LONGREPLAY: EXPLORABLE LONG VIDEO RESHOOTING WITH DYNAMIC MEMORY

Anonymous authors
Paper under double-blind review

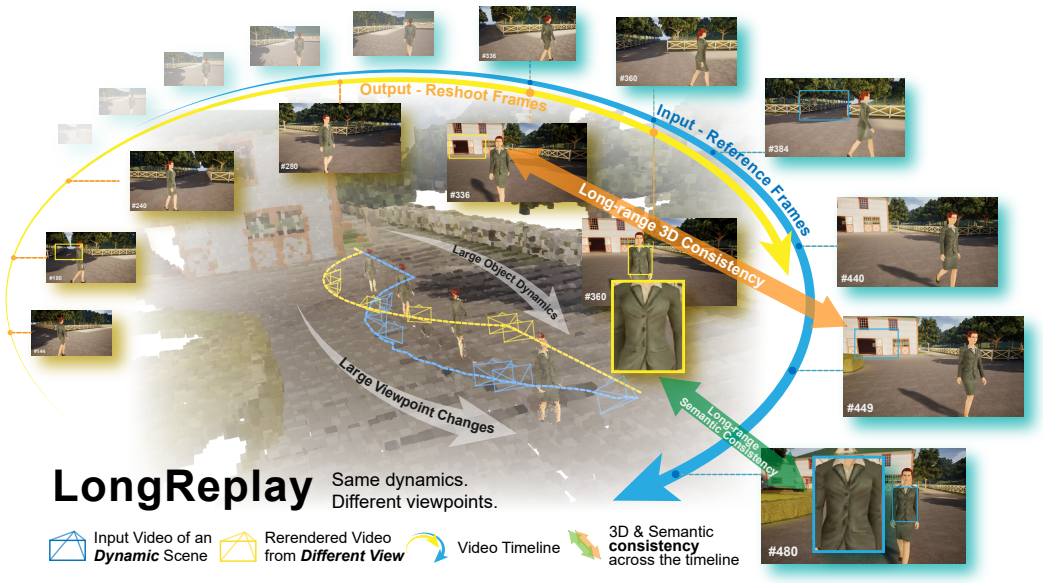

Figure 1: Given a source video and a user-specified camera trajectory, LongReplay reshoots a long novel-view video. The generated view stays synchronized with source dynamics under large camera motion, and remains consistent with the source scene and previously generated chunks.

ABSTRACT

Camera-controlled video reshooting synthesizes a novel-view video along a user-specified camera trajectory while remaining temporally synchronized with a source recording. Existing methods operate on short clips, where the source and the target occupy a single attention window. We study long-video reshooting, in which the source may span minutes and the target is generated autoregressively in overlapping chunks. This setting requires synchronization under large camera baselines and large subject motion, together with long-term 3D consistency of the static scene and semantic consistency of dynamic subjects. Our key insight is to query a dynamic memory of the source video along time, space, and semantics. We propose LongReplay, which realizes this readout in two stages with dynamic awareness. Memory selection retrieves a synchronized source chunk, multi-time spatial patches that cover the target field of view, and semantic anchors that store subject frames independently of the queried time and location. Dynamic-aware correspondence then identifies retrieved patches jointly in time and space with a 4D projection embedding, and matches tokens by semantics with TIPS descriptors. Supervising these queries requires paired long videos of the same dynamic scene under different cameras, so we construct WorldReplay, an Unreal Engine 5 pipeline that renders multiple camera trajectories in the same long take. Experiments show improved multi-view synchronization, including under large viewpoint changes, together with long-range geometric and identity consistency, relative to baselines.

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

Camera-controlled video reshooting synthesizes a novel-view video along a user-specified camera trajectory while remaining temporally synchronized with a source recording. Recent generative methods provide accurate camera control and temporal synchronization on short clips Yu et al. (2025b); Bai et al. (2025); Li et al. (2026c). Their fixed temporal windows, however, limit the duration of the input video. We study *long-video reshooting*, where a target view is generated autoregressively for a source video that may span minutes.

Long-video reshooting introduces two main challenges that a short clip does not pose. The target view must stay synchronized with source dynamics under large camera baselines and large subject motion, which rarely arise inside a short clip. Generation across chunks further requires long-term consistency: 3D consistency of the static scene with the source and with previously generated views, and semantic consistency of dynamic subjects even when the relevant evidence is distant in time or space. A direct solution is to attend to the complete source and generated history at every denoising step. The cost grows with video length, most tokens are redundant or irrelevant, and the transformer must still discover source–target correspondence from this unstructured context.

Our key insight is to query a dynamic memory of the source video along three dimensions: time, space, and semantics. Each source observation has a time t at which it was captured, a space X that it records, and a semantic identity. Synchronization is a query along time: the target frame at timestamp t should retrieve source observations with that same t. Long-term 3D consistency is a query along space: a target pixel that images a point X should retrieve stored observations of X, irrespective of when they were captured. Each such observation must keep both t and X, so the model can identify a dynamic event and reason about the state at the target time, rather than blend observations of the same X recorded at different times. Long-term semantic consistency is a query along semantics: a target token that depicts a subject should retrieve stored observations of that subject, irrespective of the (t, X) at which they were captured. Together, these queries unify temporal synchronization and long-term consistency, and also enable a compact memory that **remembers dynamics**.

Existing methods index context along only one of these coordinates. Short-clip reshooting keeps the full source clip in context and recovers synchronization by concatenating that clip or warping it into the target view Chen et al. (2026a); Yu et al. (2025b); Lin et al. (2026); Bai et al. (2025); Li et al. (2026c). That cue is time-indexed, and the window never requires a long-range query along X. Long-video generators without camera control compress or retrieve their own generated history by semantics Zhang et al. (2026); Wei et al. (2026), which is semantically indexed: it preserves identity across a long rollout, but it does not retrieve a synchronized source event at time t or observations of a point X. Video world models retrieve history by camera and scene geometry AlayaWorld Team et al. (2026); Zhao et al. (2025); Wu et al. (2025); Yu et al. (2025a); Wang et al. (2026d); Xu et al. (2026b); Shen et al. (2026); Qian et al. (2026); Yu et al. (2026b). That memory is space-indexed: it is sufficient for static revisits, but it maps every observation of a 3D point onto the same target coordinate, so states recorded at different t are superimposed. Time-shifted training that enforces static consistency can further encourage the model to disregard dynamic objects Xu et al. (2026b).

To operationalize this multi-dimensional memory readout, we separate the process into memory selection and correspondence injection. Memory selection determines which temporal, spatial, and semantic evidence enters the context, while correspondence injection encodes how each selected patch relates to the target token within attention. Therefore, we propose LongReplay, which replaces full-history attention with **hybrid memory selection** and a **dynamic-aware correspondence** on the retrieved patches. Temporal retrieval returns the source chunk at the target timestamp. Spatial retrieval returns patches that cover the target field of view and retains more than one t for each visible X, so observations of the same X from other times enter the context together. A 4D projection embedding records the source timestamp together with the projected coordinate in the target view, so two observations of the same X remain distinguishable by t, making the embedding dynamic-aware. Semantic anchors store subject frames independently of the queried (t, X), and tokens are matched by semantics. We adopt the TIPS descriptors Maninis et al. (2025); Cao et al. (2026) to provide a semantic relation between source and target patches. Warping those descriptors from the synchronized chunk into the target view implements the semantic query on the target tokens. The same warped descriptors also place the current event in the novel camera, supplying the layout that

warped RGB would otherwise provide Lin et al. (2026); Team et al. (2026); Song et al. (2026a). These embeddings are added to queries and keys, while values retain the retrieved pixels.

Each chunk is generated as image-to-video from the overlapping first frame, under dense RGB and depth warp control Yu et al. (2025b); Hu et al. (2025); Chen et al. (2026a). After generation, estimated cameras and depths are aligned to the persistent memory, and an error-aware blend prevents camera residuals from accumulating across chunks. Supervising these queries requires paired long videos of the same dynamic scene under different cameras, which existing short-clip datasets do not provide. We therefore construct WorldReplay, an Unreal Engine 5 pipeline that renders multiple camera trajectories in the same long take. Experiments show improved multi-view synchronization, including under large viewpoint changes, together with long-range geometric and identity consistency, relative to chunked reshooting baselines. In summary, our contributions are:

- We formulate long-video reshooting as a time–space–semantic readout of dynamic memory that unifies temporal synchronization and long-term consistency.
- We propose a framework, LongReplay, to realize this readout with a two-stage design. Memory selection retrieves synchronized temporal context, multi-time spatial context, and semantic anchors. Dynamic-aware correspondence injection then builds relations between target and context tokens.
- We construct WorldReplay, a scalable synchronized multi-camera pipeline for training this correspondence on long-horizon reshooting.

2 Related Work

Video Reshooting. Existing methods mainly follow three paradigms: explicit 3D/4D proxies that condition diffusion on target-view renders Yu et al. (2024); Ren et al. (2025); Yu et al. (2025b); Wang et al. (2025c); Bian et al. (2025); Hu et al. (2025); Lin et al. (2026); Zhou et al. (2026); Yang et al. (2026b); Wu et al. (2026a), implicit camera injection without reconstruction Bai et al. (2025); Luo et al. (2025); Li et al. (2026c); Liu et al. (2026a;b); Li et al. (2026a); Wang et al. (2026c), and training-free steering of frozen generators Hou & Chen (2025); Song et al. (2026b;a); Wang et al. (2025a). A further line disentangles camera pose from event time Wang et al. (2025b); Huang et al. (2025b); Fu et al. (2026). Most reshooting models remain limited to clip-length videos. Long source videos can be processed by chunking the input and registering generated content into a persistent 4D cache Lin et al. (2026), streaming spatiotemporal memory Team et al. (2026), or sparse KV retrieval over generated history Xu et al. (2026a).

Long-Term Memory in Video Generation. Existing memory designs mainly follow four paradigms. *Streaming windows* keep only recent latents or KVs (optionally a first-frame sink) and drop the rest He et al. (2025); Li et al. (2025a); Gao et al. (2026); Team et al. (2026); Yang et al. (2025; 2026a); Zhu et al. (2025). *History compression* retains a lossy trace of the full past as packed tokens, downsampled embeddings, merged KVs, or a constant recurrent/hierarchical state Zhang et al. (2025; 2026); Mao et al. (2025); Wu et al. (2026b); Hong et al. (2025); Zhu et al. (2026); Chandratreya et al. (2026); Samuel et al. (2026); Bian et al. (2026); Wu et al. (2026c). *Retrieval* stores past KVs, latents, or frames and reads a sparse subset by FOV/pose overlap, token affinity, or learned queries Xiao et al. (2025); Yu et al. (2025a); Sun et al. (2026); Yu et al. (2026a); Peng et al. (2026); Xu et al. (2026a); Wang et al. (2026d); Xu et al. (2026b); Hu et al. (2026); Cai et al. (2025); Chen et al. (2026b); Li et al. (2026b). *Explicit 3D/4D state* indexes history in geometry—surfels, point clouds, patches, or reconstructed caches—so revisit becomes a lookup Li et al. (2025b); Zhao et al. (2025); Huang et al. (2025a); Wu et al. (2025); Yu et al. (2026b); Wang et al. (2026e); Shen et al. (2026); AlayaWorld Team et al. (2026); Qian et al. (2026); Lin et al. (2026); Wang et al. (2026a). A remaining gap is robustness to dynamics.

3 Method

Let $\{I^{\mathrm{src}}[t]\}_{t=1}^{T}$ be a source video with estimated depths $\{D^{\mathrm{src}}[t]\}$ and cameras $\{C^{\mathrm{src}}[t]\}$. Given a target camera trajectory $\{C^{\mathrm{tgt}}[t]\}_{t=1}^{T}$, we synthesize a synchronized novel-view video $\{\hat{I}[t]\}_{t=1}^{T}$.

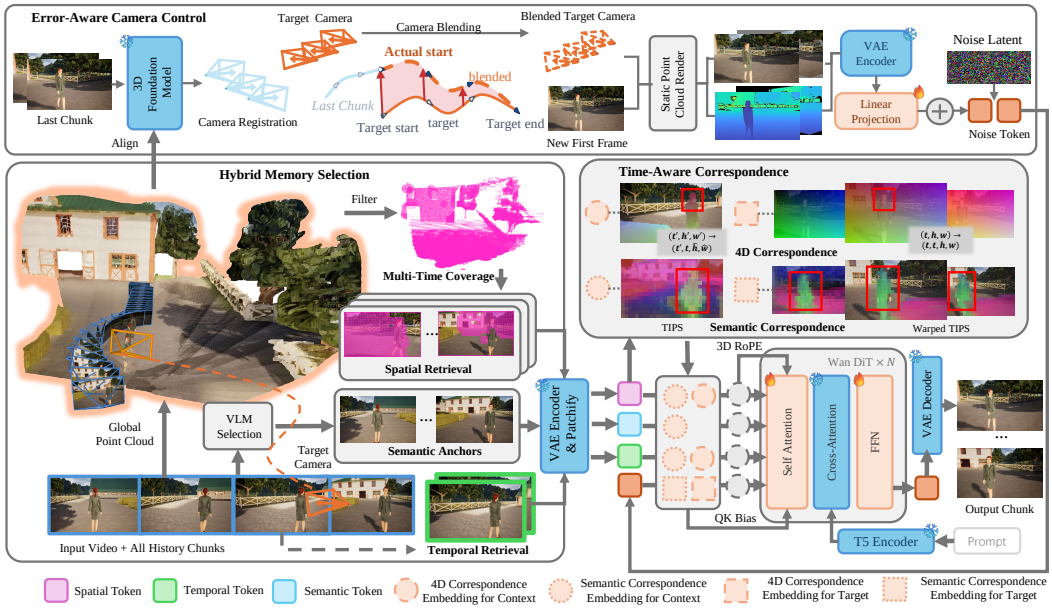

Figure 2: **Overview of LongReplay.** Given the source video and a target camera trajectory, we generate the novel-view video autoregressively in overlapping chunks. A dynamic memory is queried along time, space, and semantics. Retrieved patches are concatenated with noise tokens and equipped with a dynamic-aware correspondence. Each chunk is synthesized under dense RGB/depth warp control with an error-aware blended camera, then written back into the memory.

As shown in Fig. 2, we generate the target autoregressively in overlapping chunks of F frames. The i-th chunk covers an index set $\mathcal{T}_i$ of F consecutive frames, and consecutive chunks share a single overlapping frame. A dynamic memory $\mathcal{M}$ stores the source observations together with generated frames, depths, and cameras accumulated so far. We generate the chunk under explicit camera control (section 3.2), query $\mathcal{M}$ with hybrid memory selection (section 3.3), and encode dynamic-aware correspondence embeddings on the retrieved patches (section 3.4).

3.1 Preliminaries

Our study is built on the Wan 2.2 pretrained text/image-to-video foundation model Wan-AI (2025). This architecture comprises a 3D Variational Auto-Encoder (VAE) Kingma & Welling (2014) for latent-space mapping and a series of transformer blocks for sequence modeling. Each basic transformer block consists of a 3D spatiotemporal attention layer, a cross-attention layer, and a feed-forward network (FFN). The text prompt embedding $\boldsymbol{c}_{\text{text}}$ is obtained by a T5 encoder Raffel et al. (2020) and injected into the model through cross-attention. We adopt the Rectified Flow Lipman et al. (2023) framework to train the diffusion transformer, enabling the generation of a data sample $\boldsymbol{x}$ from an initial Gaussian sample $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Specifically, for a data point $\boldsymbol{x}$, we construct its noised version $\boldsymbol{x}_\sigma$ at noise level $\sigma \in [0, 1]$ as

$$\boldsymbol{x}_\sigma = (1 - \sigma)\boldsymbol{x} + \sigma\, \boldsymbol{z}. \tag{1}$$

The training loss is a simple mean squared error (MSE):

$$\mathcal{L}_{\text{RF}}(\theta) = \mathbb{E}_{\sigma, \boldsymbol{x}, \boldsymbol{z}} \left\| \boldsymbol{v}_\theta(\boldsymbol{x}_\sigma, \sigma, \boldsymbol{c}_I, \boldsymbol{c}_{\text{text}}) - (\boldsymbol{z} - \boldsymbol{x}) \right\|_2^2, \tag{2}$$

where the velocity $\boldsymbol{v}_\theta$ is parameterized by the network θ.

3.2 Autoregressive Camera-Controlled Video Reshooting

To ensure visual consistency between chunks, we formulate autoregressive generation as an image-to-video task. Chunk i is initialized with the overlapping frame: $I^{\text{src}}[1]$ for the first chunk and $\hat{I}[t_i]$

for subsequent chunks. We unproject this first frame with its associated camera and depth, then reproject it along the cameras of $\mathcal{T}_i$. The resulting warped RGB and depth videos, $\{\tilde{I}[t], \tilde{D}[t]\}_{t\in\mathcal{T}_i}$, are encoded by the VAE and linearly projected into the latent space. Adding these latents to the noisy video latents provides precise camera control with large freedom relative to the first frame, so the generated views can be distant from the input video.

After generating a chunk, we estimate $\{\hat{C}[t], \hat{D}[t]\}_{t\in\mathcal{T}_i}$ and align them to the world frame with a $\mathrm{Sim}(3)$ transform fitted to previously calibrated context frames. The aligned observations $(\hat{I}[t], \hat{D}[t], \hat{C}[t])$ are then appended to $\mathcal{M}$. Due to residual errors in depth estimation, $\hat{C}[t]$ cannot perfectly match $C^{\mathrm{tgt}}[t]$. If the next chunk were still conditioned on $\{C^{\mathrm{tgt}}[t]\}$, the first-frame warp would start from a camera inconsistent with $\hat{C}[t_{i+1}]$, and the discrepancy would accumulate over the rollout. To prevent this, we introduce an error-aware camera control mechanism, as shown in Fig. 2. Let $\hat{C}[t_i]$ be the estimated pose of the overlapping first frame. We form a single world-frame residual between $\hat{C}[t_i]$ and $C^{\mathrm{tgt}}[t_i]$, and add a linearly decaying copy of it to the original trajectory. The blended chunk therefore starts at $\hat{C}[t_i]$ and ends at the commanded destination $C^{\mathrm{tgt}}[t_i + F - 1]$, while preserving the relative motion of $\{C^{\mathrm{tgt}}[t]\}$.

3.3 Hybrid Memory Selection

The memory $\mathcal{M}$ stores source observations and previously generated chunks. Instead of attending to every stored frame, we select a compact context along three complementary dimensions.

Spatial retrieval. Spatial retrieval supplies scene evidence visible from the target cameras $\{C^{\mathrm{tgt}}[t]\}_{t\in\mathcal{T}_i}$. We unproject observations in $\mathcal{M}$ into a world-space point cloud and project it into each target camera, obtaining a visible set $\mathcal{P}[t] \subset \mathbb{R}^3$. The core design of spatial retrieval is to ensure that each 3D point in $\mathcal{P}[t]$ is covered by at least two context patches from different times. Since many stored views are redundant, we use a greedy algorithm to cover the visible set with the fewest views and the fewest patches. We cap the number of contributing views at F_S, and denote the selected patches for the target camera $C^{\mathrm{tgt}}[t]$ by $\mathcal{S}_{\mathrm{sp}}[t]$. Because spatiotemporal self-attention propagates information between frames, spatial patches are retrieved only on a fixed keyframe subset $\mathcal{K}_i \subset \mathcal{T}_i$ with $|\mathcal{K}_i| = 3$. Dynamic pixels are retained in the point cloud rather than manually removed, so the retrieved patches contain dynamic subjects from past or future observations. This design guarantees that each target patch is supported by multiple context patches from different times. Therefore, the video diffusion model can learn to identify dynamic parts by aggregating information from different times via the dynamic-aware correspondence in Sec. 3.4.

Temporal retrieval. To preserve the time-aligned dynamics from the source video, we directly select the synchronized source video chunk. The source frame at time t is related to the target frame of the same time index. Therefore, each target camera $C^{\mathrm{tgt}}[t]$ yields a temporally retrieved patch set, denoted by $\mathcal{S}_{\mathrm{tm}}[t]$.

Semantic anchors. Semantic anchors aim to preserve the overall semantic information of dynamic subjects from the source video. We use an off-the-shelf VLM to caption sampled video keyframes and an off-the-shelf LLM to select a fixed number F_M of keyframes as semantic anchors $\mathcal{S}_{\mathrm{sem}}$. These anchors stay fixed throughout the autoregressive reshooting process.

3.4 Dynamic-Aware Correspondence

The selected patches are concatenated token-wise with the noisy target tokens and processed by the video diffusion transformer (DiT). All memory patches retain their original coordinates through 3D RoPE, which presents them as an extended context video. We use normalized RoPE to keep frame indices in a bounded range. Noise tokens occupy $[0, F-1]$, temporal patches occupy $[F, 2F-1]$, semantic anchors occupy $[2F, 2F+F_M-1]$, and spatial patches occupy $[2F+F_M, 2F+F_M+F_S-1]$. Since the patches are sparsely sampled from the context, it is difficult for the DiT to learn correspondence between context and noise tokens from coordinates alone. We therefore introduce two explicit correspondence embeddings that are injected into queries and keys, while keeping values as visual evidence.

4D correspondence. For each context patch associated with target time t, we reproject it onto the target view $C[t]$ to establish geometric correspondence. Let a context patch reside at pixel (h', w')



Figure 3: **Qualitative comparison** under a large viewpoint change. The input video circles the subject clockwise, while the target trajectory goes counter-clockwise. The second row is the warped synchronized chunk, showing the target viewpoint and the position of the dynamic subject. The gray mask indicates the backface of the point cloud. LongReplay achieves precise camera control (aligned with warped RGB), synchronization with the source video (see yellow box), and long-term 3D consistency (see red box).

in source frame t'. Reprojecting it by its depth and camera into a target camera $C^{\mathrm{tgt}}[t]$ yields the target-view coordinate $[\hat{h}, \hat{w}, \hat{d}]^\top$, which, together with the target frame index t, gives a fractional $(t, \hat{h}, \hat{w})$ coordinate at patch resolution. For a target token at position (t, h, w), the embedding uses its own coordinate. This is similar to 3D-based memory Shen et al. (2026); Yu et al. (2026b); Qian et al. (2026). However, the context-token embedding encodes only the corresponding target time and does not encode its source time relative to the timeline of the current chunk. Furthermore, due to the necessity of normalized RoPE, the source frame index is omitted from RoPE injection. Therefore, we additionally inject the original frame indices of both the target token and the context token after normalizing them to $[0, 1]$. That is, for a context token (respectively, a target token), we encode this projected coordinate into a 4D correspondence embedding:

$$\mathbf{e}^{\mathrm{4d}}_{\mathrm{ctx}} = \mathrm{PosEmb}(Norm(t'), t, \hat{h}, \hat{w}), \quad \mathbf{e}^{\mathrm{4d}}_{\mathrm{tgt}} = \mathrm{PosEmb}(Norm(t), t, h, w). \tag{3}$$

where PosEmb is an absolute positional encoding. Tokens that refer to the same coordinate and time therefore receive similar embeddings. Tokens with the same projected coordinate but different source times remain distinguishable, allowing the transformer to aggregate observations while reasoning about dynamic changes. The projection is estimated from depth and is neither exact nor injective, so the embedding supplies a correspondence cue rather than a hard assignment. Since projection correspondence requires the context patch to have a target time index in the current chunk, only spatial and temporal patches ($\mathcal{S}_{\mathrm{sp}}[t]$, $\mathcal{S}_{\mathrm{tm}}[t]$) are injected with $\mathbf{e}^{\mathrm{4d}}_{\mathrm{ctx}}$. For semantic anchor patches ($\mathcal{S}_{\mathrm{sem}}$), we set $\mathbf{e}^{\mathrm{4d}}_{\mathrm{ctx}} = \mathbf{0}$.

Semantic correspondence. To build semantic relations for dynamics whose evidence may be distant in time and space, we extract TIPS descriptors Maninis et al. (2025); Cao et al. (2026) for each context frame t' and interpolate them to the patch grid, obtaining $\mathbf{f}^{\mathrm{tips}}[t']$. After large-scale text–image pretraining, TIPS features encode per-patch semantics that remain similar across viewpoints and time, providing correspondence beyond space and time. We convert each context patch's fea-

Table 1: Quantitative comparison of long-video reshooting. ↑: higher is better; ↓: lower is better.

	Revisit			Camera Accuracy				VBench				
Method	PSNR↑	SSIM↑	LPIPS↓	APE-t↓	APE-R↓	RPE-t↓	RPE-R↓	TF↑	MS↑	DD↑	AQ↑	IQ↑
InSpatio-World	14.73	0.515	0.573	0.133	42.28	3.63	0.615	0.951	<u>0.986</u>	1.00	<u>0.561</u>	0.647
Vista4D	<u>16.85</u>	<u>0.570</u>	<u>0.448</u>	0.125	58.13	8.15	1.850	0.950	0.986	1.00	0.543	0.656
World-in-World	13.37	0.474	0.609	<u>0.049</u>	<u>21.21</u>	<u>1.01</u>	<u>0.123</u>	0.949	0.981	1.00	0.552	**0.679**
CameraAnything	11.85	0.449	0.655	0.144	99.11	10.68	3.319	<u>0.952</u>	0.978	1.00	0.521	0.649
LongReplay	**17.73**	**0.596**	**0.387**	**0.014**	**3.02**	**0.64**	**0.062**	**0.956**	**0.991**	1.00	**0.572**	<u>0.660</u>

ture into a semantic embedding through a lightweight projection. For target tokens, we construct the semantic embedding from a warped TIPS feature: we reproject the synchronized source frame t into a point cloud, assign each point its feature from $\mathbf{f}^{\text{tips}}[t]$, and render onto the target view at patch resolution, yielding $\hat{\mathbf{f}}^{\text{tips}}[t]$.

$$\mathbf{e}^{\text{sem}}_{\text{ctx}} = \text{Proj}(\mathbf{f}^{\text{tips}}[t']), \quad \mathbf{e}^{\text{sem}}_{\text{tgt}} = \text{Proj}(\hat{\mathbf{f}}^{\text{tips}}[t]). \tag{4}$$

This warped TIPS feature provides time awareness from the synchronized video. For the first target frame (the I2V condition), we directly extract its TIPS feature instead of warping. Context and target patches depicting the same semantic content receive similar $\mathbf{e}^{\text{sem}}$, enabling identity matching even without 4D correspondence. Although a token whose region is unseen in the synchronized video has an empty warped TIPS feature, it can still refer to the corresponding semantic evidence. This is made possible by the information-transfer capacity of multilayer self-attention. As long as nearby patches of the unseen region share a similar semantic embedding, the corresponding semantic evidence is collected by these nearby patches and transferred to the unseen patch by 3D RoPE. The semantic embedding is injected into all context tokens, because $\mathbf{e}^{\text{sem}}_{\text{tgt}}$ provides a target-view layout after reprojection. This serves as a cue to maintain synchronization.

Injection into attention. For each token (context or target), let $\mathbf{h}$ denote its DiT hidden state, and let $\mathbf{e}^{\text{4d}}$, $\mathbf{e}^{\text{sem}}$ be its corresponding 4D and semantic embeddings computed above (i.e., the $_{\text{ctx}}$ or $_{\text{tgt}}$ variant as appropriate). The query and key incorporate these embeddings while the value retains only visual evidence:

$$Q = W_Q\big(\mathbf{h} + \mathbf{e}^{\text{4d}} + \mathbf{e}^{\text{sem}}\big), \quad K = W_K\big(\mathbf{h} + \mathbf{e}^{\text{4d}} + \mathbf{e}^{\text{sem}}\big), \quad V = W_V\,\mathbf{h}. \tag{5}$$

4 WorldReplay Dataset

Training the video reshooting model requires paired videos that share the same underlying dynamic scene but differ in camera trajectory. Although open-source datasets Bai et al. (2025); Chen et al. (2026a); Li et al. (2026c); Wang et al. (2026c) provide useful paired videos, their clips are typically short, often around 5 seconds, which is insufficient for long-horizon world modeling. We therefore build a scalable long multi-camera video dataset using Unreal Engine 5 Games (2022), named the WorldReplay Dataset. We first collect multiple 3D environments as "maps". Then, we place animated characters within these environments as the "main subjects" of the videos. In addition to on-the-spot animations as in MultiCamVideo Bai et al. (2025), we simulate characters randomly moving inside the scene along random navigation paths. Thus, large actor movements and long-range scene exploration can be simulated. Specifically, this feature is implemented with a NavMesh planner that computes collision-free paths. We then design multiple cameras facing the subjects and moving along random trajectories to simulate simultaneous shooting. Cameras follow a random state machine: (1) rotate around the character, (2) go somewhere else, (3) go back to the character, and (4) wait and stand by. The resulting camera trajectories are, therefore, diverse. During camera movements, we detect collisions within the scenes to prevent unnatural camera clipping. This allows us to render high-quality videos with synchronized cameras. In total, we collect 8K synchronized one-minute-long videos shot in more than 100 UE5 scenes with hundreds of different character and animation combinations. All videos are of cinematic quality with Lumen global illumination. The entire data production pipeline can be launched automatically on a headless Linux server with an NVIDIA RTX GPU, making it easy to scale up. A visualization of the dataset can be found in Fig. 6 and Fig. 7.

5 Experiments

Metrics and Dataset. For evaluation, we select 50 in-the-wild web videos from the royalty-free stock video website Pexels Pexels (2025) and 50 synthesized videos from Unreal Engine Games



Figure 4: **Qualitative comparison** on long-term consistency. Long-term semantic consistency (Left): after the subject goes out of sight and then comes back, LongReplay generates a consistent human avatar. Long-term 3D consistency (Right): LongReplay generates a consistent 3D environment, while baseline methods fail to do so. These cues cannot be retrieved from synchronized frames.

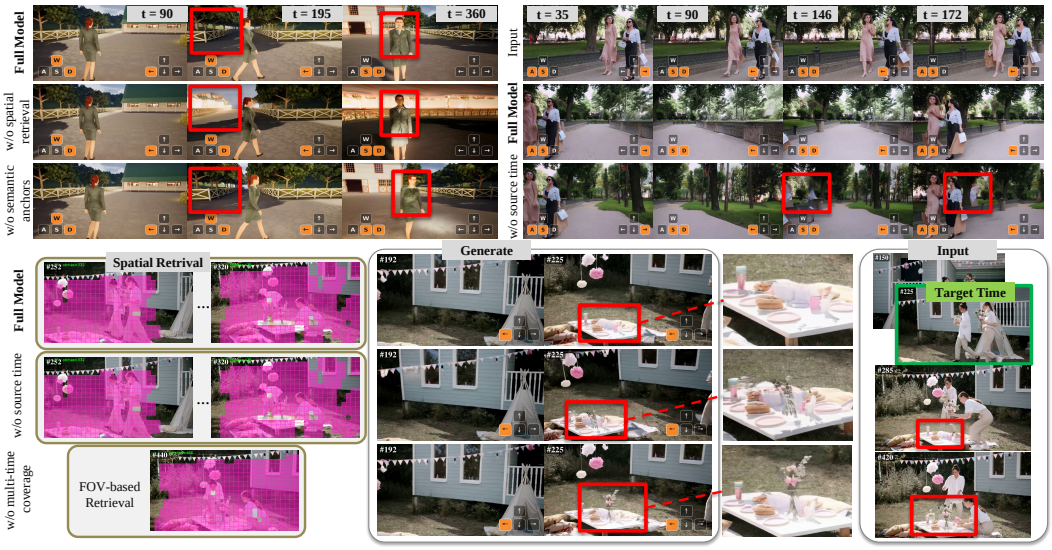

Figure 5: **Qualitative results** of ablation studies. (Top-left) Without $\mathcal{S}_{\text{sp}}[t]$, the 3D environment fails to stay consistent. Without $\mathcal{S}_{\text{sem}}$, the model fails to infer the appearance of the dynamic subject. (Top-right) Without 4D correspondence, the model fails to identify dynamics from old observations, producing ghosting artifacts. (Bottom) Without multi-time coverage, the model cannot reason about the state of dynamic objects, while our full model demonstrates dynamic awareness.

(2022) for ground-truth comparison. The input videos span 480–1,800 frames and include large motions of dynamic subjects. Target trajectories include revisits to assess whether previously observed content is preserved over long horizons. To evaluate camera trajectory accuracy, we estimate poses on the generated videos and compute absolute and relative pose error (APE/RPE) in translation and rotation. We also evaluate our method on VBench Huang et al. (2024) metrics for perceptual quality, including temporal flickering, motion smoothness, dynamic degree, aesthetic quality, and imaging

quality. We employ PSNR, SSIM, and LPIPS to measure revisit consistency against ground-truth videos from the synthesized dataset.

Baselines. We compare against current state-of-the-art video reshooting methods: Vista4D Lin et al. (2026), Inspatio-World Team et al. (2026), World-in-World Song et al. (2026a), and CameraAnything Li et al. (2026c). Other short-clip video reshooting methods Bai et al. (2025); Hu et al. (2025); Chen et al. (2026a); Yu et al. (2025b) require the first frame to be identical to the source video, limiting their ability to generate chunk by chunk. Therefore, we omit comparison with them. Please see the Appendix for more implementation details.

Camera Control and Synchronization. As shown in Table 1, our method strictly follows the target camera trajectories without drifting, achieving the lowest APE and RPE among all methods. We attribute this to the effectiveness of the first-frame warp condition and the error-aware camera control mechanism. Furthermore, as shown in Fig. 3, our method can reshoot a long video from an extreme novel viewpoint, even 180° opposite to the source video, while preserving synchronization of dynamic subjects. Baseline methods all fail to recover the backfaces of dynamic subjects, because they rely only on source-video concatenation or warped video to maintain synchronization, and therefore fail to reason about the correct correspondence across an extreme viewpoint difference from the input.

Long-range Consistency. As shown in Table 1, our method achieves the highest scores on revisit visual metrics. Fig. 4 shows further long-range consistency comparisons with baselines. Our method generates a 3D environment consistent with the input video across hundreds of frames. Vista4D Lin et al. (2026) cannot recover these details from a temporally persistent point cloud due to inevitable noise in the depth estimation. Owing to the effectiveness of the semantic anchors and semantic correspondence, LongReplay preserves the appearance of dynamic subjects that are unseen in the synchronized video, after the subject goes out of sight and then back into view. These results demonstrate that LongReplay achieves long-term consistency.

Ablation Studies. We carefully ablate key components of our design. The core of our design consists of multi-time coverage in spatial retrieval and the dynamic-aware 4D correspondence. The former guarantees that each 3D point inside the target view is supported by at least two context patches from different source indices, and the latter identifies the dynamic parts and reasons about which context patch should be consistent with the target patch. We replace our spatial retrieval with a naive FoV-based retrieval that retrieves only one context frame based on camera overlap. As shown in Fig. 5 (Bottom), without multi-time coverage, even our full model with dynamic-aware 4D correspondence cannot reason about whether the vase is on the table at the current time, and therefore can only generate the view based on the retrieved context frame. With the multi-time coverage mechanism of spatial retrieval, our model successfully identifies that the table should be empty at the current time. In contrast, the model (w/o source time) that replaces the 4D correspondence embedding with 3D correspondence as in Shen et al. (2026) can only blend multiple contexts and still generates a table with a vase. As shown in Fig. 5 (Top-right), this model also produces a ghosting effect at old observations of the dynamic subject. Compared with our full model, 4D correspondence helps the model understand dynamics and identify whether the current target token should be consistent with spatial-retrieval tokens from past or future observations. As shown in Fig. 5 (Top-left), to demonstrate the effectiveness of the memory selection mechanism, we remove context tokens from spatial retrieval. This causes an inconsistent 3D environment. Removing semantic anchors causes the dynamic subject to have a different identity when seen at different times and places. Our full model maintains long-term consistency. Please refer to the Appendix for more analysis.

6 Conclusion

To our knowledge, LongReplay is the first camera-controlled reshooting method that can generate minute-scale novel-view video while remaining synchronized with source dynamics under large camera motion and consistent with the source scene over long horizons. The key is a time–space–semantic readout of dynamic memory: hybrid selection retrieves the right evidence, and dynamic-aware correspondence tells the model how it relates to the target. WorldReplay dataset supplies the paired long takes needed to supervise this readout. Experiments confirm improved multi-view synchronization, including under extreme viewpoint changes, together with long-range geometric and identity consistency.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation and to edit the research paper to improve readability. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat datasets, support qualitative and thematic data analysis, interpret results, or assist in writing proofs. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Ruicong Liu, Xiaojie Xu, Xuangeng Chu, Zhen Li, Zhengyuan Lin, Zhixiang Wang, Zian Meng, and Zihui Gao. AlayaWorld: Long-Horizon and Playable Video World Generation. *arXiv:2607.06291*, 2026.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14834–14844, 2025.

Weikang Bian, Zhaoyang Huang, Xiaoyu Shi, Yijin Li, Fu-Yun Wang, and Hongsheng Li. Gs-dit: Advancing video generation with pseudo 4d gaussian fields through efficient dense 3d point tracking. *arXiv preprint arXiv:2501.02690*, 2025.

Yuxuan Bian, Zeyue Xue, Songchun Zhang, Shiyi Zhang, Weiyang Jin, Yaowei Li, Junhao Zhuang, Haoran Li, Jie Huang, Haoyang Huang, Nan Duan, and Qiang Xu. Echo-infinity: Learning evolving memory for real-time infinite video generation. *CoRR*, abs/2606.04527, 2026. doi: 10.48550/ARXIV.2606.04527. URL https://doi.org/10.48550/arXiv.2606.04527.

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan L. Yuille, Leonidas J. Guibas, Maneesh Agrawala, Lu Jiang, and Gordon Wetzstein. Mixture of contexts for long video generation. *CoRR*, abs/2508.21058, 2025. doi: 10.48550/ARXIV.2508.21058. URL https://doi.org/10.48550/arXiv.2508.21058.

Bingyi Cao, Koert Chen, Kevis-Kokitsi Maninis, Kaifeng Chen, Arjun Karpur, Ye Xia, Sahil Dua, Tanmaya Dabral, Guangxing Han, Bohyung Han, Joshua Ainslie, Alex Bewley, Mithun Jacob, Rene Wagner, Washington Ramos, Krzysztof Choromanski, Mojtaba Seyedhosseini, Howard Zhou, and Andre Araujo. TIPSv2: Advancing Vision-Language Pretraining with Enhanced Patch-Text Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Ishaan Preetam Chandratreya, David Charatan, Basile Van Hoorick, Sergey Zakharov, Vitor Guizilini, Phillip Isola, and Vincent Sitzmann. Millivid: Hierarchical latents for long-range consistency in video generation. *CoRR*, abs/2606.09056, 2026. doi: 10.48550/ARXIV.2606.09056. URL https://doi.org/10.48550/arXiv.2606.09056.

Dong-Yu Chen, Yixin Guo, Shuojin Yang, Tai-Jiang Mu, and Shi-Min Hu. Beyond inpainting: Unleash 3d understanding for stable camera-controlled video re-rendering. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 148–166, 2026a.

Kaijin Chen, Dingkang Liang, Xin Zhou, Yikang Ding, Xiaoqiang Liu, Pengfei Wan, and Xiang Bai. Out of sight but not out of mind: Hybrid memory for dynamic video world models. *CoRR*, abs/2603.25716, 2026b. doi: 10.48550/ARXIV.2603.25716. URL https://doi.org/10.48550/arXiv.2603.25716.

Xiao Fu, Shitao Tang, Min Shi, Xian Liu, Jinwei Gu, Ming-Yu Liu, Dahua Lin, and Chen-Hsuan Lin. Plenoptic video generation. *CoRR*, abs/2601.05239, 2026. doi: 10.48550/ARXIV.2601.05239. URL https://doi.org/10.48550/arXiv.2601.05239.

Epic Games. Unreal engine 5. https://www.unrealengine.com/en-US/unreal-engine-5, 2022. Accessed Feb 7, 2026.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite Worlds with Versatile Interactions. *arXiv:2607.07534*, 2026.

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, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv:2508.13009*, 2025.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, Kalyan Sunkavalli, Feng Liu, Zhengqi Li, and Hao Tan. RELIC: Interactive Video World Model with Long-Horizon Memory. *arXiv:2512.04040*, 2025.

Chen Hou and Zhibo Chen. Training-free camera control for video generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

JiaKui Hu, Jialun Liu, Liying Yang, Xinliang Zhang, Kaiwen Li, Shuang Zeng, Yuanwei Li, Haibin Huang, Chi Zhang, and Yanye Lu. Geometry-as-context: Modulating explicit 3d in scene-consistent video generation to geometry context. *CoRR*, abs/2602.21929, 2026. doi: 10.48550/ARXIV.2602.21929. URL https://doi.org/10.48550/arXiv.2602.21929.

Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh. *arXiv preprint arXiv:2506.05554*, 2025.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics*, 44(6):245:1–245:15, 2025a.

Zhening Huang, Hyeonho Jeong, Xuelin Chen, Yulia Gryaditskaya, Tuanfeng Y. Wang, Joan Lasenby, and Chun-Hao Huang. Spacetimepilot: Generative rendering of dynamic scenes across space and time. *CoRR*, abs/2512.25075, 2025b. doi: 10.48550/ARXIV.2512.25075. URL https://doi.org/10.48550/arXiv.2512.25075.

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.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *The 2nd International Conference on Learning Representations*, 2014.

Chunyang Li, Yuanbo Yang, Jiahao Shao, Hongyu Zhou, Katja Schwarz, and Yiyi Liao. Rerope: Repurposing rope for relative camera control. In Christopher P. Redmann, Justin Solomon, and Mirela Ben-Chen (eds.), *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference, SIGGRAPH Conference Papers 2026, Los Angeles, CA, USA, July 19-23, 2026*, pp. 64:1–64:11. ACM, 2026a. doi: 10.1145/3799902.3811226. URL https://doi.org/10.1145/3799902.3811226.

Jia Li, Han Yan, Yihang Chen, Siqi Li, Xibin Song, Yifu Wang, Jianfei Cai, Tien-Tsin Wong, and Pan Ji. I3DM: implicit 3d-aware memory retrieval and injection for consistent video scene generation. *CoRR*, abs/2603.23413, 2026b. doi: 10.48550/ARXIV.2603.23413. URL https://doi.org/10.48550/arXiv.2603.23413.

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:2506.17201*, 2025a.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. VMem: Consistent Interactive Video Scene Generation with Surfel-Indexed View Memory. In *ICCV*, 2025b.

Yixuan Li, Yanhong Zeng, Ka Leong Cheng, Jiayi Zhu, Hanlin Wang, Wen Wang, Yihao Meng, Hao Ouyang, Qiuyu Wang, Yue Yu, ZiDong Wang, Yiyuan Zhang, Yujun Shen, and Dahua Lin. Cameraanything: Refilming videos with arbitrary camera control. *CoRR*, abs/2607.24591, 2026c. doi: 10.48550/ARXIV.2607.24591. URL https://doi.org/10.48550/arXiv.2607.24591.

Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan D. Burgert, Yuancheng Xu, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, Paul Debevec, and Ning Yu. Vista4d: Video reshooting with 4d point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pp. 32671–32682. Computer Vision Foundation, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Lin_Vista4D_Video_Reshooting_with_4D_Point_Clouds_CVPR_2026_paper.html.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Jiwen Liu, Shujuan Li, Zhixue Fang, Xiaohan Li, Yan Zhou, Zijie Meng, Zhimin Zhang, Yawen Luo, Guoxin Zhang, Yu-Shen Liu, and Pengfei Wan. Omnidirector: General multi-shot camera cloning without cross-paired data. *CoRR*, abs/2606.13432, 2026a. doi: 10.48550/ARXIV.2606.13432. URL https://doi.org/10.48550/arXiv.2606.13432.

Jiwen Liu, Shujuan Li, Xiaohan Li, Zijie Meng, Xinyue Liu, Yulong Xu, Yan Zhou, and Guoxin Zhang. TARS: timestep-aware data scaling for 3d-free video re-shooting. *CoRR*, abs/2607.28261, 2026b. doi: 10.48550/ARXIV.2607.28261. URL https://doi.org/10.48550/arXiv.2607.28261.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015.

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 *SIGGRAPH Asia 2025 Conference Papers*, pp. 20:1–20:10, 2025.

Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, Dan Gnanapragasam, Mojtaba Seyedhosseini, Howard Zhou, and André Araujo. TIPS: Text-image pretraining with spatial awareness. In *International Conference on Learning Representations*, 2025.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An Interactive World Generation Model. *arXiv:2507.17744*, 2025.

Zhan Peng, Jie Ma, Huiqiang Sun, Chong Gao, Zhijie Xue, Zhiyu Pan, Zhiguo Cao, Jun Liang, and Jing Li. Compression and Retrieval: Implicit Memory Retrieval for Video World Models. *arXiv:2606.23105*, 2026.

Pexels. Pexels: Free stock photos & videos, 2025.

Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, Mengyin An, Haobo Zhang, Biao Jiang, Jiahua Wang, Haofeng Sun, Yang Liu, and Yangguang Li. Matrix-Game 3.5: Enhancing Real-Time Streaming Interactive World Models with Patch Memory. *arXiv:2608.29910*, 2026.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6121–6132, 2025.

Dvir Samuel, Issar Tzachor, Matan Levy, Micahel Green, Gal Chechik, and Rami Ben-Ari. Fast autoregressive video diffusion and world models with temporal cache compression and sparse attention. *CoRR*, abs/2602.01801, 2026. doi: 10.48550/ARXIV.2602.01801. URL https://doi.org/10.48550/arXiv.2602.01801.

István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3D human pose and shape estimation. In *Advances in Neural Information Processing Systems*, 2024.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, Sanja Fidler, Jiahui Huang, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra 2.0: Explorable generative 3d worlds. *CoRR*, abs/2604.13036, 2026. doi: 10.48550/ARXIV.2604.13036. URL https://doi.org/10.48550/arXiv.2604.13036.

Xuelun Shen, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, Cheng Wang, et al. Gim: Learning generalizable image matcher from internet videos. In *The Twelfth International Conference on Learning Representations*, 2024.

Chenxi Song, Yanming Yang, and Chi Zhang. World in world: Explore the world with world models, 2026a. URL https://arxiv.org/abs/2609.11548.

Chenxi Song, Yanming Yang, Tong Zhao, Ruibo Li, and Chi Zhang. Taming video models for 3d and 4d generation via zero-shot camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40352–40363, 2026b.

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. 2026.

InSpatio Team, Donghui Shen, Guofeng Zhang, Haomin Liu, Haoyu Ji, Hujun Bao, Hongjia Zhai, Jialin Liu, Jing Guo, Nan Wang, Siji Pan, Weihong Pan, Weijian Xie, Xianbin Liu, Xiaojun Xiang, Xiaoyu Zhang, Xinyu Chen, Yifu Wang, Yipeng Chen, Zhenzhou Fan, Zhewen Le, Zhichao Ye, and Ziqiang Zhao. Inspatio-world: A real-time 4d world simulator via spatiotemporal autoregressive modeling, 2026. URL https://arxiv.org/abs/2604.07209.

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

Wan-AI. Wan2.2-ti2v-5b: Text-to-video generation model. https://huggingface.co/Wan-AI/Wan2.2-TI2V-5B, 2025. Accessed Feb 7, 2026.

Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Qingyan Bai, Ka Leong Cheng, Yue Yu, Yixuan Li, Yihao Meng, Zichen Liu, Yanhong Zeng, Yujun Shen, and Qifeng Chen. Worlddirector: Building controllable world simulators with persistent dynamic memory. *CoRR*, abs/2607.02517, 2026a. doi: 10.48550/ARXIV.2607.02517. URL https://doi.org/10.48550/arXiv.2607.02517.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Qisen Wang, Yifan Zhao, Peisen Shen, Jialu Li, and Jia Li. Chronosobserver: Taming 4d world with hyperspace diffusion sampling. *CoRR*, abs/2512.01481, 2025a. doi: 10.48550/ARXIV.2512.01481. URL https://doi.org/10.48550/arXiv.2512.01481.

Yiming Wang, Qihang Zhang, Shengqu Cai, Tong Wu, Jan Ackermann, Zhengfei Kuang, Yang Zheng, Frano Rajic, Siyu Tang, and Gordon Wetzstein. Bullettime: Decoupled control of time and camera pose for video generation. *CoRR*, abs/2512.05076, 2025b. doi: 10.48550/ARXIV.2512.05076. URL https://doi.org/10.48550/arXiv.2512.05076.

Yukun Wang, Ruihuang Li, Jiale Tao, Shiyuan Yang, Liyi Chen, Zhantao Yang, Handz, Yulan Guo, Shuai Shao, and Qinglin Lu. Omnicamera: A unified framework for multi-task video generation with arbitrary camera control. *CoRR*, abs/2604.06010, 2026c. doi: 10.48550/ARXIV.2604.06010. URL https://doi.org/10.48550/arXiv.2604.06010.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, Yidan Xietian, Jiangbo Pei, Liang Hu, Boyi Jiang, Hua Xue, Zidong Wang, Haofeng Sun, Wei Li, Wanli Ouyang, Xianglong He, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-Game 3.0: Real-Time and Streaming Interactive World Model with Long-Horizon Memory. *arXiv:2604.08995*, 2026d.

Zun Wang, Jaemin Cho, Jialu Li, Han Lin, Jaehong Yoon, Yue Zhang, and Mohit Bansal. Epic: Efficient video camera control learning with precise anchor-video guidance. *arXiv preprint arXiv:2505.21876*, 2025c.

Zun Wang, Han Lin, Jaehong Yoon, Jaemin Cho, Yue Zhang, and Mohit Bansal. Anchorweave: World-consistent video generation with retrieved local spatial memories. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision - ECCV 2026 - 19th European Conference, Malmö, Sweden, September 8-12, 2026, Proceedings, Part LI*, volume 17051 of *Lecture Notes in Computer Science*, pp. 203–224. Springer, 2026e. doi: 10.1007/978-3-032-37356-4_12. URL https://doi.org/10.1007/978-3-032-37356-4_12.

Xuan Wei, Longbin Ji, Guan Wang, Xiangrui Liu, Zhenyu Zhang, Shuohuan Wang, Yu Sun, and Qingqi Hong. Memento: Reconstruct to remember for consistent long video generation, 2026. URL https://arxiv.org/abs/2606.14667.

Jiayi Wu, Haoming Cai, Cornelia Fermüller, Christopher A. Metzler, and Yiannis Aloimonos. Real2sam2real: Generative 3d caches as complementary context for video diffusion. *CoRR*, abs/2606.00299, 2026a. doi: 10.48550/ARXIV.2606.00299. URL https://doi.org/10.48550/arXiv.2606.00299.

Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, and Ming-Ming Cheng. Infinite-World: Scaling Interactive World Models to 1000-Frame Horizons via Pose-Free Hierarchical Memory. 2026b.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video World Models with Long-term Spatial Memory. In *NeurIPS*, 2025.

Xindi Wu, Sven Elflein, James Lucas, Olga Russakovsky, Laura Leal-Taixé, Despoina Paschalidou, Jonathan Lorraine, and Aljosa Osep. Addressable memory for video world models. *CoRR*, abs/2608.07408, 2026c. doi: 10.48550/ARXIV.2608.07408. URL https://doi.org/10.48550/arXiv.2608.07408.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term Consistent World Simulation with Memory. In *NeurIPS*, 2025.

Jiacong Xu, Hanwen Jiang, Zhixin Shu, Kalyan Sunkavalli, Vishal M. Patel, and Yiqun Mei. Wonder: Video world model done better. *CoRR*, abs/2607.26037, 2026a. doi: 10.48550/ARXIV.2607.26037. URL https://doi.org/10.48550/arXiv.2607.26037.

Tian-Xing Xu, Zi-Xuan Wang, Guangyuan Wang, Li Hu, Zhongyi Zhang, Peng Zhang, Bang Zhang, and Song-Hai Zhang. UCM: Unified Modeling of Camera Control and Memory with Time-aware Positional Encoding Warping for World Models. 2026b.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. *CoRR*, abs/2509.22622, 2025. doi: 10.48550/ARXIV.2509.22622. URL https://doi.org/10.48550/arXiv.2509.22622.

Ying Yang, Zhengyao Lv, Tianlin Pan, Haofan Wang, Binxin Yang, Hubery Yin, Chen Li, Ziwei Liu, and Chenyang Si. Stableworld: Towards stable and consistent long interactive video generation. *CoRR*, abs/2601.15281, 2026a. doi: 10.48550/ARXIV.2601.15281. URL https://doi.org/10.48550/arXiv.2601.15281.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pp. 40340–40351. Computer Vision Foundation, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yang_NeoVerse_Enhancing_4D_World_Model_with_in-the-wild_Monocular_Videos_CVPR_2026_paper.html.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In Taku Komura, Michael Wimmer, and Hongbo Fu (eds.), *Proceedings of the SIGGRAPH Asia 2025 Conference Papers, SA Conference Papers 2025, Hong Kong, December 15-18, 2025*, pp. 19:1–19:11. ACM, 2025a. doi: 10.1145/3757377.3763833. URL https://doi.org/10.1145/3757377.3763833.

Jiwen Yu, Jianxiong Gao, Jianhong Bai, Yiran Qin, Kaiyi Huang, Quande Liu, Xintao Wang, Pengfei Wan, Kun Gai, and Xihui Liu. MemLearner: Learning to Query Context Memory for Video World Models. In *ECCV*, 2026a.

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*, pp. 100–111, 2025b.

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. *arXiv preprint arXiv:2409.02048*, 2024.

Wei Yu, Runjia Qian, Yumeng Li, Liquan Wang, Songheng Yin, Sri Siddarth Chakaravarthy P, Dennis Anthony, Yang Ye, Yidi Li, Weiwei Wan, and Animesh Garg. Mosaicmem: Hybrid spatial memory for controllable video world models. *CoRR*, abs/2603.17117, 2026b. doi: 10.48550/ARXIV.2603.17117. URL https://doi.org/10.48550/arXiv.2603.17117.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame Context Packing and Drift Prevention in Next-Frame-Prediction Video Diffusion Models. In *NeurIPS*, 2025.

Lvmin Zhang, Shengqu Cai, Muyang Li, Chong Zeng, Beijia Lu, Anyi Rao, Song Han, Gordon Wetzstein, and Maneesh Agrawala. Tinyhistory: Lightweight video history embeddings via two-stage context learning, 2026. URL https://arxiv.org/abs/2512.23851.

Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video generation with updatable spatial memory. *CoRR*, abs/2512.15716, 2025. doi: 10.48550/ARXIV.2512.15716. URL https://doi.org/10.48550/arXiv.2512.15716.

Haiyang Zhou, Wangbo Yu, Chaoran Feng, Xunyu Zhou, Yonghong Tian, and Li Yuan. Uniworldview: Large-baseline view synthesis via video diffusion models. *CoRR*, abs/2608.04701, 2026. doi: 10.48550/ARXIV.2608.04701. URL https://doi.org/10.48550/arXiv.2608.04701.

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. *CoRR*, abs/2605.15178, 2026. doi: 10.48550/ARXIV.2605.15178. URL https://doi.org/10.48550/arXiv.2605.15178.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jie Zhou, and Jiwen Lu. Astra: General interactive world model with autoregressive denoising. *CoRR*, abs/2512.08931, 2025. doi: 10.48550/ARXIV.2512.08931. URL https://doi.org/10.48550/arXiv.2512.08931.

APPENDIX



Figure 6: **Overview of the WorldReplay dataset.** We visualize key features of our dataset. We record multi-camera synchronized long videos and simulate dynamic subject movement along large trajectories (Left), together with various in-place animations (Right). We randomly combine various actors and environments into clips of cinematic quality (Center).

A IMPLEMENTATION DETAILS

Training. Our approach builds upon the Wan2.2-TI2V-5B model Wan et al. (2025); Wan-AI (2025). We fine-tune its pretrained DiT weights while freezing the cross-attention layers and patch embedding layers, and introduce a camera-control adapter, a TIPS projection, and a query–key bias for correspondence injection. The input videos are resized to a resolution of 576×1024 with chunk length $F = 49$. We randomly construct 40k training pairs from our WorldReplay dataset. We sample a random chunk from a randomly selected video A and take a synchronized video B as the input video. The prefix of the selected chunk of video A and video B therefore constitutes the entire history context stored in $\mathcal{M}$. Chunk starting times t_i are sampled with a FoV-coverage-weighted scheme that prefers target windows poorly covered by $\{C^{\mathrm{src}}[t]\}$. Following Sec. 3.3, we perform spatial retrieval on the keyframe subset $\mathcal{K}_i \subset \mathcal{T}_i$ with $|\mathcal{K}_i| = 3$ and temporal retrieval of $\{\mathcal{S}_{\mathrm{tm}}[t]\}_{t\in\mathcal{T}_i}$, using rendered depth from Unreal Engine, and randomly select $F_M = 8$ frames as semantic anchors $\mathcal{S}_{\mathrm{sem}}$ instead of using VLM selection. Retrieved patches $\mathcal{S}_{\mathrm{sp}}$, $\mathcal{S}_{\mathrm{tm}}$, and $\mathcal{S}_{\mathrm{sem}}$ are concatenated with noise tokens, equipped with 4D projection and TIPS correspondences, and trained with first-frame RGB/depth warp control $\{\tilde{I}[t], \tilde{D}[t]\}_{t\in\mathcal{T}_i}$ together with warped TIPS features of the synchronized source chunk. Trainable parameters are optimized with AdamW at learning rate 1×10^{-5} and weight decay 0.01, mixed-precision bf16, per-GPU batch size 1, and gradient clipping 1.0. Our method is implemented on 8 NVIDIA A800 GPUs with DeepSpeed ZeRO-2. Training takes 4 days.

Inference. We generate $\{\hat{I}[t]\}_{t=1}^{T}$ autoregressively in overlapping chunks of $F = 49$ frames with a single-frame overlap, following Sec. 3.2. We adopt VGGT-Omega Wang et al. (2026b) to estimate $\{D^{\mathrm{src}}[t]\}$ and $\{C^{\mathrm{src}}[t]\}$ for the input video, processing the source in windows of 500 frames with 100-frame overlap and aligning consecutive windows. A global point cloud for spatial retrieval is built after a spatiotemporal downsample of stride $4\times16\times16$ (time×height×width). Each chunk $\mathcal{T}_i$ is synthesized as image-to-video from the overlapping first frame, under dense RGB/depth warp control, temporal retrieval $\{\mathcal{S}_{\mathrm{tm}}[t]\}_{t\in\mathcal{T}_i}$, spatial retrieval $\{\mathcal{S}_{\mathrm{sp}}[t]\}_{t\in\mathcal{K}_i}$ with $|\mathcal{K}_i| = 3$, and semantic anchors $\mathcal{S}_{\mathrm{sem}}$ of $F_M = 8$ that stay fixed across the rollout. We adopt Qwen3.8-27BQwen Team (2026) as VLM to select semantic anchors. We disable classifier-free guidance and sample with 50 flow-matching denoising steps (shift 5). After each chunk, VGGT-Omega is run jointly on all newly generated frames together with 60 previously calibrated context frames. The estimated cameras $\{\hat{C}[t]\}_{t\in\mathcal{T}_i}$ and depths $\{\hat{D}[t]\}_{t\in\mathcal{T}_i}$ are aligned to the world frame by a $\mathrm{Sim}(3)$ transform fitted on depth-projected correspondences, then written back into $\mathcal{M}$. Error-aware camera blending (Sec. 3.2) is applied only to the warp condition of the next chunk, while retrieval still queries $\{C^{\mathrm{tgt}}[t]\}$. Note that we still use the target camera $\{C^{\mathrm{tgt}}[t]\}$ instead of the blended target camera in Sec. 3.2 to calculate the projected coordinate in 4D correspondence and the warped TIPS in semantic correspondence, so that the projection stays stable between chunks. Since these two correspondence embeddings also provide a rough signal for camera control, using the target camera as the projection also helps to correct the current camera back onto its intended track. Inference generates each chunk in approximately 150 seconds.

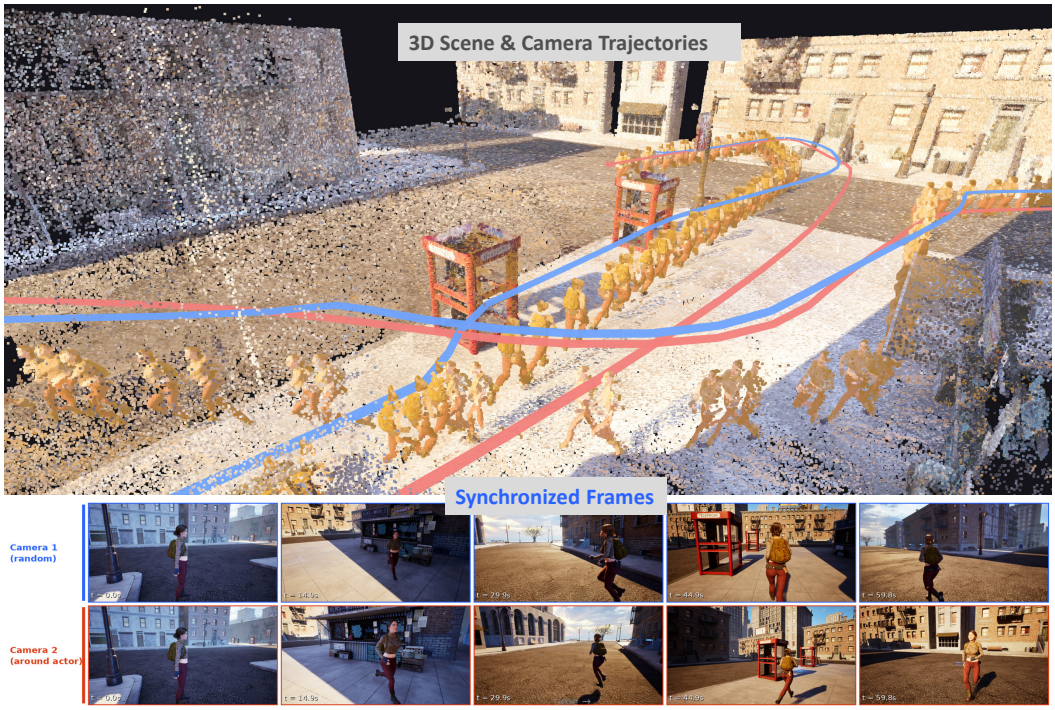

Figure 7: **Visualization** of the WorldReplay dataset. We render synchronized one-minute-long videos shot in cinematic quality with Lumen global illumination.

Baselines. Vista4D Lin et al. (2026) and Inspatio-World Team et al. (2026) use dynamic point-cloud rendering and source-video concatenation to maintain synchronization. Vista4D additionally accumulates a static point cloud to maintain long-range multi-view consistency. We adopt its open-source T2V checkpoints for chunk-by-chunk generation. Inspatio-World takes only the last chunk as its local window memory. World-in-World Song et al. (2026a) maintains multi-view synchronization only by warping the synchronized video, and its long-term 3D memory injection retrieves only from generated history rather than from the source video. CameraAnything Li et al. (2026c) is designed for short-clip video reshooting with implicit camera conditioning, and maintains multi-view synchronization only by source-video concatenation. We directly extend it to generate chunk by chunk. Wonder Xu et al. (2026a) serves as a strong baseline, though it is not open-source. It achieves long-horizon consistency by retrieving key–value (KV) states from context frames. Although this mechanism provides a solid foundation, its pooled query–key similarity relies solely on appearance features and camera embeddings. As a result, it fails to establish correspondences that are as precise and robust as ours. Indeed, minor 3D inconsistencies are still observable in its demo videos.

Training Dataset. As shown in Fig. 6, we construct our WorldReplay Dataset, a scalable, long-duration, multi-camera video dataset built on Unreal Engine 5. Featuring 8K+ synchronized 1-minute videos across 100+ UE5 environments with hundreds of character and motion combinations. Powered by NavMesh collision-free motion and an intelligent multi-camera state machine for long-range scene exploration and cinematic shot diversity—all rendered with Lumen global illumination. We provide a visualization of our WorldReplay dataset, as shown in Fig. 7. The dataset and the data generation pipeline will be open source.

B More Ablation Studies

We provide quantitative results of our ablation studies in Tab. 2. The table reports revisit photometric scores together with the multi-view synchronization metrics of Sec. C, demonstrating the importance of our proposed correspondence injection.

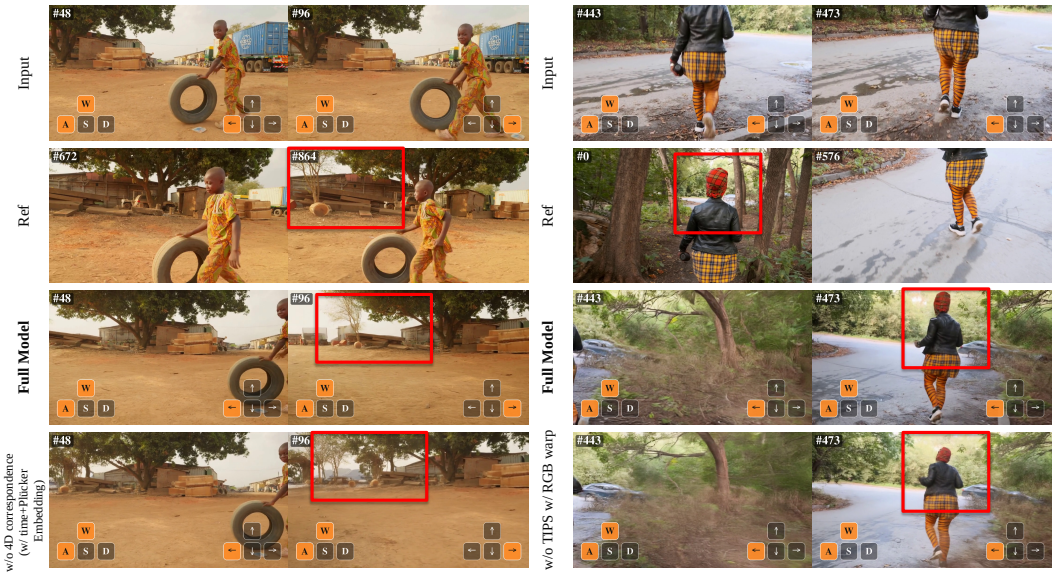

Figure 8: **Qualitative results** of an ablation study on patch-level correspondence.

Table 2: Ablation of revisit consistency and dynamic synchronization. Revisit metrics are means over synthetic videos with ground-truth views. Synchronization metrics follow Table 3. Best results are bold and second-best are underlined.

	Revisit			Cross-View Matching		Human Pose Synchronization			
Method	PSNR↑	SSIM↑	LPIPS↓	FG PCK↑	FG recall↑	Orient.↓	Action↓	$PCK_{0.15}$↑	Joint suc.↑
(a) w/o 4D correspondence (w Plücker)	17.34	0.582	0.393	99.01	<u>40.38</u>	9.3	8.8	97.0	95.0
(b) w/o warped TIPS for target	16.94	0.586	0.407	57.94	19.89	12.5	9.8	90.9	88.6
(c) w/o TIPS for ($\mathcal{S}_{\text{sp}}[t]$, $\mathcal{S}_{\text{tm}}[t]$)	<u>17.68</u>	**0.601**	**0.383**	97.86	39.44	8.6	<u>8.2</u>	98.2	97.1
(d) w/o TIPS correspondence	16.60	0.583	0.425	63.40	19.16	13.2	10.1	90.7	86.4
(e) w/o TIPS, w/ RGB warp	17.41	<u>0.600</u>	0.406	98.48	39.58	9.0	9.1	97.1	94.6
Full Model	**17.73**	0.596	<u>0.387</u>	<u>99.33</u>	**40.83**	<u>6.5</u>	7.4	<u>98.6</u>	**98.8**

We replace the 4D correspondence embedding with a time-aware camera embedding, that is, using Plücker embeddings of target and context patches in place of the projection-coordinate embedding (Model (a) in Tab. 2). As shown in Fig. 8 (Left), this model cannot recover the exact scene details from a context view at a large distance, because the camera embedding only provides a rough geometric correspondence. In contrast, the projection-coordinate embedding provides accurate patch-level correspondence, enabling our full model to achieve long-term 3D consistency.

As shown in Tab. 2, we remove the warped TIPS as the semantic embedding for target tokens (Model (b)). Model (b) fails to reason about the correct position of the dynamic subject, with low FG PCK and FG recall. This demonstrates that the warped TIPS serves as a strong cue for the model to maintain view synchronization. Removing the TIPS semantic embedding for context tokens from spatial retrieval $\mathcal{S}_{\text{sp}}[t]$ and temporal retrieval $\mathcal{S}_{\text{tm}}[t]$ causes a slight drop in performance on cross-view matching and human-pose synchronization. Removing the TIPS semantic embedding for all tokens (Model (d)) yields similar synchronization performance to Model (b).

As shown in Fig. 8 (Right), we remove the TIPS embedding of semantic correspondence while using a similar warped synchronized RGB video to maintain multi-view synchronization (Model (e) in Tab. 2), similar to Lin et al. (2026); Yu et al. (2025b); Chen et al. (2026a); Hu et al. (2025). This forces the model to build correspondence between the target patch and the semantic anchors from the original appearance latents of the VDM. This model can only roughly recover the red hat of the woman, while our full model can faithfully recover its details through strong attention. As shown in Tab. 2, Model (e) can recover the position of dynamic subjects with high FG PCK and FG recall, but fails to preserve detailed dynamic motions, especially orientation.

C Additional Evaluation of Multi-View Synchronization for Dynamic Subjects

We follow ReCamMaster Bai et al. (2025) to evaluate view synchronization with the source video. However, directly applying the state-of-the-art image matching method GIM Shen et al. (2024) fails to capture the true synchronization quality under large-baseline camera viewpoints. We therefore calculate valid correspondences only on the foreground. To assess precise motion synchronization for human-centric scenes, we additionally evaluate human-pose consistency. A correct camera change can reveal a person's back and remove appearance matches to the source. We therefore complement foreground correspondence with camera-conditioned human orientation and pose.

Metrics. We use SuperPoint and LightGlue (Lindenberger et al., 2023) with GIM weights (Shen et al., 2024) at 640×360. Matching uses full images; scoring retains source foreground points after 5-pixel mask erosion and excludes points whose reliable surface normal is at least 60° from the direction toward the target camera. A match is counted as correct if it lies within 5% of the image diagonal of its depth-and-camera projection. FG PCK and recall divide correct matches by matched and eligible points, respectively, pooling counts within each case.

NLF-L (Sárándi & Pons-Moll, 2024) recovers SMPL bodies (Loper et al., 2015) independently in source and generated frames. Orientation error compares chest-forward directions after the known source-to-target camera rotation; action error compares limb directions in body-root coordinates. 3D PCK counts up to 14 reliable major joints within 0.15 of anatomical scale after pelvis centering and camera rotation, without free rotation alignment. Scale is pelvis-to-neck length plus mean hip–knee–ankle chain length. Joint success requires both angular errors to be at most 30°. Source-only reliability and visibility checks define the same eligible set for every method; multi-person assignment uses torso color. Missing targets receive zero PCK and success; angular errors use recovered targets. People are averaged within frames before case-level averaging.

Table 3: View Synchronization. PCK, recall, and success are percentages; angular errors are in degrees. Best results are bold and second-best are underlined.

	Cross-View Matching		Human Pose Synchronization			
Method	FG PCK↑	FG recall↑	Orient. error↓	Action error↓	$PCK_{0.15}$↑	Joint success↑
CameraAnything	14.33	7.26	<u>12.2</u>	9.8	84.7	81.3
InSpatio-World	<u>96.76</u>	32.25	18.7	9.2	91.9	83.1
Vista4D	92.00	<u>35.50</u>	12.3	<u>8.5</u>	<u>95.7</u>	<u>91.9</u>
World-in-World	90.56	26.07	15.8	19.5	85.3	74.8
LongReplay	**99.34**	**40.39**	**6.7**	**7.7**	**99.0**	**99.0**

Results. LongReplay leads across all reported metrics (Table 3). For reference-defined viewpoint changes of at least 60°, it achieves 86.7% joint success versus Vista4D's 49.7% (+37.0 percentage points), with orientation error reduced from 45.2° to 13.7° (Table 4). Figure 9 shows consistency over video progress and angular tolerance; Fig. 10 illustrates correct target-view body orientation. These automatic metrics assess same-frame geometric and body consistency; they do not directly estimate event-level time lag.

Table 4: Large viewpoint changes ($\geq 60^\circ$), defined from the source body and requested cameras. The success criterion remains $30^\circ/30^\circ$.

Method	Orient. error (°)↓	Action error (°)↓	3D $PCK_{0.15}$ (%)↑	Joint success (%)↑
CameraAnything	41.2	20.2	71.5	51.7
InSpatio-World	59.6	15.6	65.8	24.8
Vista4D	45.2	<u>15.1</u>	74.1	49.7
World-in-World	<u>16.8</u>	21.1	<u>78.9</u>	<u>60.3</u>
LongReplay	**13.7**	**12.2**	**94.4**	**86.7**

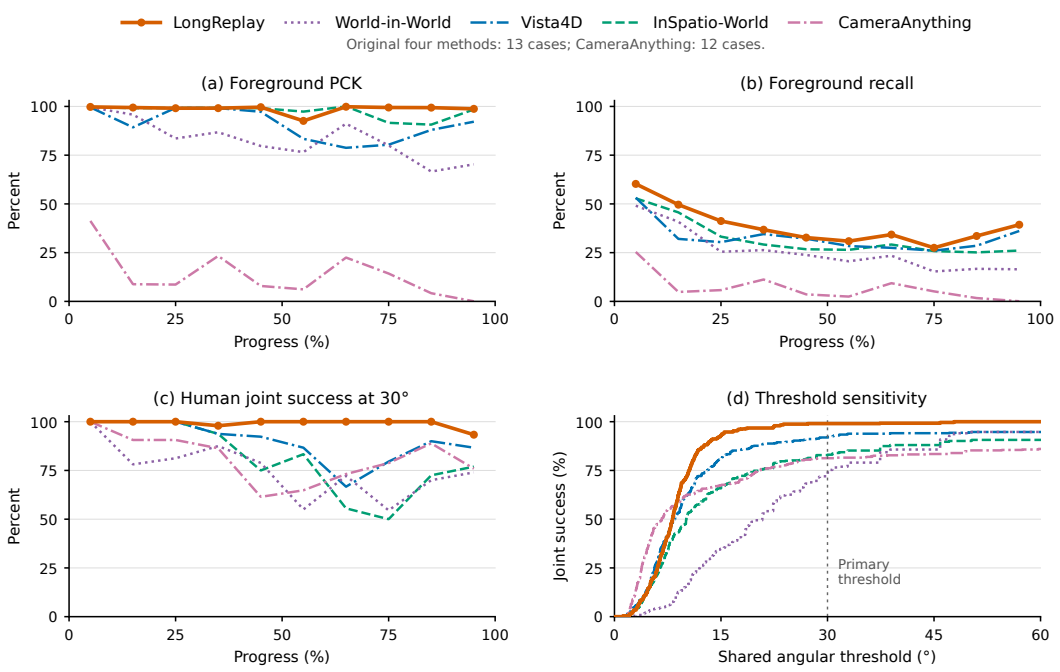

Figure 9: **Consistency over video progress and angular tolerance.** (a–c) Case-balanced scores in ten progress bins; eligible cases are shared across methods but can vary across bins. (d) Post-hoc sensitivity to a common orientation/action threshold on the same 361 person–frames; the dashed line marks the primary 30° threshold.

D LIMITATIONS

Semantic anchors are extracted only from the input source video and are not a learnable module. This design limits the model's ability to maintain cross-chunk semantic consistency for newly observed parts of dynamic subjects during generation. However, in most real-world video reshooting tasks, the semantic information is fully provided in the source video. We leave a learnable semantic-anchor selection mechanism as future work.

Our method is built on Wan2.2-TI2V-5B, whose generative capacity remains limited. Fine-grained identity cues such as faces, hands, and clothing textures can drift across chunks, even when the corresponding semantic evidence is present in $\mathcal{S}_{\text{sem}}$. When the target trajectory reveals regions that were never observed in the source video, both 3D consistency and visual quality degrade: the model must hallucinate unseen geometry rather than retrieve it from $\mathcal{M}$, and the resulting novel-view content is less stable than regions that remain in the source FoV. A stronger backbone would alleviate these failure modes; we leave this to future work.

Current inference is also not interactive. Each $F{=}49$ chunk takes about 150 seconds with 50 denoising steps, plus VAE encoding/decoding and VGGT-Omega calibration after every chunk. This latency precludes real-time, user-in-the-loop camera exploration. Reducing the number of sampling steps, using a faster decoder, and amortizing pose estimation are promising directions toward interactive long-video reshooting.

E TRAIN-TEST GAP

We trained our video diffusion model solely on synthetic data without observing generalization issues—an observation consistent with prior workBai et al. (2025); Luo et al. (2025). The main reason is that video reshooting conditions generation on an input video. Leveraging the video model's inherent generalization capability, the model adapts to different datasets naturally: having learned what synthetic data entails during training, it generates realistic content when conditioned on real video



Figure 10: **Camera-conditioned body orientation.** Gold: source skeleton rotated into the target camera; blue: body recovered from the target image. Skeletons are pelvis-centered and scale-normalized, with the same display view and no free rotation alignment. Arrows and labels show chest-forward directions and their angular errors. Actor crops are independent; the Source column shows only the expected target skeleton. LongReplay follows the requested facing direction in both examples.

at inference. Furthermore, the training objective of video rerendering mandates content consistency between input and output, which does not introduce dataset-specific biases. That said, incorporating high-quality real-world video during training could yield further quality improvements, particularly for fine details like faces and hands that are underrepresented in synthetic datasets.

F Additional Qualitative Results

We provide additional qualitative comparisons against long-video reshooting baselines on in-the-wild videos in Figs. 11–13 and on synthetic videos in Figs. 14.



Figure 11: Additional qualitative comparison on an in-the-wild video.



Figure 12: Additional qualitative comparison on an in-the-wild video.



Figure 13: Additional qualitative comparison on an in-the-wild video.

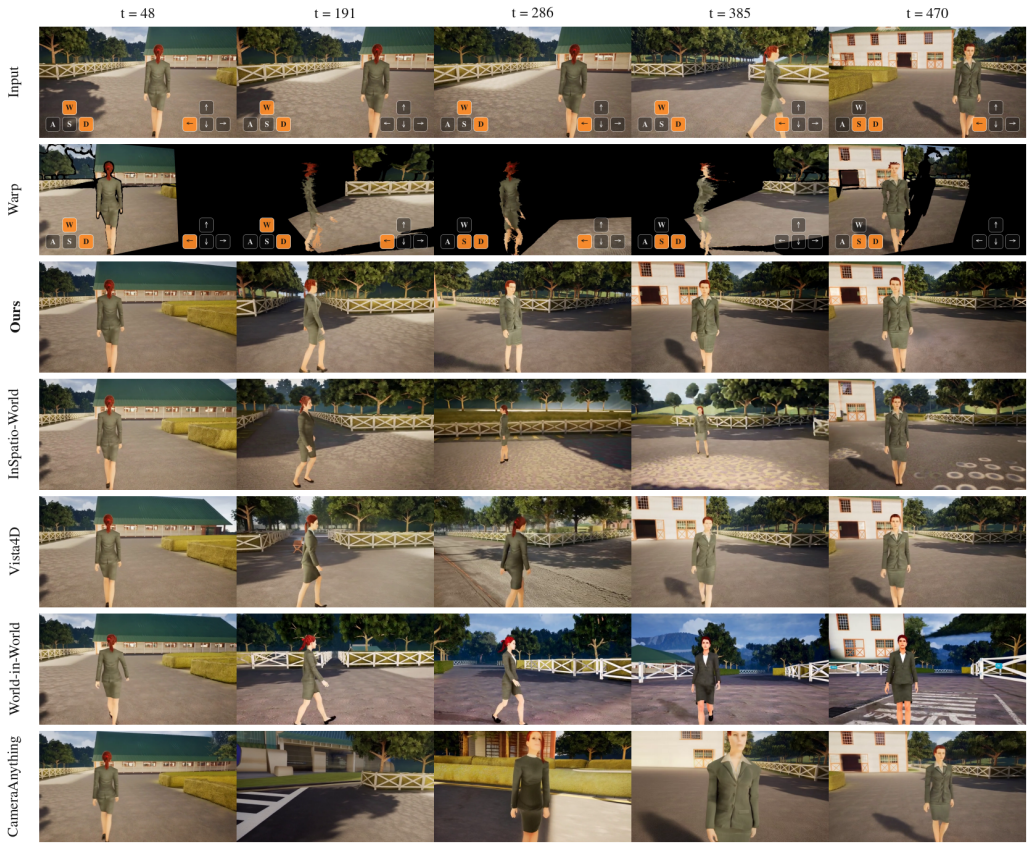

Figure 14: Additional qualitative comparison on a synthetic video.