WORLD IN WORLD: EXPLORE THE WORLD WITH WORLD MODELS

Anonymous authors
Paper under double-blind review



Figure 1: **WORLD IN WORLD: Explore the world within a video.** Video generation becomes an ongoing exploration of an evolving visual world. In WORLD IN WORLD, this exploration is grounded in the world of a given video, with the recorded event anchoring the appearance and dynamics of each new observation.

ABSTRACT

Recent advances in video world models enable interactive generation beyond fixed-length synthesis, yet extending pretrained models to new forms of control typically requires control-specific architectures or additional training. We introduce **World in World (WiW)**, a training-free visual-evidence interface that flexibly extends the controllability of frozen causal video world models. WiW leverages native self-attention, which already consumes clean visual states, as a shared interface for heterogeneous external controls without modifying the pretrained backbone. It converts source observations, target-view projections, rendered geometry, and generated history into clean visual evidence annotated with camera, temporal, and spatial-validity information. These sources provide appearance references, spatial guidance, completion cues for newly visible regions, and long-range context for revisiting generated states. We introduce correspondence-guided attention routing (CGAR) to direct queries toward geometrically corresponding visual tokens, and evidence-wise attention CFG (EWA) to regulate the influence of different evidence sources within native self-attention without additional network function evaluations for guidance. We instantiate WiW on a frozen LingBot-World 2.0 backbone and evaluate camera-controlled video rerendering on DAVIS and OpenVid-1M across diverse viewpoint changes. Beyond rerendering, WiW enables exploration within a given video's dynamic world and supports bullet-time generation, video stabilization, video editing, and cross-generation K/V sharing, demonstrating a unified and versatile framework for training-free world-model control.

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

Recent advances in video generation have improved visual quality, motion realism, and temporal coherence, and have supported the development of video world models built on autoregressive generation (Kong et al., 2025; Huang et al., 2025c; Robbyant Team et al., 2026; Gao et al., 2026b; Xu et al., 2026a; InSpatio Team et al., 2026). While most text- or image-conditioned systems generate a finite clip from conditions specified before inference, video world models support an ongoing interactive process. They continually predict subsequent observations from observed or generated visual states, allowing users to change camera positions, viewing directions, and control actions throughout a rollout (Sun et al., 2025; AlayaWorld Team et al., 2026). Video generation thus extends from producing a recording of an event to supporting exploration of an evolving visual world, with applications in interactive content creation, virtual production, game generation, and embodied-agent simulation (Li et al., 2025a; Zhang et al., 2025; He et al., 2025; Wang et al., 2026b; Qian et al., 2026). In particular, as large pretrained world models acquire stronger visual and motion priors, a key question is how to flexibly extend their controllability without retraining, enabling the same general world model to accept diverse forms of visual control and world information and make fuller use of its pretrained capabilities.

Supporting such diverse controls requires a world model to integrate complementary visual evidence from observations, geometry, and generated history, which differ in representation, spatial coverage, and temporal relevance. As viewpoints and scene states change, generation must remain synchronised with the dynamics specified by the visual evidence while maintaining correct spatial placement and occlusion relationships in the target view. Large viewpoint changes require using the model's generative prior to complete unobserved scene and object surfaces, while long-horizon revisits require recovering earlier appearance and spatial layout. The challenge is therefore to make relevant evidence available and use it at the appropriate locations and times throughout a rollout.

Existing approaches have explored camera conditioning, source-video rerendering, geometric control, and long-term memory (Zhou et al., 2026; InSpatio Team et al., 2026; Li et al., 2026b; Yu et al., 2025b; Bai et al., 2025; Xu et al., 2026a; Xiao et al., 2025). Many rely on control-specific pathways or representations, so supporting new evidence types or combinations often requires redesign or additional training. This motivates a shared interface through which the same pretrained world model can use heterogeneous current and historical visual evidence without learning a separate pathway for each control, which is the focus of our paper.

Our key observation is that causal video world models equipped with a clean-state cache already have a shared entry point for visual information: their native self-attention. During autoregressive generation, the model reads the initial observation and recently finalised outputs as clean visual states, while camera and temporal encodings specify their viewpoints and temporal positions. This suggests that external control information can be converted into the same representation: clean visual states associated with camera poses, event times, and valid spatial regions, made accessible through the existing attention layers. Instead of adapting the model to each new control, we express different controls in a visual representation that the pretrained model already processes. Under this view, extending world-model control becomes a visual evidence construction and orchestration problem, rather than a model adaptation problem.

Based on this idea, we introduce WORLD IN WORLD (WiW), a training-free visual-evidence interface for controlling frozen causal video world models (Figure 1). WiW converts various visual conditions, such as source observations, target-view projections, rendered geometry, and generated history, into clean visual states annotated with camera, temporal, and spatial-validity information. These states can be directly accessed through the model's native self-attention, allowing different sources of evidence to guide generation where and when they are reliable. Our method is entirely training-free, requiring neither updates to the pretrained world model nor learned control-specific modules. Under this unified interface, flexible world-model control reduces to three complementary problems: constructing useful visual evidence, localising the relevant evidence, and regulating its influence on generation.

To provide the frozen world model with the complementary information required for controllable long-horizon rollouts, we first construct multiple forms of visual evidence for different control requirements. Our core idea is to let each evidence source contribute the information it can provide most reliably, while expressing all of them through the same clean-state interface that the pretrained

model already understands. Concretely, source-video observations provide appearance and content references from the recorded event. Since these observations do not explicitly determine where their content should appear under the target camera, *target-view scene evidence* uses depth-based projection to place observed appearance in the requested view and provide spatial guidance within geometrically supported regions. When large camera motions expose subject surfaces that are absent from the source observations, *rendered geometry evidence* supplies target-view shape and appearance proposals for the corresponding event state, helping the pretrained model complete newly visible subject surfaces. Finally, because a finite rolling cache eventually removes access to earlier generated states, we maintain a rollout-wide history and retrieve relevant and diverse archived states as *generated evidence* when the camera returns. By assigning different control requirements to complementary evidence sources, WiW provides the frozen backbone with appearance references, spatial guidance, completion cues, and long-range visual context through a single shared visual interface.

To ensure that the model reads the right evidence and uses it with appropriate strength, we further regulate how visual evidence participates in native self-attention. Our core idea is to separately control *where to read* and *how strongly to use* the available evidence, addressing errors in evidence localisation and differences in evidence reliability. For *where to read*, appearance-based attention can become ambiguous under large viewpoint changes, dynamic motion, and repeated textures. We therefore introduce *correspondence-guided attention routing* (CGAR), which uses persistent point identities and camera geometry to route current queries towards geometrically corresponding source-video tokens when valid matches are available. For *how strongly to use*, different evidence sources can vary in reliability across spatial regions and denoising stages. We introduce *evidence-wise attention CFG* (EWA), which compares native and evidence-conditioned attention responses to selectively amplify compatible and complementary information from each evidence channel. Both mechanisms operate directly within the model's native self-attention, and EWA reuses responses from the same denoising forward pass, introducing no additional network function evaluations (NFE) for guidance.

Through this formulation, WiW provides a unified, training-free framework for extending the control capabilities of frozen world models through visual evidence, without introducing control-specific training or adaptation. Notably, beyond exploring worlds freely generated by a world model, WiW enables exploration *within the world of a given video*: users can navigate the recorded dynamic world along new camera trajectories while preserving the appearance and temporal progression of the original event. We instantiate WiW on the publicly released causal-fast checkpoint of LingBot-World 2.0 (Gao et al., 2026b), with all pretrained parameters frozen, and evaluate camera-controlled video rerendering on DAVIS and OpenVid-1M (Pont-Tuset et al., 2017; Nan et al., 2024) across diverse viewpoint changes. Beyond rerendering, WiW supports applications including bullet-time generation, video stabilization, video editing, and K/V sharing between two generation cases produced by the same frozen model, demonstrating the versatility of the proposed interface.

Our main contributions are:

- We introduce WiW, a training-free visual-evidence interface for flexibly extending the control capabilities of frozen causal video world models.
- We develop complementary visual evidence that enables event-synchronised, spatially aligned, and long-horizon-consistent exploration of dynamic video worlds.
- We introduce correspondence-guided attention routing (CGAR) and evidence-wise attention CFG (EWA) to localise and regulate visual evidence in native self-attention.
- We demonstrate flexible exploration of given video worlds across diverse camera trajectories and downstream applications using a single frozen world-model backbone.

2 Related work

2.1 World models

A video world model predicts the visual observations an observer would receive while moving through an environment, given an initial observation and a requested camera or action. In static environments, a key requirement is to maintain consistent scene structure and appearance as the viewpoint changes and previously seen regions are revisited. One line of work uses explicit scene

representations, such as point clouds, meshes, or Gaussians, to organize observations in a shared coordinate system and render them from the target camera. These representations provide a clear spatial reference, but require completing unobserved regions and maintaining the scene state as exploration continues (Kong et al., 2025; Yu et al., 2024b;a; Liang et al., 2024; Wang et al., 2024; Yu et al., 2024c; Chen et al., 2025c; Huang et al., 2025e; Ren et al., 2025; Bahmani et al., 2025; Huang et al., 2026; Gao et al., 2026a; Li et al., 2025c; Team HY-World et al., 2026). Another line of work directly predicts target-view observations with video models, using learned generative priors to handle viewpoint changes and newly visible regions (Sun et al., 2024; Zhao et al., 2024; Wang et al., 2025a; Zhu et al., 2026; Höllein & Nießner, 2026). When the environment changes over time, generation must remain consistent with both the target viewpoint and the evolving event state. Previous work has studied this problem through dynamic scene generation, source-video re-observation, and continuous world modeling (Chen et al., 2025f; Dai et al., 2025; Xiang et al., 2026b; Zheng et al., 2026; Bai et al., 2025; InSpatio Team et al., 2026; Paliwal et al., 2026; Xu et al., 2026b; Lin et al., 2026; Hu et al., 2025a; Lu et al., 2025). Autoregressive methods further extend scene generation to continuous rollouts, requiring models to maintain consistent appearance, spatial relationships, and event states over long explorations (Chen et al., 2024; Huang et al., 2025d;c; Mao et al., 2025; Chen et al., 2025b; AlayaWorld Team et al., 2026; Gao et al., 2026b). Pretrained video models have appearance and motion priors that generalize to new scenes, but their native conditioning interfaces are usually determined by the inputs used during training. Warp-as-History (Wang & He, 2026) uses temporally aligned warped observations as visual history and further improves camera control through LoRA finetuning. WiW studies how source observations, projected views, rendered geometry, and retrieved history constrain exploration of a recorded event. CGAR localises source evidence, while EWA regulates each channel's additional contribution within a frozen backbone.

2.2 Conditioning mechanisms for video world models

Existing video world models usually use dedicated conditioning mechanisms to introduce different types of control into generation. For camera control, prior methods represent camera trajectories as rays, poses, positional encodings, projected features, or rendered geometric proxies, and use corresponding conditioning modules to guide generation (Jang et al., 2026; Li et al., 2026a; Xiang et al., 2026a; Lee et al., 2025; Liu et al., 2026; Wu et al., 2025; Gu et al., 2025; Kim et al., 2025). Source-video rerendering extends control from camera motion to re-observation of dynamic content. These methods often train additional modules for source-video inputs to bring the original event's appearance and dynamics into the target view (Bai et al., 2025; Chen et al., 2025e; Seo et al., 2025; InSpatio Team et al., 2026; Wang et al., 2026a). Similarly, the concurrent work Wonder supports image- and video-conditioned world generation by jointly training a rendered control field, a sparse memory, and a distilled causal student to combine multiple conditions (Xu et al., 2026a). Such task-specific conditioning pathways often require different input interfaces and training procedures. Beyond current observations and external controls, continuous exploration also requires access to scene information generated earlier. Since causal models typically read only a limited recent context, prior work uses explicit geometric memory, persistent states, or historical information retrieval to recover scene content beyond the current context and constrain subsequent generation (Li et al., 2025b; Wang et al., 2025b; Wei et al., 2026; Huang et al., 2025b; Xiao et al., 2025; Yu et al., 2025a; Joo et al., 2026; Yi et al., 2026). Memory therefore helps maintain long-term consistency and can also serve as a visual condition alongside current observations and geometry. WiW follows this idea by including historical information in a unified visual conditioning framework, allowing scene evidence from different sources to jointly guide subsequent generation.

3 Method

We propose World in World (WiW), a unified visual-evidence interface for extending the control capabilities of frozen causal video world models. We illustrate WiW through camera-controlled video rerendering: given a source video and target camera trajectory C, we rerender the dynamic event from requested viewpoints while preserving its appearance and temporal progression. To this end, we convert source observations, geometry, and generated history into clean visual states with camera, temporal, and spatial-validity information, which the frozen model reads through native self-attention without additional training or learned control-specific adapters.

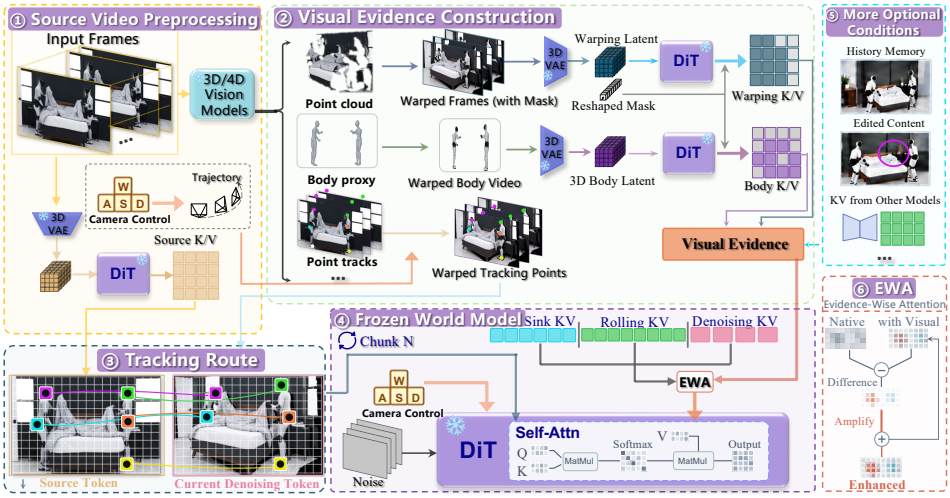

Figure 2: **Overview of the WORLD IN WORLD pipeline.** Target-view projections, rendered geometry, and retrieved history provide complementary visual evidence. Correspondence-guided attention routing connects queries to matching source-video tokens, while evidence-wise attention CFG regulates each evidence channel's contribution. Auxiliary evidence is removed after each chunk; finalized outputs enter the rolling cache, and their K/V features are archived upon eviction.

Figure 2 summarizes the overall pipeline. We first define the shared evidence interface and explain how visual evidence participates in native self-attention (Section 3.1). Building on this interface, we project source observations into the target view to provide layout references (Section 3.2), use rendered geometry to guide completion of newly exposed subject surfaces (Section 3.3), and retrieve generated history for consistent long-horizon revisits (Section 3.4). Together, these sources provide complementary references for generation. Attention routing and evidence-wise attention CFG then localize relevant evidence and regulate its influence, respectively (Section 3.5).

3.1 CONTROL THROUGH VISUAL EVIDENCE

When generating the current chunk, our causal video backbone reads retained generated anchors and recently finalized states through native self-attention. We use this existing pathway to convert multiple forms of visual evidence, including source-video observations, target-view scene projections, rendered geometry, and generated history, into the same type of states that the model already reads. These sources differ in representation, valid spatial extent, and temporal coverage. Their shared representation must therefore specify evidence content, camera and temporal information, spatial support, and channel activation across denoising stages.

To represent this information consistently, at denoising step s, converter Ψ_b maps the raw evidence E_b of channel b and target camera trajectory C to:

$$\Psi_b(E_b; C, s) = \left(X_b,\, P_b,\, \boldsymbol{\omega}_b,\, \gamma_b^{(s)}\right). \tag{1}$$

Here, X_b denotes the visual content of the evidence; P_b specifies the camera condition and event-time indices used to encode evidence channel b; $\omega_{bj} \in [0, 1]$ specifies the spatial support of evidence token j; and $\gamma_b^{(s)} \in \{0, 1\}$ determines whether the channel is active at step s.

To encode external image evidence into attention features, we feed the clean visual states corresponding to X_b into the frozen video backbone at diffusion timestep $t = 0$, using the channel-specific camera condition and event-time alignment specified by P_b. A single network forward pass extracts the keys K_b^ℓ and values V_b^ℓ from each attention layer, which we cache as *clean K/V*. During target-chunk generation, queries from the current denoising states read these cached features through self-attention, allowing external visual evidence to guide generation. External image evidence introduced below uses this procedure to obtain attention features.

These K/V features participate in attention through either the native block or temporary auxiliary blocks. The native attention cache accommodates up to 18 latent-frame states. Once the cache is filled, it retains the first six generated states as fixed anchors, eight recent-history states, and four states in the current chunk. Source-video K/V is accessed through a separate correspondence-routing branch, while the other auxiliary evidence blocks are attached temporarily and removed after the current chunk is finalized. All blocks reuse the frozen backbone's query/key/value (Q/K/V) projections, while preserving its native text- and camera-conditioning pathways. When visual evidence enters attention, its temporal relationship to the current generation states needs to be specified. We therefore set the temporal coordinates of rotary position embeddings (RoPE) and apply the corresponding rotations to attention queries and keys, incorporating relative temporal information into attention. Spatial support weights regulate the contribution of evidence tokens by scaling their unnormalized attention weights by ω_{bj}, while channel activation $\gamma_b^{(s)}$ further determines the denoising stages in which each evidence source participates.

3.2 Target-View Scene Evidence

Source-video observations provide appearance references for the recorded event, but they remain in the source-camera view and do not explicitly determine where their content should appear in the target image. Camera motion changes projected positions and occlusion relationships, while ambiguous cross-view appearance matches can cause visible structures to drift. To provide an explicit spatial layout reference, we project source observations into the target view, allowing the model to read observed appearance aligned with the requested camera.

To construct this reference for target frame m and temporally aligned source view r, we use source RGB $I^{\mathrm{src}}_{r,m}$ and estimate its depth $D^{\mathrm{src}}_{r,m}$ with an off-the-shelf depth estimator. Source-camera intrinsics and poses are obtained during depth preprocessing. We back-project source pixels into 3D and reproject them from source camera $C^{\mathrm{src}}_{r,m}$ to target camera $C_m = (A_m, T_m)$, where A_m denotes camera intrinsics and T_m is the camera-to-world pose:

$$\left(X^{\mathrm{proj}}_{r,m}, M^{\mathrm{proj}}_{r,m}, Z^{\mathrm{proj}}_{r,m}\right) = \mathrm{Proj}\left(I^{\mathrm{src}}_{r,m}, D^{\mathrm{src}}_{r,m}; C^{\mathrm{src}}_{r,m} \to C_m\right). \tag{2}$$

Projection uses a shared coordinate system and scale, with scale normalization referenced to the depth estimator's output. The outputs $X^{\mathrm{proj}}_{r,m}$, $M^{\mathrm{proj}}_{r,m}$, and $Z^{\mathrm{proj}}_{r,m}$ are target-view RGB, a binary visibility mask, and target-camera depth, respectively. Following the shared interface, we encode the projected RGB with target-camera and event-time information. We convert the visibility mask into token-level support weights $\boldsymbol{\omega}_{\mathrm{proj}}$ using the visible-pixel fraction within each token, so that tokens with greater visible coverage receive higher weights.

3.3 Rendered Geometry Evidence

Target-view projection reorganizes content already present in the source observations, but its coverage remains limited by source-view visibility. When camera motion exposes subject surfaces not observed in the source video, these regions lack direct shape and appearance references, and their generation may deviate from the subject's structure or current pose. To provide evidence for completing such regions, we use a renderable subject representation to produce shape and appearance proposals from the target camera at the corresponding event time. Given subject state G_m at event time m and target camera C_m, rendering yields:

$$(X^{\mathrm{geo}}_m, M^{\mathrm{geo}}_m, Z^{\mathrm{geo}}_m) = \mathrm{Render}(G_m, C_m). \tag{3}$$

Here, X^{geo}_m, M^{geo}_m, and Z^{geo}_m denote RGB, a geometric support mask, and target-camera depth, respectively. The rendering is encoded through the shared interface, and its support mask is converted into token-level support weights. Taking human geometry as an example, we instantiate $G_m = (\mathcal{A}, \Theta_m)$ using an avatar reconstructed with LHM++ (Qiu et al., 2025) and driven by per-frame SMPL-X (Pavlakos et al., 2019) parameters. The resulting renderings provide target-view appearance, support, and depth cues; reconstruction and alignment details are provided in Appendix C.

3.4 Retrieved Generated Evidence

Scene projections and geometry renderings provide observed and reconstruction-based evidence for the current target view. As generation progresses, previously generated appearance and layout

also become important references for subsequent frames. The native rolling cache retains fixed generated anchors and a bounded recent history, so states corresponding to an earlier region may have been evicted when the camera returns, leaving their generated content inaccessible to the model. To recover these references, we maintain a rollout-wide history bank and retrieve finalized states relevant to the current view.

The history bank accumulates evidence as the rolling cache is updated. For each evicted finalized state, we archive its per-layer clean K/V with the associated camera pose and temporal index. Since the bank grows throughout generation, we select a bounded number of relevant states to participate in attention for each target chunk. Specifically, we rank historical states by target-view surface coverage and viewing-direction compatibility, then select a diverse top-K_{ret} subset to reduce redundant references. The model reads the retrieved K/V directly through temporary auxiliary attention blocks, recovering historical appearance and layout beyond the rolling cache to maintain scene consistency.

3.5 Attention Routing and Guidance

The construction and retrieval steps above provide complementary evidence for the source event, target-view layout, subject geometry, and generated history. However, making evidence accessible does not ensure that queries locate the relevant positions within it. Viewpoint changes and dynamic motion alter the appearance of a surface, while repeated textures can make distinct locations look similar, introducing ambiguity into appearance-based attention. We therefore first use geometric correspondences to match current generation tokens with source-video tokens and guide evidence reading, then regulate each channel's additional influence through its attention response.

Correspondence-guided attention routing. We introduce correspondence-guided attention routing (CGAR) as a dedicated attention branch for reading clean K/V features from the unwarped source video. Within this branch, geometric correspondences determine which source video tokens each query can access. We track points with persistent identities in the source video and combine depth and camera information to establish token-level correspondences between the current view and the source video. For a geometrically matched query i and source-video key j, we add the logarithm of correspondence weight $\beta_{ij} > 0$ to the routing attention score:

$$\ell_{ij}^{\mathrm{corr}} = \frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d_h}} + \log \beta_{ij}. \tag{4}$$

Here, d_h is the attention-head dimension, and $\beta_{ij} > 0$ controls the contribution of the correspondence. We combine the routing response with the other attention-block responses through joint normalization, allowing current queries to read geometrically matched source-video evidence.

Evidence-wise attention CFG. The evidence sources constructed above participate in video generation through self-attention during denoising. Different auxiliary evidence channels provide constraints on target-view layout, subject structure, and generated history, and the information they provide also differs in reliability and importance. Their influence on generation therefore needs to be controlled independently. Using all evidence with the same strength makes it difficult to balance adherence to different conditions with the model's generative prior. Inspired by classifier-free guidance (CFG) (Ho & Salimans, 2022) and NAG (Chen et al., 2025a), we propose evidence-wise attention CFG (EWA). While NAG guides attention using positive and negative text prompts, EWA adjusts each visual evidence channel's guidance against a shared native response to strengthen useful control signals while preserving visual quality.

Specifically, at the same attention layer and for the same queries, let $\mathbf{o}_0$ denote the response obtained by attending only to the native block, and let $\mathbf{o}_e$ denote the jointly normalized response obtained by attending to both the native block and evidence e. To form the correction, we subtract the component of $\mathbf{o}_e$ along $\mathbf{o}_0$ and scale the remainder by their nonnegative cosine similarity and the channel-specific guidance strength g_e. The updated response is:

$$\widetilde{\mathbf{o}}_e = \mathbf{o}_e + g_e \, \max(0, \cos(\mathbf{o}_e, \mathbf{o}_0)) \left[\mathbf{o}_e - \mathrm{Proj}_{\mathbf{o}_0}(\mathbf{o}_e)\right]. \tag{5}$$

Here, $\cos(\mathbf{o}_e, \mathbf{o}_0)$ is the cosine similarity between the two responses. The projection of $\mathbf{o}_e$ onto $\mathbf{o}_0$ is $\mathrm{Proj}_{\mathbf{o}_0}(\mathbf{o}_e) = (\|\mathbf{o}_e\|_2 / \|\mathbf{o}_0\|_2) \cos(\mathbf{o}_e, \mathbf{o}_0)\, \mathbf{o}_0$. When the two responses differ sharply in direction, the nonnegative cosine term adaptively reduces the correction to limit the impact of extreme updates on visual quality. We add the channel-specific corrections $\widetilde{\mathbf{o}}_e - \mathbf{o}_e$ to the joint response of all active

attention blocks. We bound the combined output magnitude using the norm of $\mathbf{o}_0$ as a reference to balance evidence guidance with the native output scale. EWA reuses attention-block outputs and normalization results from the same denoising forward pass, adding no network function evaluations (NFE) for guidance. Detailed parameter settings are provided in Appendix C.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Implementation. We instantiate WORLD IN WORLD on the publicly released causal-fast checkpoint of LingBot-World 2.0 (Gao et al., 2026b), keeping all pretrained parameters frozen. At inference time, visual evidence is constructed from source observations, target-view projections, rendered geometry, and retrieved history, and exposed to the model through its native self-attention. No additional training or learned control-specific adapters are required.

Baselines. Following prior work on camera-controlled video rerendering, we evaluate on 400 video–trajectory pairs from DAVIS (Pont-Tuset et al., 2017) and OpenVid-1M (Nan et al., 2024), and compare with ReCamMaster, TrajectoryCrafter, WorldForge, InSpatio-World, UniWorld-View, and CameraAnything (Bai et al., 2025; Yu et al., 2025b; Song et al., 2026; InSpatio Team et al., 2026; Zhou et al., 2026; Li et al., 2026b). All baselines use their official implementations, released checkpoints, and recommended configurations. Details on frame counts, camera conventions, output alignment, and metric preprocessing are provided in Appendix B.

Metrics. We evaluate image fidelity using PSNR, SSIM, and LPIPS. To assess video quality, temporal consistency, and dynamics, we use the official VBench implementation (Huang et al., 2023) to report seven dimensions: Aesthetic Quality, Imaging Quality, Temporal Flickering, Motion Smoothness, Subject Consistency, Background Consistency, and Dynamic Degree. We additionally report *Overall*, computed as the average score of these seven dimensions. Following InSpatio-World (InSpatio Team et al., 2026), we evaluate camera control using translation error (TransError) and rotation error (RotError). We independently estimate camera trajectories from the generated videos using Depth Anything 3 and ViPE (Lin et al., 2025; Huang et al., 2025a). For each estimator, we compare the estimated trajectories with the prescribed target trajectories. The reported TransError and RotError are obtained by averaging the corresponding errors across the two estimators.

4.2 COMPARISON RESULTS

We compare WORLD IN WORLD with prior state-of-the-art methods quantitatively and qualitatively. Table 1 reports camera-controlled video rerendering results on DAVIS and OpenVid-1M across VBench, camera trajectory errors, and image-quality metrics. WORLD IN WORLD achieves the highest overall VBench score, the lowest camera trajectory errors, and the best or tied-best image-quality results among the compared methods. Figure 3 presents qualitative comparisons under the same source video and target camera path. WORLD IN WORLD better preserves subject appearance and scene structure, follows the requested viewpoints, and plausibly completes newly exposed regions under diverse camera motions. Additional applications of the shared visual-evidence interface are presented in Appendix D.

4.3 ABLATION STUDIES

We examine how the choice and use of visual evidence contribute to controllable generation. In *w/o CGAR*, we omit the additional attention response from the source-video branch. Table 2 shows that the full method achieves the lowest camera errors and the best or tied-best VBench and image-quality results. Compared with target-view warping alone, the full method improves PSNR by 2.06 dB under the same backbone. Accurate camera control requires a clear spatial reference. Removing target-view warping causes the largest degradation across all metrics, with rotation and translation errors rising to approximately $3.4\times$ and $10.9\times$ their full-method values. We combine the source and requested camera trajectories to construct the Plücker condition. Omitting the source-camera motion treats the source camera as static and increases camera errors. Additional qualitative ablations in Appendix E examine spatial evidence, camera conditioning, evidence routing, rendered geometry, and historical retrieval.

Table 1: Quantitative comparison for camera-controlled video rerendering on DAVIS and OpenVid-1M. **Bold** and underline denote the best and second-best results, respectively.

	VBench ↑								Camera errors ↓		Image quality		
Method	Aesth.	Img.	Flick.	Smooth.	Subj.	Bg.	Dyn.	Overall	TransError	RotError (°)	PSNR ↑	SSIM ↑	LPIPS ↓
ReCamMaster	54.283	66.290	**95.178**	97.503	88.941	92.154	92.500	83.836	0.1243	7.8626	15.14	0.444	0.378
TrajectoryCrafter	51.322	61.466	94.169	97.423	87.032	90.727	**98.750**	82.984	0.0908	5.3254	19.97	0.625	0.218
WorldForge	49.915	62.356	94.262	96.934	84.683	90.017	98.500	82.381	0.0800	4.5021	19.65	0.589	0.218
InSpatio-World	53.162	66.076	93.622	96.764	87.070	91.188	96.250	83.447	0.0824	4.4011	20.69	0.613	0.183
UniWorld-View	54.237	67.534	93.444	96.627	88.946	91.777	97.500	84.295	0.0687	4.1958	22.47	0.771	**0.122**
CameraAnything	54.365	67.526	94.489	96.908	87.399	91.817	88.750	83.036	0.1864	3.1868	15.16	0.454	0.342
Ours	**56.556**	**68.004**	93.711	**97.749**	**88.948**	**92.627**	**98.750**	**85.192**	**0.0686**	**2.8326**	**23.15**	**0.787**	**0.122**



Figure 3: Qualitative comparisons under diverse camera motion. All methods receive the same source video and target camera path. Representative frames illustrate subject and scene consistency, viewpoint control, and completion of newly exposed regions.

Table 2: Ablations for camera-controlled video rerendering on DAVIS. We evaluate EWA, CGAR, target-view warping, and source-camera Plücker conditioning with VBench, camera-control, and image-quality metrics. **Bold** and underline denote the best and second-best results, respectively.

	VBench ↑					Camera errors ↓		Image quality		
Variant	Subj.	Bg.	Smooth.	Aesth.	Flick.	RotError (°)	TransError	PSNR ↑	SSIM ↑	LPIPS ↓
w/o EWA and CGAR	**88.076**	91.747	96.054	50.688	92.857	2.0346	0.0569	21.49	0.668	0.157
w/o EWA	88.056	91.706	96.055	**50.720**	92.867	1.9944	0.0566	21.52	0.669	0.156
w/o CGAR	88.055	91.728	96.072	50.713	92.891	1.7828	0.0539	22.88	0.720	0.138
w/o Target-View Warp	83.378	89.540	94.351	50.240	90.792	6.1158	0.5819	13.26	0.331	0.455
w/o Plücker	88.066	91.692	96.073	50.669	92.881	1.8811	0.0569	22.90	0.722	0.139
Only Target-View Warp	88.035	91.662	95.885	50.655	92.844	2.2781	0.0879	20.92	0.601	0.179
Full method	**88.076**	**91.749**	**96.074**	**50.720**	**92.902**	**1.7823**	**0.0533**	**22.98**	**0.727**	**0.133**

5 Conclusion

We presented World in World, a training-free interface extending the controllability of frozen causal video world models. WiW represents source observations, target-view projections, rendered geometry, and generated history as camera- and time-labelled clean visual states read through native self-attention. CGAR localizes source-video evidence, while EWA regulates auxiliary contributions using same-pass attention responses without extra denoising-network evaluations for guidance. For camera-controlled video rerendering on DAVIS and OpenVid-1M, WiW achieves the highest average VBench score across seven dimensions and the lowest camera trajectory errors among compared methods. Qualitative results further demonstrate downstream applications of the shared interface. These findings show how constructing, selecting, and regulating visual evidence extends pretrained world models to explore dynamic worlds depicted in videos.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, grammar correction, and improvements to the clarity and readability of the manuscript. All AI-assisted text was reviewed, modified where necessary, and verified by the authors. The authors take full responsibility for the final content of the paper and the correctness of the accompanying implementation.

REFERENCES

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Mingliang Zhai, Ruicong Liu, Xiaojie Xu, Xuangeng Chu, Zhen Li, Zhengyuan Lin, Zhixiang Wang, Zian Meng, and Zihui Gao. AlayaWorld: Interactive long-horizon world modeling—full technical report. *arXiv preprint arXiv:2607.18367*, 2026.

Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B. Lindell, Zan Gojcic, Sanja Fidler, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra: Generative 3D scene reconstruction via video diffusion model self-distillation. *arXiv preprint arXiv:2509.19296*, 2025.

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

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 (NeurIPS)*, volume 37, 2024.

Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. Normalized attention guidance: Universal negative guidance for diffusion models, 2025a. URL https://arxiv.org/abs/2505.21179.

Junyi Chen, Haoyi Zhu, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, and Tong He. Deepverse: 4D autoregressive video generation as a world model. *arXiv preprint arXiv:2506.01103*, 2025b.

Luxi Chen, Zihan Zhou, Min Zhao, Yikai Wang, Ge Zhang, Wenhao Huang, Hao Sun, Ji-Rong Wen, and Chongxuan Li. Flexworld: Progressively expanding 3D scenes for flexiable-view synthesis. *arXiv preprint arXiv:2503.13265*, 2025c.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video Depth Anything: Consistent depth estimation for super-long videos. *arXiv preprint arXiv:2501.12375*, 2025d. URL https://arxiv.org/abs/2501.12375.

Yipeng Chen, Zhichao Ye, Zhenzhou Fang, Xinyu Chen, Xiaoyu Zhang, Jialing Liu, Nan Wang, Guofeng Zhang, and Haomin Liu. Postcam: Camera-controllable novel-view video generation with query-shared cross-attention. *arXiv preprint arXiv:2511.17185*, 2025e.

Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, and Ziwei Liu. 4DNeX: Feed-forward 4D generative modeling made easy. *arXiv preprint arXiv:2508.13154*, 2025f.

Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3D prediction. *arXiv preprint arXiv:2509.21657*, 2025.

Sensen Gao, Zhaoqing Wang, Qihang Cao, Dongdong Yu, Changhu Wang, and Jia-Wang Bian. Pixworld: Unifying 3D scene generation and reconstruction in pixel space. *arXiv preprint arXiv:2607.05373*, 2026a.

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

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, Wenping Wang, and Yuan Liu. Diffusion as shader: 3D-aware video diffusion for versatile video generation control. *arXiv preprint arXiv:2501.03847*, 2025.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-Game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025. doi: 10.48550/arXiv.2508.13009. URL https://arxiv.org/abs/2508.13009.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Lukas Höllein and Matthias Nießner. World reconstruction from inconsistent views. *arXiv preprint arXiv:2603.16736*, 2026.

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

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *CVPR*, 2025b.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixé, and Sanja Fidler. ViPE: Video pose engine for 3D geometric perception. *arXiv preprint arXiv:2508.10934*, 2025a.

Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3R: 3D scene generation meets feed-forward reconstruction. *arXiv preprint arXiv:2601.04090*, 2026.

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*, 2025b.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2World: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv:2505.14357*, 2025c.

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, 2025d.

Yuanhui Huang, Weiliang Chen, Wenzhao Zheng, Xin Tao, Pengfei Wan, Jie Zhou, and Jiwen Lu. Terra: Explorable native 3D world model with point latents. *arXiv preprint arXiv:2510.14977*, 2025e.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. VBench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*, 2023.

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

Wonbong Jang, Shikun Liu, Soubhik Sanyal, Juan Camilo Perez, Kam Woh Ng, Sanskar Agrawal, Juan-Manuel Perez-Rua, Yiannis Douratsos, and Tao Xiang. Rays as pixels: Learning a joint distribution of videos and camera trajectories. In *International Conference on Machine Learning (ICML)*, 2026.

Minseok Joo, Dogyun Park, Taehoon Lee, Kyujin Lee, and Hyunwoo J. Kim. Retrieve what's missing: Coverage-maximizing retrieval for consistent long video generation. *arXiv preprint arXiv:2606.02479*, 2026.

Geonung Kim, Janghyeok Han, and Sunghyun Cho. VideoFrom3D: 3D scene video generation via complementary image and video diffusion models. *arXiv preprint arXiv:2509.17985*, 2025.

Lingdong Kong, Yu Yang, Jianbiao Mei, Youquan Liu, Ao Liang, Dekai Zhu, Dongyue Lu, Wei Yin, Xiaotao Hu, Mingkai Jia, Junyuan Deng, Kaiwen Zhang, Yang Wu, Tianyi Yan, Shenyuan Gao, Song Wang, Linfeng Li, Liang Pan, Yong Liu, Jianke Zhu, Wei Tsang Ooi, Steven C. H. Hoi, and Ziwei Liu. 3D and 4D world modeling: A survey. *arXiv preprint arXiv:2509.07996*, 2025.

Zihang Lai, Eldar Insafutdinov, Edgar Sucar, and Andrea Vedaldi. CoWTracker: Tracking by warping instead of correlation. *arXiv preprint arXiv:2602.04877*, 2026. URL https://arxiv.org/abs/2602.04877.

JoungBin Lee, Jaewoo Jung, Jisang Han, Takuya Narihira, Kazumi Fukuda, Junyoung Seo, Sunghwan Hong, Yuki Mitsufuji, and Seungryong Kim. 3D scene prompting for scene-consistent camera-controllable video generation. *arXiv preprint arXiv:2510.14945*, 2025.

Chunyang Li, Yuanbo Yang, Jiahao Shao, Hongyu Zhou, Katja Schwarz, and Yiyi Liao. ReRoPE: Repurposing RoPE for relative camera control. *arXiv preprint arXiv:2602.08068*, 2026a.

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-GameCraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2025a. doi: 10.48550/arXiv.2506.17201. URL https://arxiv.org/abs/2506.17201.

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*, 2025b.

Xinyang Li, Tengfei Wang, Zixiao Gu, Shengchuan Zhang, Chunchao Guo, and Liujuan Cao. FlashWorld: High-quality 3D scene generation within seconds. *arXiv preprint arXiv:2510.13678*, 2025c.

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

Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N. Plataniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3D scenes from a single image. *arXiv preprint arXiv:2412.12091*, 2024.

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

Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan Burgert, Yuancheng Xu, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, Paul Debevec, and Ning Yu. Vista4D: Video reshooting with 4D point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Xuanyi Liu, Deyi Ji, Liqun Liu, Lanyun Zhu, Xuhang Chen, Qianxiong Xu, Peng Shu, Huan Yu, Jie Jiang, Feng Gao, and Siwei Ma. CamGeo: Sparse camera-conditioned image-to-video generation with 3D geometry priors. In *Proceedings of the International Conference on Machine Learning*, 2026.

Dongyue Lu, Ao Liang, Tianxin Huang, Xiao Fu, Yuyang Zhao, Baorui Ma, Liang Pan, Wei Yin, Lingdong Kong, Wei Tsang Ooi, and Ziwei Liu. See4D: Pose-free 4D generation via auto-regressive video inpainting. *arXiv preprint arXiv:2510.26796*, 2025. Accepted to Eurographics 2026.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. OpenVid-1M: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

Avinash Paliwal, Adithya Iyer, Shivin Yadav, Muhammad Ali Afridi, and Midhun Harikumar. Reshoot-Anything: A self-supervised model for in-the-wild video reshooting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2026.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985, 2019.

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

Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, Mengyin An, Haobo Zhang, Biao Jiang, Jiahua Wang, Haofeng Sun, Yang Liu, and Yangguang Li. Matrix-Game 3.5: Enhancing real-time streaming interactive world models with patch memory. *arXiv preprint arXiv:2608.29910*, 2026. doi: 10.48550/arXiv.2608.29910. URL https://arxiv.org/abs/2608.29910.

Lingteng Qiu, Peihao Li, Heyuan Li, Qi Zuo, Xiaodong Gu, Yuan Dong, Weihao Yuan, Rui Peng, Siyu Zhu, Xiaoguang Han, Guanying Chen, and Zilong Dong. LHM++: An efficient large human reconstruction model for pose-free images to 3D. *arXiv preprint arXiv:2506.13766*, 2025. URL https://arxiv.org/abs/2506.13766.

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.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026. doi: 10.48550/arXiv.2601.20540. URL https://arxiv.org/abs/2601.20540.

Junyoung Seo, Jisang Han, Jaewoo Jung, Siyoon Jin, Joungbin Lee, Takuya Narihira, Kazumi Fukuda, Takashi Shibuya, Donghoon Ahn, Shoukang Hu, Seungryong Kim, and Yuki Mitsufuji. Vid-CamEdit: Video camera trajectory editing with generative rendering from estimated geometry. *arXiv preprint arXiv:2506.13697*, 2025.

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

Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. DimensionX: Create any 3D and 4D scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 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. *arXiv preprint arXiv:2512.14614*, 2025.

Team HY-World, Chenjie Cao, Xuhui Zuo, Zhenwei Wang, Yisu Zhang, Junta Wu, Zhenyang Liu, Yuning Gong, Yang Liu, Bo Yuan, Chao Zhang, Coopers Li, Dongyuan Guo, Fan Yang, Haiyu Zhang, Hang Cao, Jianchen Zhu, Jiaxin Lin, Jie Xiao, Jihong Zhang, Junlin Yu, Lei Wang, Lifu Wang, Lilin Wang, Linus, Minghui Chen, Peng He, Penghao Zhao, Qi Chen, Rui Chen, Rui Shao, Sicong Liu, Wangchen Qin, Xiaochuan Niu, Xiang Yuan, Yi Sun, Yifei Tang, Yifu Sun, Yihang Lian, Yonghao Tan, Yuhong Liu, Yuyang Yin, Zhiyuan Min, Tengfei Wang, and Chunchao Guo. HY-World 2.0: A multi-modal world model for reconstructing, generating, and simulating 3D worlds. *arXiv preprint arXiv:2604.14268*, 2026.

Haiping Wang, Yuan Liu, Ziwei Liu, Wenping Wang, Zhen Dong, and Bisheng Yang. VistaDream: Sampling multiview consistent images for single-view scene reconstruction. *arXiv preprint arXiv:2410.16892*, 2024.

Hanyang Wang, Fangfu Liu, Jiawei Chi, and Yueqi Duan. VideoScene: Distilling video diffusion model to generate 3D scenes in one step. *arXiv preprint arXiv:2504.01956*, 2025a.

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.

Jiahao Wang, Luoxin Ye, TaiMing Lu, Junfei Xiao, Jiahan Zhang, Yuxiang Guo, Xijun Liu, Rama Chellappa, Cheng Peng, Alan Yuille, and Jieneng Chen. EvoWorld: Evolving panoramic world generation with explicit 3D memory. *arXiv preprint arXiv:2510.01183*, 2025b.

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

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, Yidan Xietian, Jiangbo Pei, Liang Hu, Boyi Jiang, Hua Xue, Zidong Wang, Haofeng Sun, Wei Li, Wanli Ouyang, Xianglong He, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-Game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026b. doi: 10.48550/arXiv.2604.08995. URL https://arxiv.org/abs/2604.08995.

Zhengxuan Wei, Xu Guo, Xinghui Li, Xunzhi Xiang, Min Wei, Yiran Zhu, Qiulin Wang, Xintao Wang, Pengfei Wan, Xiangwang Hou, and Qi Fan. Geometry-aware implicit memory for video world models. *arXiv preprint arXiv:2606.02436*, 2026.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3D representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025.

Chendong Xiang, Jiajun Liu, Jintao Zhang, Xiao Yang, Zhengwei Fang, Shizun Wang, Zijun Wang, Yingtian Zou, Hang Su, and Jun Zhu. Geometry-aware rotary position embedding for consistent video world model. *arXiv preprint arXiv:2602.07854*, 2026a.

Xunzhi Xiang, Zixuan Duan, Yabo Chen, Zhengxuan Wei, Guiyu Zhang, Zixiao Gu, Zhe Gao, Haibin Huang, Chi Zhang, Qi Fan, and Xuelong Li. VideoWeave: Unlocking geometric consistency in video generation via joint geometry-video modeling. *arXiv preprint arXiv:2606.14162*, 2026b.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

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*, 2026a.

Youcan Xu, Jiaxin Shi, Zhen Wang, Wensong Song, Feifei Shao, Chen Liang, Jun Xiao, and Long Chen. RealCam: Real-time novel-view video generation with interactive camera control. *arXiv preprint arXiv:2605.06051*, 2026b.

Jung Yi, Minjae Kim, Paul Hyunbin Cho, Wooseok Jang, Sangdoo Yun, and Seungryong Kim. WorldKV: Efficient world memory with world retrieval and compression. *arXiv preprint arXiv:2605.22718*, 2026.

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T. Freeman, and Jiajun Wu. WonderWorld: Interactive 3D scene generation from a single image. *arXiv preprint arXiv:2406.09394*, 2024a.

Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T. Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, and Charles Herrmann. WonderJourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6658–6667, 2024b.

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. *arXiv preprint arXiv:2506.03141*, 2025a.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b. doi: 10.1109/ICCV51701.2025.00017.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024c.

Zhu Yu, Jingnan Gao, Runmin Zhang, Lingteng Qiu, Zhengyi Zhao, Rui Peng, Yichao Yan, Kejie Qiu, Siyu Zhu, Zilong Dong, Si-Yuan Cao, and Hui-Liang Shen. Towards consistent video geometry estimation, 2026. URL https://arxiv.org/abs/2605.30060.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, and Yahui Zhou. Matrix-Game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025. doi: 10.48550/arXiv.2506.18701. URL https://arxiv.org/abs/2506.18701.

Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan, Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim Hee Lee, and Lijuan Wang. GenXD: Generating any 3D and 4D scenes. *arXiv preprint arXiv:2411.02319*, 2024.

Sixiao Zheng, Minghao Yin, Wenbo Hu, Xiaoyu Li, Ying Shan, and Yanwei Fu. VerseCrafter: Dynamic realistic video world model with 4D geometric control. *arXiv preprint arXiv:2601.05138*, 2026.

Haiyang Zhou, Wangbo Yu, Chaoran Feng, Xunyu Zhou, Yonghong Tian, and Li Yuan. UniWorld-View: Large-baseline view synthesis via video diffusion models. *arXiv preprint arXiv:2608.04701*, 2026.

Hanxin Zhu, Cong Wang, Peiyan Tu, Jiayi Luo, Tianyu He, Xin Jin, and Zhibo Chen. GTA: Advancing image-to-3D world generation via geometry then appearance video diffusion. *arXiv preprint arXiv:2605.12957*, 2026.

APPENDIX CONTENTS

A More Results 16
B Evaluation Details 16
C Implementation Details 17
D Additional Applications 18
E Additional Ablation Visualizations 18
F Limitations and Future Directions 20

A MORE RESULTS

We further compare WORLD IN WORLD with InSpatio-World and CameraAnything under camera trajectories spanning rotations of $120°$, $240°$, and $360°$. These settings evaluate whether the methods maintain generation quality over a wider range of viewpoints. As shown in Figure A1, WORLD IN WORLD produces better visual quality than both baselines across the evaluated rotation ranges. This advantage remains evident in the $240°$ and $360°$ cases, supporting the effectiveness of our method under substantial camera motion. At large viewpoint changes, quality scores are indicative, since the source video provides limited reference for evaluating target-view motion and scene consistency, and reconstruction metrics such as PSNR are computed only in visible warped regions.

Runtime efficiency. All timings are measured on a single NVIDIA A100 80GB GPU. Table A2 reports inference time per generated frame.

For our method, inference includes warping along the target trajectory, VAE encoding and DiT forward passes to extract evidence K/V, denoising, and final decoding. Depth estimation and our CGAR point tracking are reported separately as preprocessing costs and are excluded from the table.

For the timing evaluation, depth estimation uses Video Depth Anything and takes approximately 0.084 s/frame, with similar costs across methods that require depth. Source-camera parameter estimation is included in this preprocessing cost. Our CGAR point tracking adds approximately 0.098 s/frame.

B EVALUATION DETAILS

Evaluation data and camera paths. For the main comparison (Table 1), we randomly select 20 videos from DAVIS and 20 from OpenVid-1M. Each video is evaluated along ten random camera trajectories covering translation and rotation, yielding 400 video–trajectory pairs in total. We use DAVIS for the ablation (Table 2). For the large-rotation tests, we select 20 videos from OpenVid-1M and evaluate each along target camera paths spanning $120°$, $240°$, and $360°$ (Table A1).

Baseline protocol and metrics. ReCamMaster and TrajectoryCrafter generate 81 and 49 frames, respectively, as required by their official implementations, while the remaining methods use their officially recommended frame counts. All methods receive the same source video and continuous target camera path, with camera parameters converted to the convention required by each method. We use uniform temporal sampling and interpolation to match each model's required frame count. All methods share the same metric preprocessing pipeline.

Our rerendering task generates target views along the source video's timeline, retaining its duration. With no captured target-view ground truth, we use warped source frames at corresponding times as approximate references, with identical reference images and validity masks for all methods. PSNR, SSIM, and LPIPS assess adherence to guided appearance and motion states within these visible regions; newly exposed regions lack references for these metrics. VBench assesses generated-video quality, while trajectory errors assess camera controllability. All methods use the same scale and coordinate alignment procedure for trajectory evaluation.

Table A1: Quantitative comparison under large camera rotations of 120°, 240°, and 360°. We compare InSpatio-World, CameraAnything, and our method using VBench, camera-control errors, and image-quality metrics. Overall denotes the arithmetic mean of the seven reported VBench dimensions. **Bold** denotes the best results, including ties.

	VBench ↑								Camera Errors ↓		Image Quality		
Method	Aesth.	Img.	Flick.	Smooth.	Subj.	Bg.	Dyn.	Overall	Trans. Err.	Rot. Err. (°)	PSNR ↑	SSIM ↑	LPIPS ↓
InSpatio-World	63.98	67.70	91.74	95.77	87.94	**89.80**	100.00	85.28	0.25270	25.50	19.56	0.665	0.200
CameraAnything	59.57	64.29	92.19	95.73	86.86	88.23	100.00	83.84	0.27720	17.93	12.21	0.441	0.427
Ours	**64.58**	**67.90**	**92.20**	**95.91**	**88.00**	**89.80**	100.00	**85.48**	**0.16130**	**8.59**	**20.60**	**0.751**	**0.142**
ours w/o EWA	62.07	67.58	91.59	95.01	87.03	88.55	100.00	84.55	0.16434	9.26	19.59	0.697	0.181
ours w/o Rendered Geometry	62.50	66.43	91.72	95.15	87.16	88.58	100.00	84.51	0.20308	10.96	20.05	0.732	0.167



Figure A1: Qualitative comparisons under large camera rotations. We compare InSpatio-World, CameraAnything, and our method along camera trajectories spanning 120°, 240°, and 360°. All methods receive the same source video and target camera trajectory in each case. Our method produces better visual quality across those shown rotation ranges.

C Implementation Details

Evidence schedule and guidance. We use the four-step distilled causal-fast checkpoint of LingBot-World 2.0. Evidence channels are active during the first three denoising steps of each chunk and inactive during the final step. The default EWA guidance strengths are $g_{\text{geo}} = 2$ for rendered geometry evidence and $g_{\text{proj}} = 0.5$ for target-view scene evidence.

We use depth estimates from DepthCrafter, Video Depth Anything, and ViGeo to construct target-view scene evidence (Hu et al., 2025b; Chen et al., 2025d; Yu et al., 2026). For Video Depth Anything, we use the ViT-L encoder and apply a shared depth normalization across the entire source sequence. We examine sensitivity to the depth-estimator choice in Appendix E.

Token support weights. For target-view scene evidence, a token's support weight is the fraction of visible warped pixels in its 16×16 pixel region: $\omega_{\text{proj},j} = v_j/256$, where v_j is the visible-pixel count. The weight ranges from 0 for no visible support to 1 for full coverage.

To complement depth-based projections when camera motion exposes previously unobserved subject surfaces, the rendered-geometry branch uses target-view renderings of a reconstructed subject (Section 3.3). For example, if the subject is human, we represent the subject state as $G_m = (\mathcal{A}, \Theta_m)$, where $\mathcal{A}$ is an avatar reconstructed with LHM++ (Qiu et al., 2025) from sharp, minimally occluded full-body source-video crops, and Θ_m contains the per-frame SMPL-X (Pavlakos et al., 2019) parameters that drive the avatar. We align the avatar and body model to the projected scene's coordinate frame and depth scale using depth correspondences in the source view. LHM++ provides target-view RGB and support masks for observed and inferred unseen surfaces, while the aligned SMPL-X geometry provides depth for resolving occlusion between subjects. We further compare this depth with the projected scene depth to remove unreliable background projections near subject boundaries. We disable this branch for non-human subjects or when geometry renderings are unavailable.

CGAR weights. We obtain the persistent point tracks used by CGAR with the pretrained CoW-Tracker model (Lai et al., 2026). The tracks are combined with depth and camera information to

Table A2: Runtime comparison for camera-controlled video rerendering. Parameter counts refer to the pretrained video backbone. Training-free indicates no additional control-specific training relative to the pretrained base model. Resolution is reported as width × height.

Method	Parameters	Training-free	Resolution	Inference time (s/frame)
InSpatio-World	1.3B	No	832×480	0.634
UniWorld-View	14B	No	832×480	2.636
ReCamMaster	1.3B	No	832×480	6.733
TrajectoryCrafter	5B	No	672×384	6.804
WorldForge	14B	Yes	800×496	22.975
CameraAnything	14B	No	832×480	37.864
Ours	14B	Yes	832×464	1.389

associate target queries with source-video tokens. For query i and source token j, let n_{ij} count the pixels in the source token's 16×16 region that match the query. We set $\beta_{ij} = 2n_{ij}/256$, giving a weight of 2 for a fully matched token and 0 for no matches. Only pairs with $n_{ij} > 0$ participate in routing attention, whose response is jointly normalized with the other active attention blocks.

For the source-video branch, we encode the source video with the VAE and feed the unwarped latents into the frozen backbone at diffusion timestep $t = 0$ to extract clean K/V.

History retrieval and capacity. We rank archived states by target-view surface coverage and viewing-direction compatibility. By default, we consider only candidates with target-view surface coverage above 40% and select a diverse subset of up to four top-ranked states to reduce redundant references. If none qualify, no retrieved-history evidence is attached. The history-bank capacity is set to fit the available GPU/CPU memory. When the archive exceeds this budget, we uniformly subsample stored states along the timeline to balance memory use and historical coverage. On revisits, we use depth-based source-video projections at the current event time as the main reference where visible support is available, and use retrieved history to guide regions unobserved in the source video. Tracking the off-screen evolution of generated dynamic objects remains challenging for current memory. We plan to explore 3D proxies to represent and update their states during these intervals.

To limit text-induced subject duplication in newly revealed regions, we condition the first chunk on the full source-video description and use a scene-focused version with explicit person and animal descriptions removed for subsequent chunks. The source-video evidence remains available throughout the rollout to guide subject appearance and event dynamics.

D Additional Applications

Our framework expresses diverse control requirements through a shared visual-evidence interface. Figure A2 demonstrates this interface beyond camera-controlled rerendering, including event-time control, appearance and motion editing, and cross-model communication.

In the cross-model communication example, Model A and Model B are two instances of the same frozen model running different generation cases. They communicate by passing cached K/V as visual evidence. Across these applications, different evidence sources guide generation through the same native attention mechanism, while camera and temporal information specify how the evidence relates to the requested output. This extends the model's control capabilities without modifying the frozen backbone or requiring task-specific training.

E Additional Ablation Visualizations

Ablation settings. *Only Target-View Warp* uses only target-view scene evidence on the frozen LingBot-World 2.0 backbone, with all other evidence branches, EWA, and CGAR disabled. In *w/o EWA*, we set $g_e = 0$ for all evidence channels, so no additional amplification is applied.

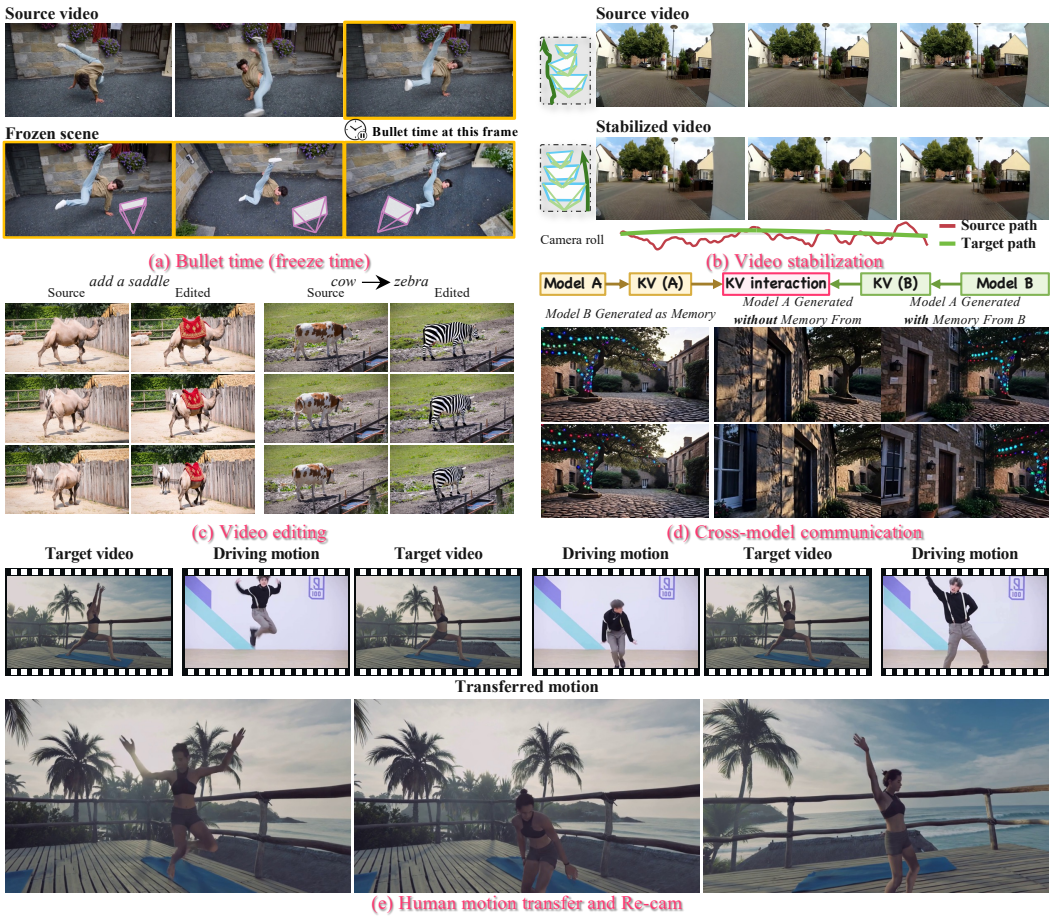

Figure A2: **Additional applications of the shared visual-evidence interface.** We demonstrate bullet-time rendering, video stabilization, video editing, K/V sharing between generation cases, and human motion transfer. In the K/V-sharing example, Model A and Model B are two instances of the same frozen model running different generation cases and communicate by sharing cached K/V as visual evidence.

We further analyze the results in Table 2 and provide qualitative comparisons to examine how spatial evidence, camera conditioning, and auxiliary evidence affect generation. Figure A3 studies target-view warping and source-camera Plücker conditioning. Target-view warping provides an explicit spatial reference that helps preserve subject structure and scene layout under viewpoint changes, while source-camera conditioning retains source-view camera information for cross-view consistency. Figure A4 examines how different forms of evidence are used when direct observations are ambiguous or insufficient. Under large viewpoint changes, appearance-based attention can struggle to locate the corresponding source content, while different evidence channels may vary in reliability across spatial regions. CGAR mitigates this ambiguity by routing current queries toward geometrically corresponding source-video tokens, while EWA adjusts the guidance strength of each evidence channel.

Removing CGAR or EWA produces only small changes in the reported VBench scores (Table 2). However, the evaluated VBench dimensions only characterize visual quality and temporal consistency within each generated video, without explicitly testing correspondence with the source video or agreement with the prescribed camera trajectory (Huang et al., 2023). A generated subject can therefore remain temporally coherent while differing in appearance from the source. As shown in Figure A4(a), the ablated results exhibit visible deviations from the source video in subject appearance and background structure, while the full method better preserves both. These qualitative

comparisons complement the VBench scores by revealing differences in adherence to the source evidence.

The remaining cases illustrate complementary evidence for regions or states that are not directly available from the current observations. In Figure A4(b), removing rendered body geometry causes the highlighted dancer to deviate from the source pose: the raised, bent-knee configuration is no longer maintained. With rendered body geometry, the generated dancer better preserves this configuration under the changed viewpoint. This comparison shows that geometry renderings help maintain pose fidelity when completing subject regions with limited source-view support. Figure A5 further shows that rendered body geometry helps preserve subject orientation and motion under large camera rotations.

Figure A4(c) examines a long-horizon revisit after the relevant generated states have been evicted from the rolling cache. Without historical retrieval, the previously generated round wall mirror above the desk is replaced by a tall oval mirror, and the desk and chair disappear from the revisited view. With retrieval, the generated view retains the round mirror and the desk-and-chair arrangement established earlier in the rollout. These results demonstrate that historical retrieval helps preserve object appearance and scene layout when the camera returns to previously visited regions.

We further examine the robustness of WORLD IN WORLD to the choice of depth estimator. Figure A6 compares generated target views using depth estimates from DepthCrafter, Video Depth Anything, and ViGeo. Across these estimators, the generated results exhibit comparable visual quality, with similar subject appearance and scene structure. This qualitative consistency suggests that WORLD IN WORLD is robust to the choice of depth estimator across the tested cases, without relying on a particular estimator to preserve the source content under viewpoint changes.

We further examine the effectiveness of historical retrieval by comparing WORLD IN WORLD with a variant without memory and WorldKV (Yi et al., 2026). We port WorldKV to LingBot-World 2.0, using the same backbone and other settings as our method. Figure A7 compares the generated views at frame 64 with the revisited views at frame 220 across four scenes, evaluating whether previously generated appearance and layout are preserved when the camera returns. In the Stonehenge example (Figure A7(a)), the building established during the earlier visit disappears in both the memory-free and WorldKV results, whereas WORLD IN WORLD retains the building and its appearance. The Great Wall and Lakeside House examples (Figure A7(b,c)) further reveal changes in the wall–walkway arrangement and the house facade in the baseline revisits, while our method better preserves these structures relative to the earlier generated views. In the Courtyard example (Figure A7(d)), the previously generated tree is replaced by a building without memory and is absent from the WorldKV revisit. With our historical retrieval, the tree and the surrounding path layout are better preserved. These comparisons show that retrieved generated evidence helps maintain object appearance and scene structure beyond the rolling cache, yielding more consistent revisits.

F LIMITATIONS AND FUTURE DIRECTIONS

The historical K/V memory is maintained with a bounded capacity to keep long-horizon inference tractable. For very long rollouts, once the memory reaches its capacity, stored K/V states are uniformly subsampled along the temporal axis to maintain a compact archive. While this preserves broad coverage of the generation history, aggressive compression over extremely long sequences may reduce the fidelity of retrieved historical evidence. A promising direction is to augment the current retrieval mechanism with a more explicit and structured long-term memory, enabling selective preservation and consolidation of important visual states without continuously increasing the memory footprint.



Figure A3: Qualitative ablations of spatial evidence and camera conditioning. **(a)** Target-view warping preserves subject structure and scene layout under viewpoint changes. **(b)** Source-camera Plücker conditioning maintains cross-view structure using source-view camera information.

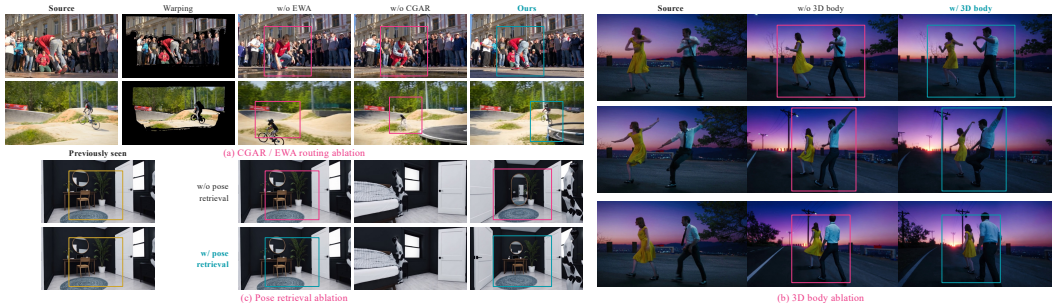

Figure A4: Qualitative ablations of evidence use and availability. **(a)** Correspondence-guided attention routing (CGAR) and evidence-wise attention (EWA) help preserve subject appearance and background structure. **(b)** Rendered body geometry guides the completion of newly exposed subject regions. **(c)** The camera revisits a previously generated region after earlier states have been removed from the rolling cache. Historical retrieval helps preserve the appearance and layout established during the earlier visit.



Figure A5: Qualitative ablation of 3D body guidance under large camera rotations. **(a)** A chef observed along a 240° camera orbit. **(b)** A walking subject observed along a 360° camera orbit. The three rows show distinct viewpoints along each trajectory. Each comparison includes the temporally aligned source frame, the result without 3D body guidance, and our result. Rendered body geometry helps preserve subject orientation and motion as the camera reveals different sides of the body.



Figure A6: Sensitivity to the depth estimator used for target-view evidence construction. We compare warped images and generated outputs using DepthCrafter, Video Depth Anything, and ViGeo under otherwise fixed inference settings.



Figure A7: Qualitative comparisons of memory for long-horizon revisits. We compare a variant without memory, WorldKV, and our method. Each example shows the input, an earlier generated view (frame 64), and a revisited view (frame 220). Highlighted regions show that our historical retrieval better preserves object appearance and scene layout established during the earlier visit, while the compared alternatives exhibit missing objects or altered structures.