GEOMETRY TELLS WHERE; HISTORY TELLS WHAT: APPEARANCE MEMORY IN CAUSAL VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video generators now demonstrate impressive capabilities in creating controlled qualitative content. However, camera-controlled video generators still produce substantial appearance drift when revisiting earlier viewpoints. In this paper, we build a competitive geometry- and camera-conditioned causal backbone and use it to investigate what information should be preserved and where recalled memory should enter the generator. We find that geometry alone is insufficient, even when it covers almost the entire view: appearance is recovered reliably only when previously generated RGB image is recalled and temporally aligned with the frame currently being generated. Crucially, simply providing the correct recalled RGB is not enough: when its temporal position is reset to the beginning of the sequence, rather than current ,its benefit largely disappears, performing no better than having no memory. Based on these findings, we introduce CHaRM, a training-free appearance memory that stores generated RGB frames indexed by camera pose. At revisit time, CHaRM reuses the stored frames when the camera returns to a stored pose, or otherwise warps nearby stored frames into the new view, gives them the temporal position of the frame being generated, and injects them next to the KV cache of recent frames. As a result, CHaRM cuts self-revisit LPIPS (consistency with previously generated views) from 0.504 to 0.099 on exact revisits, and brings return LPIPS (fidelity to ground truth on the way back) to the level the backbone reaches in a single uninterrupted pass, i.e., with no loss of quality from revisiting.

1 INTRODUCTION

Video generative models have progressed from producing short, isolated clips to acting as interactive visual simulators. Given a camera trajectory or user actions, recent camera-controlled generators (Wang et al., 2024; He et al., 2025) synthesize long, explorable sequences, while autoregressive or causal variants (Chen et al., 2024; Yin et al., 2025a; Huang et al., 2026) display the ability to stream frames in real time. Yet, current generators fall short when the camera returns to a place it has already seen, as they fail to effectively retrieve relevant memory even when the complete KV cache is available. Camera conditioning and geometric cues constrain where surfaces lie, but they leave what they look like largely undetermined: texture, color, lighting, and fine detail are resampled on each visit. The result is appearance drift – revisited views keep a plausible layout but not the same content – which undermines the idea of a coherent, persistent world.

Several recent works add explicit memory to bidirectional video generators, either by rendering a 3D proxy of the scene (a point cloud, a 3D cache, or surfels) into the target view (Yu et al., 2025b; Ren et al., 2025; Wang et al., 2023; Li et al., 2025), by retrieving relevant past frames as additional context (Xiao et al., 2026; Yu et al., 2025a), or by compressing the context so that more history fits (Cai et al., 2026; Zhang et al., 2026). Most of them require training or fine-tuning large bidirectional backbones. Recent causal video generators pose a different memory problem. They generate progressively from an ordered history represented by a KV cache. Recalling an earlier observation therefore requires deciding not only what to retrieve, but also where to place it in the generator's temporal history: which temporal position should a recalled frame receive? This interaction between explicit memory and causal temporal context remains largely unexplored.

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.

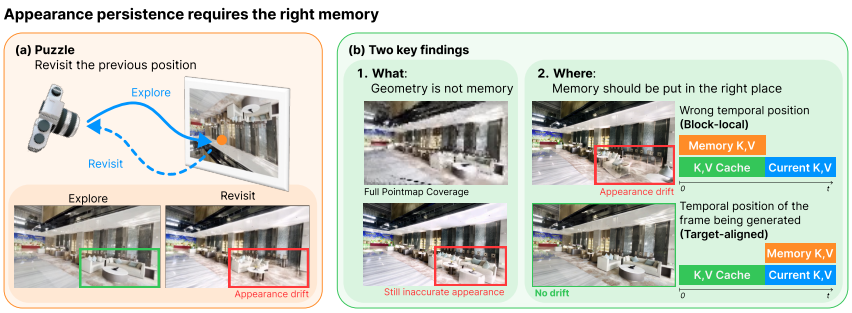


Figure 1: Diagnosing appearance drift on revisit. (a) Returning to the same camera pose can produce appearance drift. (b) Two findings isolate the memory requirements: geometric support does not preserve earlier appearance (WHAT), and recalled appearance must be placed at the right temporal position (WHERE).

Exploring these questions requires a causal generator that (i) sees earlier frames only through its KV cache, where recalled content would enter; (ii) accepts camera poses and scene geometry, so that the camera can return to exactly the poses it visited; and (iii) samples in a few steps, so that it must rely on the cache rather than regenerate its history. None of the released models we evaluated has all three (Table 1).

We therefor build a competitive task-matched 1.3B causal backbone following Self Forcing (Huang et al., 2026) and Causal Forcing (Zhu et al., 2026b). The same line underlies subsequent causal-generation and distillation work (Zhao et al., 2026a; Yin et al., 2025b; Zheng et al., 2026; Yu et al., 2026; Liu et al., 2026; Yang et al., 2026; Kim, 2026; Yesiltepe et al., 2026; Li et al., 2026), making it a representative backbone for this study. Because the released Causal Forcing model provides neither geometry conditioning nor camera control, we extend it with continuous camera-pose inputs and a geometry canvas obtained by projecting the 3D point cloud into the target camera. This provides the generator with explicit geometric information about previously observed scene content while retaining the established Causal Forcing distillation recipe (Section 4.1).

Using this backbone, we study the two memory questions separately. First, geometric information alone does not preserve appearance. Even when the projected point-cloud canvas covers almost the entire revisited view, appearance still drifts, whereas recalling the generator's own previously generated RGB provides a much stronger appearance memory. Second, the effectiveness of this RGB memory depends critically on its temporal placement. Every frame attended to by the generator carries a temporal coordinate through its rotary position encoding. Recalled frames are effective when assigned the temporal position of the frame currently being generated (target-aligned). Resetting their positions to the beginning of the sequence (block-local) largely eliminates this benefit, although the RGB content is identical (Figure 1; Secs. 5.3–5.4).

Based on these findings, we introduce CHARM (Clean History Appearance Retrieval Memory), a training-free, read-only appearance memory for causal video generators; "clean" denotes RGB frames previously produced by the generator itself, rather than ground-truth images. When the camera revisits a stored pose, CHARM directly recalls the frames generated during the first visit; for novel but nearby viewpoints, it instead warps the nearest stored observations into the target view. In both cases, recalled frames are assigned the temporal position of the frame currently being generated and injected alongside, rather than in place of, the existing KV cache (Sec. 4.2). On exact revisits, CHARM reduces self-revisit LPIPS, which compares each returning frame with the frame generated at the same pose during the first visit, from 0.504 to 0.099 in every scene and at all three seeds. It also brings return LPIPS back to the backbone's single-pass level, so the gain in self-consistency does not cost fidelity.

As a result, this paper makes three main contributions: (i) we identify what causal video generators should remember. Geometry alone is insufficient to preserve appearance on revisits, even with near-complete view coverage, whereas recalling previously generated RGB substantially improves consistency, (ii) we identify how recalled appearance should enter a causal generator. Recalled RGB is effective when aligned with the temporal position currently being generated; resetting its temporal position largely eliminates the benefit of memory, and (iii) we introduce CHARM, a training-free appearance memory for causal video generation. Without retraining its frozen 1.3B backbone,

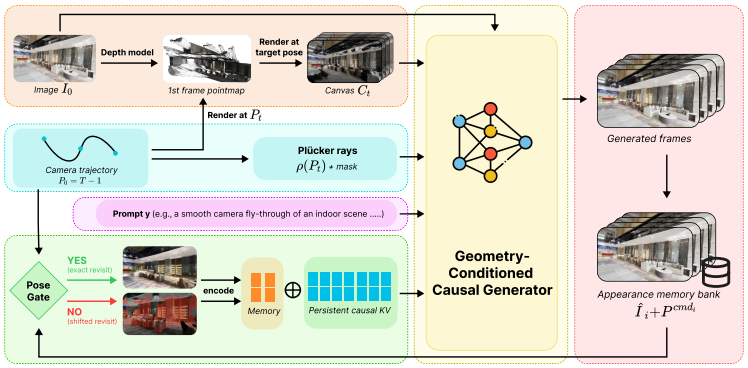


Figure 2: **CHARM pipeline.** When the camera returns to a stored pose, the stored generated frames are reused directly; otherwise, the nearest stored frames are warped into the new view, keeping only visible pixels. Either route is re-encoded at the target block's rotary position and appended next to the KV cache, without writing back to memory.

Table 1: **Memory and control mechanisms of selected public variants.** Our 1.3B backbone combines persistent causal KV with geometry, pose control, and four-step sampling. "–" denotes an interface absent from the cited variant; parentheses name substitutes. PKV: persistent causal KV; BC-AR/S-AR/C-AR: block-/segment-/chunk-level autoregression; TA: target-aligned temporal position.

System	Paradigm	Geometry input	Camera input	Sampler	Recall
Causal Forcing/minWM	PKV	–	–	few-step	–
Warp-as-History	chunk repack	warped history	pose	few-step	warped history
HY-WorldPlay	chunk stream	–	actions	streaming	–
SANA-WM	GDN hybrid	– (pose rays)	6-DoF pose	4-step (var.)	–
Lyra 2.0	window, 3D cache	depth-warped ctx.	pose	4-step (var.)	visibility
Wonder	sparse-attn. KV	3D lattice	traj. video	4-step	sparse KV
GeoStream	sink + local KV	3D reprojection	view proj.	few-step	– (implicit)
WorldCam	anchored latents	– (Plücker)	6-DoF pose	64-step	latent pool
VMem	image window	surfel memory	pose	multi-step	surfels
Ours	**PKV**	**PCD canvas**	**pose**	**4-step**	**TA RGB**

CHARM reduces self-revisit LPIPS from 0.504 to 0.099 and achieves lower return and self-revisit LPIPS than the evaluated 8B–14B baselines. We release our geometry- and camera-conditioned causal backbone and checkpoint.

2 Related Work

Video Generation with Camera and Geometry. CameraCtrl (He et al., 2025) injects camera parameters through control branches or attention, while GEN3C (Ren et al., 2025) and ViewCrafter (Yu et al., 2025b) use explicit 3D support for novel views. GeoStream (Zhao et al., 2026b) re-projects a self-refreshing 3D cache from generated frames. Directing the World (Wang et al., 2026a) renders point-cloud videos from the first frame with block-aligned Plücker camera injection; neither it nor GeoStream reports an explicit appearance-recall path. GeoStream argues that conditioning on geometry rebuilt from the model's own outputs requires training on such self-generated inputs. CHARM needs no such training for its direct route (Sec. 5.5), and our shifted-revisit results support its warp route (Sec. 5.6). Warp-as-History (Wang & He, 2026) is the precedent for target-aligned, visibility-filtered warped history, but on repacked chunks rather than a persistent KV cache, where we isolate the time position.

Video model memory. WorldMem (Xiao et al., 2026), Memory Forcing (Huang et al., 2025), Context-as-Memory (Yu et al., 2025a), VMem (Li et al., 2025), SPMem (Wu et al., 2026), and Lyra 2.0 (Shen et al., 2026) retrieve frames, surfels, or 3D caches; several train a memory pathway. Wonder (Xu et al., 2026) retrieves KV chunks through trained sparse attention, while WorldCam (Nam et al., 2026) retrieves pose-indexed latents in a 64-step sampler. Our training-free intervention keeps a frozen bank of generated RGB indexed by input camera pose. Because the bank is never written back, recall cannot accumulate its own drift.

Causal video generation Self Forcing (Huang et al., 2026) and Causal Forcing (Zhu et al., 2026b) establish the Wan2.1-derived persistent-KV, block-causal line adopted here, while Causal Forcing++ (Zhao et al., 2026a) and subsequent causal-generation and distillation work (Yin et al., 2025b; Zheng et al., 2026; Yu et al., 2026; Liu et al., 2026; Yang et al., 2026; Kim, 2026; Yesiltepe et al., 2026; Li et al., 2026) further develop the same family. This recurring use motivates our choice of the line as a representative backbone for the fixed-cache diagnosis. WorldPlay (Sun et al.), HY-WorldPlay (Team HunyuanWorld, 2025), SANA-WM (Zhu et al., 2026a), Warp-as-History/Helios (Wang & He, 2026; Yuan et al., 2026), Lyra (Shen et al., 2026; Zhang et al., 2026), and Infinite Worlds (Gao et al., 2026) use streaming, chunk-level, hybrid-recurrent, mixed bidirectional/causal, or packed-window history. These behavioral references do not provide the persistent-KV setting in which later blocks access prior content through a fixed cache rather than repacking or recomputation.

3 Problem Setup and Motivating Analysis

Experimental setup. We study autoregressive video generator that takes scene geometry as input. It generates the video block by block, and each new block can access earlier frames only through the keys and values it attends to. When generating block u, the generator can learn about the past from three sources. Geometry ($\mathcal{S}_{\text{geo}}$) tells it where surfaces lie in the new view. A memory of stored frames ($\mathcal{M}_{\text{app}}$) tells it what those surfaces looked like. The KV cache ($\mathcal{K}_{\text{AR}}$) holds the keys and values of the most recent frames and keeps the video smooth from one block to the next. Together with the text prompt y and the target camera poses P_u, these three sources determine the generated block:

$$\hat{x}_u = G_\theta\big(z_u;\ y,\ P_u,\ \underbrace{\mathcal{S}_{\text{geo}}(P_u)}_{\text{where}},\ \underbrace{\mathcal{M}_{\text{app}}(P_u)}_{\text{what}},\ \underbrace{\mathcal{K}_{\text{AR},u}}_{\text{local continuity}}\ \big), \tag{1}$$

where z_u is the initial noise of the block. Geometry is given to the generator as a condition, like the camera pose. The KV cache and the memory are instead used as attention context: the retrieved frames are encoded into keys and values and concatenated with those in the KV cache. Without memory, $\mathcal{M}_{\text{app}}$ is empty and the equation describes the base generator. This leaves two questions. First, should information from previously generated frames be provided as geometry or as memory? Second, each frame in the attention context has a temporal position given by rotary position embedding (RoPE) (Su et al., 2024); which temporal position should the retrieved frames be assigned?

We study both questions on a revisit task: the camera first follows a trajectory (first visit) and then returns to poses it has visited, either exactly (exact revisit) or moved toward the scene so that no pose repeats (shifted revisit). The memory stores each first-visit frame with its camera pose. Since these are the generator's own outputs, retrieving them reproduces what the generator showed before, not the real scene. A stored frame is retrieved directly if its pose matches the requested one (direct route, always the case in an exact revisit); otherwise, the nearest stored frame is warped into the requested view (warp route). By default, the geometry canvas $\mathcal{S}_{\text{geo}}$ renders a point cloud reconstructed from the first frame only.

Evaluation protocol. We evaluate 55 disjoint DL3DV-Eval scenes at one primary seed and one fixed prompt; the headline exact-revisit result is repeated at three preset seeds. Exact-revisit return and replay protocols use the direct route. The full point-cloud canvas renders the pointcloud of generated frames (Sec. 5.3); the oracle GT-colored canvas is a non-deployable diagnostic whose colors come from real future frames. Single-pass and return LPIPS use real GT. Self-revisit and out-of-order LPIPS use the model's own earlier frames as reference, so exact replay is their optimum. Shifted-revisit scores use warped references because no real frame exists at the query poses (Appendix Table 7).

4 Method

We first describe the causal backbone used throughout the paper (Sec. 4.1). CHARM then implements the two findings of Section 1: it stores the generator's own previously generated RGB (Sec. 4.2) and recalls it at the temporal position currently being generated (Sec. 4.3). CHARM op-

erates entirely at inference time, adding no parameters or fine-tuning. Figure 2 summarizes the full pipeline.

4.1 GEOMETRY-CONDITIONED CAUSAL BACKBONE

Our experiments require a causal generator that can revisit controlled viewpoints while retaining its past only through a persistent KV cache. Concretely, at each generation step, the backbone takes the recent causal history together with a target camera pose and a geometric projection of the scene into that camera, and generates the next block of frames. No released model provides this combination of persistent KV cache, explicit geometry and camera control, and few-step sampling (Table 1).

We therefore extend Causal Forcing (Zhu et al., 2026b) with camera-pose and geometry conditioning. Camera pose is represented by Plücker rays, while geometry is provided as a target-view canvas rendered from the VGGT-Ω (Wang et al., 2026b) point cloud reconstructed from the first frame. Both are injected into the input tokens and, through gated residuals, into the first 20 of 30 transformer blocks. This gives us a camera-controllable causal backbone in which geometry, appearance memory, and temporal placement can be studied separately.

Training follows the Causal Forcing pipeline. First, we train the geometry-conditioned autoregressive teacher on DL3DV, jointly updating the DiT backbone and the added camera and canvas modules. Second, we fit an ODE student to the teacher's causal trajectories. Third, we distill it into the deployed four-step generator using asymmetric DMD (Yin et al., 2024b;a). Only the geometry and camera conditioning are introduced by us; the two distillation stages retain the established Causal Forcing recipe.

After training, the backbone is frozen for all experiments. CHARM is added only at inference time, ensuring that all memory comparisons use exactly the same generator checkpoint. Additional architecture and training details are provided in Appendix A.1.

4.2 POSE-INDEXED MEMORY BANK AND RECALL ROUTES

During the first visit, CHARM stores every generated frame together with the camera pose it was generated for. When the camera later requests a pose, CHARM first checks whether that pose matches a stored one. When the return begins, the bank is the set of generated RGB and input camera poses, $\mathcal{M}_{\text{app}} = \{(\hat{I}_i, P_i^{\text{cmd}}) \mid 0 \le i < T_{\text{out}}\}$, where T_{out} is the number of first-visit frames; it is frozen and has no write-back, and the frames retrieved from it for a requested pose P_u form $\mathcal{M}_{\text{app}}(P_u)$ in Eq. 1. A query matches a stored pose when

$$g(q,i) = \mathbf{1}[d_c(q,i) \le \epsilon_c]\, \mathbf{1}[d_R(q,i) \le \epsilon_R], \tag{2}$$

where d_c is normalized center distance and d_R is the rotation angle. Among the stored frames that match, direct recall uses the one with the smallest pose error $e(q,i)$. If none matches, the nearest stored frame is used only as the source for warping, never for direct reuse. The source pose is the command supplied to the generator, not a realized-pose estimate.

We fix $\epsilon_c = 0.02$ (2% of the bank's radial scale) and $\epsilon_R = 2^\circ$ without tuning. The check compares only camera poses: the requested pose against the stored ones. Unlike retrieval based on field-of-view overlap or surface visibility (Li et al., 2025; Xiao et al., 2026; Wang & He, 2026), it never looks at image content. An exact revisit requests a stored pose exactly ($d_c = d_R = 0$), so it always matches, and the exact-revisit results in Table 3 therefore do not depend on the threshold values.

The direct route encodes the stored generated RGB with the causal VAE and admits a two-latent block only if both latent poses and all contributing RGB frames pass the pose gate. A match at the block's final pose alone is not enough. When no stored pose matches, we reconstruct a point cloud from the nearest stored keyframe with VGGT-Ω and render it into the requested view, keeping only pixels that are visible from there. We inject a block only if the rendering covers more than 90% of the view; a fixed selection rule, given in Appendix A.2, decides which tokens enter. No real future frame, GT-video pointcloud, or target RGB enters the generator or CHARM at deployment.

4.3 TARGET-ALIGNED INJECTION

The stored frames are images, but the generator reads its history as keys and values in the KV cache. We therefore convert the recalled frames into keys and values by passing them once through the frozen generator, as if it had just generated them for the current block; we call the result pseudo-history. For a target block beginning at latent index u and spanning $B = 2$ latent frames, the selected interval is re-noised to $\tau_{\text{ctx}} = 50$ and passed once through the frozen generator. The pass uses the positive text, target camera/canvas conditions, and the detached KV cache; it neither writes the persistent buffer nor advances cache indices. With L tokens per latent frame, pseudo-history rotary positions start at uL, not at zero:

$$(\tilde{K}_u^\ell, \tilde{V}_u^\ell) = \text{CaptureKV}_\ell((\tilde{z}_u, \dots, \tilde{z}_{u+B-1}), \tau_{\text{ctx}}; y, c^{\text{cam}}_{u:u+B-1}, c^{\text{can}}_{u:u+B-1}, \mathcal{K}_{\text{AR},u}; \text{RoPEStart} = uL). \tag{3}$$

RoPE (rotary positional embedding) encodes token positions for self-attention (Su et al., 2024). We start the recalled memory at the target block's token position uL, rather than at zero. Starting it at zero would give the memory the positions of the first frames of the video, far from the current block. We call the assigned start the memory's temporal position.

After route admission, the denoising context is

$$K_u^\ell = [K^\ell_{\text{AR},u}; \tilde{K}_u^\ell; K^\ell_{\text{self},u}], \qquad V_u^\ell = [V^\ell_{\text{AR},u}; \tilde{V}_u^\ell; V^\ell_{\text{self},u}]. \tag{4}$$

Here $(K^\ell_{\text{AR},u}, V^\ell_{\text{AR},u})$ is the detached KV cache of earlier blocks, $(\tilde{K}_u^\ell, \tilde{V}_u^\ell)$ is captured memory, and $(K^\ell_{\text{self},u}, V^\ell_{\text{self},u})$ is the target block; semicolons concatenate tokens. The KV cache is never removed. Warped frames are injected only in the first 2 of the 4 denoising steps, directly reused frames are injected in all 4. Because the unconditional branch of classifier-free guidance sees only the negative prompt, guidance strengthens text, geometry, and recalled frames together. CHARM raises peak memory by +29% (29.9 to 38.7 GiB). Captured KV is transient, held one block at a time, and never accumulated across the rollout.

5 EXPERIMENTS

In the following, we ask whether revisit drift occurs (Q1), whether geometry or the generator's own frames preserve earlier appearance (Q2), whether the time position of recalled frames matters (Q3), whether CHARM restores exact-revisit recall (Q4), whether the position dependence persists on shifted revisits (Q5) and how CHARM compares with larger external systems. All internal experimental settings use the same fixed persistent-KV architecture, prompt, trajectory and seed unless stated otherwise.

5.1 EVALUATION PROTOCOLS

In an exact revisit, the camera returns to poses it visited earlier in the same video. In the single-pass protocol, the camera moves along an 81-frame path once and never returns; its LPIPS is the reference level. In round-trip, the camera then retraces the same 80 poses in reverse (161 frames). In out-of-order, the camera instead jumps back and forth along the first-visit path, visiting the points at 100%, 62%, 86%, 28%, 70% and 0% of it, so earlier poses are revisited in a shuffled order. Appendix A.5 gives the full protocol definition.

The Q2 and Q3 experiments are separate controlled reruns, not the headline run in Table 3, so their numbers differ slightly (direct-RGB self-revisit LPIPS is 0.097 in both reruns and 0.099 in the headline run). Confidence intervals resample scenes, and win counts use unrounded values. The single-pass run renders its canvas from the first frame only and follows the ground-truth camera poses without modification (Table 13 in Appendix A.9).

5.2 Q1: REVISIT DRIFT WITHOUT EXPLICIT APPEARANCE RECALL

Without CHARM, exact-revisit return gives 0.600 real-GT LPIPS and 0.504 generated-reference LPIPS, versus 0.447 for a single pass without revisit. The gap persists despite the ordinary KV cache, showing that short-term causal continuity alone does not ensure faithful return. Q2–Q3 examine what memory should hold and where it should be put.

Table 2: **What memory should hold (Q2) and where it should enter (Q3).** Input LPIPS and coverage describe the injected content; return (vs. real GT) and self-revisit (vs. the model's own first-visit frames) describe the output. The two Q3 columns inject identical frames and differ only in their time index: reset to zero, as if they were the first frames of the video (block-local), or equal to that of the block being generated (target-aligned). Separate reruns from Table 3.

	Q2: what to store		Q3: same RGB, position only	
	no memory	full point-cloud canvas	block-local	target-aligned
Input LPIPS ↓	–	0.691	8.2×10^{-7}	8.2×10^{-7}
Coverage ↑	0.000	0.992	1.000	1.000
Return LPIPS ↓	0.589	0.544	0.589	**0.445**
Self-revisit LPIPS ↓	0.490	0.431	0.485	**0.097**

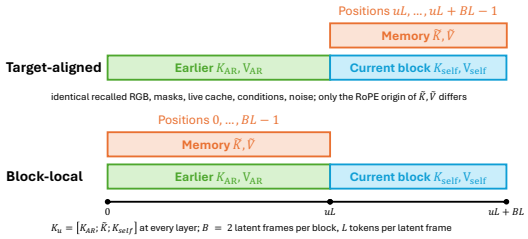


Figure 3: **Temporal-position intervention (Q3).** The recalled frames receive either the time index of the block being generated (rotary position uL, target-aligned) or a time index reset to zero, the position of the video's first frames (block-local). Recalled RGB, masks, KV cache, conditions, and scheduler noise are fixed; only the memory's temporal position changes.



Figure 4: **Memory content and position comparison at one repeated pose (Q2, Q3).** Real and generated first-visit frames provide the fidelity and replay references. Four settings from Table 2 share one generated first visit; additional settings appear in Appendix Figure 15.

5.3 Q2: Does geometry preserve earlier appearance?

We render the merged first-visit point cloud into the query view and feed it through the canvas input that the backbone already uses; it covers almost the whole view. This comparison changes several things at once (what is stored, how much of the view is covered, and how it enters the model), so it compares working setups rather than isolating one factor. Input LPIPS is measured against the generated target. The canvas improves self-revisit LPIPS only from 0.490 to 0.431 and wins 42/55 scenes, while return wins are 47/55. Direct generated RGB through target-aligned KV reaches 0.097 self-revisit LPIPS and wins all 55 scenes on both return and self-revisit LPIPS. Nearest-source warp and the same-content merged-point-cloud KV control are reported in Appendix A.15 and Appendix A.16.

As a non-deployable diagnostic, the *GT colored canvas* is rendered from real frames sampled along the first-visit trajectory, so it carries real colors instead of generated ones. Its coverage is 0.992, compared with 0.583 for the default canvas rendered from the first frame alone, yet self-revisit LPIPS stops at 0.416, so full coverage with real colors still does not yield replay (overlay in Appendix Figure 17; details in Appendix A.19).

5.4 Q3: Target-aligned versus block-local position

The matched position rerun changes only the temporal position of the recalled frames (the capture RoPE start) and keeps every other input fixed (Figure 3). Target alignment improves return by 0.145 [0.130,0.161] LPIPS and self-revisit by 0.388 [0.357,0.418], winning all 55 scenes with no ties. Block-local is close to no history (0.589, 0.485 versus 0.589, 0.490). This isolates temporal position rather than retrieval quality and establishes that retrieval alone is insufficient on this interface. Figure 3 shows the intervention and Figure 4 the qualitative comparison; per-scene views and PSNR/SSIM are in Appendix Figure 12 and Appendix A.14. Content matters as well as position: under the same admission mask and compute, injecting shuffled sources or only the latest frame is worse than no CHaRM (Appendix A.21).

5.5 Q4: Does CHaRM restore exact-revisit recall?

The matched intervention fixes the prefix, trajectory, prompt, canvas, KV cache, and noise, changing only target-aligned history injection. It improves all 55 scenes on real-GT return and generated-

Table 3: **CHaRM exact- and shifted-revisit interventions.** (a) Exact-revisit LPIPS (lower is better). In (a), Δ is no-CHARM minus CHARM LPIPS, so positive values favor CHARM; wins count scenes where CHARM is better. (b) Shifted-revisit proxy LPIPS scores (no real frame exists at the query poses; Sec. 5.6): the same recalled frames are injected at the temporal position of the block being generated (target-aligned), with their time index reset to zero, as if they were the first frames of the video (block-local), or not at all (no memory). Scene means over 55 scenes (lower is better). Fidelity is scored against the warped real first-visit frame, recall against the warped generated first-visit frame. Block-local comes from a separate run whose own no-memory control scores 0.725 fidelity and 0.678 recall; paired comparisons are in the text.

(a) Exact revisit					(b) Shifted revisit		
Protocol	CHARM	No CHARM	Δ	Wins	Condition	Fidelity ↓	Recall ↓
Return	**0.445**	0.600	+0.154	55/55	Target-aligned	**0.673**	**0.542**
Self-revisit	**0.099**	0.504	+0.404	55/55	Block-local	0.729	0.681
Out-of-order	**0.144**	0.536	+0.392	55/55	No memory	0.724	0.675



Figure 5: **Qualitative exact revisit.** First-visit and return views share the same camera pose. Columns show the captured first-visit reference and each method's outputs under its native protocol.

reference protocols at three seeds (Table 14): return LPIPS falls from 0.600 to 0.445, matching the 0.447 single-pass level, and self-revisit from 0.504 to 0.099. Figure 5 and Appendix Figure 13 show outputs.

Because recalled frames enter only the conditional branch of classifier-free guidance, guidance could amplify their effect; with guidance disabled, CHARM still wins all 55 scenes on the three exact-revisit protocols (Appendix A.10).

The return trajectory revisits poses in the reverse of their generation order. To test recall in a different order, we use the out-of-order protocol. In it, all the 80 queries matches a stored pose, so all of them use direct recall. Recall follows query command order rather than generation order, demonstrating access to earlier appearances beyond reverse traversal. The last query returns to the first frame, which is the real input image rather than a generated one, so we score the remaining 79 pairs. Appendix Figure 8 provides the qualitative view.

A simpler alternative to CHARM is to copy the stored pixels directly. These exact-revisit protocols are diagnostics that isolate recall; on them, replaying the stored pixels would reproduce the generated reference by construction. Replayed pixels leave the generator without state, however: when the camera then leaves the stored poses, the replay produces a frame jump at the departure (boundary-pair LPIPS 0.62 and a 51° camera jump in all 55 scenes), whereas CHARM recalls the same pixels through the generator and continues without a seam (Appendix A.7).

5.6 Q5: DOES TEMPORAL-POSITION DEPENDENCE PERSIST ON SHIFTED REVISITS?

The shifted-reverse protocol uses 55 scenes and a fixed prompt. The camera first generates 81 frames along the path, then moves forward for 16 frames along its last viewing direction. On the way back, it keeps each first-visit rotation but is moved a distance l closer to the scene along that view's axis, so no return pose exactly matches a stored one.

No GT exists at these poses, so each frame is scored against the first-visit frame warped to its pose: the real one for *fidelity* and the generated one for *recall*. The same recalled frames enter at the temporal position of the block being generated (target-aligned), with their time index reset to zero, as if they were the first frames of the video (block-local), or not at all (Table 3(b)). Target-aligned memory improves fidelity in 52/55 scenes and recall in all 55; block-local is no better than no memory.

Table 4: **A 1.3B causal backbone achieves lower return and self-revisit LPIPS than the evaluated 8B–14B baselines** (55 scenes; native protocols). Single-pass/return LPIPS uses real GT; self-revisit/out-of-order uses each baselines' own generated history. WorldScore reports percentages on 32 static scenes at native resolutions. Abbreviations follow Table 1. Native protocol differ across baselines, so the table ranks revisit performance without attributing it to a single factor. †VMem: translation-jittered out-of-order trajectory. $^{\text{ref}}$minWM: temporally resampled, not pose-matched. ‡CHARM is inactive without revisits (single-pass and WorldScore).

				LPIPS ↓				WorldScore ↑	
Method	Size	Paradigm	Resolution	single-pass	return	self-revisit	out-of-order	camera	3D
Ours w/ CHARM	1.3B	PKV + BC-AR	480×832	0.447	**0.445**	**0.099**	0.144	**77.58**	81.74
Ours w/o CHARM	1.3B	PKV + BC-AR	480×832	0.447‡	0.600	0.504	0.536	**77.58**	81.74
Lyra 2.0	14B	S-AR + 3D cache	480×832	**0.411**	0.475	0.323	0.397	75.93	84.28
Warp-as-History	14B	C-AR + local noncausal	384×640	0.491	0.587	0.482	0.507	71.18	72.68
HY-WorldPlay	8B	streaming chunk	480×832	0.570	0.586	0.367	0.339	62.54	**89.39**
VMem	1.3B	image window	576×576	0.757	0.753	0.153	**0.130**†	43.24	53.08
minWM	1.3B	Action2V BC-AR	480×832	0.612$^{\text{ref}}$	–	–	–	–	–

The position rerun shares the source bank, warped references, masks, token counts, KV cache, conditions, seed. Block-local comes from a separate run. Because VAE re-encoding is not bit-identical across runs, two runs of the same setting already differ slightly; we measure this run-to-run noise with an 11-scene repeated run of the same setting. Block-local gives no gain over no memory, while target-aligned beats block-local in 52/55 scenes on fidelity and 55/55 on recall. On the 53 scenes where the warp route is active, the gap is +0.055 [+0.044, +0.067] (50/53 scenes), well outside the run-to-run noise. Thus position dependence recurs on warped earlier frames, extending the direct-route finding to the evaluated shifted-revisit protocol. Same-trajectory external context and the earlier 25-rollout build-up are in Appendices A.6 and A.8.

5.7 Revisit performance against other baselines

Table 4 compare CHARM with the evaluated 8B–14B systems under their native protocols. Lyra 2.0 (Shen et al., 2026) is the closest external comparison, matching trajectories, seed, resolution, frame counts, and evaluator pairing. Its 14B packed-window generator with a 3D cache has better single-pass fidelity, yet CHARM has lower revisit errors: stronger single-pass generation does not by itself ensure appearance persistence. The temporal-position and CHARM effects themselves are established by the matched internal interventions. WorldScore (Duan et al., 2025) provides complementary camera-control context (Appendix A.4).

Replay consistency and scene fidelity should be read together. VMem combines low generated-reference replay errors with much higher real-GT single-pass and return LPIPS; its out-of-order score uses a translation-jittered trajectory. CHARM improves recall while bringing return LPIPS back to the backbone's single-pass level (0.445 vs. 0.447), preserving scene fidelity under this protocol.

6 Limitations and Conclusion

Limitations. Our experiments cover static scenes from 2 datasets (DL3DV, WorldScore). Cross-backbone validation is currently constrained by causal architecture availability: among public baselines we evaluated, none expose the combination of persistent causal KV, continuous camera control, explicit geometric conditioning, and few-step sampling required for the matched address intervention.

Conclusion. Our experiments establish that appearance recall on the evaluated persistent-KV architecture requires the generator's own earlier RGB at the temporal position of the frame being generated. Geometry supplies spatial support, generated RGB preserves earlier appearance, and the correct temporal position makes them usable. CHARM turns this diagnosis into training-free memory while retaining causal execution. Our findings therefore characterize the memory design of the fixed-cache causal regime instantiated here. Testing the diagnosis across temporal architectures, together with independent query views, dynamic scenes (e.g., human characters), and recursive write-back, remains future work.

REPRODUCIBILITY STATEMENT

The generator interface is specified in Section 4.1; its gated residuals, training schedule, learning rates, checkpoint selection, corpus, and compute are in Appendix A.1. Section 5.1 defines the protocols; resolution, guidance, denoising levels, and fixed-prompt scope are in Appendix A.9 with preprocessing, seed, bootstrap, and route-study provenance. Appendix A.3 records external adaptations and parameter conventions. Sections 4.2 and 4.3 specify the bank, gate, routing, capture operator, and CHARM overhead. The backbone checkpoint, the three-stage training scripts, the CHARM runners, and the scene manifests are released with the code.

AI USE STATEMENT

We used AI assistants for supporting tasks: writing evaluation, job-submission scripts, proofreading manuscript text under author direction. The authors defined research questions, hypotheses, experimental design, protocol and experimental-setting definitions, choice of metrics, and the interpretation of results. The authors take responsibility for all claims and for every number reported here.

REFERENCES

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, et al. Mixture of contexts for long video generation. In *International Conference on Learning Representations*, volume 2026, pp. 92075–92093, 2026.

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

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 27713–27724, 2025.

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

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 100433–100464, 2025.

Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. *arXiv preprint arXiv:2510.03198*, 2025.

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

Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. In *International Conference on Machine Learning (ICML)*. IEEE, 2026.

Haodong Li, Shaoteng Liu, Zhe Lin, and Manmohan Chandraker. Rolling sink: Bridging limited-horizon training and open-ended testing in autoregressive video diffusion. *arXiv preprint arXiv:2602.07775*, 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*, pp. 25690–25699, 2025.

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

Jisu Nam, Yicong Hong, Chun-Hao Paul Huang, Feng Liu, JoungBin Lee, Jiyoung Kim, Siyoon Jin, Yunsung Lee, Jaeyoon Jung, Suhwan Choi, et al. Worldcam: Interactive autoregressive 3d gaming worlds with camera pose as a unifying geometric representation. *arXiv preprint arXiv:2603.16871*, 2026.

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.

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

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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. In *Forty-third International Conference on Machine Learning*.

Team HunyuanWorld. HY-World 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. Technical report, Tencent Hunyuan, 2025. URL https://3d-models.hunyuan.tencent.com/world/world1_5/HYWorld_1.5_Tech_Report.pdf.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv: Arxiv-2305.16291*, 2023.

Haoyuan Wang, Yabo Chen, Haibin Huang, Chi Zhang, and Xuelong Li. Directing the world: Fast autoregressive video generation with compositional human-camera control. *arXiv preprint arXiv:2606.27964*, 2026a.

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.

Yifan Wang and Tong He. Warp-as-history: Generalizable camera-controlled video generation from one training video. *arXiv preprint arXiv:2605.15182*, 2026.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *Advances in Neural Information Processing Systems*, 38:49371–49393, 2026.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *Advances in Neural Information Processing Systems*, 38:49632–49652, 2026.

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

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, YINGCONG CHEN, Yao Lu, et al. Longlive: Real-time interactive long video generation. In *International Conference on Learning Representations*, volume 2026, pp. 90037–90059, 2026.

Hidir Yesiltepe, Tuna 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*, pp. 40256–40265, 2026.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 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*, 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*, pp. 22963–22974, 2025a.

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

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.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–18, 2025b.

Yonghao Yu, Lang Huang, Runyi Li, Zerun Wang, and Toshihiko Yamasaki. Video-mirai: Autoregressive video diffusion models need foresight. *arXiv preprint arXiv:2606.03971*, 2026.

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

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*, 2026a.

Yizhou Zhao, Yifan Wang, Xiaoyuan Wang, Yushu Wu, Hao Zhang, Moayed Haji-Ali, Rameen Abdal, Ashkan Mirzaei, Yanyu Li, Willi Menapace, et al. Geostream: Toward precise camera controlled streaming video generation. *arXiv preprint arXiv:2606.15162*, 2026b.

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rcm: A unified teacher-forcing and self-forcing open recipe for autoregressive diffusion distillation in streaming video generation and interactive world models. *arXiv preprint arXiv:2606.25473*, 2026.

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

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. In *International Conference on Machine Learning (ICML)*, 2026b.