CLOSING THE LOOP: TRAINING-FREE REVISIT CONSISTENCY FOR AUTOREGRESSIVE GENERATIVE RENDERING

Anonymous authors
Paper under double-blind review

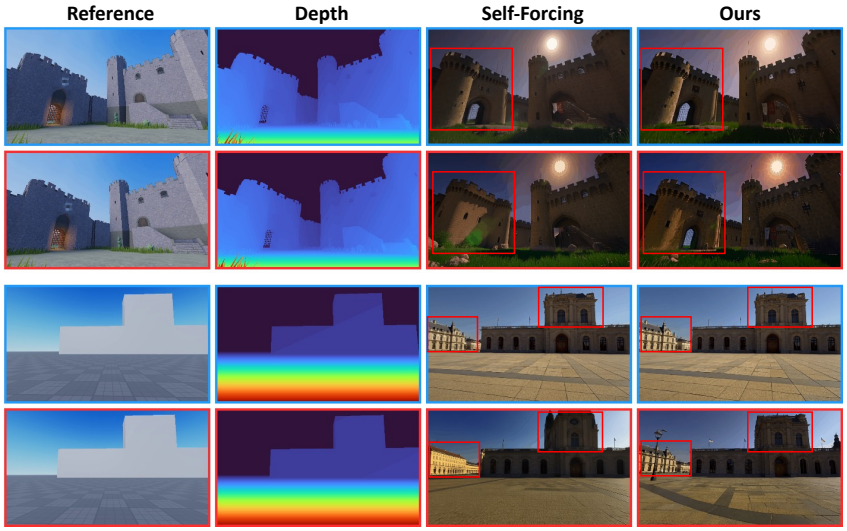


Figure 1: **Revisit-consistent autoregressive generative rendering.** Each pair of rows shows a revisit clip: conditioned on depth streamed from a 3D engine, an autoregressive video generator renders a viewpoint at the first visit (blue) and again when the camera returns (red), far beyond its bounded KV-cache horizon. Our method re-renders the same structures on return.

ABSTRACT

Recent conditional video generation models have shown promising potential to transform 3D engine renderings, such as depth maps and untextured geometry, into photorealistic videos for diverse gaming and immersive content creation. These applications require long-horizon autoregressive generation that continuously synthesizes new frames while preserving a persistent 3D world. Autoregressive generators synthesize video chunk by chunk with a bounded KV cache, so when the camera revisits a location after its context has been evicted from the KV cache, the model often regenerates inconsistent appearance, even though the conditioning renderings (e.g., depth) remain perfectly aligned since they originate from the same underlying 3D geometry. We address this revisit inconsistency without any post-training by exploiting correspondences the 3D engine already provides: temporal correspondence retrieves pose-matched historical latent chunks into the KV cache as loop-closure memory, while spatial correspondence from camera pose and depth reprojection biases token-level attention toward geometrically corresponding regions of the retrieved chunks. We demonstrate our method on loop-closure trajectories mined from the TartanAir and TartanGround datasets to mirror complex real-world application scenarios, where it outperforms existing training-free baselines on revisit consistency without losing video quality.

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

Generative rendering aims to lift low-fidelity outputs from 3D engines into realistic visual content. Given engine-provided conditions such as depth maps, untextured geometry, semantic buffers, or coarse simulations, a generative model can synthesize photorealistic images or videos while preserving the controllability of the underlying graphics pipeline. This capability is especially attractive for gaming, virtual production, simulation, and immersive content creation, where users require precise control over camera motion, scene layout, object trajectories, and interactions, while also seeking visual realism that would otherwise demand costly asset authoring, material design, lighting, and rendering in traditional graphics pipelines.

Prior works (Wang et al., 2018a; Mallya et al., 2020; Cai et al., 2024; Gomez-Nogales et al., 2026; Cohen-Bar et al., 2026; Abu Alhaija et al., 2025; NVIDIA, 2025; Jiang et al., 2025; fal.ai & Lightricks, 2026) have explored this direction by combining 3D or simulator-derived conditions with image- or video-generation models (Isola et al., 2017; Wang et al., 2018b; Rombach et al., 2022; Peebles & Xie, 2023; Zhang et al., 2023; Wan Team, 2025; HaCohen et al., 2024). These methods demonstrate the promise of generative rendering, but they are primarily designed for short or offline video clips, where all historical context remains within the generation window or can be propagated through an offline pipeline.

However, many practical applications demand long-horizon streaming generation. In an interactive game or virtual environment, for instance, the camera moves continuously through a large scene, leaves a region, and revisits it much later; the generator must synthesize frames chunk by chunk while maintaining a persistent world over horizons far exceeding the native context window of the base video model. Recent distillation methods for autoregressive video diffusion make such streaming feasible through causal attention and a bounded KV cache (Chen et al., 2024; Yin et al., 2025; Huang et al., 2025; Yang et al., 2025; Zhu et al., 2026; Zhao et al., 2026). Yet bounded memory introduces a new failure mode for generative rendering, illustrated in Fig. 2: under a plain sliding window (Fig. 2(a)), the chunk depicting a revisited location has long been evicted by the time the camera returns. Motivated by StreamingLLM (Xiao et al., 2024), some works (Shin et al., 2026; Yesiltepe et al., 2026) add a fixed attention sink (Fig. 2(b)) that preserves the first chunk, mitigating color drift and improving consistency when the foreground remains unchanged throughout the video. But the first chunk in general does not coincide with the revisited view, so this sink still cannot anchor a loop closure. In both cases the model must regenerate the location from scratch and may produce different textures, colors, or structures (Fig. 1), even though the conditioning depth and camera trajectory are perfectly consistent since they originate from the same underlying 3D geometry.

We refer to this failure as *revisit inconsistency*. Prior work on world-consistent video generation combats forgetting by maintaining explicit 3D state, geometric view memories, hierarchical latents, or learned context-querying and memory modules that can be reused when content reappears (Mallya et al., 2020; Ren et al., 2025; Garcin et al., 2026; Li et al., 2025; Chandratreya et al., 2026; Chen et al., 2026; Yu et al., 2026). These approaches underscore that memory beyond a local history is essential, but they typically depend on costly curated long-horizon datasets and additional model training. In this work, we ask: *how can we leverage the temporal and spatial correspondences already available from a 3D engine to improve revisit consistency in a pretrained autoregressive video generator, without expensive video-model training and data curation?*

To this end, we propose a correspondence-guided memory framework for long-horizon autoregressive generative rendering (Fig. 2(c)). Temporally, we retrieve historical latent chunks whose camera poses match the current chunk and reintroduce them into the clean KV cache as loop-closure memory. This mechanism gives the model access to the actual previously generated view, extending its usable memory beyond the default cache window. Spatially, reinstating a chunk makes its earlier appearance available, but under viewpoint changes the current tokens may not attend to the right regions of it. Using the metric depth and camera poses provided by the engine, we reproject each query token of the current chunk onto the token grid of the retrieved chunk to find its corresponding 2D position. We then add a Gaussian bias to the query's attention logits over the retrieved chunk that peaks at this position and decays with distance from it, softly steering attention toward geometrically corresponding tokens. Unlike warping and blending historical features, which can introduce blur and artifacts, the bias only guides where the model attends and leaves synthesis to the generator.

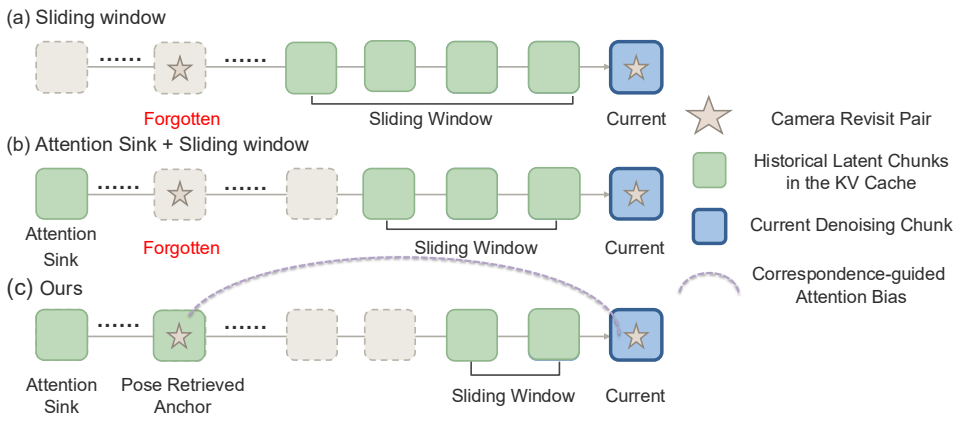


Figure 2: KV cache strategies under camera revisits. Naive sliding-window caching (a), or with an attention sink (b), forgets chunks from previously visited viewpoints. Ours (c) retrieves the pose-matched historical chunk into the KV cache and applies a spatial correspondence-guided attention bias that links current-frame tokens to corresponding regions in the retrieved frame.

Our experiments show that the proposed correspondence-guided memory substantially improves revisit consistency in streaming generative rendering. On a loop-closure benchmark we construct from TartanGround (Patel et al., 2025) and TartanAir (Wang et al., 2020), photorealistic navigation trajectories that repeatedly leave and return to the same locations, our method consistently outperforms existing training-free cache-management schemes on pixel, semantic, and geometric-correspondence metrics adapted from MilliVid (Chandratreya et al., 2026), without degrading overall video quality as measured by VBench-Long (Huang et al., 2026). We also showcase integrating our method with a live game engine, demonstrating revisit-consistent generative rendering directly from engine-streamed depth and camera poses.

2 Related Work

2.1 Generative Rendering

Generative rendering combines the controllability of graphics engines with the realism of generative models. Early video-to-video synthesis methods translate structured conditions, such as semantic layouts or rendered guidance, into photorealistic videos using conditional GANs (Wang et al., 2018a; Mallya et al., 2020). Recent diffusion-based methods improve visual quality by conditioning image or video diffusion models on 3D or simulator-derived signals. Generative Rendering (Cai et al., 2024) uses dynamic mesh correspondences with a pretrained 2D diffusion model to synthesize temporally consistent frames from low-fidelity animated mesh renderings. Coarse-to-Real (Gomez-Nogales et al., 2026) generates realistic populated urban videos from coarse 3D simulations, while RealMaster (Cohen-Bar et al., 2026) lifts 3D-engine rendered videos into photorealistic videos while preserving renderer-specified geometry and dynamics. Large conditional video models such as Cosmos-Transfer and Wan-VACE further demonstrate controllable video generation from spatial controls that could be extracted from graphics engines, including depth, segmentation, edges, and rendered videos (Abu Alhaija et al., 2025; NVIDIA, 2025; Wan Team, 2025; Jiang et al., 2025). These methods mainly target short or offline clips, whereas we study long-horizon autoregressive generative rendering with bounded KV memory and revisit inconsistency.

2.2 Consistent Video Generation

Long-term consistency has been studied in video world models through explicit memory, retrieval, and compact historical state. Seoul World Model (Seo et al., 2026) grounds world simulation in a real metropolis by retrieving relevant street-view context from a large database. VMem (Li et al., 2025) maintains a surfel-indexed view memory, enabling previously generated views to be stored

and reused in an interactive scene. Context-as-Memory (Yu et al., 2025a) retrieves previously generated frames whose camera views overlap the current one and conditions on them as context, and MemLearner (Yu et al., 2026) replaces such rule-based retrieval with learned query tokens that adaptively gather useful context. Matrix-Game 3.0 (Wang et al., 2026b) uses camera-pose-based memory buffers for real-time streaming world modeling. Hybrid Memory (Chen et al., 2026) separates memory for static backgrounds and dynamic objects to preserve content after it leaves the visible field. RELIC (Hong et al., 2025) introduces compact long-horizon memory for interactive video world models. MilliVid (Chandratreya et al., 2026) uses hierarchical latents for long-range consistency, and PERSIST (Garcin et al., 2026) maintains a persistent 3D state rather than relying only on pixel histories. These approaches highlight the importance of memory beyond local context, but generally require curated long-horizon data and costly model training. In contrast, our method is training-free: it reinstates pose-matched history in a frozen distilled generator, using correspondences supplied by the 3D engine.

Another related line improves long-video generation through post-training, distillation, or inference-time memory design. Memorize-and-Generate (Zhu et al., 2025) trains a memory block to compress historical information for real-time generation. MotionStream (Shin et al., 2026) combines Self-Forcing, sliding-window causal attention, and attention sinks for interactive motion-controlled video generation. VideoSSM (Yu et al., 2025b) augments autoregressive generation with hybrid state-space memory, while Helios (Yuan et al., 2026) uses compressed historical context in a multi-stage real-time generation pipeline. Deep Forcing (Yi et al., 2025) proposes training-free KV-cache management with deep sinks and participative compression, Infinity-RoPE (Yesiltepe et al., 2026) modifies RoPE and cache handling for autoregressive self-rollout, and LongLive-RAG (Hu et al., 2026) retrieves historical latents for long video generation. MemRoPE (Kim et al., 2026) maintains a fixed set of evolving memory tokens that continuously compress past keys via exponential moving averages. Our work is complementary to these approaches. Instead of learning or compressing generic video history, we exploit correspondences already available from the 3D engine: temporal correspondence retrieves loop-closure chunks into the KV cache, while spatial correspondence biases attention toward geometrically matched historical tokens.

3 PRELIMINARIES

3.1 AUTOREGRESSIVE VIDEO GENERATION WITH SELF-FORCING

Autoregressive video diffusion models generate long videos in latent chunks: a causal generator denoises the current chunk while attending to a fixed-size sliding window of previously generated chunks through a key–value (KV) cache, so memory and per-chunk compute remain constant as the video grows. Self-Forcing (Huang et al., 2025) distills a bidirectional video diffusion teacher into such a causal, few-step student by making training mirror this inference process. Instead of conditioning on ground-truth history, the student rolls out from pure Gaussian noise; after each chunk is denoised, a clean-context pass writes its keys and values into a clean KV cache that later chunks attend to, exposing the student to its own accumulated errors and reducing the train–test mismatch of teacher forcing. Supervision follows distribution-matching distillation (Yin et al., 2024b;a): a frozen teacher provides the target score, a jointly trained fake-score network estimates the score of the student's samples, and their difference drives the student update, optionally with a GAN discriminator on teacher features.

3.2 CAUSAL DISTILLATION OF DEPTH-CONDITIONED WAN-VACE

We instantiate this recipe on Wan-VACE (Jiang et al., 2025), which conditions the Wan text-to-video (T2V) DiT (Wan Team, 2025) on depth: normalized monocular depth maps are encoded by the Wan VAE and processed by VACE control blocks, whose outputs are injected as residuals into selected backbone layers. Since the pretrained model is bidirectional, we causalize both the main branch and the VACE control branch, each with its own KV cache, to obtain Causal Wan-VACE. Training then follows Sec. 3.1: the teacher and fake-score networks remain bidirectional Wan-VACE, and a single depth-control context is shared by all networks. We use this self-forced, depth-conditioned Causal Wan-VACE as our baseline model.

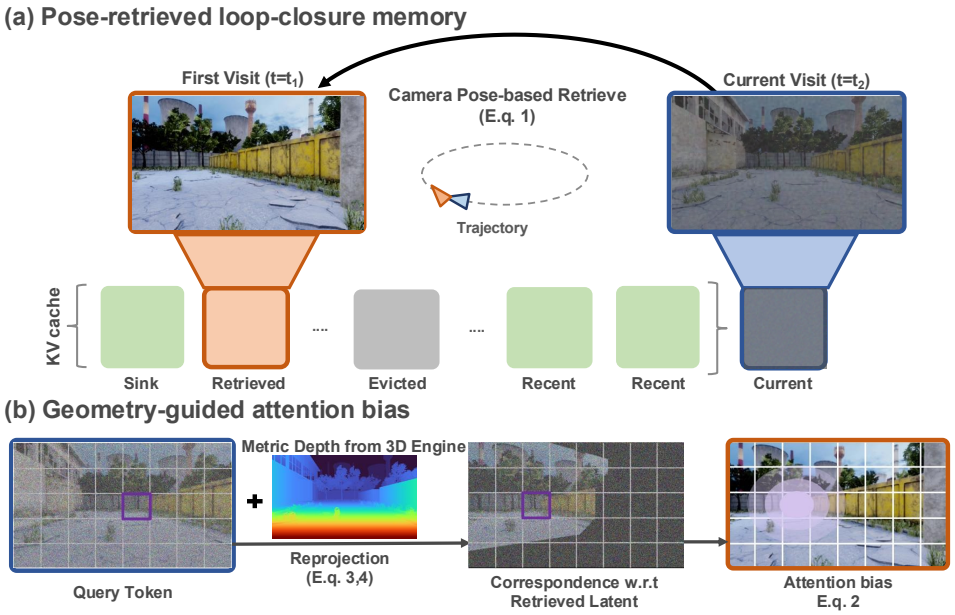


Figure 3: **(a)** When the current pose ($t{=}t_2$) returns near a previously generated viewpoint ($t{=}t_1$), pose-based retrieval (Eq. 1) reinstates the evicted chunk into the bounded KV cache. **(b)** Each query token is reprojected into the retrieved view via metric depth and relative pose (Eq. 3 and Eq. 4), and a Gaussian bias at the predicted correspondence is added to the attention logits (Eq. 2); no cached features are warped.

4 METHOD

Our goal is long-horizon revisit consistency: when the camera leaves a region and returns after the corresponding chunks have been evicted from the clean KV cache, the regenerated content should be consistent with what was generated before. We address this with two complementary, purely inference-time mechanisms that operate on the frozen causal student of Sec. 3.2. First, *pose-retrieved loop-closure memory* (Sec. 4.2) decides *what* the model can see: it detects that the current camera pose revisits a previously generated viewpoint and reinstates the corresponding historical latent chunk into the clean cache. Second, a *geometry-guided attention bias* (Sec. 4.3) decides *where* the model looks: using the known correspondence from the 3D engine, each current query token is softly steered toward its spatially corresponding token in the retrieved chunk.

4.1 SETUP AND ASSUMPTIONS

The student generates latent chunks $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ of F latent frames each ($F{=}3$ in our implementation), and self-attention operates on an $h \times w$ token grid per latent frame (30×52 for 832×480 outputs in Wan-VACE). We assume that each latent frame f is annotated with pinhole intrinsics $\mathbf{K}_f$, rescaled to the token grid, and a rigid camera pose, written as a camera-to-world rotation and center $(\mathbf{R}_f, \mathbf{c}_f)$. We also assume per-frame planar metric depth D_f, i.e., distance along the optical axis in scene units consistent with the camera translations. Both annotations are natural byproducts of the generative-rendering setting we target, where a 3D engine supplies the depth control for the video generation model together with the corresponding camera parameters and metric depth.

4.2 POSE-RETRIEVED LOOP-CLOSURE MEMORY

Cache layout. As shown in Fig. 3(a), we partition the clean-cache budget of M latent frames into three slots: a persistent *anchor* holding the first chunk, one *pose-retrieved* chunk, and a *recent* window of the latest chunks; we use $M{=}4F{=}12$, with F anchor, F retrieved, and $2F$ recent frames. The anchor provides a fixed global reference, the recent window preserves local temporal continuity, and the retrieved slot supplies view-specific long-term memory. Clean latents and poses of evicted chunks are retained outside the attention cache so that any historical chunk can be retrieved. **Loop-**

closure retrieval. Let $(\mathbf{c}_n, \mathbf{v}_n)$ denote the camera center and unit viewing direction associated with the current chunk n, whose first latent frame is f_n, and $(\mathbf{c}_j, \mathbf{v}_j, f_j)$ those of a historical chunk j. We reinstate the admissible candidate closest in pose to the current chunk:

$$\begin{aligned} s(j) \;&=\; \|\mathbf{c}_j - \mathbf{c}_n\|_2/E \;+\; w_v\big(1 - \mathbf{v}_j^\top \mathbf{v}_n\big), \\ \mathcal{A}_n \;&=\; \big\{\, j \;:\; \|\mathbf{c}_j - \mathbf{c}_n\|_2/E \,\le\, \tau_c,\;\; \mathbf{v}_j^\top \mathbf{v}_n \,\ge\, \cos\tau_v,\;\; f_n - f_j \,\ge\, \Delta \,\big\}, \end{aligned} \tag{1}$$

where E is the diagonal of the bounding box of all camera centers in the clip and w_v=0.5 weights the angular term. The admissible set $\mathcal{A}_n$ encodes three gates: a normalized return distance below τ_c=0.2, viewing directions within τ_v=45°, and a temporal gap of at least Δ=8 latent frames. If $\mathcal{A}_n$ is empty, the retrieved slot simply extends the recent window. Each gate serves a distinct purpose: the distance and angle constraints ensure that long-range memory is injected only at genuine revisits, while the temporal gap prevents the immediately preceding chunks already present in the recent window from being trivially selected as spurious loop closures.

Temporal RoPE adjustment. Wan-VACE applies 3D RoPE, which assigns each latent frame a temporal position and each location on the $h \times w$ grid a row and column position. Because RoPE rotates queries and keys by angles proportional to these positions, attention between two latent frames only depends on their relative position. If the anchor or a reinstated chunk kept its true temporal position, its offset to the current chunk would equal the time elapsed since it was generated. Such relative positions between chunks lie far outside the training distribution and can cause visible artifacts (Liu et al., 2025; Yi et al., 2025). Following these works, we set the positions of the cached latents as follows. The recent window and the current chunk keep their true positions, $[f_n{-}2F, f_n)$ and $[f_n, f_n{+}F)$, while the anchor and the retrieved chunk, whatever their true positions, are placed at $[f_n{-}4F, f_n{-}3F)$ and $[f_n{-}3F, f_n{-}2F)$, directly before the recent window. The cache thus always occupies the contiguous block $[f_n{-}M, f_n)$ right before the current chunk, so offsets between the current chunk and any cached latent frame stay below $M{+}F$, within the range seen during training. When $\mathcal{A}_n$ is empty, the retrieved slot instead holds the next-older recent chunk at its true positions. Finally, since Causal Wan-VACE keeps separate caches for the main and control branches, the depth latents of each cached chunk use the same RoPE positions as its video latents, keeping the two caches aligned.

4.3 Geometric Correspondence as an Attention Prior

Reinstating a chunk restores access to the earlier appearance, but the model must still establish correspondences between current queries and retrieved keys across a potentially large viewpoint change, a matching problem that the self-attention mechanism in the video DiT would otherwise have to solve implicitly. Since the geometric correspondence between the two views is known from the 3D engine, we inject it explicitly as a soft prior on the attention logits (Fig. 3(b)).

General form. Consider a query token p on the current (target) frame t and a key token k on a retrieved (source) frame s. Their positions on the $h \times w$ token grid are $\mathbf{u}_p$ and $\mathbf{u}_k$, and their query and key features in a given self-attention head are $\mathbf{q}_p, \mathbf{k}_k \in \mathbb{R}^d$. Suppose a geometric warp predicts the source-frame location $\hat{\mathbf{u}}_s(p)$ at which the content observed at $\mathbf{u}_p$ appeared in frame s (constructed below). We penalize each key by its distance to this predicted correspondence through a Gaussian bias $B_s(p,k)$, added to the standard scaled dot-product logit:

$$B_s(p,k) \;=\; -\,\frac{\|\mathbf{u}_k - \hat{\mathbf{u}}_s(p)\|_2^2}{2\sigma^2}, \qquad A_{pk} \;=\; \frac{\mathbf{q}_p^\top \mathbf{k}_k}{\sqrt{d}} \;+\; \lambda\, B_s(p,k), \tag{2}$$

where σ sets the spatial bandwidth of the prior in grid units and λ its strength (σ=6, λ=1); logits of all other key columns are unchanged. Because $B_s(p,k) \le 0$ and vanishes exactly at the predicted correspondence, the bias acts multiplicatively on the unnormalized attention weights as a factor $\exp\big(\lambda B_s(p,k)\big) \in (0,1]$: the key at $\hat{\mathbf{u}}_s(p)$ keeps its weight, while keys farther away are exponentially suppressed. The prior is therefore a Gaussian-shaped suppression mask rather than an additive bonus; it narrows *where* the model looks but cannot force content into place, leaving *what* to copy to the learned attention.

Reprojection warp. We obtain $\hat{\mathbf{u}}_s(p)$ by lifting the query to 3D with its planar metric depth $z_p = D_t(\mathbf{u}_p)$ and reprojecting it through the camera poses:

$$\mathbf{X}_t \;=\; z_p\, \mathbf{K}_t^{-1} \tilde{\mathbf{u}}_p, \qquad \mathbf{X}_s \;=\; \mathbf{R}_s^\top \big(\mathbf{R}_t\, \mathbf{X}_t + \mathbf{c}_t - \mathbf{c}_s\big), \qquad \hat{\mathbf{u}}_s(p) \;=\; \pi\big(\mathbf{K}_s\, \mathbf{X}_s\big), \tag{3}$$

where $\tilde{\mathbf{u}}_p$ denotes homogeneous grid coordinates and $\pi(\cdot)$ perspective projection: the query is unprojected into the target camera, expressed in world coordinates through the camera-to-world pose, re-expressed in the source camera, and projected. Because each token is displaced according to its own depth, foreground and background move by different amounts under camera translation, as parallax requires. When the camera center is fixed ($\mathbf{c}_s = \mathbf{c}_t$), depth cancels under perspective projection and Eq. 3 reduces to the rotation–zoom homography $\pi(\mathbf{K}_s\mathbf{R}_s^\top\mathbf{R}_t\mathbf{K}_t^{-1}\tilde{\mathbf{u}}_p)$.

Occlusion-aware visibility. The reprojected point may nevertheless be occluded in the source view. At disocclusions, a newly exposed background token would otherwise be steered toward the foreground surface that covered it, producing duplicated structure. We therefore accept a correspondence only if it passes a source-depth visibility test,

$$\big[\mathbf{X}_s\big]_z \;\le\; (1+\varepsilon)\, D_s\big(\hat{\mathbf{u}}_s(p)\big), \tag{4}$$

where $[\mathbf{X}_s]_z$ is the point's depth in the source camera and ε=0.1 an occlusion tolerance; rejected correspondences receive no bias.

5 Experiments

5.1 Evaluation Setup

Evaluation datasets. We mine loops from TartanGround (Patel et al., 2025), a photorealistic Unreal Engine dataset for ground-robot navigation that provides the camera poses and planar metric depth of Sec. 4.1. Its trajectories naturally revisit places after gaps far beyond any cache horizon, under parallax-heavy translational motion. A first-visit/return pair qualifies if the camera returns within 2 m and 45° after an excursion of at least 100 frames and 5 m (ruling out pauses and in-place turns), yielding 71 loops (*TartanGround-Revisit*). We also build *TartanAir-Revisit* from TartanAir (Wang et al., 2020): since its 3D scenes are not public and its drone trajectories are too jittery for video generation, we render smooth revisits from its 360° panoramas with two scripted patterns, a *rotation* that yaws away by 90–180° and returns, and a *zoom* that narrows the field of view from 90° to 35° and back, yielding 285 clips over 72 environments.

Evaluation metrics. We evaluate both revisit consistency and overall video quality. Similar to MilliVid (Chandratreya et al., 2026), we measure consistency by comparing each loop pair's two generated frames with L1 error and DINOv2 (Oquab et al., 2024) cosine similarity. Most diagnostic is geometric correspondence: we detect SuperPoint (DeTone et al., 2018) keypoints per frame, match them with LightGlue (Lindenberger et al., 2023), and count matches with confidence above 0.5; a high count means that structures from the first visit are re-rendered where a matcher can confidently re-localize them, the object permanence that loop closure demands. For overall quality, we report VBench-Long (Huang et al., 2024; 2026), omitting its dynamic-degree dimension, which saturates at 1.0 for all methods since motion is controlled by the conditioning depth sequence.

Baseline methods. All methods share the same Causal Wan-VACE base model (Jiang et al., 2025) of Sec. 3.2, depth control, prompts, random seeds, and KV cache size. We compare against the base model with a plain rolling window (**Self-Forcing**) and three training-free cache-management schemes applied to the same student: **Infinity-RoPE** (Yesiltepe et al., 2026), **Deep Forcing** (Yi et al., 2025), and **MemRoPE** (Kim et al., 2026).

Table 1: Loop-closure consistency and overall video quality on TartanGround-Revisit. Consistency metrics are macro-averaged over each clip's genuine loop-closure revisit pairs; quality is measured by VBench-Long. Best in **bold**. ↑/↓ indicate higher/lower is better.

Method	Loop-closure consistency			Video quality (VBench-Long)					
	Keypoint matches ↑	DINO similarity ↑	L1 error ↓	Subject consistency ↑	Background consistency ↑	Motion smoothness ↑	Aesthetic quality ↑	Imaging quality ↑	Mean ↑
Self-Forcing (Huang et al., 2025)	23.26	0.6087	0.1148	0.8472	0.9371	**0.9781**	0.4627	0.3952	0.7241
Infinity-RoPE (Yesiltepe et al., 2026)	26.89	0.6235	0.1103	0.8239	0.9302	0.9745	0.4542	0.4073	0.7180
MemRoPE (Kim et al., 2026)	27.07	0.6244	0.1089	0.8227	0.9297	0.9751	0.4537	0.3991	0.7161
Deep Forcing (Yi et al., 2025)	43.90	0.6848	0.1063	0.8586	0.9421	0.9754	0.4738	0.4572	0.7414
Ours	**49.08**	**0.6904**	**0.1052**	**0.8592**	**0.9428**	0.9754	**0.4740**	**0.4586**	**0.7420**

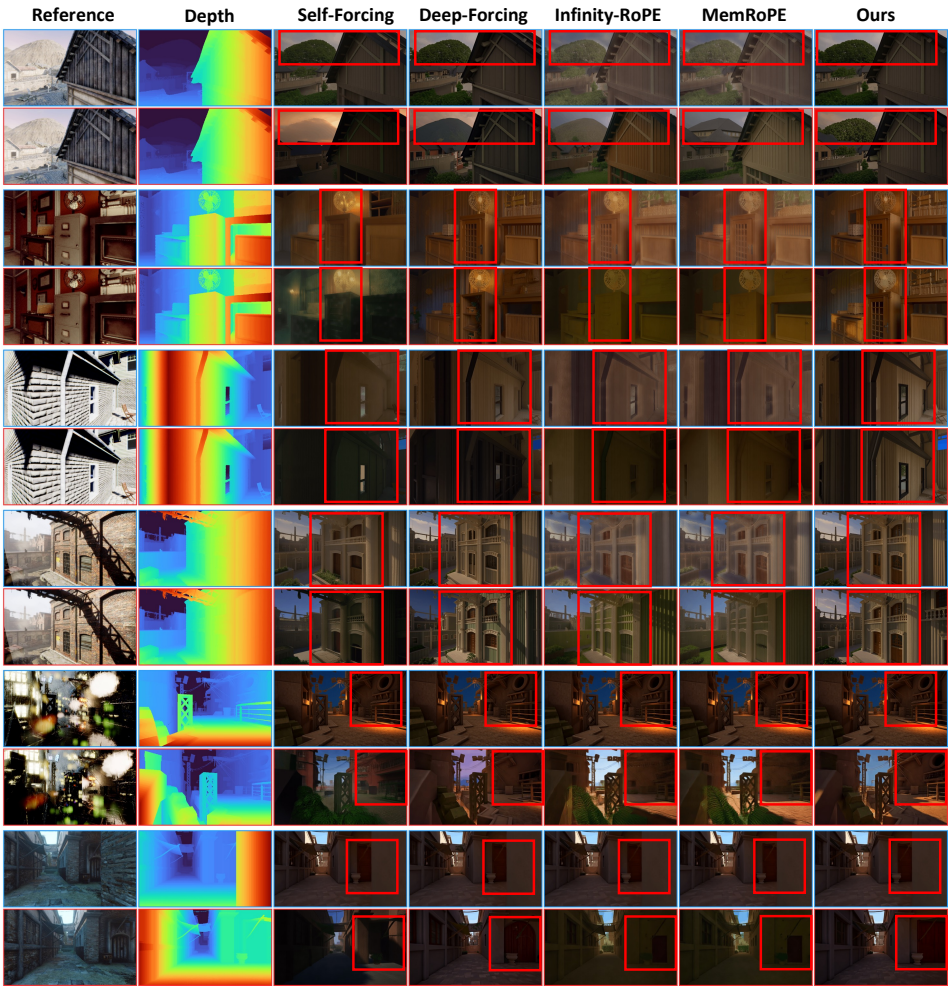


Figure 4: **Qualitative comparison on loop-closure revisits.** Each pair of rows shows the first visit (blue) and the return (red) of one clip; the Reference and Depth columns show the conditioning source. Red boxes mark the same scene region at both visits. Baselines regenerate altered structures on return; our method re-renders that content consistently.

Table 2: Loop-closure consistency and overall video quality on TartanAir-Revisit.

Method	Loop-closure consistency			Video quality (VBench-Long)					
	Keypoint matches ↑	DINO similarity ↑	L1 error ↓	Subject consistency ↑	Background consistency ↑	Motion smoothness ↑	Aesthetic quality ↑	Imaging quality ↑	Mean ↑
Self-Forcing (Huang et al., 2025)	195.65	0.7623	0.0862	0.9163	0.9402	0.9894	0.4592	0.4768	0.7564
Infinity-RoPE (Yesiltepe et al., 2026)	204.14	0.7622	0.0792	0.8901	0.9360	0.9868	0.4446	0.4487	0.7412
MemRoPE (Kim et al., 2026)	207.95	0.7637	0.0769	0.8882	0.9357	0.9867	0.4428	0.4478	0.7402
Deep Forcing (Yi et al., 2025)	267.87	0.8291	0.0664	0.9271	**0.9475**	**0.9897**	0.4682	0.5357	0.7737
Ours	**284.92**	**0.8407**	**0.0654**	**0.9279**	0.9468	0.9896	**0.4701**	**0.5428**	**0.7754**

5.2 Quantitative and Qualitative Results

Tables 1 and 2 report results on TartanGround-Revisit and TartanAir-Revisit. Every training-free cache-management scheme improves revisit consistency over the base model, confirming that memory management is the right lever for this problem; our method pushes further by selecting memory with camera geometry and achieves the best score on every consistency metric, roughly doubling the base model's high-confidence keypoint matches on TartanGround-Revisit. The same ordering holds on TartanAir-Revisit. On video quality, our method attains the best VBench-Long mean on both benchmarks, showing that our consistency gains do not come at the cost of visual fidelity. Qualitatively (Fig. 4), baselines may regenerate altered content on return, such as changed facades and inserted structures, while our method re-renders the boxed regions consistently across the loop.

Table 3: Component ablation on both benchmarks. Each row adds one component to the base student distilled with Self-Forcing (SF): AS denotes the attention sink, PR the pose-retrieved loop-closure memory (Sec. 4.2), and GB the geometry-guided attention bias (Sec. 4.3).

Configuration	TartanGround-Revisit			TartanAir-Revisit		
	Keypoint↑	DINO↑	L1↓	Keypoint↑	DINO↑	L1↓
SF	23.26	0.6087	0.1148	195.65	0.7623	0.0862
SF + AS	39.40	0.6761	0.1078	257.05	0.8110	0.0721
SF + AS + PR	47.84	**0.6926**	**0.1052**	283.30	**0.8410**	0.0655
SF + AS + PR + GB (**Ours**)	**49.08**	0.6904	**0.1052**	**284.92**	0.8407	**0.0654**



Figure 5: **Qualitative ablation.** Columns cumulatively add our components to Self-Forcing (SF): attention sink (AS), pose-retrieved memory (PR), and geometry-guided attention bias (GB). Rows pair the first visit (blue) with the return (red); red boxes mark the same region.

Table 3 adds our components one at a time to the Self-Forcing (SF) baseline. The attention sink (AS) yields a large gain by stabilizing global appearance against drift. Pose retrieval (PR) is the dominant revisit-specific factor: access to the actual earlier view is what loop closure requires. The geometric bias (GB) further improves keypoint matches; it refines *where* each token reads within the retrieved chunk, exactly the placement property the matcher-based metric measures. Fig. 5 shows the progression qualitatively. Qualitative video results are available on our supplementary website.

5.3 Integration with a Live Game Engine

We integrate our approach with a live game engine that streams the annotations assumed in Sec. 4.1, namely camera poses and metric depth, directly to the generator. Fig. 6 shows two representative cases on loop-closure camera paths: a castle with complex architecture (top) and a scene built from untextured geometric primitives (bottom). In both cases, the generator must dress the geometry with plausible appearance. Compared with baseline methods that regenerate altered structures (e.g., an added dome), our method re-renders the same structures on return.



Figure 6: **Integration with a live game engine** on loop-closure paths (blue: first visit; red: return). On return, baselines alter structures; ours re-renders the same structures.

6 Conclusion

We presented a training-free approach to revisit-consistent streaming generative rendering. Exploiting correspondences a 3D engine already provides, our method retrieves pose-matched historical chunks into the bounded KV cache as loop-closure memory and steers attention toward geometrically corresponding tokens via a depth-reprojection Gaussian bias. The main limitation is the reliance on engine-supplied camera poses and metric depth; estimated geometry from online reconstruction methods like VGGT (Wang et al., 2025) and VGGT-Ω (Wang et al., 2026a) would extend the method to real video, at the cost of correspondence noise that our soft attention bias is designed to tolerate.

AI USE STATEMENT

In this work, we used generative AI tools only for tasks with recommended disclosure: (i) editing the paper to improve readability, including polishing grammar and wording and revising paragraphs for clarity and concision; (ii) formatting references and LaTeX, for example, laying out tables and figures; and (iii) identifying relevant literature, searching for information, and analyzing existing work. We have reviewed all AI-assisted work: the authors checked every AI-suggested edit to ensure that the technical content and claims were unchanged. 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 involves no human subjects or personal data: we use only publicly available datasets of synthetic 3D environments (TartanAir and TartanGround) and build on publicly available pretrained video models. As with any video generator, outputs could be misused to create misleading content; our method is training-free, targets rendering scenes supplied by a 3D engine, and adds no generative capability beyond the underlying model.

REPRODUCIBILITY STATEMENT

Our method is training-free; the underlying model and its causal distillation are described in Sec. 3. Sec. 4 specifies the loop-closure retrieval rule, the RoPE positions of the cached latents, and the geometric attention bias, and Sec. 5.1 details how both revisit benchmarks are built from public datasets, including the loop-mining criteria and the scripted camera patterns. We will release our code and the benchmark trajectories upon acceptance.

REFERENCES

Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, Dieter Fox, Yunhao Ge, Jinwei Gu, Ali Hassani, Michael Isaev, Pooya Jannaty, Shiyi Lan, Tobias Lasser, Huan Ling, Ming-Yu Liu, Xian Liu, Yifan Lu, Alice Luo, Qianli Ma, Hanzi Mao, Fabio Ramos, Xuanchi Ren, Tianchang Shen, Xinglong Sun, Shitao Tang, Ting-Chun Wang, Jay Wu, Jiashu Xu, Stella Xu, Kevin Xie, Yuchong Ye, Xiaodong Yang, Xiaohui Zeng, and Yu Zeng. Cosmos-Transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025.

Shengqu Cai, Duygu Ceylan, Matheus Gadelha, Chun-Hao Paul Huang, Tuanfeng Yang Wang, and Gordon Wetzstein. Generative rendering: Controllable 4D-guided video generation with 2D diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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. *arXiv preprint arXiv:2606.09056*, 2026.

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

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. *arXiv preprint arXiv:2603.25716*, 2026.

Dana Cohen-Bar, Ido Sobol, Raphael Bensadoun, Shelly Sheynin, Oran Gafni, Or Patashnik, Daniel Cohen-Or, and Amit Zohar. RealMaster: Lifting rendered scenes into photorealistic video. *arXiv preprint arXiv:2603.23462*, 2026.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 224–236, 2018.

fal.ai and Lightricks. LTX-2.3 quality render-to-real. `https://fal.ai/models/fal-ai/ltx-2.3-quality/render-to-real`, 2026. Accessed 2026-07-08.

Samuel Garcin, Thomas Walker, Steven McDonagh, Tim Pearce, Hakan Bilen, Tianyu He, Kaixin Wang, and Jiang Bian. Beyond pixel histories: World models with persistent 3D state. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Gonzalo Gomez-Nogales, Yicong Hong, Chongjian Ge, Marc Comino-Trinidad, Peiye Zhuang, Dan Casas, and Yi Zhou. Coarse-to-Real: Generative rendering for populated dynamic scenes. *arXiv preprint arXiv:2601.22301*, 2026.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

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 preprint arXiv:2512.04040*, 2025.

Qixin Hu, Shuai Yang, Wei Huang, Song Han, and Yukang Chen. LongLive-RAG: A general retrieval-augmented framework for long video generation. *arXiv preprint arXiv:2606.02553*, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3268–3285, 2026. doi: 10.1109/TPAMI.2025.3633890.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Youngrae Kim, Qixin Hu, C.-C. Jay Kuo, and Peter A. Beerel. MemRoPE: Training-free infinite video generation via evolving memory tokens. *arXiv preprint arXiv:2603.12513*, 2026.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. VMem: Consistent interactive video scene generation with surfel-indexed view memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

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 (ICCV)*, pp. 17627–17638, 2023.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling Forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

Arun Mallya, Ting-Chun Wang, Karan Sapra, and Ming-Yu Liu. World-consistent video-to-video synthesis. In *European Conference on Computer Vision*, 2020.

NVIDIA. World simulation with video foundation models for physical AI. *arXiv preprint arXiv:2511.00062*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Manthan Patel, Fan Yang, Yuheng Qiu, Cesar Cadena, Sebastian Scherer, Marco Hutter, and Wenshan Wang. TartanGround: A large-scale dataset for ground robot perception and navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

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*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Junyoung Seo, Hyunwook Choi, Minkyung Kwon, Jinhyeok Choi, Siyoon Jin, Gayoung Lee, Junho Kim, JoungBin Lee, Geonmo Gu, Dongyoon Han, Sangdoo Yun, Seungryong Kim, and Jin-Hwa Kim. Grounding world simulation models in a real-world metropolis. *arXiv preprint arXiv:2603.15583*, 2026.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. MotionStream: Real-time video generation with interactive motion controls. In *International Conference on Learning Representations (ICLR)*, 2026.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025.

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)*, 2026a.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems*, 2018a.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018b.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A dataset to push the limits of visual SLAM. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916, 2020.

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.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations (ICLR)*. OpenReview.net, 2024.

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. *arXiv preprint arXiv:2509.22622*, 2025.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep Forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo 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 (CVPR)*, pp. 6613–6623, 2024b.

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*, 2025.

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 *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025a. doi: 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. *arXiv preprint arXiv:2606.31734*, 2026.

Yifei Yu, Xiaoshan Wu, Xinting Hu, Tao Hu, Yangtian Sun, Xiaoyang Lyu, Bo Wang, Lin Ma, Yuewen Ma, Zhongrui Wang, and Xiaojuan Qi. VideoSSM: Autoregressive long video generation with hybrid state-space memory. *arXiv preprint arXiv:2512.04519*, 2025b.

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.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

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

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 video generation. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Tianrui Zhu, Shiyi Zhang, Zhirui Sun, Jingqi Tian, and Yansong Tang. Memorize-and-Generate: Towards long-term consistency in real-time video generation. *arXiv preprint arXiv:2512.18741*, 2025.