WORLD RERENDERER: REAL-TIME CAMERA-CONTROLLED WORLD RE-RENDERING VIA VIDEO DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

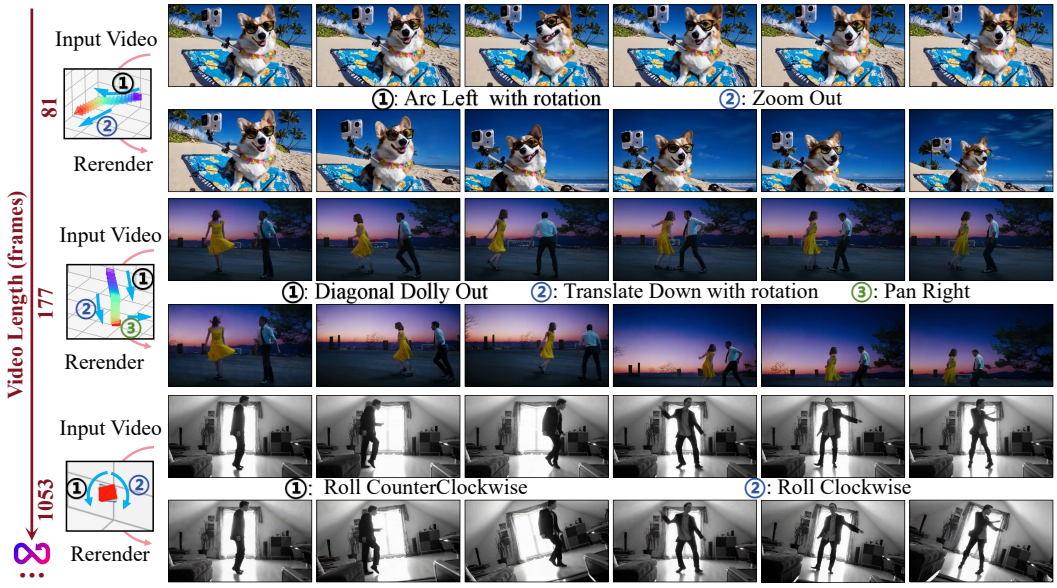


Figure 1: **World Rerenderer for real-time world re-rendering.** Given a streaming input video and online user-specified camera trajectories (numbered arrows), World Rerenderer enables real-time streaming re-rendering of the dynamic scene from the requested viewpoints, preserving synchronization and content consistency. Although trained only on short clips, its length-agnostic streaming design imposes no fixed upper bound on rollout length—in principle supporting indefinitely long horizons; empirically, we show qualitative rollouts at 81, 177, and 1053 frames here.

ABSTRACT

Real-time camera-controlled world re-rendering aims to synthesize target views from a streaming source video under online user-specified camera trajectories. This task poses two key challenges: maintaining synchronization and content consistency across views, and achieving low-latency rendering with long-horizon stability. Existing approaches are predominantly designed for offline full-sequence processing and rely on bidirectional attention or costly depth-based geometric preprocessing, making them unsuitable for low-latency streaming scenarios. To address these limitations, we develop a teacher-to-student framework that distills a high-fidelity bidirectional re-rendering teacher into a few-step causal autoregressive student. We introduce *Cross-Frame Injection* to interleave each incoming source frame with its synchronized target latent, thereby establishing direct cross-view correspondence and improving content consistency. To ensure stable long-horizon generation, we further propose *Shared Relative RoPE*, which bounds temporal indices within the pretrained range, and *Cycle Streaming Tuning*, which exposes the student to extended cyclic trajectories during streaming training and provides a closed-loop consistency signal. Together, these designs enable low-latency online video re-rendering with precise camera control and sustained spatiotemporal consistency over long streaming rollouts.

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.

1 INTRODUCTION

Interactive video world models (Mao et al., 2026; Parker-Holder et al., 2024; Li et al., 2025a) have attracted increasing attention as a promising paradigm for understanding and interacting with the physical world. By modeling how dynamic environments evolve in response to external actions, such models provide a unified foundation for a wide range of applications, including embodied intelligence (Kim et al., 2026), autonomous driving (Hu et al., 2023), and interactive content creation (Li et al., 2025a; Zhang et al., 2025b). However, many existing interactive world models (Mao et al., 2026) are initialized from text prompts or static images, limiting their ability to ground predictions in an ongoing real-world scene. In contrast, interactive *video-conditioned* world models (Chen et al., 2025) continuously integrate incoming video observations into an evolving world state and respond to user actions in real time, enabling users or agents to interact with a dynamically evolving environment. This capability is particularly valuable for live broadcasting and filmmaking (Luo et al., 2025; Bai et al., 2025). For example, in a live sports broadcast, a user steers a virtual camera around the players as the footage streams in, and expects to instantly view the action from the requested viewpoint with minimal latency. Realizing this vision demands real-time responsiveness, precise camera control, and faithful preservation of the incoming stream *simultaneously*.

This motivates the task of real-time camera-controlled *world re-rendering*: given a streaming source video that captures the observed world and online user-specified camera trajectories, the model autoregressively synthesizes the corresponding target views, rendering each new frame as the source stream advances. This setting poses two fundamental challenges: **1) Synchronization and content consistency across views**: the model must faithfully preserve the content and dynamics of the source stream while plausibly hallucinating regions that become newly visible from the target viewpoints, all while precisely following the user's camera controls. **2) Low-latency rendering and long-horizon stability**: The target stream must be rendered with low latency and maintain spatiotemporal consistency over long rollouts.

Despite the challenges outlined above, existing approaches rarely satisfy these requirements simultaneously. Camera-controlled video re-rendering methods (Bai et al., 2025; Yu et al., 2025a; Park et al., 2026) have achieved promising results in novel-view synthesis. However, they are primarily designed for offline full-sequence processing: they either relying on explicit geometric priors such as depth-based warping (Yu et al., 2025a) or implicit camera embeddings processed with bidirectional attention (Bai et al., 2025; Park et al., 2026). Both paradigms require the complete source context to be available before starting to generate, making them poorly suited to low-latency interaction with incrementally arriving video streams. More recently, Inspatio-World (Team et al., 2026a) introduces a few-step causal autoregressive generator, together with depth-based reference warping for autoregressive block-wise generation. Nevertheless, it still assumes a pre-observed reference video and relies on offline models (*e.g.*, depth estimation models (Lin et al., 2025)) to construct geometric priors. This pre-initialization requirement constitutes a fundamental limitation for real-time video-conditioned world modeling, where both scene observations and user controls must be handled online. Furthermore, the underlying depth-based warping mechanism introduces an additional limitation: as the target viewpoint deviates substantially from the reference view, the reprojected visible regions become increasingly sparse and provide weaker source guidance, leading to degraded fidelity to the source content, particularly during large viewpoint transitions.

To address these limitations, we propose *World Rerenderer*, a implicit framework for real-time camera-controlled world re-rendering. Our framework follows a teacher-to-student pipeline. We first train a high-fidelity bidirectional camera-controlled video-to-video (re-rendering) teacher, and then distill it into a few-step causal autoregressive student for low-latency streaming inference (Yin et al., 2024; 2025; Huang et al., 2025b). Unlike Inspatio-World (Team et al., 2026a), which relies on time-consuming offline pre-processing to align the source content, our framework operates directly on the incoming source stream. Specifically, we introduce a *Cross-Frame Injection* mechanism that interleaves each incoming source frame with its corresponding target latent as a synchronized contextual pair. This design establishes direct cross-view correspondence at every time step, improving temporal synchronization and content consistency with the source stream.

However, we find that direct distillation (Yin et al., 2024) is insufficient to ensure stable long-horizon streaming inference. We attribute this to two causes. **First**, the temporal RoPE indices grow with the global video time. Once the rollout exceeds the horizon covered by pretraining, the model must

extrapolate to unseen positions, producing a positional train–test mismatch (Yesiltepe et al., 2025). **Second**, the student is typically trained on short video clips, whereas streaming deployment requires it to autoregressively generate substantially longer sequences. The resulting discrepancy in sequence length can cause progressive degradation over long rollouts. To address these two issues, we introduce two complementary designs. 1) **Shared Relative RoPE (SR-RoPE)**. The key of SR-RoPE is to assign a shared temporal index to each synchronized source-target frame pair. Rather than increasing with global video time, the temporal indices are defined relative to the current cache window and follow the same bounded schedule during training and inference. This keeps all active temporal indices within the pre-trained range and thereby avoids the extrapolation-induced mismatch. 2) **Cycle Streaming Tuning**. Starting from each short synchronized training clip, we construct a long cyclic trajectory by extending the original video and camera motion with their temporal reversals, causing the camera to return to previously observed viewpoints. We expose the model to these long cyclic sequences through efficient streaming training, thereby narrowing the gap between short-clip training and long-horizon autoregressive inference. Moreover, the cyclic construction enriches the camera trajectories and provides a closed-loop consistency signal by requiring the model to reproduce previous content when the camera revisits an earlier viewpoint.

World Rerenderer achieves streaming rendering at 10 FPS on H20 GPU and substantially reduces startup latency by eliminating time-consuming offline preprocessing, yielding a 13–29$\times$ reduction relative to Inspatio-World. Through extensive experiments and ablation studies on diverse real and synthetic videos, we demonstrate competitive visual quality and camera accuracy, together with improved source consistency and long-horizon stability enabled by our core designs.

2 Related Work

Camera-Controlled Video Generation. With the goal of steering viewpoint changes through camera conditions, camera-controlled video generation has become an important branch of controllable video synthesis (Ma et al., 2025). Early studies primarily explored this capability in text-to-video (He et al., 2024; Kuang et al., 2024; Bahmani et al., 2025; Wang et al., 2025b; 2026a) or image-to-video (Yu et al., 2025b; Li et al., 2025c; Feng et al., 2024; Xu et al., 2024; Xiao et al., 2025; Cheng et al., 2025; Lyu et al., 2026) settings. More recently, growing attention has shifted toward camera-controlled video-to-video generation (re-rendering) (Van Hoorick et al., 2024; Bai et al., 2025; Luo et al., 2025; Li et al., 2026). This challenging task necessitates maintaining rigorous temporal synchronization and content consistency with the source video while precisely following the provided camera trajectory. Current methods generally follow two technical routes. Warping-based methods (Xiao et al., 2025; Yu et al., 2025a; Lu et al., 2025; Zhang et al., 2025a; Jeong et al., 2025; Lin et al., 2026) typically employ depth estimation to lift 2D frames into 3D point clouds, from which geometrically aligned proxy videos are rendered from target viewpoints. These proxies then serve as auxiliary conditions to guide the generative process. Conversely, implicit synthesis methods (Bai et al., 2025; Park et al., 2026; Van Hoorick et al., 2024) directly encode camera parameters into the model, often leveraging large-scale synthetic datasets for training. Despite their impressive visual quality, both paradigms are predominantly designed for offline, full-sequence modeling, rendering them computationally prohibitive for real-time interactive control.

Autoregressive Video Diffusion Models. Traditional video diffusion models (Yang et al., 2024; Wang et al., 2025a) typically rely on full-sequence, bidirectional modeling, which limits their scalability for long-horizon generation and real-time streaming. To overcome these constraints, a growing line of research (Chen et al., 2024; Jin et al., 2025; Zhu et al., 2026; Yuan et al., 2026) integrates autoregressive (AR) prediction with diffusion modeling. Early attempts like MAGI-1 (Teng et al., 2025) and CausVid (Yin et al., 2025) transition from bidirectional to causal architectures, enabling chunk-wise rollout. To mitigate the inherent challenges of temporal drift and error accumulation in AR sampling, recent works have introduced advanced training and inference strategies. For instance, Self-Forcing (Huang et al., 2025b) addresses the train-test mismatch by conditioning the model on its own generated outputs. LongLive (Yang et al., 2025; Chen et al., 2026) leverages attention-sink mechanisms and KV-cache management to maintain global consistency during extended rollouts. However, these methods are primarily optimized for prompt-based generation. While effective for visual quality, they often lack the flexibility required for fine-grained, user-driven control.

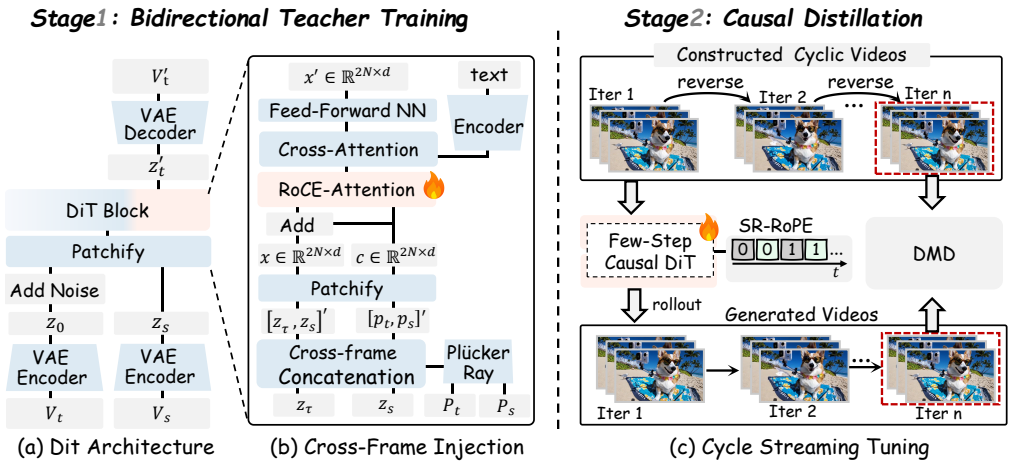


Figure 2: **Overview of the two-stage training pipeline of World Rerenderer.** *Stage 1:* We train a bidirectional camera-controlled V2V teacher, where Cross-Frame Injection interleaves synchronized source–target latents and camera ray embeddings. *Stage 2:* We distill the teacher into a few-step causal DiT via distribution matching distillation (DMD), enhanced with Cycle Streaming Tuning and Shared Relative RoPE (SR-RoPE) for stable long-horizon streaming generation.

Interactive World Models. Interactive video world models (Team et al., 2025; Wang et al., 2026b; Gao et al., 2026) simulate dynamic environments that evolve in response to user actions. Recent work explores diverse control modalities, including camera poses (Team et al., 2026b; Sun et al., 2025) and motion trajectories (Zhao et al., 2025; Shin et al., 2025), audio (Su et al., 2026), and natural-language instructions (Yesiltepe et al., 2025; Li et al., 2025b). However, many existing models are initialized from text prompts or static images and generate subsequent dynamics without continuously incorporating new observations from an ongoing scene. In this work, we bridge this gap by focusing on generating novel views of input videos through interactive camera control, enabling an immersive "walk-through" experience. Closest to our task, Inspatio-World (Team et al., 2026a) supports real-time re-rendering but relies on offline geometric preprocessing of a pre-observed reference video. World Rerenderer instead operates directly on incoming video observations and online camera controls, enabling low-latency interaction without offline scene reconstruction.

3 Approach

As illustrated in Figure 2, World Rerenderer follows a two-stage training pipeline (Yin et al., 2025; 2024). We first train a bidirectional camera-controlled V2V teacher (Sec. 3.2), and then distill it into a few-step causal student for low-latency streaming inference (Sec. 3.3). We further introduce *Shared Relative RoPE* and *Cycle Streaming Tuning* to improve long-horizon autoregressive stability.

3.1 Preliminary

3D Rotary Position Embedding (3D-RoPE). Following RoPE (Su et al., 2024), the query (q) and key vectors (k) in self-attention layers are grouped into complex-valued pairs, denoted as $\bar{q}, \bar{k} \in \mathbb{C}^{N\times(d_{\text{head}}/2)}$, and rotated by

$$\bar{q}' = \bar{q} \odot \mathbf{R}, \qquad \bar{k}' = \bar{k} \odot \mathbf{R}, \tag{1}$$

where $\mathbf{R} \in \mathbb{C}^{N\times(d_{\text{head}}/2)}$ is the rotary matrix, $\odot$ denotes element-wise multiplication. The resulting attention logit between n-th query token and m-th key token is

$$A(n,m)' = \text{Re}\left[\bar{q}'_n \bar{k}'^{*}_m\right] = \text{Re}\left[\bar{q}_n \left(\bar{k}^{*}_m \odot e^{\text{i}\theta_c(n-m)}\right)\right], \tag{2}$$

where $\text{Re}[\cdot]$ extracts the real part and $(\cdot)^*$ denotes complex conjugation, and θ_c is the frequency related to the c-th rotary channel of matrix $\mathbf{R}$. Current video diffusion transformers (Yang et al., 2024; Wang et al., 2025a) extend this construction to the frame, height, and width axes by partitioning the rotary channels into three groups. Accordingly, the 3D rotary matrix is constructed as $\mathbf{R} = \mathbf{R}_f \oplus \mathbf{R}_h \oplus \mathbf{R}_w$, where $\mathbf{R}_l \in \mathbb{C}^{N\times(d_{\text{head}}/6)}$ denotes the axis-specific rotation matrix for $l \in \{f, h, w\}$, and $\oplus$ denotes channel-wise concatenation.

3.2 CAMERA-CONTROLLED V2V TEACHER MODEL.

In this section, we fine-tune the pretrained Wan-I2V (Wang et al., 2025a; He et al., 2024) bidirectional diffusion model into a high-fidelity camera-controlled V2V teacher. Given a fixed-length source video $V_s \in \mathbb{R}^{F\times C\times H\times W}$ and a user-specified target camera trajectory P_t, the teacher generates a re-rendered video $V_t \in \mathbb{R}^{F\times C\times H\times W}$ that follows P_t while preserving the content and dynamics of V_s. Our teacher contains two key components: Cross-Frame Injection for synchronized source conditioning and RoCE (Park et al., 2026) for camera conditioning.

Cross-Frame Injection. Let $z_s = \mathcal{E}(V_s)$ and $z_0 = \mathcal{E}(V_t)$ denote the source and target latents encoded by the frozen video VAE, respectively. Following the rectified-flow formulation (Labs, 2024), we sample $\epsilon \sim \mathcal{N}(0, I)$ and $\tau \in [0, 1]$ to obtain the noisy target latent $z_\tau = (1-\tau)z_0 + \tau\epsilon$. Existing V2V models (Bai et al., 2025; Luo et al., 2025) commonly concatenate the complete source and target latents as two temporal prefixes. Under this organization, the distance between corresponding source and target frames depends on the video length, making their cross-view correspondence difficult to model. We instead interleave the source and target latents at the frame level:

$$\widetilde{z}_\tau = [z_\tau, z_s]' = [z_\tau^1, z_s^1, z_\tau^2, z_s^2, \ldots, z_\tau^f, z_s^f], \tag{3}$$

where f is the temporal length in the latent space. Each noisy target frame is therefore placed adjacent to its synchronized clean source frame, forming a local cross-view context pair. We further assign the same temporal and spatial 3D-RoPE coordinates to the source and target tokens within each pair. Consequently, their relative positional relationship remains unchanged with the input length, allowing self-attention to directly associate each target frame with the corresponding source observation. This organization also provides a natural source–target unit for the subsequent causal streaming distillation. We validate this in our experiments in Table 2.

Rotary Camera Encoding (RoCE). To provide more precise camera control, we adopt RoCE (Park et al., 2026), which injects camera information as relative phase shifts in the self-attention layers. Specifically, let $\mathbf{c} \in \mathbb{R}^{2N\times d}$ denote the Plücker-ray tokens derived from the target and source camera trajectories, P_t and P_s. RoCE maps these tokens to camera-dependent phase shifts $\phi_{\text{qk}} = [\mathbf{0}, \text{MLP}_{\text{qk}}(\mathbf{c})]$ and constructs the camera rotation matrix $\widetilde{\mathbf{R}}_{\text{qk}} = e^{\text{i}\phi_{\text{qk}}}$. The zero-padding leaves the temporal rotary channels unchanged, such that only the spatial components encode camera information. The standard RoPE transformation in Eq. 1 is then modified as:

$$\bar{q}' = \bar{q} \odot \widetilde{\mathbf{R}}_{\text{qk}} \odot \mathbf{R}, \qquad \bar{k}' = \bar{k} \odot \widetilde{\mathbf{R}}_{\text{qk}} \odot \mathbf{R}. \tag{4}$$

Consequently, the attention interaction between tokens n and m becomes $A'(n,m) = \text{Re}\big[\bar{q}_n \big(\bar{k}_m^* \odot e^{\text{i}(\theta_c(n-m)+\phi_{\text{qk}}(n,\mathbf{c})-\phi_{\text{qk}}(m,\mathbf{c}))}\big)\big]$, thereby jointly encoding relative positions and camera poses. We refer the reader to (Park et al., 2026) for the full derivation.

3.3 AUTOREGRESSIVE DIT FOR STREAMING RERENDERING

Although the bidirectional teacher achieves high-fidelity re-rendering, its full-sequence attention and multi-step denoising preclude low-latency streaming. We therefore distill it into a few-step causal student that renders the target stream chunk by chunk. Following the Cross-Frame Injection design in Eq. 3, we define the i-th synchronized chunk as $Z^i = [z_\tau^i, z_s^i]$, which pairs the noisy target latent to be rendered with the clean source latent arriving online. For each chunk, the student denoises only the target component while conditioning on the source component and a bounded KV cache of previously generated chunks. We optimize the student using Self-Forcing-style distribution matching distillation (Huang et al., 2025b; Yin et al., 2024). During training, the student

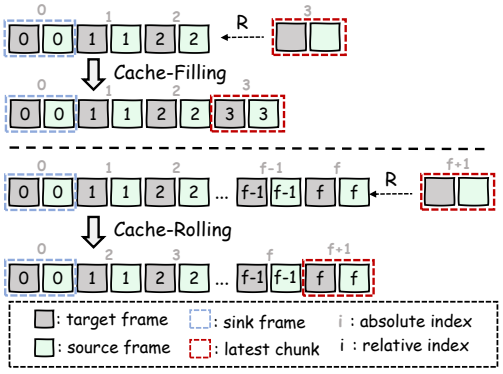


Figure 3: **Shared Relative RoPE.** Each source–target pair shares the same temporal index. Here, $f = f_{\max}$ denotes the maximum temporal index, and $\mathbf{R}$ denotes the rotary operation in Eq. 1.

performs few-step autoregressive rollout on its own generated history, and the DMD objective (Yin et al., 2024) is evaluated on the generated target frames. This closely reproduces the inference-time dynamics and reduces the exposure bias caused by conditioning on ground-truth history. Nevertheless, two challenges arise when generation extends beyond the training horizon: temporal RoPE indices enter unseen positional ranges, and prediction errors accumulate in the self-generated context. We address these challenges with *Shared Relative RoPE* and *Cycle Streaming Tuning*, respectively.

Cycle Streaming Tuning. Causal distillation on short clips does not expose the student to the progressively degraded self-generated contexts encountered during long inference. Naively distilling an entire long rollout is also impractical: the teacher is reliable only within its short training horizon, while backpropagating through the complete sequence incurs prohibitive memory cost (Yang et al., 2025). We address both limitations with *Cycle Streaming Tuning* (CST), which constructs long synchronized training streams from short V2V samples and applies teacher supervision only to a local segment at the end of each rollout. As shown in Figure 2 (c), given a training sample $\mathcal{V} = (V_s, V_t, P_s, P_t)$ of F frames, we synthesize an extended stream by alternating between the original sequence and its temporally reversed counterpart:

$$\mathcal{V}_n^{\text{long}} = \bigoplus_{j=0}^{n} \mathcal{V}^{(j)}, \qquad \mathcal{V}^{(j)} = \begin{cases} \mathcal{V}, & j \text{ is even}, \\ \mathcal{V}^{\text{rev}}, & j \text{ is odd}, \end{cases} \qquad n \sim \mathcal{U}\{0, \dots, n_{\max}\}, \tag{5}$$

where $\bigoplus$ denotes temporal concatenation with duplicated turning frames removed; the reversed sequence follows the temporal order $\{F-1, F-2, \dots, 1\}$. The source video, target video, and their camera trajectories are reversed jointly to preserve frame-level synchronization. Thus, n controls the number of direction reversals: $n = 0$ retains the original sample, while larger n produces progressively longer forward–backward streams that repeatedly revisit previously observed viewpoints.

The student autoregressively processes the entire extended stream to build a long self-generated history in its KV cache. We divide the stream into fixed-length segments and stop gradients through all preceding segments, using them only to construct the streaming context. Distillation is applied exclusively to the final segment, whose length remains same as the teacher's training horizon. CST therefore exposes the student to realistic long-horizon context degradation while retaining short-horizon teacher supervision and bounded training memory. Beyond this efficient train-long–test-long optimization, the cyclic construction provides a natural closed-loop consistency signal: when the camera revisits a previous viewpoint, the student must recover the corresponding appearance despite conditioning on its accumulated predictions. It also increases camera-trajectory diversity without requiring additional long-video data.

Shared Relative RoPE. A naive causal implementation assigns temporal indices according to the global frame number. Once a rollout exceeds the pretrained range $[0, f_{\max}]$, the model must evaluate RoPE phases at unseen indices, resulting in quality degradation and eventual overflow (Yesiltepe et al., 2025). Training-free re-anchoring (Yesiltepe et al., 2025) at inference can alleviate overflow, but leaves a positional train–test mismatch as the model is still optimized using absolute indices. Moreover, existing relative schedules (Bian et al., 2026) are designed for a single video stream. Naively assigning consecutive indices to our interleaved source and target tokens would double the effective temporal span and assign different coordinates to synchronized cross-view tokens.

We therefore propose *Shared Relative RoPE*, which applies the same bounded, cache-relative schedule during both training and inference while respecting the source–target pair structure. As shown in Figure 3, temporal indices increase normally during the *Cache-Filling Phase*. Once the cache is full, the schedule enters the *Cache-Rolling Phase*: the oldest non-sink chunk is evicted, the newest chunk is anchored at $f_{\max}$, and the remaining cached chunks are shifted backward while the sink chunk remains fixed at the initial temporal index. The temporal RoPE coordinates are therefore determined by each chunk's relative position within the current cache window, rather than its global frame index. Consequently, regardless of how long the global rollout becomes, all active RoPE indices remain within $[0, f_{\max}]$, while the pairwise temporal offsets within the rolling window are preserved. Only the temporal component of 3D-RoPE is rescheduled; the spatial coordinates and RoCE camera phases remain unchanged. In practice, the KV cache stores RoCE-conditioned but RoPE-free keys, *e.g.*, $\bar{k} \odot \widetilde{\mathbf{R}}_{\text{qk}}$ in Eq. 4. The cache-relative RoPE rotations are applied only immediately before attention computation. This design preserves the camera conditioning of previous

Table 1: **Quantitative comparison with SOTA methods on short and long video clips.** Our method improves visual quality and maintains accurate camera control while significantly reducing inference time. Best results in **bold** and second best are underlined.

Method	Latency (s) ↓			FPS	CLIP-V	VBench (Huang et al., 2024) ↑						Camera Acc. ↓	
	Init.	Gen.	Total	↑	↑	Sub. Cons	Bg. Cons.	Aes. Qual.	Img. Qual.	Tem. Flick.	Motion Smooth	Trans Err	Rot Err
Short Video Clips													
ReCamMaster (2025)	0	426	426	0.19	84.71	89.49	93.08	52.58	53.40	97.10	98.86	0.035	2.615
CamCloneMaster (2025)	0	1010	1010	0.08	90.67	90.37	93.07	54.08	56.13	96.63	99.01	0.0364	3.753
TrajectoryCrafter (2025)	74	495	569	0.14	90.51	91.58	94.04	57.64	58.35	**97.42**	99.06	0.047	3.506
ReDirector (2026)	30	472	502	0.16	92.74	93.30	94.68	56.60	59.84	96.85	99.13	0.0203	**1.176**
NeoVerse (2026)	12	50	62	1.31	91.75	93.16	**94.69**	**57.79**	**63.68**	96.35	99.09	0.0217	1.743
Inspatio-World (2026)	35	1.4	36.4	2.13	**93.01**	92.86	94.40	55.27	59.29	96.63	98.95	0.024	1.665
Ours	1	1.76	**2.76**	**10**	92.74	**93.70**	93.95	55.48	60.43	96.81	**99.26**	**0.0178**	1.530
Long Video Clips (177 frames)													
ReCamMaster (2025)	0	1771	1771	0.10	65.64	70.88	88.73	38.55	31.63	98.40	98.88	0.148	12.181
CamCloneMaster (2025)	0	4140	4140	0.04	78.18	80.41	89.27	48.34	49.27	**98.61**	98.98	0.136	4.826
TrajectoryCrafter (2025)	94	1736	1830	0.096	86.22	91.63	94.41	55.63	58.28	97.14	98.74	0.133	4.692
ReDirector (2026)	45	2105	2150	0.082	93.86	93.82	**95.24**	60.16	60.70	97.53	99.18	**0.034**	2.660
NeoVerse (2026)	50	157	207	0.855	93.31	92.70	94.74	**60.19**	**65.26**	96.84	99.01	0.069	**0.998**
Inspatio-World (2026)	79	1.4	80.4	1.96	**95.56**	**94.09**	95.16	59.92	61.54	97.56	99.20	0.055	1.731
Ours	1	1.76	**2.76**	**10**	94.34	93.75	94.31	59.21	62.19	97.54	**99.33**	0.053	1.603

frames directly in the KV cache, avoiding an auxiliary cache for historical camera conditions. The complete procedure for our distillation incorporating our core designs is detailed in Algorithm 1.

4 Experiments

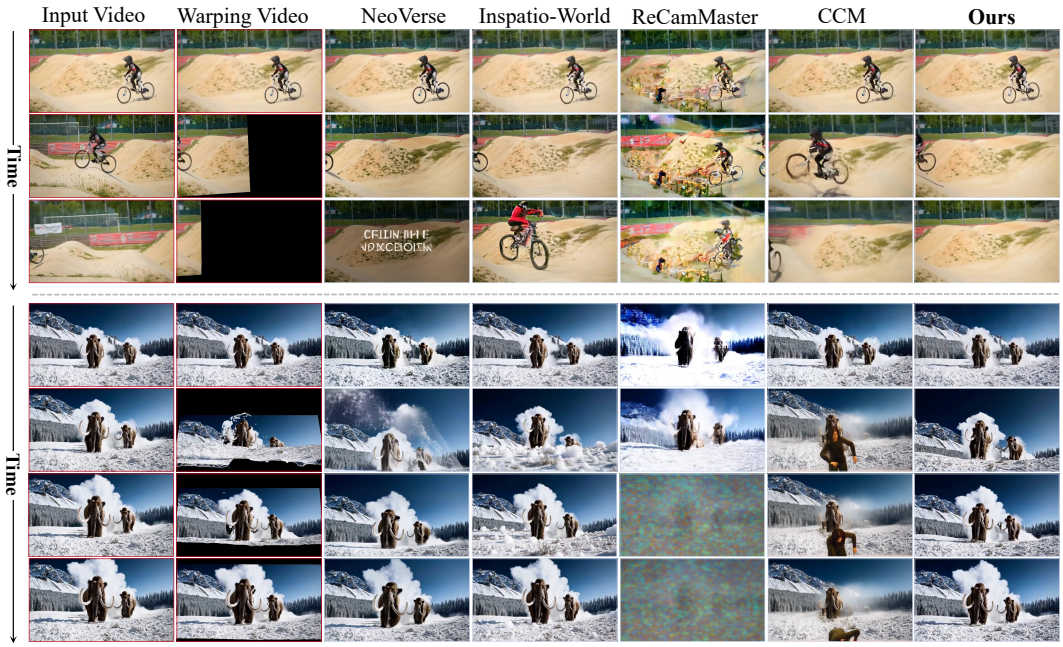


Figure 4: **Qualitative comparison with state-of-the-art methods.** Two examples of video re-rendering are shown: short clips (*top*) and long sequences (*bottom*).

Implementation details. We first fine-tune the bidirectional teacher model on the MultiCamVideo dataset (Bai et al., 2025) for 25K steps at a resolution of 384 × 672 with a learning rate of 0.0001

Table 2: **Ablation study in 177 frames.** Rows (A–B) compare condition-injection strategies under the full pipeline; rows (C–F) ablate the two long-horizon designs on top of Cross-Frame Injection.

Exp.	Condition Injection	SR-RoPE	CST	CLIP-V ↑	VBench (Huang et al., 2024) ↑				Camera Acc. ↓	
					Sub. Cons	Bg. Cons.	Aes. Qual.	Img. Qual.	Trans Err	Rot Err
(A)	Temporal Concat.	✓	✓	90.65	91.45	93.48	56.72	56.55	0.080	2.931
(B)	Channel-wise Concat.	✓	✓	84.11	89.13	91.59	56.46	57.37	0.094	2.920
(C)	Cross-Frame (Ours)	×	×	92.70	91.14	93.78	58.80	58.29	0.091	7.105
(D)	Cross-Frame (Ours)	✓	×	93.83	93.35	94.26	58.74	60.63	0.074	3.636
(E)	Cross-Frame (Ours)	×	✓	93.80	90.73	93.72	59.19	60.31	0.068	2.487
(F)	Cross-Frame (Ours)	✓	✓	94.34	93.75	94.31	59.21	62.19	0.053	1.603

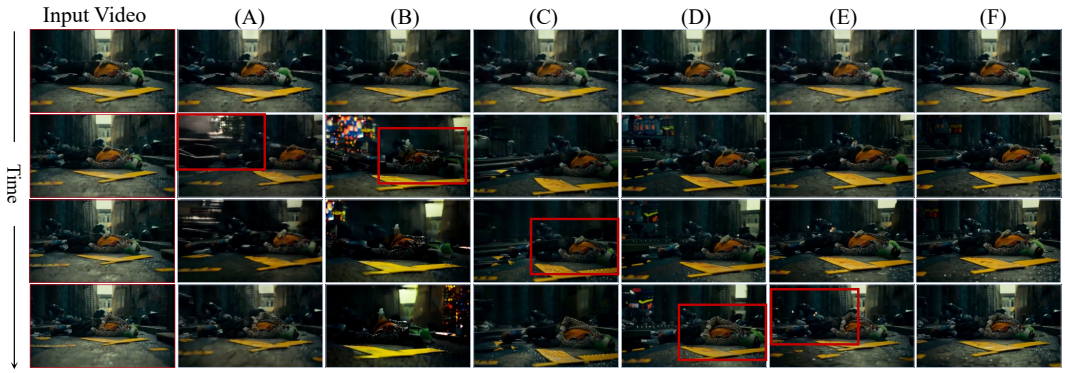


Figure 5: **Qualitative ablation study on 177-frame videos.** Red boxes highlight failure cases. Rows (A)–(F) correspond to the configurations in Table 2.

and a batch size of 8. During teacher fine-tuning, we update only the self-attention layers and keep all remaining modules frozen, as in prior camera-controlled V2V models (Bai et al., 2025; Park et al., 2026). The causal student uses three denoising steps (times steps $t_0 = 1000, t_1 = 755, t_2 = 522, t_3 = 0$), and all quantitative experiments use a KV cache size of 24 (12 frame pairs). We set $n_{max} = 2$. All the models are trained on four H20 GPUs.

Baselines. We compare our method with state-of-the-art video re-rendering generation approaches. ReCamMaster (Bai et al., 2025) and ReDirector (Park et al., 2026) use implicit camera conditioning, while CamCloneMaster (CCM) (Luo et al., 2025) transfers camera motion from a reference video. TrajectoryCrafter (Yu et al., 2025a), NeoVerse (Yang et al., 2026), and Inspatio-World (Team et al., 2026a) use explicit geometric warping. Inspatio-World also employs a few-step causal generator, making it a particularly relevant comparison for streaming efficiency.

Evaluation protocol. We construct two test sets with diverse real and synthetic content. The short set contains 50 videos (33–81 frames) from DAVIS (Pont-Tuset et al., 2017) and prior works (Bai et al., 2025; Brooks et al., 2024); the long set contains 18 clips of 177 frames. Applying 8 distinct camera trajectories to each source video yields 400 short and 144 long video–camera pairs, covering varied content and motions while explicitly probing autoregressive stability over extended rollouts. For metrics, we report view synchronization with CLIP-V (Bai et al., 2025), video fidelity with the official VBench suite (Huang et al., 2024), and camera control with translation and rotation error (Trans/Rot Err). We further report time-to-first-output (latency) and FPS on H20 GPUs to assess real-time interactivity; per-metric details and the user study are in the supplementary.

4.1 Results

Quantitative comparisons. Table 1 summarizes the results on both test sets. In terms of real-time interactivity, World Rerenderer produces its first output in 2.76 s, comprising 1 s of initialization and 1.76 s of first-chunk generation, and subsequently sustains streaming rendering at 10 FPS.

This reduces time-to-first-output by factors of approximately 13–29 relative to Inspatio-World, the only other streaming-capable baseline (36.4/80.4 s, dominated by offline reference preprocessing). Compared with offline methods, which require full-sequence processing before producing output (62–4,140 s), the latency reduction reaches three orders of magnitude. Our throughput is also approximately $5\times$ that of Inspatio-World (1.96–2.13 FPS). In terms of quality, our model achieves the best subject consistency, motion smoothness, and translation accuracy on short clips and maintains competitive performance on 177-frame sequences. By comparison, some implicit baselines degrade substantially over longer rollouts; for example, ReCamMaster's CLIP-V score drops from 84.71 to 65.64. Beyond sequence length, warping-based methods (Yu et al., 2025a; Yang et al., 2026) are also sensitive to large viewpoint deviations: as the target view moves away from the source, the reprojected source covers fewer target-view regions, weakening guidance and potentially compromising source fidelity (Figure 4). Overall, World Rerenderer combines low startup latency and high streaming throughput with strong source fidelity and long-horizon consistency.

Qualitative comparisons. Figure 4 compares video re-rendering under diverse camera motions and sequence lengths. Warping-based methods (Yang et al., 2026; Team et al., 2026a) exhibit ghosting and structural distortions in regions where the reprojected source becomes sparse or misaligned (warping video in Figure 4). Implicit methods (Bai et al., 2025; Luo et al., 2025) remain plausible on short clips but tend to lose view synchronization and produce blurry or drifting content as the sequence extends beyond the training horizon. In contrast, World Rerenderer preserves sharper scene details, maintains temporal synchronization with the source video, and follows the specified camera trajectories in both short and long sequences. Minute-scale results are shown in A.2.

4.2 Ablation Studies

We ablate the core designs on 177-frame sequences in Table 2, with qualitative visualizations shown in Figure 5. Additional ablation results are provided in the supplementary material.

Condition injection (A–B vs. F). Replacing Cross-Frame Injection with temporal concatenation (A) (Bai et al., 2025) places the entire source sequence in a fixed prefix, weakening local cross-view correspondence as the sequence grows and reducing synchronization and consistency. Channel-wise concatenation (B) (Shin et al., 2025) further lacks explicit temporal alignment between corresponding source and target frames, resulting in poorer view synchronization and camera accuracy. By contrast, Cross-Frame Injection (F) maintains a synchronized source anchor for every target frame and achieves the strongest overall performance across the reported metrics.

Long-horizon designs (C–E vs. F). The directly distilled student (C) degrades over long rollouts as temporal indices move beyond the pretrained range and errors accumulate in the self-generated history. This degradation manifests as reduced camera accuracy and visible temporal inconsistencies in Figure 5. Adding SR-RoPE alone (D) improves consistency and camera control by keeping all active temporal indices within the pretrained range. Adding Cycle Streaming Tuning alone (E) mainly improves image quality and camera accuracy by exposing the student to extended self-generated contexts and revisited viewpoints. Combining SR-RoPE and CST (F) yields the best overall results, confirming that the two designs address complementary bottlenecks: positional extrapolation and long-horizon error accumulation.

5 Conclusion

We presented *World Rerenderer*, a real-time framework for camera-controlled world re-rendering from continuously arriving video streams. Our teacher-to-student framework combines Cross-Frame Injection for synchronized source–target conditioning, Shared Relative RoPE for bounded temporal indexing, and Cycle Streaming Tuning for stable long-horizon autoregressive generation. This design enables the model to preserve source content and dynamics while following online camera controls without offline scene reconstruction. Experiments on short and 177-frame videos demonstrate competitive visual quality and camera accuracy, while achieving 10 FPS streaming rendering and substantially lower startup latency than existing methods. These results establish World Rerenderer as a practical step toward interactive video-conditioned world models and immersive real-time "walk-through" experiences. Limitations and future directions are discussed in the Appendix.

AI USE STATEMENT

In this work, we used generative AI tools (GPT-5.6 and Qwen3.8-Max) for language editing and polishing of the manuscript text. Additionally, we used generative AI tools for grammar checking and sentence-level paraphrasing. All AI-assisted output was manually reviewed, verified for technical accuracy, and revised by the authors before inclusion; no AI-generated text was adopted without author verification. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics statement. This work does not involve ethical issues and poses no ethical risks.

Reproducibility Statement. We make the following efforts to ensure the reproducibility of our proposed method: (1) We provide training and inference details in the appendix (Sec. A.1). (4) We provide more analysis and discussion in the appendix (Sec. A.2).

REFERENCES

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22875–22889, June 2025.

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

Yuxuan Bian, Zeyue Xue, Songchun Zhang, Shiyi Zhang, Weiyang Jin, Yaowei Li, Junhao Zhuang, Haoran Li, Jie Huang, Haoyang Huang, et al. Echo-infinity: Learning evolving memory for real-time infinite video generation. *arXiv preprint arXiv:2606.04527*, 2026.

Ollin Boer Bohan. Taehv: Tiny autoencoder for hunyuan video. https://github.com/madebyollin/taehv, 2025.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

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.

Yabo Chen, Yuanzhi Liang, Jiepeng Wang, Tingxi Chen, Junfei Cheng, Zixiao Gu, Yuyang Huang, Zicheng Jiang, Wei Li, Tian Li, et al. Teleworld: Towards dynamic multimodal synthesis with a 4d world model. *arXiv preprint arXiv:2601.00051*, 2025.

Yukang Chen, Luozhou Wang, Wei Huang, Shuai Yang, Bohan Zhang, Yicheng Xiao, Ruihang Chu, Weian Mao, Qixin Hu, Shaoteng Liu, et al. Longlive-2.0: An nvfp4 parallel infrastructure for long video generation. *arXiv preprint arXiv:2605.18739*, 2026.

Tianle Cheng, Zeyan Zhang, Kaifeng Gao, and Jun Xiao. Adaptive begin-of-video tokens for autoregressive video diffusion models. *arXiv preprint arXiv:2511.12099*, 2025.

Wanquan Feng, Jiawei Liu, Pengqi Tu, Tianhao Qi, Mingzhen Sun, Tianxiang Ma, Songtao Zhao, Siyu Zhou, and Qian He. I2vcontrol-camera: Precise video camera control with adjustable motion strength. *arXiv preprint arXiv:2411.06525*, 2024.

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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

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*, 2025a.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 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.

Hyeonho Jeong, Suhyeon Lee, and Jong Chul Ye. Reangle-a-video: 4d video generation as video-to-video translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11164–11175, 2025.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=66NzcRQuOq.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.

Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. *Advances in Neural Information Processing Systems*, 37:16240–16271, 2024.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

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, 2025a.

Runjia Li, Moayed Haji-Ali, Ashkan Mirzaei, Chaoyang Wang, Arpit Sahni, Ivan Skorokhodov, Aliaksandr Siarohin, Tomas Jakab, Junlin Han, Sergey Tulyakov, et al. Egoedit: Dataset, real-time streaming model, and benchmark for egocentric video editing. *arXiv preprint arXiv:2512.06065*, 2025b.

Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen, Lin Cheng, and Xi Li. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28785–28796, October 2025c.

Yixuan Li, Yanhong Zeng, Ka Leong Cheng, Jiayi Zhu, Hanlin Wang, Wen Wang, Yihao Meng, Hao Ouyang, Qiuyu Wang, Yue Yu, et al. Cameraanything: Refilming videos with arbitrary camera control. *arXiv preprint arXiv:2607.24591*, 2026.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan Burgert, Yuancheng Xu, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, et al. Vista4d: Video reshooting with 4d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32671–32682, 2026.

Dongyue Lu, Ao Liang, Tianxin Huang, Xiao Fu, Yuyang Zhao, Baorui Ma, Liang Pan, Wei Yin, Lingdong Kong, Wei Tsang Ooi, et al. See4d: Pose-free 4d generation via auto-regressive video inpainting. *arXiv preprint arXiv:2510.26796*, 2025.

Yawen Luo, Xiaoyu Shi, Jianhong Bai, Menghan Xia, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Camclonemaster: Enabling reference-based camera control for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–10, 2025.

Hengye Lyu, Zisu Li, Yue Hong, Yueting Weng, Jiaxin Shi, Hanwang Zhang, and Chen Liang. Dit as real-time rerenderer: Streaming video stylization with autoregressive diffusion transformer. In *European Conference on Computer Vision*, pp. 256–273. Springer, 2026.

Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Bingyuan Wang, Qinghe Wang, Xuanhua He, Hongfa Wang, et al. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 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 (CVPR)*, pp. 7752–7761, June 2026.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pp. 3481–3490. PMLR, 2018.

Byeongjun Park, Byung-Hoon Kim, Hyungjin Chung, and Jong Chul Ye. Redirector: Creating any-length video retakes with rotary camera encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11163–11173, June 2026.

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, et al. Genie 2: A large-scale foundation world model. *URL: https://deepmind. google/discover/blog/genie-2-a-large-scale-foundation-world-model*, 2, 2024.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

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. *arXiv preprint arXiv:2511.01266*, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

Yaofeng Su, Yuming Li, Zeyue Xue, Jie Huang, Siming Fu, Haoran Li, Ying Li, Zezhong Qian, Haoyang Huang, and Nan Duan. Omniforcing: Unleashing real-time joint audio-visual generation. *arXiv preprint arXiv:2603.11647*, 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.

HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, et al. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025.

InSpatio Team, 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*, 2026a.

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*, 2026b.

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.

Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. *European Conference on Computer Vision (ECCV)*, 2024.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 3(4):6, 2025a.

Qinghe Wang, Yawen Luo, Xiaoyu Shi, Xu Jia, Huchuan Lu, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–10, 2025b.

Zhen Wang, Youcan Xu, Jun Xiao, and Long Chen. Flowmotion: Training-free flow guidance for video motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38143–38153, 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.

Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Trajectory attention for fine-grained video motion control. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=2z1HT5lw5M.

Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024.

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.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40340–40351, June 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Hidir Yesiltepe, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. *arXiv preprint arXiv:2511.20649*, 2025.

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.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22963–22974, 2025.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 100–111, October 2025a.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2025b. doi: 10.1109/TPAMI.2025.3613256.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026.

David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E. Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2050–2062, 2025a. doi: 10.1109/CVPR52734.2025.00197.

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*, 2025b.

Kesen Zhao, Jiaxin Shi, Beier Zhu, Junbao Zhou, Xiaolong Shen, Yuan Zhou, Qianru Sun, and Hanwang Zhang. Real-time motion-controllable autoregressive video diffusion. *arXiv preprint arXiv:2510.08131*, 2025.

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.

A APPENDIX

A.1 MORE EXPERIMENTAL SETTINGS

Teacher Model Training. Following ReDirector (Park et al., 2026), we additionally adopt two data strategies. **(i) Identity retakes**: we include pairs in which the target video and camera trajectory coincide with the source, $\{\mathbf{V}_s, \mathbf{P}_s\} = \{\mathbf{V}_t, \mathbf{P}_t\}$; because the paired tokens then carry identical 3D-RoPE and RoCE coordinates, this supervises exact content preservation and tightens the cross-view alignment learned by Cross-Frame Injection. **(ii) Time reversal**: we augment each clip with its temporally reversed counterpart, which doubles the diversity of camera trajectories and allows the target viewpoint to deviate from the source already at the first frame. Note that this reversal augmentation acts on the teacher's training data and is distinct from the cyclic trajectories used in Cycle Streaming Tuning (Sec. 3.3), which operate on the distillation stage.

Self-Forcing-style Distillation. We distill the multi-step teacher into a few-step causal student via a two-phase recipe. In the first phase (causal architecture adaptation), we follow the multi-condition causal initialization of Team et al. (2026a): during training, the ground-truth previous block is injected into the current denoising block via channel concatenation, while at inference the model conditions on its own self-generated previous block. In the second phase, we optimize the causal student with distribution matching distillation (DMD) (Yin et al., 2024), combining the DMD loss with an adversarial loss (Goodfellow et al., 2020) and R1 regularization (Mescheder et al., 2018). The DMD loss, generator loss, and discriminator loss are weighted 1.0, 0.1, and 0.05, respectively; the generator and the critic (fake score function) are updated at a 1:5 ratio; and the R1 regularization weight is set to 1e2. In addition, we adopt the self-rollout strategy (Zhao et al., 2025), updating the KV cache with fully denoised self-generated chunks rather than one-step estimates, to further bridge the gap between autoregressive training and inference. The same hyperparameters are retained during Cycle Streaming Tuning, and the distillation stage runs for 3,000 iterations.

The full training pipeline is summarized in Algorithm 1. Here, H denotes the teacher's supervision horizon measured in latent blocks, $\mathrm{sg}[\cdot]$ denotes stop-gradient, and Ψ denotes the forward noising operation. G_θ^{KV} denotes the key/value embeddings returned by the same generator during a forward pass on the clean generated block. The source and target latents are interleaved through Cross-Frame Injection, while the previous generated block p provides an additional condition through channel concatenation. Gradients are retained only at the sampled denoising step within the final supervised segment. All remaining denoising steps are completed without gradients to construct the subsequent self-generated context. SR-RoPE applies bounded, shared temporal indices to synchronized source–target pairs, while the cache retains the sink block and evicts the oldest non-sink blocks as needed.

Prompt Setting. All experiments use a fixed pair of positive and negative prompts. In practice, we find that detailed video captions degrade causal streaming generation, because such captions can leak the content of future chunks into the current one. We therefore replace per-video captions with a single task-level instruction shared by all samples:

Positive Prompt Template

Re-render the provided source video from the novel target camera trajectory while preserving the original scene dynamics and appearance, maintaining strict temporal and multi-view consistency, photorealistic rendering, 4k resolution.

The negative prompt is likewise shared across all videos:

Negative Prompt Template

Overly vivid colors, overexposed, static, blurry details, subtitles, style, artwork, frame, still, overall grayish, worst quality, low quality, JPEG compression artifacts, ugly, incomplete, extra fingers, poorly drawn hands, poorly drawn faces, deformed, disfigured, malformed limbs, fused fingers, motionless frame, cluttered background, three legs, many people in the background, walking upside down.

Table 3: **Effect of input trajectories on short and long video clips.**

Camera Pose	CLIP-V ↑	VBench (Huang et al., 2024) ↑						Camera Acc. ↓	
		Sub. Cons	Bg. Cons.	Aes. Qual.	Img. Qual.	Tem. Flick.	Motion Smooth	Trans Err	Rot Err
Short Video Clips									
ViPE (Huang et al., 2025a)	92.67	93.34	93.93	55.23	60.16	96.76	99.26	0.0185	**1.444**
ViPE Frontend (Huang et al., 2025a)	**92.74**	**93.70**	**93.95**	**55.48**	**60.43**	**96.81**	**99.26**	**0.0178**	1.530
Long Video Clips (177 frames)									
ViPE (Huang et al., 2025a)	**94.56**	93.69	94.22	59.10	61.95	97.54	99.33	**0.050**	**1.475**
ViPE Frontend (Huang et al., 2025a)	94.34	**93.75**	**94.31**	**59.21**	**62.19**	**97.54**	**99.33**	0.053	1.603

Metric details. For view synchronization, following ReCamMaster (Bai et al., 2025), we compute CLIP-V as the average CLIP similarity between source and target frames at the same timestamp, where higher values indicate tighter synchronization with the incoming stream. For video fidelity, we use the official VBench implementation (Huang et al., 2024) and report six dimensions that separate appearance preservation (subject and background consistency), perceptual quality (aesthetic and imaging quality), and short-term temporal stability (temporal flickering and motion smoothness). For camera accuracy, we report TransErr and RotErr (Bai et al., 2025; Park et al., 2026), which measure the relative translation and rotation errors for every frame pair, with the camera poses of generated videos estimated by ViPE (Huang et al., 2025a). For efficiency, we report latency as the time-to-first-output, which comprises initialization (e.g., the offline geometric preprocessing required by Inspatio-World) plus the model's generation of the first output chunk. For fairness, the throughput (FPS) of every method is measured on the same two-H20 setup. In our full streaming pipeline the two GPUs operate in parallel: one estimates the camera pose of the incoming source stream with the ViPE frontend (Huang et al., 2025a), while the other performs generation. Note that this same two-GPU budget brings the baselines no benefit: offline and bidirectional methods must first complete full-sequence preprocessing (e.g., depth estimation) or full-sequence generation before emitting any frame, so their latency is algorithmically inherent and cannot be reduced by additional hardware. When measuring throughput, we replace the original Wan VAE with a lightweight Tiny-VAE (Boer Bohan, 2025) to accelerate encoding and decoding.

A.2 ADDITIONAL ABLATION EXPERIMENTS AND QUALITATIVE RESULTS

Impact of Input Trajectories. Our main experiments use camera trajectories estimated by the ViPE frontend, which processes the source stream causally in an online manner without subsequent offline backend optimization. To assess the effect of using frontend-only trajectory estimation, we compare against the original ViPE pipeline (Huang et al., 2025a), which includes offline optimization and serves as an offline reference. As shown in Table 3, the two trajectory estimators yield highly similar re-rendering quality. On short clips, ViPE frontend slightly improves CLIP-V from 92.67 to 92.74 and increases subject consistency, imaging quality, and motion smoothness. It also reduces the translation error from 0.0185 to 0.0178. On 177-frame clips, the original ViPE obtains a slightly higher CLIP-V score and lower camera errors, while ViPE frontend achieves marginally better subject consistency, background consistency, aesthetic quality, and imaging quality. These small differences indicate that the performance of World Rerenderer is not sensitive to the choice of trajectory estimator. More importantly, ViPE frontend provides comparable camera trajectories without accessing future frames, making it suitable for the online streaming setting used by our method.

Impact of prompt type. Table 4 and Figure 6 compare detailed per-video captions with our fixed task-level instruction. On short clips the two settings perform comparably, and captions even hold a slight edge on CLIP-V and subject consistency (93.10 vs. 92.74 and 93.91 vs. 93.70), since a caption that describes the entire clip provides a useful content prior when the generation horizon is short. The picture reverses on long sequences: because a detailed caption describes content that has not yet arrived in the stream, it leaks future information into the current chunk and prompts the causal student to hallucinate it prematurely, drifting away from the incoming source. Accordingly, the fixed



Figure 6: The visualization of ablation on positive prompt type. Red boxes indicate low quality content in new views.

Table 4: **Effect of positive prompt type.**

Prompt	CLIP-V ↑	VBench ↑ Sub. Cons.	Bg. Cons.	Aes. Qual.	Img. Qual.	Tem. Flick.	Mot. Smth.
Short Video Clips							
Caption	**93.10**	**93.91**	93.93	**55.70**	59.38	**96.86**	99.23
Fixed	92.74	93.70	**93.95**	55.48	**60.43**	96.81	**99.26**
Long Video Clips							
Caption	94.07	93.39	94.04	59.08	60.09	97.48	99.29
Fixed	**94.34**	**93.75**	**94.31**	**59.21**	**62.19**	**97.54**	**99.33**

Table 5: **Effect of max number of direction reversals (n_{max}) on short and long video clips.**

n_{max}	CLIP-V ↑	VBench (Huang et al., 2024) ↑ Sub. Cons	Bg. Cons.	Aes. Qual.	Img. Qual.	Tem. Flick.	Motion Smooth	Camera Acc. ↓ Trans Err	Rot Err
Short Video Clips									
$n_{max} = 0$	92.31	93.26	93.88	55.28	59.46	96.89	99.26	0.023	1.851
$n_{max} = 1$	92.44	93.25	93.82	55.31	59.23	**96.89**	**99.28**	0.0200	1.689
$n_{max} = 2$	**92.74**	**93.70**	**93.95**	**55.48**	**60.43**	96.81	99.26	**0.0178**	1.530
Long Video Clips (177 frames)									
$n_{max} = 0$	93.83	93.35	94.26	58.74	60.63	**97.88**	99.37	0.074	3.636
$n_{max} = 1$	93.96	93.62	94.21	58.74	60.89	97.82	**99.37**	0.074	2.712
$n_{max} = 2$	**94.34**	**93.75**	**94.31**	**59.21**	**62.19**	97.54	99.33	**0.053**	**1.603**

prompt outperforms captions on *all* metrics for 177-frame clips (e.g., CLIP-V 94.34 vs. 94.07 and imaging quality 62.19 vs. 60.09), and Figure 6 shows that caption-conditioned rollouts degrade in newly exposed regions over time (red boxes), whereas the fixed instruction remains tightly synchronized with the source stream. We therefore adopt the fixed task-level prompt in all experiments.

Impact of the maximum number of direction reversals. We vary $n_{\max}$, the maximum number of direction reversals used to construct the cyclic training stream, while keeping all other settings fixed. As reported in Table 5, increasing $n_{\max}$ from 0 to 2 consistently improves long-horizon performance: on 177-frame clips, CLIP-V rises from 93.83 to 94.34, imaging quality from 60.63 to 62.19, and TransErr/RotErr drop from 0.074/3.636 to 0.053/1.603. The short-clip results follow the same trend on content preservation and camera accuracy, with the best CLIP-V, subject/background consistency, imaging quality, and TransErr obtained at $n_{\max} = 2$; temporal flickering and motion smoothness vary only marginally across the three settings. These gains confirm that revisiting earlier viewpoints supplies a useful closed-loop consistency signal while exposing the student to longer self-generated contexts. Importantly, extending the stream is cheap in memory: because gradients are stopped through all preceding segments, a larger $n_{\max}$ adds only no-grad forward passes. Under our compute budget we therefore set $n_{\max} = 2$; scaling it further offers a straightforward time-for-stability trade-off for long-horizon streaming.

Ablation of the core designs on 81-frame clips. We additionally repeat the condition-injection and long-horizon ablations on the 81-frame evaluation set, without camera-pose metrics. As shown in Table 6, Cross-Frame Injection substantially outperforms temporal concatenation and channel-wise concatenation: the full model (F) improves CLIP-V over (A) and (B) by 1.92 and 5.99 points, respectively, and also gives the strongest subject and background consistency and image quality among the injection variants. Starting from Cross-Frame Injection, SR-RoPE alone (D) gives the highest CLIP-V (92.77), while CST alone (E) gives a small gain in aesthetic quality. Combining

Table 6: **Ablation study in 81 frames.** Rows (A–B) compare condition-injection strategies under the full pipeline; rows (C–F) ablate the two long-horizon designs on top of Cross-Frame Injection.

Exp.	Condition Injection	SR-RoPE	CST	CLIP-V ↑	VBench (Huang et al., 2024) ↑					
					Sub. Cons	Bg. Cons.	Aes. Qual.	Img. Qual.	Tem. Flick.	Motion Smooth
(A)	Temporal Concat.	✓	✓	90.82	91.94	93.35	54.52	55.82	97.18	98.99
(B)	Channel-wise Concat.	✓	✓	86.75	92.01	93.02	55.77	58.49	96.90	99.26
(C)	Cross-Frame (Ours)	×	×	92.22	92.22	93.65	55.57	57.26	**97.16**	99.12
(D)	Cross-Frame (Ours)	✓	×	**92.77**	93.53	93.88	55.49	59.36	97.01	**99.28**
(E)	Cross-Frame (Ours)	×	✓	92.48	91.79	93.43	**55.65**	58.19	96.90	98.98
(F)	Cross-Frame (Ours)	✓	✓	92.74	**93.70**	**93.95**	55.48	**60.43**	96.81	99.26

both designs (F) provides the best subject and background consistency and image quality, while retaining near-best motion smoothness (99.26 vs. 99.28 for D). The short-clip results thus show metric-dependent trade-offs, while the 177-frame results more clearly demonstrate the benefits of combining the two long-horizon designs.

Additional results. Figures 9–11 provide additional re-rendering results across diverse scenes and camera trajectories. These examples confirm that World Rerenderer preserves synchronization and content fidelity even under large viewpoint deviations—including close-in and rapidly panning trajectories—while precisely following the requested camera motion over the full rollout. We further provide minute-scale rollouts in Figure 12–14: the re-rendered stream remains temporally stable and consistent with the source over horizons far beyond both the training clips and the 177-frame quantitative benchmark, offering qualitative evidence for the long-horizon stability claimed in the main paper.

A.3 USER STUDY

User study. Since automatic metrics cannot fully capture perceptual preference, we conduct a forced-choice user study following prior work (Lin et al., 2026). We randomly sample 30 video–camera pairs from each of the short (400-pair) and long (144-pair) evaluation sets, and recruit 35 participants. In each trial, participants are shown the source video, a point-cloud rendering of the requested trajectory as the camera-motion reference, and the re-rendered videos of *all* compared methods simultaneously—anonymized and in randomized order (labeled A–G in Figure 8). For each of three questions: (i) source content preservation ("Which option best preserves the input video's content (identity, scene, and motion)?"), (ii) camera control accuracy ("Which option's camera trajectory best matches the camera visualization?"), and (iii) overall video fidelity ("Which option has the best overall quality (video fidelity, geometric coherence, and motion quality)?")—participants select the single best option among all methods. We report each method's preference rate as the percentage of trials in which it is selected. As shown in Figure 7, World Rerenderer is consistently preferred on source content preservation and camera control accuracy. On overall fidelity, NeoVerse is slightly preferred, which is expected given its bidirectional architecture and larger Wan2.1-14B backbone; nevertheless, our causal streaming model is favored on the two task-critical dimensions.

A.4 DISCUSSION

Limitations. While World Rerenderer demonstrates strong real-time re-rendering capabilities, we identify three main limitations. First, training on synthetic data leads to appearance-level overfitting: because the synthetic streams repeatedly expose consistent identity and texture cues, the model occasionally imposes these learned appearance priors on regions that the source never reveals—for instance, hallucinating a cartoon face (Figure 15(a)). Second, visual artifacts emerge under extremely rapid or large-scale camera rotations (Figure 15(b)). We attribute this to two factors: the limited diversity of extreme-motion trajectories in the synthetic training set, and the absence of explicit geometric priors (*e.g.*, depth), which makes it challenging for an implicit model to plausibly hallucinate heavily disoccluded regions. Third, our Cross-Frame Injection mechanism interleaves source and target frames, effectively doubling the sequence length in the attention layers. This

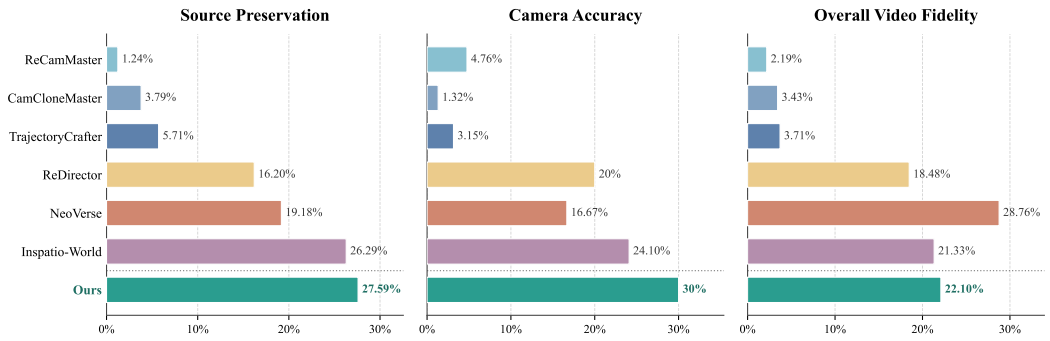


Figure 7: **User study results.** Preference rates over all trials. World Rerenderer is consistently preferred on source content preservation and camera control accuracy—the two task-critical dimensions, while NeoVerse (bidirectional, Wan2.1-14B) is slightly preferred on overall video fidelity.

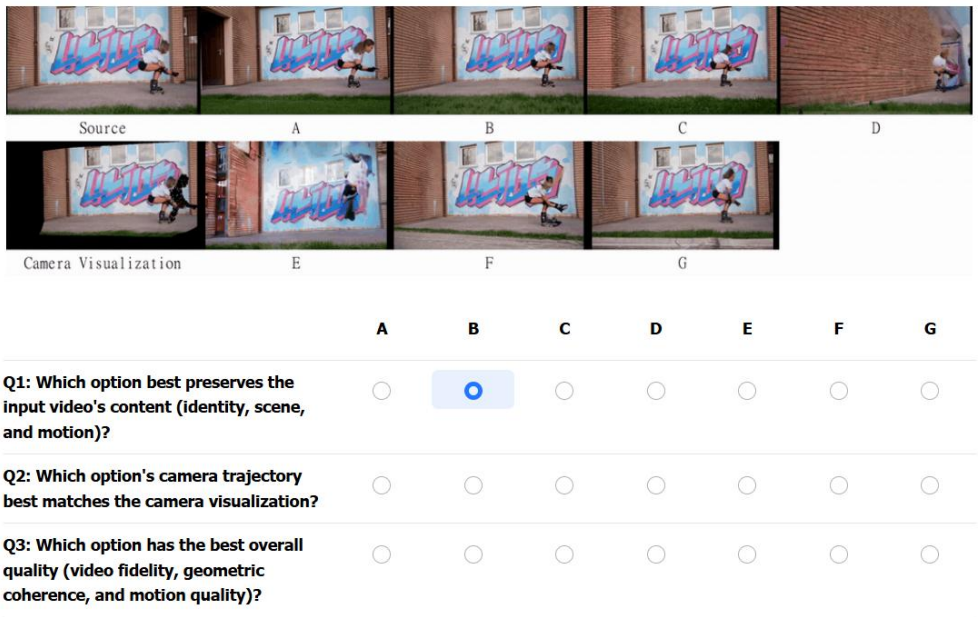


Figure 8: **User study interface.** Each trial displays the source video and the point-cloud camera visualization as references (left), followed by the re-rendered results of all compared methods in anonymized, randomized order (A–G). Participants select the single best option for each of the three questions: source content preservation (Q1), camera control accuracy (Q2), and overall fidelity (Q3).

doubles the KV-cache size and increases the memory footprint during both teacher distillation and autoregressive inference, trading some computational efficiency for strict cross-view consistency.

Future directions. Our future work will directly address these limitations. To mitigate appearance-level overfitting, we plan to incorporate self-supervised learning on real-world videos (Yu et al., 2025a; Yang et al., 2026) to regularize the training process and reduce the model's reliance on synthetic appearance priors. To improve robustness under extreme viewpoints without sacrificing real-time inference speed, an exciting direction is to introduce lightweight geometric guidance (e.g., via teacher distillation or auxiliary depth supervision) that enhances disocclusion reasoning during training *without* adding computational overhead during streaming inference. Finally, scaling up the diversity and volume of the training data—particularly with real-world, multi-view sequences featuring extreme camera motions—will further improve the model's generalization and long-horizon stability.



Figure 9: **Qualitative results of World Rerenderer across diverse camera trajectories.** The first row is the input video; each following row shows World Rerenderer's re-rendering under a different user-specified trajectory (visualized at left)



Figure 10: **Qualitative results of World Rerenderer across diverse camera trajectories.** The first row is the input video; each following row shows World Rerenderer's re-rendering under a different user-specified trajectory (visualized at left)



Figure 11: **Qualitative results of World Rerenderer across diverse camera trajectories.** The first row is the input video; each following row shows World Rerenderer's re-rendering under a different user-specified trajectory (visualized at left)

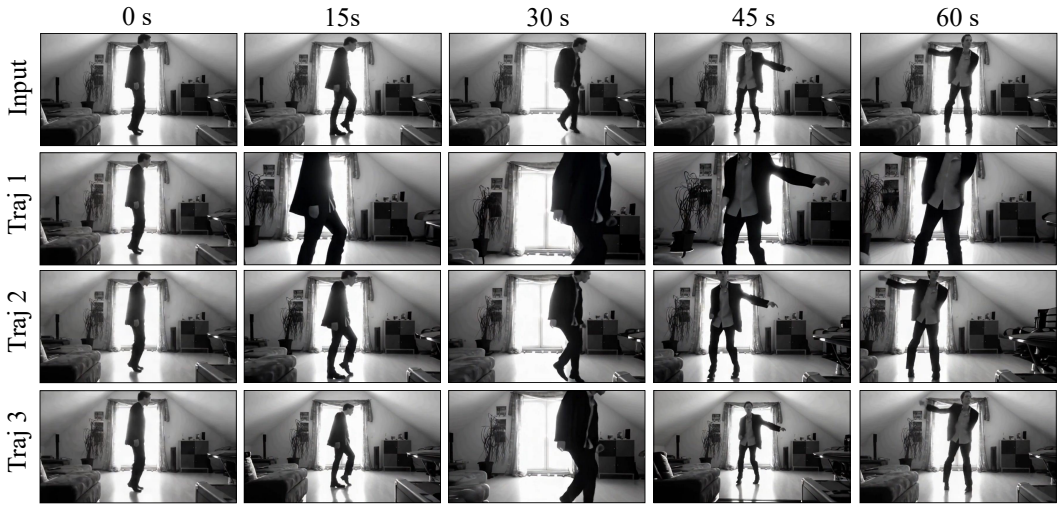


Figure 12: **Minute-scale qualitative results across diverse camera trajectories.** The first row is the input video; each following row shows World Rerenderer's re-rendering under a different user-specified trajectory over a 60-second horizon. As minute-long camera trajectories are impractical to visualize in a static figure, we refer the reader to the supplementary videos for the full rollouts and their trajectory visualizations.

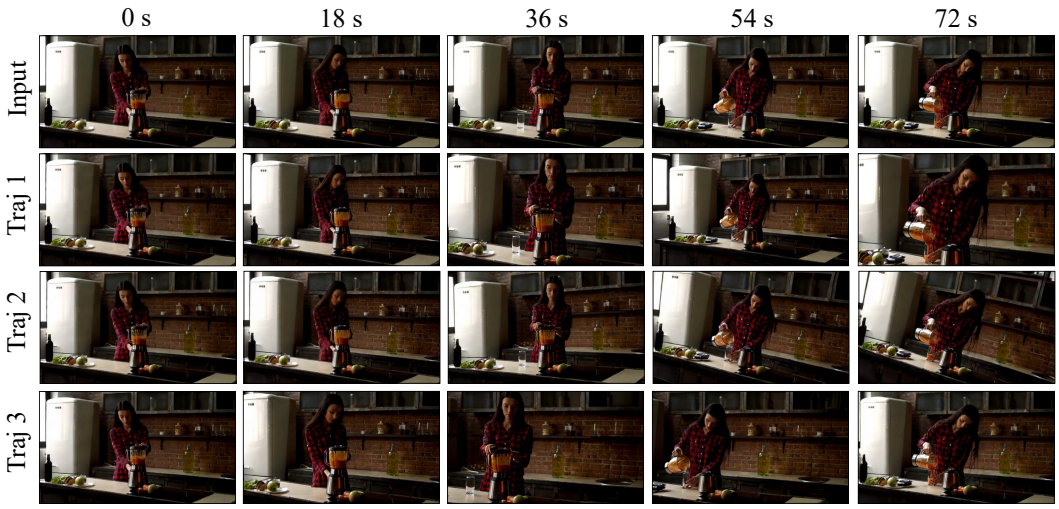


Figure 13: **Minute-scale qualitative results across diverse camera trajectories.** The first row is the input video; each following row shows World Rerenderer's re-rendering under a different user-specified trajectory over a 60-second horizon. As minute-long camera trajectories are impractical to visualize in a static figure, we refer the reader to the supplementary videos for the full rollouts and their trajectory visualizations.

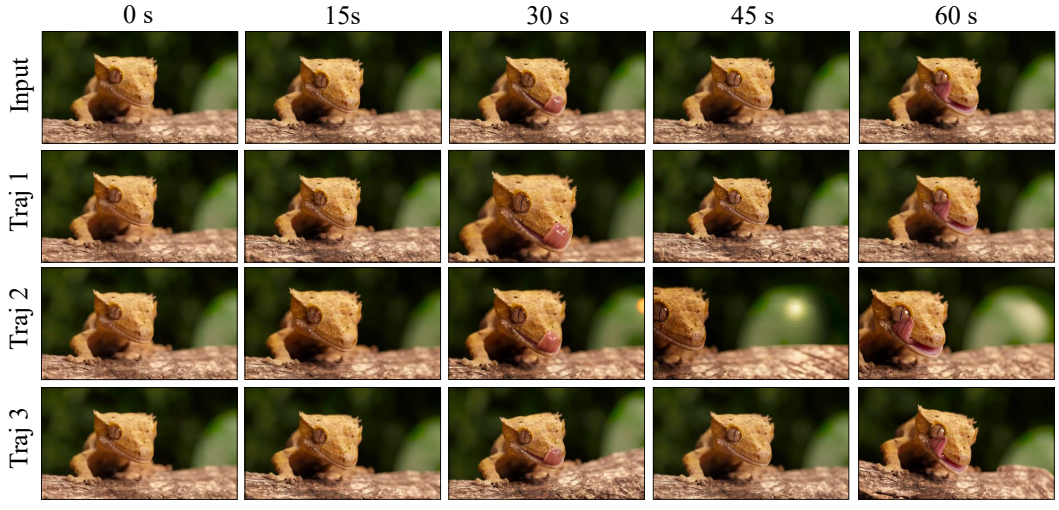
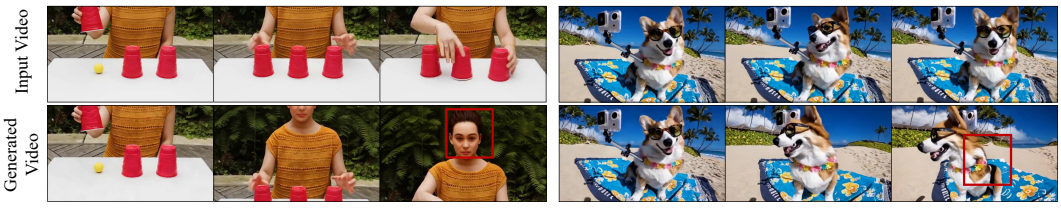


Figure 14: **Minute-scale qualitative results across diverse camera trajectories.** The first row is the input video; each following row shows World Rerenderer's re-rendering under a different user-specified trajectory over a 60-second horizon. As minute-long camera trajectories are impractical to visualize in a static figure, we refer the reader to the supplementary videos for the full rollouts and their trajectory visualizations.



(a) Failure cases of overfiting to appearance (b) Failure cases in large-scale camera pose

Figure 15: **Failure cases.** (a) Overfitting to appearance: the model imposes appearance priors learned from synthetic training data on regions unobserved in the source (hallucinated face, red box). (b) Artifacts under large-scale camera poses: without explicit geometric guidance, heavily occluded regions are implausibly rendered (red box).

Algorithm 1 Self-Forcing-Style Distillation with Cycle Streaming Tuning

Require: Training dataset $\mathcal{D}$
Require: Causal generator G_θ (returns KV embeddings via G_θ^{KV})
Require: Timesteps $t_0 > \cdots > t_T = 0$; maximum reversals $n_{\max}$; teacher supervision length H

while not converged **do**
  Sample $\mathcal{V} = (V_s, V_t, P_s, P_t) \sim \mathcal{D}$
  Sample $n \sim \mathcal{U}\{0, \ldots, n_{\max}\}$
  Construct $\mathcal{V}_n^{\mathrm{long}}$ according to Eq. 5
  Obtain source latent blocks $z_s^{1:L}$ and camera conditions $c^{1:L}$ from $\mathcal{V}_n^{\mathrm{long}}$
  $\mathcal{S} \leftarrow \{L - H + 1, \ldots, L\}$ ▷ Final segment for supervision
  Initialize outputs $\mathcal{X}_\theta \leftarrow [\,]$, KV cache $\mathcal{K} \leftarrow [\,]$, and previous-block condition $p \leftarrow \varnothing$
  Sample $s \sim \mathcal{U}\{0, \ldots, T-1\}$
  Disable gradient computation
  for $i = 1, \ldots, L$ **do**
    Prepare the bounded cache and shared relative temporal indices using SR-RoPE
    Sample $z_{t_0}^i \sim \mathcal{N}(0, I)$
    for $j = 0, \ldots, T-1$ **do**
      if $i \in \mathcal{S}$ **and** $j = s$ **then**
        Enable gradient computation
        $\hat{z}_0^i \leftarrow G_\theta(z_{t_j}^i, z_s^i, p; c^i, t_j, \mathcal{K})$
        Append $\hat{z}_0^i$ to $\mathcal{X}_\theta$
        Disable gradient computation
      else
        $\hat{z}_0^i \leftarrow G_\theta(z_{t_j}^i, z_s^i, p; c^i, t_j, \mathcal{K})$
      end if
      if $j < T-1$ **then**
        Sample $\epsilon \sim \mathcal{N}(0, I)$
        $z_{t_{j+1}}^i \leftarrow \Psi(\mathrm{sg}[\hat{z}_0^i], \epsilon, t_{j+1})$
      end if
    end for
    $\bar{z}_0^i \leftarrow \mathrm{sg}[\hat{z}_0^i]$ ▷ Fully denoised rollout output
    $\mathbf{kv}^i \leftarrow G_\theta^{\mathrm{KV}}(\bar{z}_0^i, z_s^i, p; c^i, 0, \mathcal{K})$ ▷ After camera encoding, before positional RoPE
    Append $\mathbf{kv}^i$ to $\mathcal{K}$
    $p \leftarrow \bar{z}_0^i$
  end for
  Enable gradient computation
  Update θ using DMD and adversarial generator losses on $\mathcal{X}_\theta$, conditioned on the corresponding source and camera segment
end while