-training paradigm for interactive video world models, implemented with the Mixture-of-Bidirectional-and-Autoregressive (MOBA) attention. MIX FORCING jointly trains a video diffusion model with both teacher forcing and bidirectional training, yielding a strong causal initialization for few-step post-training. In self-attention, MIX FORCING leverages teacher forcing to learn causal next-chunk prediction for autoregressive rollout, while using bidirectional attention to preserve pretrained visual priors. To support online text interaction, MIX FORCING further adopts a chunk-wise cross-attention design that conditions each video chunk on a shared environment prompt and its own local prompt, without exposing future instructions to the current chunk. Following causal pre-training on a 14B image-to-video model, we apply consistency and distribution matching distillation to obtain a four-step autoregressive generator. Experiments on our internal benchmark and public benchmarks, including WBench and SANA-WM-Bench, show that our model achieves state-of-the-art visual quality and timely event following with precise camera control. Notably, our model supports high-fidelity long-horizon generation without quality degradation even over one-hour rollouts[1].

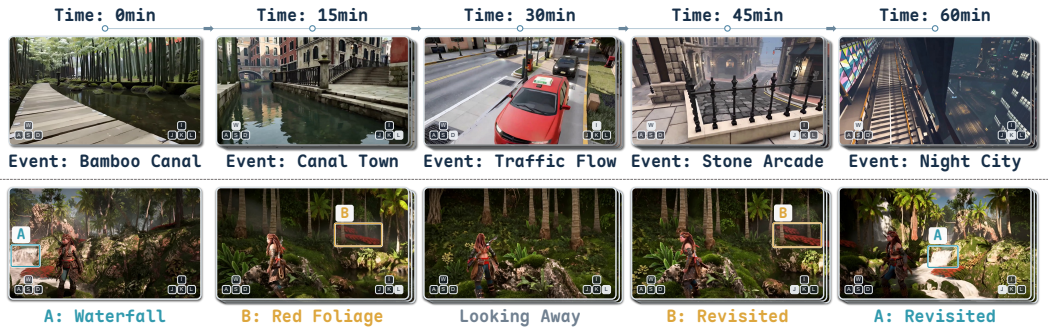


Figure 1: **Long-horizon interactive video rollout with camera control and memory.** Top: a single one-hour autoregressive rollout generated by our model trained with MIX FORCING, following successive prompt updates under keyboard camera control. Bottom: a memory case, where the camera looks away and later returns, and previously observed regions (A: waterfall; B: red foliage) remain consistent when revisited.

1 INTRODUCTION

Interactive video world models [10, 14, 19, 33, 34, 47, 49, 53], which predict future states conditioned on historical context and user inputs, have shown great potential for world simulation, video gaming, and embodied intelligence. Since users act based on past observations, each video chunk must be generated without access to future frames or future instructions, which calls for causal autoregressive generation.

[1]Visual comparison results: https://mixforcing.github.io.

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.

Recent works [14, 34, 53] build upon pretrained video diffusion models to leverage their strong visual priors. However, these priors are learned for bidirectional generation, where all frames are jointly denoised with access to both past and future frames. Some methods [16, 43] directly distill bidirectional models into few-step causal generators via ordinary differential equation (ODE) initialization and distribution matching distillation (DMD), but the trajectories of the bidirectional teacher depend on future information, preventing the causal student from converging to the target flow map. To address this, other methods [50, 54] first adapt pretrained models into many-step causal models with teacher forcing before few-step distillation, making this causal pre-training stage the foundation of the final generator.

However, teacher forcing departs from the pretraining setting of bidirectional video diffusion models in two aspects. In attention, each chunk attends only to its history rather than the full video; in noise level, each noisy chunk is predicted from clean historical context rather than jointly denoised with other chunks at a shared noise level, which encourages velocity prediction to over-rely on clean context. Both deviations disrupt pretrained visual priors and degrade autoregressive generation quality. Moreover, causal pre-training methods typically condition the whole video on a single global prompt, which exposes future events to the current chunk. Streaming methods [40] support prompt updates by recaching the history under each new prompt, which rewrites the context of past frames.

In this paper, we introduce **MIX FORCING**, a causal pre-training paradigm for interactive video world models, implemented with the **Mixture-of-Bidirectional-and-Autoregressive (MOBA)** attention block. The key idea of MIX FORCING is to mix teacher forcing and bidirectional training: teacher forcing learns causal next-chunk prediction for autoregressive rollout, while bidirectional training retains the pretraining setting in both attention and noise level, preserving pretrained visual priors during causal adaptation. In addition, MIX FORCING conditions each teacher forcing chunk on a shared environment prompt and its own local prompt, enabling online text interaction without exposing future instructions to the current chunk or recomputing the KV cache of past chunks.

Based on MIX FORCING, we build an interactive video world model with camera control and online text interaction (Fig. 1). Starting from a 14B image-to-video model, we perform causal pre-training with MIX FORCING, and then apply consistency distillation (CD) [32] and DMD to obtain a four-step generator for KV-cached autoregressive rollout (Fig. 2). Experiments on our internal benchmark and public benchmarks show that our model achieves state-of-the-art visual quality and timely event following with precise camera control, and sustains high-fidelity generation over one-hour rollouts. Ablation studies further show that MIX FORCING provides a stronger initialization for few-step post-training than teacher forcing alone.

Our contributions are summarized as follows:

- We propose the MOBA attention block, which mixes teacher forcing and bidirectional training to preserve pretrained visual priors during causal adaptation, with chunk-wise prompt conditioning for online text interaction.
- We build an interactive video world model through causal pre-training with MIX FORCING and few-step post-training with CD and DMD, supporting camera control, online text interaction, and four-step KV-cached autoregressive rollout.
- Experiments on internal and public benchmarks show that our model achieves state-of-the-art visual quality and timely event following with precise camera control.

2 RELATED WORK

Interactive and Controllable Video World Models. Interactive world models predict visual states conditioned on user actions [1, 3, 36], while controllable video generation also supports text instructions and camera trajectories. Prior work incorporates camera geometry through ray representations [12, 13, 51], trajectory embeddings [2, 37], or relative positional encodings [20], enabling geometry-aware viewpoint control in video generation. Vid2World [15] adapts pretrained video diffusion for autoregressive action control, while LongLive [40] supports streaming prompt updates by recaching the history under each new prompt. MotionStream [30] supports interactive motion control during streaming video generation. Other video world models support interactive or long-horizon generation [4, 10, 14, 33, 34, 45, 49, 53]. In this work, we aim to use causal pre-training for interactive video world models, maintaining the visual priors of pretrained video diffusion models while enabling online text interaction through chunk-wise conditioning.

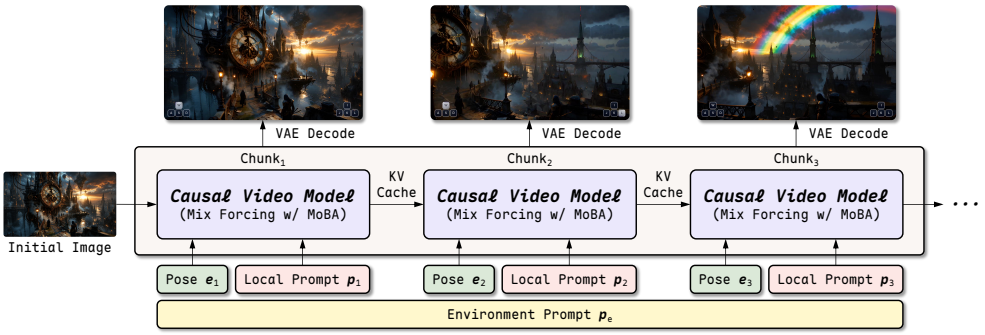


Figure 2: **Pipeline of our interactive video world model.** Given an initial image and an environment prompt $\mathbf{p}_e$, an interactive world is simulated by our four-step causal video model. Based on the KV-cached history context and user inputs, including the camera pose $\mathbf{e}_i$ and the local prompt $\mathbf{p}_i$, each latent chunk is generated autoregressively and decoded into a video clip by the VAE decoder.

Causal Autoregressive Video Generation. Autoregressive video models generate frames or chunks sequentially [8, 35], with some using KV-cached streaming inference [19, 29]. FramePack [46] compresses historical frames for long rollouts, while diffusion-forcing approaches assign varying noise levels across video frames [5, 6, 38]. To adapt pretrained bidirectional video diffusion models for online video rollout, prior methods initialize few-step causal students from probability-flow ODE trajectories [18, 31, 43] and subsequently improve them through distribution matching distillation [42]. Self Forcing [16] addresses rollout errors using self-generated history; Self-Forcing++ [7] and Self Gradient Forcing [55] extend this approach through teacher supervision on long self-generated rollouts and context-gradient reconstruction, respectively. Resampling Forcing [11] and Rolling Forcing [23] use self-resampling and joint denoising, respectively. Causal Forcing [54] and Causal Forcing++ [50] improve few-step distillation using a causal teacher and causal consistency distillation, respectively. Reward Forcing [25] uses reward-weighted distribution matching distillation for streaming video generation. In MIX FORCING, we jointly optimize teacher forcing and bidirectional attention during causal pre-training to retain pretrained visual priors before few-step post-training.

3 METHODOLOGY

In this section, we present MIX FORCING, a causal pre-training paradigm for interactive video world models. We first review flow matching and teacher forcing for interactive video modeling (Sec. 3.1). We then introduce the MOBA attention block that implements MIX FORCING, which mixes teacher forcing and bidirectional training to preserve pretrained visual priors, with chunk-wise prompt conditioning for online text interaction (Sec. 3.2). Finally, as shown in Fig. 2, we apply MIX FORCING to build an interactive video world model with camera control and online text interaction (Sec. 3.3).

3.1 PRELIMINARY

Conditional Video Generation via Flow Matching. Flow matching [9, 22, 24] learns a velocity field that transports a Gaussian distribution to the data distribution. For video generation, it operates in the latent space of a pretrained variational autoencoder (VAE), where a video $\mathbf{v}$ is encoded into a latent $\mathbf{z} = \mathcal{E}(\mathbf{v})$ by the encoder $\mathcal{E}$. Given a clean latent $\mathbf{z}$, Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, a timestep $t \in [0, 1]$, and conditioning signals $\mathbf{y}$ (e.g., text prompts, camera poses, or an initial image), the noisy latent is constructed as $\mathbf{z}^t = (1-t)\boldsymbol{\epsilon} + t\mathbf{z}$. The model is trained to predict the velocity by minimizing:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\mathbf{z},\boldsymbol{\epsilon},\mathbf{y},t}\left[\left\|v_\theta(\mathbf{z}^t, t, \mathbf{y}) - (\mathbf{z} - \boldsymbol{\epsilon})\right\|_2^2\right], \tag{1}$$

where v_θ denotes the velocity predicted by the model with parameters θ.

Interactive Video Modeling via Teacher Forcing. While bidirectional video diffusion models attend to both past and future frames, interactive video models require autoregressive generation, where each video chunk is conditioned on historical context and current user inputs. Therefore, teacher forcing is used to train this next-chunk prediction process, where each noisy chunk attends only to itself and its clean history. Let $\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$ denote the sequence of video latent chunks, where $\mathbf{z}_i \in \mathbb{R}^{H\times W\times C}$ is the i-th chunk, and let $\mathbf{y}$ denote the user inputs (e.g., camera poses

and text prompts). Each chunk is autoregressively generated from its history $\mathbf{z}_{<i}$ and the user inputs, which can be formulated as:

$$p_\theta(\mathbf{z}_{1:N} \mid \mathbf{y}) = \prod_{i=1}^{N} p_\theta(\mathbf{z}_i \mid \mathbf{z}_{<i}, \mathbf{y}) \tag{2}$$

However, teacher forcing departs from the pretraining setting of bidirectional video diffusion models. In attention, each noisy chunk attends only to itself and its history instead of the full video; in noise level, it is predicted from clean historical latents instead of being jointly denoised with other chunks at a shared timestep. Both deviations can disrupt pretrained visual priors and degrade generation quality.

3.2 MIXTURE-OF-BIDIRECTIONAL-AND-AUTOREGRESSIVE ATTENTION

As shown in Fig. 3, the MOBA attention block is designed for causal pre-training. Its core is to mix teacher forcing and bidirectional training in self-attention, preserving pretrained visual priors during causal adaptation. In addition, MOBA adopts chunk-wise prompt conditioning in cross-attention to enable online text interaction. We instantiate the user inputs $\mathbf{y}$ in Sec. 3.1 as camera poses and text prompts. Specifically, $\mathbf{e}_i$ denotes the camera pose of the i-th chunk, whose representation is detailed in Sec. 3.3. The text prompts consist of a shared environment prompt $\mathbf{p}_e$ describing the scene, a local prompt $\mathbf{p}_i$ describing the event in the i-th chunk, and a global prompt $\mathbf{p}_g$ describing the entire video.

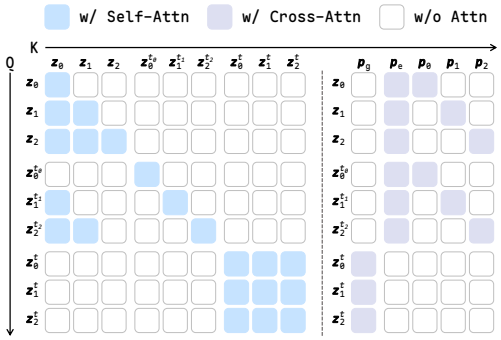


Figure 3: **MOBA attention block.** For self-attention (left), each noisy TF chunk attends only to itself and its clean history, while the appended BID noisy latents attend only to one another bidirectionally. For cross-attention (right), TF chunks attend to the shared environment prompt $\mathbf{p}_e$ and their corresponding local prompt $\mathbf{p}_i$, while BID chunks attend to the global prompt $\mathbf{p}_g$.

MOBA Self-Attention for Preserving Visual Priors. To address the deviations discussed in Sec. 3.1, we mix teacher forcing and bidirectional training in self-attention. Teacher forcing learns causal next-chunk prediction for autoregressive rollout. Bidirectional training jointly denoises all latents at a shared timestep with full attention, keeping both attention and noise level consistent with pretraining and thus preserving pretrained visual priors. Formally, for each training sample, we draw a training mode $b \in \{\mathrm{TF}, \mathrm{BID}\}$ with equal probability, which defines the following mixture of factorizations (text conditioning is omitted here and described in the next paragraph):

$$p_\theta^b(\mathcal{Z} \mid \mathbf{e}_{1:N}) = \begin{cases} \prod_{i=1}^{N} p_\theta(\mathbf{z}_i \mid \mathbf{z}_{<i}, \mathbf{e}_{\le i}), & b = \mathrm{TF} \\ p_\theta(\mathbf{z}_{1:N} \mid \mathbf{e}_{1:N}), & b = \mathrm{BID} \end{cases} \tag{3}$$

In practice, instead of sampling b, we append the noisy latents of BID to the TF training sequence and optimize both modes in a single forward pass. As shown in the left panel of Fig. 3, a sparse self-attention mask lets each TF noisy chunk attend only to itself and its clean history, while the appended BID latents attend to one another bidirectionally. Since the two modes are isolated by the mask, this concatenation optimizes the expected objective over b with lower gradient variance.

MOBA Cross-Attention for Online Text Interaction. Online interaction requires the model to update visual content as user instructions arrive. Existing causal pre-training methods [50, 54] typically condition all chunks on a global prompt describing the entire video, which reveals future events to the current chunk in advance. Streaming methods such as LongLive [40] support prompt updates by recaching the recent history under each new prompt, so that past frames are re-encoded with a prompt different from the one they were generated with. In contrast, we condition each teacher forcing chunk on the environment prompt $\mathbf{p}_e$, which describes the persistent scene, and its own local prompt $\mathbf{p}_i$, which describes the current event, while bidirectional latents remain conditioned on the

global prompt $\mathbf{p}_g$, consistent with pretraining. Formally, the cross-attention output for the i-th noisy latent chunk is:

$$\mathbf{o}_i^b = \begin{cases} \mathrm{CrossAttn}(\mathbf{z}_i^{t_i}, \mathbf{p}_e \oplus \mathbf{p}_i), & b = \mathrm{TF} \\ \mathrm{CrossAttn}(\mathbf{z}_i^{t}, \mathbf{p}_g), & b = \mathrm{BID} \end{cases} \tag{4}$$

where $\oplus$ denotes concatenation along the token dimension. In practice, we concatenate the global, shared environment, and all local prompt embeddings into a single sequence. As shown in the right panel of Fig. 3, a sparse cross-attention mask lets both clean and noisy latents of chunk i under TF attend only to $\mathbf{p}_e$ and $\mathbf{p}_i$, while BID latents attend only to $\mathbf{p}_g$. At inference, since each past chunk is encoded with its own local prompt, a prompt update only affects subsequent chunks, and the KV cache of past chunks is reused without recomputation.

MoBA Training Objective. We optimize both modes with the flow matching loss in Eq. (1). Under TF, each chunk is noised independently as $\mathbf{z}_i^{t_i} = (1 - t_i)\boldsymbol{\epsilon}_i + t_i \mathbf{z}_i$ and predicted from its clean history:

$$\mathcal{L}_{\mathrm{TF}} = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \left\| v_\theta\big(\mathbf{z}_i^{t_i}, t_i \mid \mathbf{z}_{<i}, \mathbf{e}_{\le i}, \mathbf{p}_e, \mathbf{p}_i\big) - (\mathbf{z}_i - \boldsymbol{\epsilon}_i)\right\|_2^2\right]. \tag{5}$$

In practice, following [19], we perturb the clean history latents with low-level noise with 50% probability to mitigate error accumulation during autoregressive rollout. Under BID, all chunks share a timestep t and are jointly denoised:

$$\mathcal{L}_{\mathrm{BID}} = \mathbb{E}\left[\frac{1}{N}\left\| v_\theta\big(\mathbf{z}_{1:N}^{t}, t \mid \mathbf{e}_{1:N}, \mathbf{p}_g\big) - (\mathbf{z}_{1:N} - \boldsymbol{\epsilon}_{1:N})\right\|_2^2\right]. \tag{6}$$

The overall objective is

$$\mathcal{L}_{\mathrm{MoBA}} = \lambda_{\mathrm{TF}}\mathcal{L}_{\mathrm{TF}} + \lambda_{\mathrm{BID}}\mathcal{L}_{\mathrm{BID}}, \tag{7}$$

which, with equal weights, is proportional to the expectation over the mode sampling in Eq. (3).

3.3 Mix Forcing for Interactive Video World Models

Causal Pre-Training. We initialize our interactive video world model from a pretrained bidirectional video model and apply Mix Forcing for causal pre-training. For camera control, we represent the camera pose $\mathbf{e}_i$ of chunk i with Plücker embeddings [12, 13], which encode the viewing ray of each pixel as six-dimensional coordinates. Following [10, 34], we inject them through adaptive layer normalization (AdaLN) [39] without disrupting the pretrained visual priors. During causal pre-training, we jointly train our model with camera-labeled and unlabeled data. For unlabeled data, we use an all-zero dummy Plücker embedding as the condition input. As a result, a bidirectional video diffusion model is converted into an interactive video model supporting camera control and multi-step autoregressive rollout with KV cache.

Few-Step Post-Training. To enable real-time interaction, we further compress our causal pre-trained model into a few-step generator. Let $\mathbf{z}_i^{t_i}$ denote the i-th latent chunk at timestep t_i, and $\mathcal{H}_i = (\mathbf{z}_{<i},\, \mathbf{e}_{\le i},\, \mathbf{p}_e,\, \mathbf{p}_{\le i})$ denote its causal context. We first use consistency distillation (CD) [32] to initialize a four-step generator G_θ by minimizing:

$$\mathcal{L}_{\mathrm{CD}} = \mathbb{E}\Big[d\Big(G_\theta(\mathbf{z}_i^{t_i}, t_i \mid \mathcal{H}_i),\; G_{\theta^-}(\tilde{\mathbf{z}}_i^{t_i+\Delta t}, t_i + \Delta t \mid \mathcal{H}_i)\Big)\Big] \tag{8}$$

where $d(\cdot,\cdot)$ is a distance metric, $\tilde{\mathbf{z}}_i^{t_i+\Delta t}$ is obtained by integrating the probability-flow ODE using the causal-pretrained teacher model from time t_i to $t_i + \Delta t$, and the target parameters θ^- are updated as an exponential moving average (EMA) of the student parameters θ.

We then refine this four-step generator through distribution matching distillation (DMD) [41] to enhance visual fidelity during autoregressive rollout. Let $\hat{\mathbf{z}}$ denote a video generated under user inputs $\mathcal{U} = (\mathbf{e}_{1:N}, \mathbf{p}_e, \mathbf{p}_{1:N})$, and $\hat{\mathbf{z}}^t$ denote its noised version at timestep t. We further optimize G_θ with the following objective:

$$\mathcal{L}_{\mathrm{DMD}} = \mathbb{E}_{\hat{\mathbf{z}}^t, \hat{\mathbf{z}}, \mathcal{U}, t}\left[\frac{1}{2}\left\|\hat{\mathbf{z}} - \mathrm{sg}\Big[\hat{\mathbf{z}} - \Big(s_{\mathrm{fake}}^{\phi}(\hat{\mathbf{z}}^t, t \mid \mathcal{U}) - s_{\mathrm{real}}(\hat{\mathbf{z}}^t, t \mid \mathcal{U})\Big)\Big]\right\|_2^2\right] \tag{9}$$

where s_{real} and s_{fake}^{ϕ} represent the real and fake score models, respectively, and $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator. During DMD training, the parameters ϕ of the fake score model are optimized with the standard diffusion loss on generated videos, while the real score model is kept fixed.

4 EXPERIMENTS

In this section, we first detail the experimental setup, including benchmarks, evaluation metrics, and implementation details (Sec. 4.1). We then compare our model with state-of-the-art video world models quantitatively (Sec. 4.2) and qualitatively (Sec. 4.3). Finally, we conduct an ablation study to validate the design of MOBA and the effect of MIX FORCING on few-step post-training (Sec. 4.4).

4.1 EXPERIMENTAL SETUP

Benchmarks. We construct an internal benchmark (InBench) to evaluate event-following ability and camera-control precision. We use GPT-Image 2.5 [27] to generate 96 initial images and organize InBench into two subsets: (i) an interactive subset covering environmental changes and character interactions (32 cases); and (ii) a roaming subset covering human characters and virtual objects (64 cases). Each case is evaluated with a video of 8 to 35 seconds, and each interactive case contains 3 to 5 prompt updates. We also evaluate on WBench [44] and SANA-WM-Bench [53].

Evaluation Metrics. Following WBench and SANA-WM-Bench, we report aesthetic quality (AQ), imaging quality (IQ), temporal flickering (TFlk), and dynamic degree (DD) from VBench [17], HPSv3 [26], and rotation error (Rot, in radians), translation error (Trans), and camera-motion consistency (CMC) between the input and estimated camera trajectories. For long rollouts, IQ_{a-b} denotes imaging quality from a to b seconds, and sustained imaging quality (SIQ) averages the first and last windows. For online text interaction, we use a vision-language model as a judge [52] to measure Event Adherence (EA), the completion of each event, and On-Time Accuracy (OT), whether a correct response appears within a given time after each prompt update (see Appendix).

Implementation Details. We train on 100k self-collected clips (974 hours), 70% of which have camera poses estimated by COLMAP [28], while the rest use the dummy Plücker embedding (Sec. 3.3). For causal pre-training, we initialize from Wan2.1-I2V-14B-480P, train on variable-length 480×832 videos at 16 FPS with $\lambda_{\text{TF}} = \lambda_{\text{BID}} = 1$, and use AdamW with a learning rate decaying from 5e-5 to 1e-5; the chunk sizes for teacher forcing and chunk-wise cross-attention are 1 and 4 latent frames, respectively. We randomly perturb the clean context latents using low-level noise with 50% probability [19]. For post-training, CD uses LPIPS [48] as the distance metric with a learning rate of 2e-6, and DMD follows LingBot-World-Fast with LingBot-World-Base [34] as score models. All three stages use 256 H20 GPUs, with 15k, 8k, and 1.5k steps for pre-training, CD, and DMD.

4.2 QUANTITATIVE COMPARISON

Evaluation on InBench. Tab. 1 shows that our model achieves the best average quality, event-following, and camera-control results among all methods. The highest average quality score of 75.1 indicates that MIX FORCING preserves pretrained visual priors during causal pre-training, and the event-following scores (EA=94.4, $\text{OT}_{1.5\text{s}}$=96.8) validate that chunk-wise cross-attention enables effective responses to online prompts. The lowest camera errors further demonstrate precise camera control during autoregressive rollout.

Evaluation on Public Benchmarks. We report SANA-WM-Bench results in Tab. 2. On both the simple and hard 60-second splits, our model achieves the highest average *Quality* (74.8 and 74.7) and SIQ (66.7 and 66.4) scores, whereas several baselines show clear quality drops in later windows, e.g., the IQ of LingBot-World-Fast decreases from 63.2 to 52.0 on the simple split. These results demonstrate that our model maintains visual fidelity throughout minute-long autoregressive generation; one-hour qualitative results and average quality scores measured every five minutes are further provided in Figure. S12 in the Appendix. Although competitive, our camera errors are not the lowest on SANA-WM-Bench. We partly attribute this gap to a prompt mismatch: the benchmark describes the entire video with a single global prompt, whereas our model is trained with an environment prompt and chunk-wise local prompts. Although local prompts do not explicitly specify camera motion, they indicate what should appear in the current chunk and thus help the pose condition control viewpoints more precisely, as supported by the reduced camera errors with chunk-wise prompts in Tab. 4. We further evaluate on the Navi-158 subset of WBench (Table. S1 in the Appendix), where our model achieves the highest average *Quality* (78.5) and *Setting* adherence (80.3) scores, indicating strong visual fidelity and faithful reflection of the specified world settings, while its *Interaction* scores show a gap similar to that of the camera errors on SANA-WM-Bench.

Table 1: **Quantitative comparison on InBench.** The quality, event-following, and camera-error groups are evaluated on the full InBench, the interactive subset, and the roaming subset, respectively. Metrics follow Sec. 4.1, and Avg denotes the average of the quality metrics. EA denotes event adherence, and $OT_{1.0s}$ and $OT_{1.5s}$ denote on-time accuracy within 1.0 and 1.5 seconds. The best results are in bold, and the top three are highlighted from dark to light blue.

Method	Quality ↑						Event Following ↑			Camera Error ↓		
	AQ	IQ	TFlk	DD	HPSv3	Avg	EA	$OT_{1.0s}$	$OT_{1.5s}$	Rot	Trans	CMC
HY-WorldPlay	59.4	62.4	96.8	55.6	58.4	66.5	0.0	0.0	0.0	1.795	0.553	1.968
JoyAI-Echo-WM	62.7	62.5	95.4	74.6	62.2	71.5	45.2	5.6	26.2	0.928	0.463	1.126
Matrix-Game 3.0	60.6	65.2	95.3	87.3	60.7	73.8	0.0	0.0	0.0	0.441	0.329	0.694
SANA-WM-Streaming	61.0	56.9	96.6	79.4	59.1	70.6	41.3	4.8	23.0	2.022	0.466	2.067
LingBot-World-Fast	63.8	61.7	96.3	73.0	67.4	72.4	86.5	38.1	87.3	1.065	0.433	1.312
MIX FORCING (Ours)	63.2	63.8	96.0	81.0	71.3	**75.1**	**94.4**	**88.9**	**96.8**	**0.296**	**0.316**	**0.539**

Table 2: **Quantitative comparison on SANA-WM-Bench.** 60-second generation under (a) simple and (b) hard camera trajectories. $IQ_{0–10}$ and $IQ_{50–60}$ denote imaging quality over the first and last 10 seconds. Other metrics follow Sec. 4.1. The best results are in bold, and the top three are highlighted from dark to light blue.

(a) Simple trajectory splits.

Method	Quality ↑	Temporal IQ ↑			Camera Error ↓		
		$IQ_{0–10}$	$IQ_{50–60}$	SIQ	Rot	Trans	CMC
WorldPlay	61.7	68.7	61.6	65.0	1.087	1.552	2.240
Echo-WM	69.4	67.6	64.4	66.0	0.902	**1.450**	**2.010**
Matrix-Game 3	70.4	65.3	65.5	65.4	**0.888**	1.575	2.118
SANA-Streaming	69.7	65.2	59.4	62.2	1.374	1.768	2.558
LingBot-Fast	70.1	63.2	52.0	57.1	1.964	1.616	2.836
MIX FORCING (Ours)	**74.8**	68.6	64.9	**66.7**	0.965	1.459	2.038

(b) Hard trajectory splits.

Method	Quality ↑	Temporal IQ ↑			Camera Error ↓		
		$IQ_{0–10}$	$IQ_{50–60}$	SIQ	Rot	Trans	CMC
WorldPlay	65.5	68.7	63.5	66.0	1.500	1.550	2.525
Echo-WM	72.9	67.4	64.4	65.9	1.529	1.589	2.597
Matrix-Game 3	70.2	65.2	66.7	65.9	**0.914**	1.520	**2.101**
SANA-Streaming	71.1	64.6	59.9	62.2	1.637	1.647	2.687
LingBot-Fast	69.9	63.4	54.3	58.5	1.871	1.683	2.887
MIX FORCING (Ours)	**74.7**	68.5	64.5	**66.4**	1.214	**1.456**	2.246

User Study. To complement these automatic metrics, we further conduct a pairwise user study. Participants compare our model with one baseline at a time in a method-blinded, side-by-side setting, rating visual quality on all cases, interactive following on the interactive subset, and motion clarity on the roaming subset. We collect 318 pairwise trials from 18 sessions, with details provided in the Appendix. As shown in Fig. 4, participants consistently prefer our model over each baseline in terms of visual quality, interactive following, and motion clarity.

4.3 QUALITATIVE COMPARISON

Camera-Controlled Roaming. Fig. 5(a) illustrates the qualitative comparison on the InBench roaming subset. During camera movement, especially when the viewpoint changes, the video model needs to generate the visual content from a novel perspective while keeping its structure and appearance stable across frames. We observe that existing state-of-the-art methods suffer from progressive blurring and structural distortion, recovering sharpness only after the camera stops moving. In contrast, the enlarged regions highlighted by yellow boxes show that our model better maintains scene consistency, motion clarity, and fine-grained appearance throughout the viewpoint transition.

Online Prompt Interaction. Fig. 5(b) illustrates the qualitative comparison on the InBench interactive subset. Although JoyAI-Echo-WM and LingBot-World-Fast support prompt updates, we observe that they respond with noticeable delays, consistent with their lower $OT_{1.0s}$ in Tab. 1, or disrupt existing scene content. In contrast, with the chunk-wise cross-attention in MIX FORCING, the enlarged regions highlighted by red boxes show that our model responds to each local prompt and generates the correct content while maintaining scene consistency.

4.4 ABLATION STUDY

For the ablation study, we sample a subset of our internal training data with each video clip containing 161 frames and follow the training settings described in Sec. 4.1. The evaluation is carried out on InBench, where we measure generation quality on the full benchmark, event-following ability on the interactive subset, and camera error on the roaming subset.

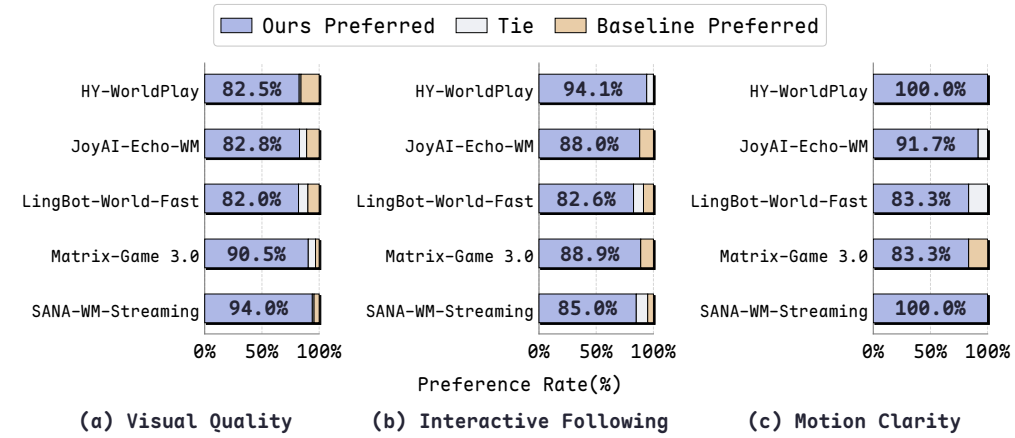
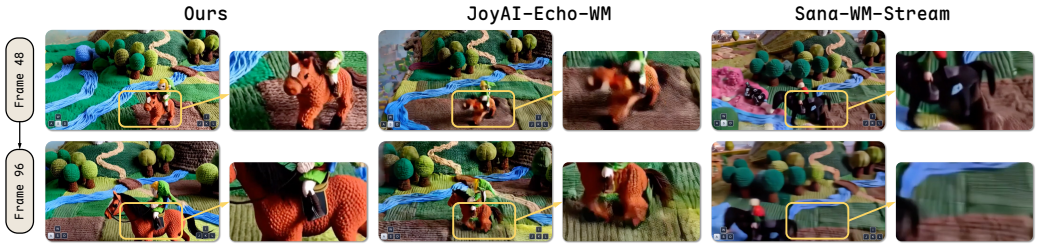


Figure 4: **Pairwise user study on InBench.** Each bar compares our model with one baseline and shows the percentages of judgments preferring ours, rating a tie, and preferring the baseline.



(a) Camera-controlled roaming.



(b) Online prompt interaction.

Figure 5: **Qualitative comparisons with state-of-the-art methods.** (a) Camera-controlled roaming: our model preserves structure and fine details during viewpoint changes, whereas other methods suffer from blurring or distortion. (b) Online prompt interaction: our model follows each event prompt with a consistent scene, whereas other methods respond with delays or break scene consistency.

Table 3: **Ablation of MoBA self-attention design.** Combinations of teacher forcing (TF), diffusion forcing (DF), and bidirectional (BID) training with different training chunk sizes. IQ_{0-2} and IQ_{8-10} denote imaging quality over the first and last 2 seconds. The best results are in bold, and the top three are highlighted from dark to light blue.

Attention Pattern	Chunk	Quality ↑	Temporal IQ ↑			Camera Error ↓		
			IQ_{0-2}	IQ_{8-10}	SIQ	Rot	Trans	CMC
TF	1	68.9	51.5	48.3	49.9	1.098	0.445	1.425
TF+DF	1	65.6	52.8	50.9	51.9	0.655	0.249	0.871
TF+DF+BID	1	67.7	52.1	48.8	50.5	0.751	0.258	0.986
TF+BID	1	**70.8**	55.9	61.0	**58.4**	**0.449**	**0.231**	**0.647**
TF+BID	4	69.1	55.2	58.8	57.0	0.884	0.323	1.165
TF+BID	1+4	69.1	54.0	55.9	54.9	0.882	0.331	1.162

Self-Attention Design for Causal Pre-Training. Using teacher forcing (TF) as the baseline, we evaluate the effects of adding diffusion forcing (DF) [5] and bidirectional (BID) training. As shown in

Tab. 3, TF+BID (row 4) improves generation quality, achieves the highest SIQ of 58.4, and maintains high final-window quality (from IQ_{0-2}=55.9 to IQ_{8-10}=61.0), while IQ drops from 51.5 to 48.3 under TF, indicating that bidirectional attention can mitigate quality degradation. However, adding DF (rows 2 and 3) can reduce generation quality. We attribute this degradation to a mismatch between training and inference, where DF uses noisy historical chunks during training, while KV-cached inference leverages fully denoised historical context. Note that DF replaces the clean history with independently noised history, which matches neither the clean KV cache at inference nor the shared noise level of pretraining. In contrast, MoBA keeps the clean history in TF for inference consistency and recovers the pretraining setting in a separate BID branch. We further study the effect of training chunk size under TF+BID. As chunk size=1 (row 4) enables fine-grained causal modeling by assigning an independent noise level to each latent frame and outperforms the other settings (rows 5 and 6) in both generation quality and camera control accuracy, we combine TF and BID training with chunk size=1 in MoBA.

Table 4: **Ablation of MoBA cross-attention design.** Global versus chunk-wise prompt conditioning, both built on the best self-attention setting in Tab. 3. Metrics follow Sec. 4.1. The best results are in bold and highlighted in blue.

Prompt	Quality ↑	Event Following ↑		Camera Error ↓		
		EA	$OT_{1.5s}$	Rot	Trans	CMC
Global	70.8	23.4	18.3	0.449	0.231	0.647
Chunk-Wise	**71.1**	**50.0**	**62.1**	**0.319**	**0.139**	**0.431**

Cross-Attention Design for Causal Pre-Training. As shown in Tab. 4, we study the effect of global and chunk-wise prompts in cross-attention. Chunk-wise prompts improve event-following ability, increasing EA from 23.4 to 50.0 and $OT_{1.5s}$ from 18.3 to 62.1. Chunk-wise prompts also reduce all three camera errors, consistent with our analysis in Sec. 4.2.

Table 5: **Ablation of Mix Forcing for few-step post-training.** Four-step models obtained from TF-only and Mix Forcing pre-training under the same CD and DMD pipeline, with the 50-step Mix Forcing model as a reference. Metrics follow Sec. 4.1. The best results are in bold and highlighted in blue.

Pre-training	Post-training	Steps	Quality ↑	Event Following ↑		Camera Error ↓		
				EA	$OT_{1.5s}$	Rot	Trans	CMC
Mix Forcing	None	50	**71.1**	50.0	62.1	**0.319**	**0.139**	**0.431**
TF only	CD → DMD	4	64.1	33.7	19.8	1.077	0.446	1.331
Mix Forcing	CD → DMD	4	70.9	**55.1**	**64.3**	0.435	0.234	0.593

Mix Forcing for Few-Step Post-Training. As shown in Tab. 5, we study the effect of pre-training with Mix Forcing on few-step post-training. Under the same CD and DMD pipeline, the four-step model pretrained with Mix Forcing (row 3) outperforms the TF-only baseline (row 2), indicating that Mix Forcing provides a stronger initialization. Compared with the 50-step model (row 1), the four-step model retains comparable generation quality and achieves higher event-following scores.

5 Conclusion

We present Mix Forcing, a causal pre-training paradigm for interactive video world models. Implemented with the MoBA attention block, it mixes teacher forcing and bidirectional training to preserve pretrained visual priors, and uses chunk-wise prompts for online text interaction. Combined with CD and DMD, our four-step model achieves state-of-the-art visual quality and timely event following with camera control over one-hour rollouts.

Limitations. Our implementation targets video DiT architectures with separate self-attention and cross-attention modules, and its extension to unified attention [21] remains to be verified. Moreover, our model is designed for chunk-wise local prompts, and its camera control is less precise when only a single global prompt is available, as observed on public benchmarks. We leave these directions for future work.

AI USE STATEMENT

We use generative AI tools to assist with paper writing and finding related works. We do not use generative AI tools to conceive the proposed research idea. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our user study involves human participants who compared anonymized generated videos through a browser-based interface, and no personally identifiable information was collected. As a realistic interactive video generator, our model could be misused to create deceptive content, such as fabricated videos of real people or events. We encourage responsible use, including disclosure and watermarking of generated content, and all generated videos in this paper are used solely for research purposes.

REFERENCES

[1] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

[2] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844. IEEE, 2025.

[3] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

[4] Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *International Conference on Learning Representations*, volume 2025, pp. 37546–37593, 2025.

[5] 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.

[6] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

[7] 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.

[8] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *International Conference on Learning Representations*, volume 2025, pp. 44730–44745, 2025.

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[10] 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.

[11] 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. In *European Conference on Computer Vision*, pp. 324–344. Springer, 2026.

[12] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for text-to-video generation, 2024. URL https://arxiv.org/abs/2404.02101.

[13] Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13416–13426. IEEE, 2025.

[14] Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Cyrus Wu, Wei Li, Xuchen Song, Yang Liu, Eric Li, and Yahui Zhou. Matrix-Game 2.0: An open-source, real-time, and streaming interactive world model, 2025. URL https://arxiv.org/abs/2508.13009.

[15] Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv:2505.14357*, 2025.

[16] 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, 2026.

[17] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

[18] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577, 2022.

[19] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

[20] Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *Advances in Neural Information Processing Systems*, 38:15984–16009, 2026.

[21] Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

[22] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[23] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, volume 2026, pp. 91177–91196, 2026.

[24] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[25] Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34385–34397, 2026.

[26] Yuhang Ma, Yunhao Shui, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[27] OpenAI. ChatGPT Images 2.5 System Card, 2026. URL https://deploymentsafety.openai.com/chatgpt-images-2-5.

[28] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[29] Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

[30] Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. In *International Conference on Learning Representations*, volume 2026, pp. 7381–7406, 2026.

[31] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[32] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 2023. URL https://proceedings.mlr.press/v202/song23a.html.

[33] 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, 2025. URL https://arxiv.org/abs/2512.14614.

[34] Robbyant Team, 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.

[35] Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

[36] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, volume 2025, pp. 73754–73776, 2025.

[37] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024. doi: 10.1145/3641519.3657519. URL https://arxiv.org/abs/2312.03641.

[38] Desai Xie, Zhan Xu, Yicong Hong, Hao Tan, Difan Liu, Feng Liu, Arie Kaufman, and Yang Zhou. Progressive autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 6312–6322. IEEE, 2025.

[39] Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32, 2019.

[40] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

[41] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yin_One-step_Diffusion_with_Distribution_Matching_Distillation_CVPR_2024_paper.html.

[42] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024.

[43] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

[44] 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, 2026. URL https://arxiv.org/abs/2605.25874.

[45] Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactorly: Creating new games with generative interactive videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11590–11599. IEEE, 2025.

[46] Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *Advances in Neural Information Processing Systems*, 38:30546–30566, 2026.

[47] Qihang Zhang, Lin Li, Luyao Zhang, Shuai Yang, Yiming Luo, Shuaiting Li, Ruilin Wang, Junke Wang, Jiahao Shao, Gangwei Xu, et al. Native video-action pretraining for generalizable robot control. *arXiv preprint arXiv:2607.08639*, 2026.

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

[49] 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*, 2026.

[50] 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*, 2026.

[51] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. CamI2V: Camera-controlled image-to-video diffusion model, 2024. URL https://arxiv.org/abs/2410.15957.

[52] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

[53] 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, 2026. URL https://arxiv.org/abs/2605.15178.

[54] 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*, 2026.

[55] 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.