POINTWORLD: EXPLICIT 3D REPRESENTATION CALIBRATION FOR INTERACTIVE WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing interactive world models accumulate geometric inconsistencies during long-sequence generation, making it difficult to maintain coherence over thousands of frames. Recent methods mostly provide visual guidance using RGB conditions rendered from 3D representations, but they still suffer from imperfect rendering and reconstruction conflicts that hinder effective geometric conditioning. We find that directly embedding 3D representations can provide geometric guidance while bypassing errors introduced during RGB rendering. Building on this finding, we propose **PointWorld** that explicitly uses 3D point clouds as geometric memory, rather than relying on rendered RGB proxies, for long-horizon consistency. The core is an explicit 3D representation calibration mechanism that selects relevant points from dynamically reconstructed point clouds, encodes their geometric information, and directly injects the resulting features into the video generation model. To provide reliable 3D conditioning, **PointWorld** further incorporates a spatiotemporal retrieval strategy into dynamic point cloud reconstruction to mitigate reconstruction conflicts. Extensive experiments demonstrate that **PointWorld** extends consistent interactive generation, maintaining long-term consistency across sequences of up to 2,000 frames and outperforming recent methods in both quantitative and qualitative comparisons. Our code will be made publicly available.



Figure 1: **Teaser Demonstration**. Our PointWorld directly injects geometric information through explicit 3D representation calibration, complemented by rendered RGB detail conditioning, to maintain scene consistency over more than **2,000** frames of keyboard-controlled generation. Compared with the baselines, PointWorld better preserves previously observed scene structures during camera revisits and reduces color drift. The frame number is annotated in the top-left of each image.

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.

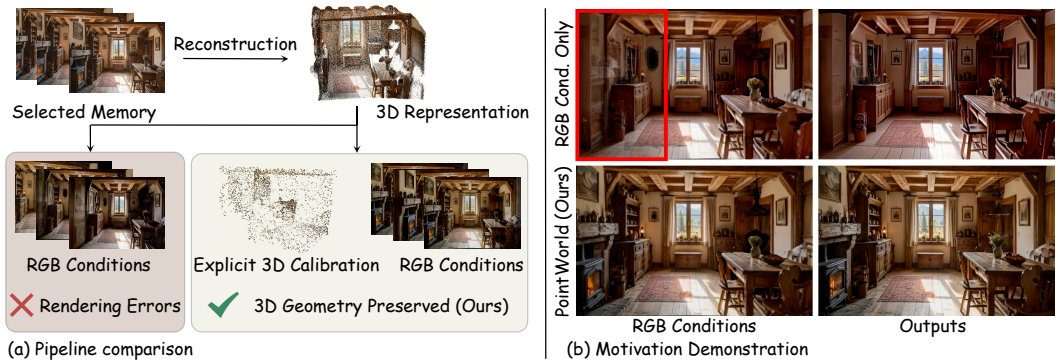

Figure 2: (a) **Pipeline Comparisons.** PointWorld complements rendered RGB conditions with explicit 3D calibration to provide geometric information directly to the generator. (b) **Motivation Demonstration.** RGB conditions and corresponding outputs for RGB-only conditioning (top) and PointWorld (bottom). The red box highlights geometric conflicts in the RGB conditions. Relying solely on RGB conditions can lead to outputs that deviate from historical scene geometry, whereas PointWorld better preserves geometric consistency.

1 Introduction

Interactive world modeling has emerged as a foundational paradigm with broad applications in virtual reality, embodied AI agents, human-computer interaction, and autonomous systems (Li et al., 2025b; Mao et al., 2025b; Guo et al., 2025; Li et al., 2025c; Liao et al., 2025; Yang et al., 2025a;b; 2026; Yu et al., 2025; Xiao et al., 2026; Wu et al., 2025; Hong et al., 2025b). In these domains, the ability to generate temporally coherent, high-fidelity sequences in response to user interactions is critical for immersive experiences, long-horizon planning, and consistent environment understanding. However, current interactive video synthesis systems remain tightly coupled with the output consistency of the underlying video generation models. This dependency significantly constrains user experiences, as inconsistencies can degrade perception, interrupt task flow, and disrupt environmental continuity.

Previous methods attempt to use the inherent temporal agreement in auto-regressive sequences (Huang et al., 2026; Liu et al., 2025; Zhu et al., 2026b; Huang et al., 2025a; HunyuanWorld, 2025; Mao et al., 2025a; Wang et al., 2026b; He et al., 2025; Li et al., 2025a; Ye et al., 2025; Shin et al., 2025; Yesiltepe et al., 2025; Zhang et al., 2025; Team et al., 2026b; Gao et al., 2026) to stabilize long-term generation. However, these approaches often accumulate errors over extended horizons, diverging from spatial and semantic consistency. Some using 2D memory mechanisms (Wu et al., 2026a) are bounded by a finite memory cache capacity. Compressing memory frames to expand capacity inevitably sacrifices critical visual information, leading to loss of details and coherence over time.

Recent methods (Wu et al., 2026b; Zhao et al., 2026) incorporate explicit 3D representations to maintain a global spatial state and provide visual guidance through RGB conditions rendered from these representations (Huang et al., 2025b; Wan et al., 2025b; Chen et al., 2025; Yu et al., 2024). These methods reformulate interactive synthesis as a video-to-video inpainting problem, indirectly guiding video models to understand spatial structure. Nevertheless, video generation models cannot guarantee consistency across generated frames, making 3D reconstruction from historical frames susceptible to conflicts among these frames and depth estimation errors. Rendering RGB conditions from the reconstructed 3D representations is likewise affected by depth estimation errors, further compromising the resulting conditions. As shown in the top row of Figure 2(b), rendered RGB conditions contain geometric conflicts, and relying solely on these conditions can lead to outputs that deviate from historical geometric structure. These generated frames then contribute to subsequent reconstruction, allowing errors to accumulate over continued generation.

Based on these observations, we argue that imperfect rendering and reconstruction conflicts distort the geometric information, limiting their ability to provide effective geometric guidance to the generator. To effectively provide geometric information to the video generation model, as illustrated in Figure 2(a), we propose to leverage explicit 3D representation embeddings. The core is a 3D representation calibration mechanism that selects relevant points from the current 3D point cloud, explicitly encodes their 3D structure, and directly injects the resulting features into the video generation

model. Direct point cloud encoding bypasses errors introduced during rendering while preserving the geometric structure of the point cloud. Figure 2(b) illustrates PointWorld's improved preservation of historical scene geometry, while ablation experiments further validate the contribution of direct geometric conditioning. Since 3D encoding alone cannot fully provide detailed information for the target view, we also supply the model with rendered RGB conditions. The 3D encoding provides explicit geometric guidance, while the rendered RGB conditions supply complementary detailed information aligned with the target view, jointly supporting consistent video generation.

To ensure the reliability of the 3D representations, we design a spatiotemporal memory retrieval strategy that selects relevant historical frames based on camera poses, temporal order, and content similarity. These frames are used to dynamically reconstruct point clouds, reducing reconstruction conflicts. By combining 3D representation calibration with the spatiotemporal memory retrieval strategy, PointWorld extends consistent generation to sequences exceeding 2,000 frames, representing a notable improvement over previous interactive world models.

Extensive experiments demonstrate that our PointWorld outperforms previous methods, such as LingBot-World 2.0 (Gao et al., 2026) and HY-World 1.5 (HunyuanWorld, 2025), qualitatively as shown in Figure 1. We also show that our PointWorld can perform well in quantitative comparisons in the experiment section. To sum up, our main contributions can be summarized as follows.

- We propose an **explicit 3D representation calibration mechanism** that selects relevant points from dynamically reconstructed point clouds, encodes their 3D structure, and injects the features into the video generation model. These features provide geometric guidance, while rendered RGB conditions add visual details, jointly improving long-term geometric consistency.
- We design a **spatiotemporal memory retrieval strategy** selecting historical frames by camera pose, temporal order, and content similarity to dynamically reconstruct point clouds, reducing reconstruction conflicts and providing geometric conditioning for 3D representation calibration.
- We present **PointWorld**, an interactive world model that enables consistent generation over sequences exceeding 2,000 frames. Extensive quantitative and qualitative experiments demonstrate that PointWorld outperforms recent methods in long-term consistency.

2 Related Work

Interactive World Models. Interactive world models (Li et al., 2025b; Zhang et al., 2025; Li et al., 2025c; He et al., 2025; Ball et al., 2025; Shen et al., 2026; Zhang et al., 2026b; Qian et al., 2026; Zhu et al., 2026a; Team et al., 2026a) continuously generate video conditioned on historical frames and user inputs, with long-term consistency closely tied to how historical information is stored and used. Existing methods employ video-context memory, explicit 3D spatial memory, and latent 3D spatial memory. Video-context methods retain historical information in video frames, latent features, or attention caches, maintaining continuity through context organization and autoregressive generation, including LingBot-World (Team et al., 2026b), HY-World 1.5 (HunyuanWorld, 2025), Yume-1.5 (Mao et al., 2025a), Matrix-Game 3.5 (Qian et al., 2026), MotionStream (Shin et al., 2025), Infinite-World (Wu et al., 2026a), RELIC (Hong et al., 2025a), and LingBot-World 2.0 (Gao et al., 2026). Explicit 3D spatial memory methods organize historical frames into 3D scene representations and guide video generation using renderings from target viewpoints, as exemplified by SPMem (Wu et al., 2026b) and Spatia (Zhao et al., 2026). MagicWorld (Li et al., 2025a) combines rendered 3D point cloud conditions with retrieval of historical video latents. Latent 3D spatial memory methods, such as Mirage (Wang et al., 2026a), use 3D geometric priors to project video latent features onto target-view latent grids, constituting a form of rendering in latent space.

3D Foundation Models. 3D foundation models (Chen et al., 2026; Zhang et al., 2026a) learn general geometric priors to recover depth, camera parameters, and scene structure from images, providing a foundation for constructing spatial representations in world models. Existing methods span pairwise reconstruction, joint multi-view prediction, and online state updates. DUSt3R (Wang et al., 2024) and MASt3R (Leroy et al., 2024) use 3D point maps to connect geometric reconstruction with image correspondence estimation. VGGT (Wang et al., 2025b), π^3 (Wang et al., 2025d), and Depth Anything 3 (Lin et al., 2025) extend geometric prediction to multi-view inputs. Another line of work processes sequential observations through a persistent state to incrementally estimate geometry, as exemplified by CUT3R (Wang et al., 2025c).

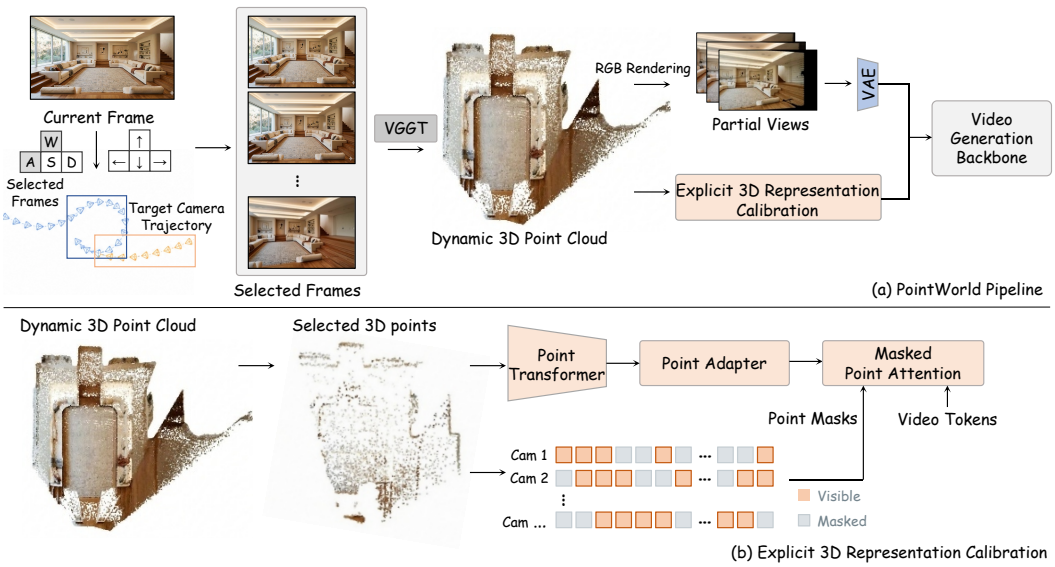

Figure 3: **Method Demonstration.** (a) PointWorld reconstructs dynamic point clouds from retrieved historical frames, providing geometric guidance through direct 3D calibration and visual details through rendered RGB partial views. (b) The calibration module encodes relevant points, selects point features within each target field of view using per-frame Point Masks, and injects them into the generator through Masked Point Attention.

3 METHODOLOGY

PointWorld takes a single image as input and performs continuous interactive video synthesis through chunk-wise generation. As illustrated in Figure 3(a), the system first constructs a target camera trajectory from keyboard commands and retrieves relevant historical frames. These frames are jointly reconstructed using VGGT (Wang et al., 2025b) to obtain a shared 3D point cloud for the current chunk. PointWorld then conditions video generation through two complementary pathways: One directly encodes the 3D structure of the point cloud, selects point features according to per-frame visibility, and integrates them with the corresponding video features; the other one renders RGB partial views along the target camera trajectory to provide detailed information.

3.1 KEYBOARD-BASED INTERACTIVE CONTROL

Users control camera motion through keyboard inputs at the beginning of each generation chunk. Specifically, W/S and A/D control forward/backward and left/right translations, respectively, while the arrow keys $\uparrow$ / $\downarrow$ and $\leftarrow$ / $\rightarrow$ control pitch and yaw rotations. Each key press corresponds to a fixed translation or rotation increment, and the accumulated inputs determine the camera motion for the current chunk. Camera rotations use fixed angular increments, whereas translations follow the relative coordinate scale of the scene reconstructed by VGGT rather than absolute physical units. Let I_{prev} denote the last frame of the previous chunk. Starting from this frame and its camera pose, we construct the camera-to-world transformation matrices for all frames in the current chunk according to the user commands, with C_t denoting the camera pose of a target frame. The resulting trajectory is used for historical frame retrieval, RGB partial-view rendering, and per-frame point cloud visibility estimation, providing consistent camera-motion control throughout generation.

3.2 EXPLICIT 3D POINT CLOUD CALIBRATION

Conveying scene information solely through rendered RGB conditions exposes geometric guidance to both reconstruction imperfections and rendering errors. To address this issue, PointWorld introduces an explicit 3D point cloud calibration mechanism that directly injects geometric information from the point cloud into the video generation model. As illustrated in Figure 3(b), the mechanism first selects relevant 3D points and encodes them through a Point Transformer (PTv3 (Wu et al., 2024)) and an

MLP-based Point Adapter. Masked Point Attention then integrates information from points visible in the current view into the corresponding video features.

Relevant 3D Point Selection. Each inference step uses a single point cloud jointly reconstructed from the selected historical frames, denoted by P_{mem}. Let I_{ref} denote the reference image corresponding to the last frame of the historical sequence, and let C_{ref} denote its camera pose. The reference point cloud is obtained through back-projection:

$$P_{ref} = \pi^{-1}(I_{ref}, C_{ref}), \tag{1}$$

where π^{-1} denotes the back-projection operator that maps an image into 3D space using the estimated depth and camera intrinsics.

To select geometric structures relevant to the current generation, we voxelize P_{mem} and P_{ref} using a shared grid with voxel size v_{size}. For each voxel containing reference points, we consider its local neighborhood with offsets $[-1, 0, 1]$ along each coordinate axis and retain points from the historical point cloud that fall within these neighborhoods. The selected point set is defined as

$$S = \{p \in P_{mem} \mid p \text{ lies within neighborhood voxels of } P_{ref}\}. \tag{2}$$

This selection process preserves historical geometry within and around the reference region while reducing interference from irrelevant points. The resulting point set S is shared across all frames in the current chunk.

Point Cloud Geometry Encoding. We use a Point Transformer (Wu et al., 2024), denoted by f_{PT}, to encode the 3D structure of the selected point set:

$$z_{calib} = f_{PT}(S). \tag{3}$$

The encoded features are mapped into a feature space suitable for interaction with video features through an MLP-based Point Adapter f_{PA}:

$$Z = f_{PA}(z_{calib}). \tag{4}$$

The point set and its encoded features are shared within the current chunk. This direct encoding pathway conveys geometric information to the generator without RGB rendering, thereby avoiding additional rendering errors along this conditioning pathway.

Per-Frame Visibility Selection and Interaction. The fields of view of different target cameras cover different regions of the shared point cloud. To focus geometric interaction on the current target field of view, we construct a per-frame binary Point Mask based on whether each point lies within the corresponding camera's field of view:

$$M \in \{0,1\}^{T_{\text{latent}} \times N_p},$$
$$M_{t,j} = \begin{cases} 1, & p_j \text{ lies within the target camera's field of view for latent frame } t, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Here, T_{latent} is the number of video latent frames, N_p is the number of shared point features, and p_j is the 3D position corresponding to the j-th point feature.

For latent frame t, we use the mask to directly select the point features participating in the interaction, obtaining the visible point feature sequence:

$$Z_t = \big[Z_j \mid M_{t,j} = 1\big]. \tag{6}$$

We construct queries Q_t from the video features of the current latent frame and keys K_t and values V_t from the selected visible point features Z_t. Point attention is computed as

$$\text{PointAttn}_t = \text{softmax}\left(\frac{Q_t K_t^\top}{\sqrt{d}}\right) V_t, \tag{7}$$

where d is the feature dimension of the queries and keys. The Point Mask selects point features before attention computation, allowing each latent frame to interact only with points visible in its

target view. Different frames share the same geometric features while using their respective visibility masks, thereby receiving view-specific geometric guidance.

Complementary RGB Detail Conditioning. Since point cloud encoding alone cannot fully provide fine-grained details in the target view, we retain rendered RGB conditions. Along the target camera trajectory, the reconstructed point cloud P_{mem} shared by the current chunk is rendered into each target view:

$$V^t = \pi(P_{mem}, C_t), \tag{8}$$

where C_t denotes the camera pose of the target video frame, π denotes 3D-to-2D projection, and V^t is the corresponding RGB partial view. The partial-view sequence is encoded by a VAE and concatenated with the video latent tensor x along the channel dimension, providing detail conditioning aligned with the target views. The RGB conditions and directly injected point cloud features complement each other to support geometrically consistent video generation.

3.3 Spatiotemporal Memory Retrieval Strategy

The directly injected geometric information depends on the reliability of the reconstructed point cloud. Since historical frames may contain inconsistent scene content, reconstructing from the entire history can introduce geometric conflicts. We therefore design a spatiotemporal memory retrieval strategy that selects historical frames based on temporal order, spatial proximity, and content quality to dynamically construct a 3D representation for the current chunk.

Candidate Frame Construction. Let $\mathcal{M}$ denote the memory pool containing historical frames and their camera poses. We use the camera pose C_m of the middle frame in the current target trajectory as the retrieval reference. Based on camera-center distances and relative rotations between historical cameras and C_m, we retrieve historical frames with the closest camera poses to form a spatially relevant candidate set.

Historical Frame Selection. We select candidate frames based on temporal order, spatial proximity, and content quality. Specifically, we favor earlier frames to limit accumulated errors, frames with camera poses closest to C_m for spatial relevance, and frames with higher cosine similarity to the initial image I_{init} in pretrained DINOv2 (Oquab et al., 2023) feature space, using content preservation as a proxy for quality. Since the initial image serves as a fixed reference, the features and similarity scores can be computed and cached during generation, reducing subsequent retrieval overhead. The resulting dynamic point cloud supports both explicit geometric injection and RGB partial-view rendering, providing a shared scene representation for the two conditioning pathways.

Reference Jump. As part of the retrieval strategy, we further introduce Reference Jump to incorporate earlier historical references. Specifically, using the camera pose C_{prev} of the last frame in the previous chunk as a reference, we retrieve candidate historical frames with the closest camera poses and select the earliest among them. The selected frame and its camera pose replace the first-frame input of the partial-view sequence for the current chunk. After generation, the corresponding first output frame is discarded to maintain video continuity, as the last frame of previous chunk naturally serves as the first frame of the current chunk. This allows the model to use historical references to improve content consistency.

4 Experiments

4.1 Experimental Setup

Implementation Details. PointWorld is trained based on the Wan-Fun-2.1-Camera-1.3B (Wan et al., 2025a) backbone. During training, we set the batch size, number of iterations, learning rate, and training resolution to 16, 20,000, 5×10^{-5}, and $832 \times 480 \times 81$, respectively. The model is trained on $8\times$ 80G GPUs for approximately 4 days. During 3D Point Selection, v_{size} is set to $1/256$ of the diagonal length of the point cloud's bounding box.

Datasets. We mainly sample around 5,000 video clips from the SpatialVID (Wang et al., 2025a) dataset for training. We re-annotate the camera parameters using VGGT (Wang et al., 2025b). During training, memory videos and point clouds reconstructed from memory videos are pre-constructed. Specifically, 0–3 chunks (0, 81, 161, or 241 frames) are randomly selected as memory videos. When

Table 1: Quantitative comparison of our PointWorld with recent methods under VBench (Huang et al., 2024). The best performances are in **bold** and the second-best performances are underlined.

Method	Overall Scores	Temporal Quality				Visual Quality	
		Temporal Flickering ↑	Motion Smoothness ↑	Subject Consistency ↑	Background Consistency ↑	Imaging Quality ↑	Aesthetic Quality ↑
Yume 1.5	0.8049	0.9625	0.9886	0.7820	0.8527	0.7100	0.5336
HY-World 1.5	0.8015	0.9511	0.9891	0.7862	0.8602	0.7294	0.4928
Matrix-Game 3.0	0.7843	0.9124	0.9730	0.7720	0.8369	**0.7388**	0.4724
LingBot-World 2.0	0.7968	0.9611	0.9821	0.7603	0.8494	0.7132	0.5150
PointWorld (ours)	**0.8302**	**0.9789**	**0.9928**	**0.8687**	**0.8945**	0.7014	**0.5449**

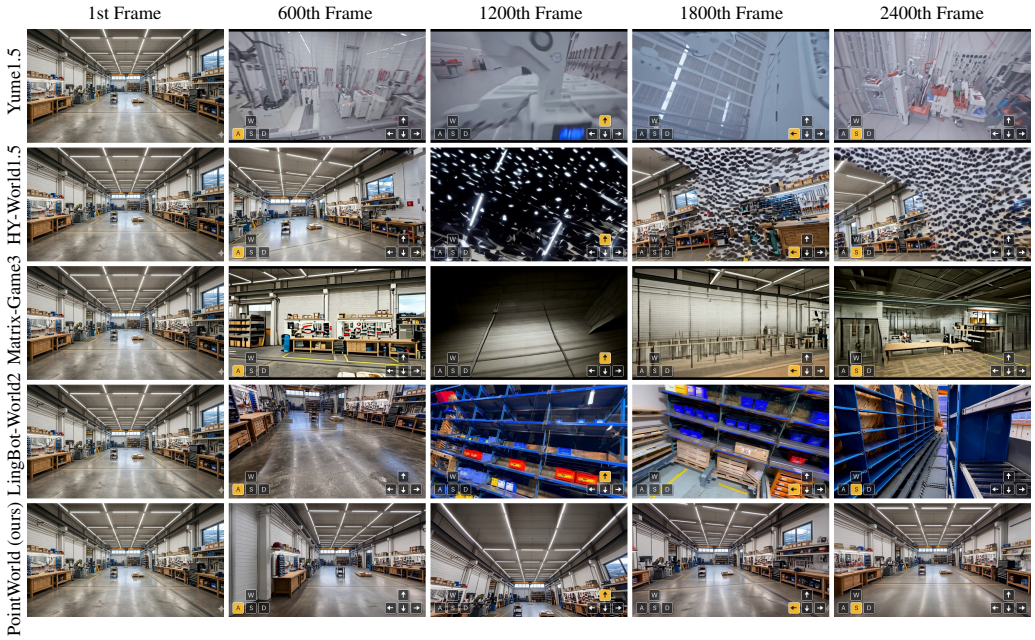

Figure 4: Qualitative comparisons. PointWorld can successfully extend consistent interactive generation, maintaining long-term coherence over 2,000 frames.

reconstructing point clouds, only frame selection based on DINOv2 feature similarity is applied, which allows the model to handle more complex memory video distributions.

Testing and Evaluation. We use a test set of 50 scene images with diverse styles. These images cover both indoor and outdoor scenes, with styles ranging from realistic to fantasy. We manually design 10 different trajectories, each including 30 keyboard interactions (2400 frames). As different models have different interactive mechanisms, we adjust their motion amplitude parameters to align the movement trajectories as consistently as possible for fair comparison. In addition, for Yume1.5, whose inference memory usage increases with the size of the memory video, memory videos exceeding 400 frames are truncated to limit GPU memory usage to 80G. We evaluate generated videos using six VBench (Huang et al., 2024) metrics: temporal flickering, motion smoothness, subject consistency, and background consistency for temporal quality; imaging quality and aesthetic quality for visual quality. The overall score is the mean of these six metrics.

4.2 Comparisons with State-of-the-Art Methods

Quantitative comparisons. Table 1 presents quantitative comparisons with recent state-of-the-art methods, including Yume1.5 (Mao et al., 2025a), HY-World1.5 (HunyuanWorld, 2025), Matrix-Game 3.0 (Wang et al., 2026b), and LingBot-World 2.0 (Gao et al., 2026). Our method achieves the best performance across the four temporal quality metrics, demonstrating its advantage in long-horizon generation. In addition, our method achieves competitive results on the two visual quality metrics, confirming its effectiveness in producing temporally coherent and visually plausible sequences.



Figure 5: Image matching results. We use MASt3R (Leroy et al., 2024) to visualize correspondences between input frames (top) and generated frames (bottom). Since we align target camera trajectories as closely as possible across methods and compare frames at identical indices, fewer matches may indicate scene inconsistencies or inaccurate camera control.



Figure 6: **Visualization of 3D interactions.** P. M. denotes Point Mask. For each of the two scenes, we show interaction points overlaid on the output, the points alone, and the generated output. Brighter points indicate stronger interactions.

Qualitative comparisons. Qualitative comparisons with the same methods are shown in Figure 4. For each scene, we visualize the generated frames at the first frame, 600th frame, 1200th frame, 1800th frame, and 2400th frame. As observed, our method preserves subject consistency during camera revisits, exhibiting less color drift and reduced geometric distortion compared to recent methods. These results demonstrate the effectiveness of PointWorld in maintaining long-term consistency over extended sequences.

4.3 Image Matching Results

We use MASt3R (Leroy et al., 2024) to perform image matching on the generated video frames to further evaluate the long-term consistency. The number of matched points can reflect the consistency between two given frames. As shown in Figure 5, by visualizing the matched points between input frames and generated frames, our method demonstrates strong preservation of the subject structure. We also provide a quantitative comparison in Table 2, reporting the average number of matched points between the first frame and the 1200th, 1600th, 2000th and 2400th frames across 50 scenes for different methods. These results further demonstrate the advantage of PointWorld in long-term consistent generation.

4.4 Visualization of 3D Interactions

We visualize cross-attention between encoded 3D point cloud features and video latent features x to examine the attention strength associated with each point during geometric conditioning. We project the corresponding 3D positions into the current target view and encode attention strength in grayscale, with brighter points indicating stronger attention. As shown in Figure 6, these points lie on scene structures such as door frames, walls, and counters, indicating that directly injected point cloud features provide geometry-related information for generation. Comparing results with and without Point Mask reveals that, without the mask, some projected interaction points are misaligned with structures in the output, and the generated results exhibit structural distortions or ghosting. With Point Mask, the interaction distribution better aligns with scene structures, and the generated geometry becomes more plausible. These results validate the effectiveness of explicit 3D representation calibration and Point Mask.

Table 2: Image matching results. Average numbers of matched points between the input frame and each specified generated frame across 50 scenes.

Method	Number of Matched Points↑			
	1200th frame	1600th frame	2000th frame	2400th frame
Yume 1.5	86.96	99.56	85.96	57.08
HY-World 1.5	491.04	562.54	357.06	430.68
Matrix-Game 3.0	368.08	230.84	262.68	451.02
LingBot-World 2.0	188.30	109.28	64.42	70.50
PointWorld (ours)	**621.68**	**725.08**	**950.56**	**1062.24**



Figure 7: Ablation study. With the addition of components, the geometric structure in the output image becomes clearer, and the color drift becomes less.

4.5 Ablation Study

We evaluate the roles of explicit 3D point cloud calibration and memory frame selection through four ablation variants. All variants use the same 50-scene test set, camera trajectories, input images, and resolutions. Figure 7 focuses on generation results during camera revisits to assess the preservation of previously observed scene structure and content. The first variant (Partial Views Only) uses only rendered RGB partial views as conditioning and selects historical frames based on DINOv2 feature similarity. The second variant (+ Explicit 3D Point Cloud Calibration) introduces direct point cloud encoding and geometric interaction through per-frame Point Masks to evaluate the role of explicit 3D conditioning. The third variant (+ Memory Frame Selection) further selects historical frames based on temporal order, spatial proximity, and content quality, and enables Reference Jump, forming the complete PointWorld. This variant evaluates how the retrieval strategy supports the reliability of 3D conditioning. The fourth variant (− Explicit 3D Point Cloud Calibration) removes the 3D calibration branch from the complete model while retaining the retrieval strategy and RGB conditioning. Comparison with the complete model further assesses the contribution of direct geometric injection beyond rendered conditioning under the same retrieval strategy.

As components are progressively introduced, geometric structures become clearer and color drift decreases. Specifically, comparisons between the first and second variants, as well as between the third and fourth variants, demonstrate that explicit 3D point cloud calibration helps the model generate clearer geometric structures. The comparison between the second and third variants demonstrates the effectiveness of the spatiotemporal memory retrieval strategy. Together, these four ablation experiments demonstrate the effectiveness of PointWorld.

5 Conclusions

We present PointWorld, which provides explicit geometric guidance for interactive video generation by directly encoding and injecting dynamic 3D representations, mitigating geometric inconsistencies and error accumulation over long generation horizons. This geometric conditioning complements rendered RGB detail conditioning, while spatiotemporal memory retrieval reduces reconstruction conflicts. Experiments demonstrate that PointWorld maintains long-term consistency over more than 2,000 frames of interactive generation and outperforms recent methods in both quantitative and qualitative comparisons.

AI USE STATEMENT

In this work, we used generative AI tools to assist with literature discovery, language editing, text condensation, and figure and table preparation. The authors reviewed and revised the AI-assisted material for accuracy and consistency with the implementation and experimental results. The authors take full responsibility for the final content of this work, including its text, claims, references, and visual artifacts.

REFERENCES

Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models. 2025. URL https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/.

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*, 2025.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. In *International Conference on Learning Representations*, volume 2026, pp. 50694–50718, 2026.

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.

Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*, 2025.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, et al. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025a.

Yining Hong, Beide Liu, Maxine Wu, Yuanhao Zhai, Kai-Wei Chang, Linjie Li, Kevin Lin, Chung-Ching Lin, Jianfeng Wang, Zhengyuan Yang, et al. Slowfast-vgen: Slow-fast learning for action-driven long video generation. In *International Conference on Learning Representations*, volume 2025, pp. 76609–76636, 2025b.

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*, 2025a.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent

video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)*, 44(6): 1–15, 2025b.

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.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Team HunyuanWorld. Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. *arXiv preprint*, 2025.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Guangyuan Li, Siming Zheng, Shuolin Xu, Jinwei Chen, Bo Li, Xiaobin Hu, Lei Zhao, and Peng-Tao Jiang. Magicworld: Interactive geometry-driven video world exploration. *arXiv preprint arXiv:2511.18886*, 2025a.

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2(3):6, 2025b.

Ruihuang Li, Caijin Zhou, Shoujian Zheng, Jianxiang Lu, Jiabin Huang, Comi Chen, Junshu Tang, Guangzheng Xu, Jiale Tao, Hongmei Wang, et al. Hunyuan-game: Industrial-grade intelligent game creation model. *arXiv preprint arXiv:2505.14135*, 2025c.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12037–12047, 2025.

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.

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

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model. *arXiv preprint arXiv:2512.22096*, 2025a.

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*, 2025b.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, et al. Matrix-game 3.5: Enhancing real-time streaming interactive world models with patch memory. *arXiv preprint arXiv:2608.29910*, 2026.

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.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. *arXiv preprint arXiv:2511.01266*, 2025.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026a.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026b.

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

Yuhao Wan, Lijuan Liu, Jingzhi Zhou, Zihan Zhou, Xuying Zhang, Dongbo Zhang, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. Geoworld: Unlocking the potential of geometry models to facilitate high-fidelity 3d scene generation. *arXiv preprint arXiv:2511.23191*, 2025b.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025b.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. IEEE, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20697–20709, 2024.

Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y Chen, Yuqing Yang, and Bohan Zhuang. Latent spatial memory for video world models. *arXiv preprint arXiv:2606.09828*, 2026a.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025d.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026b.

Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, et al. Infinite-world: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. *arXiv preprint arXiv:2602.02393*, 2026a.

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, 2026b.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4840–4851, 2024.

Yanhao Wu, Haoyang Zhang, Tianwei Lin, Lichao Huang, Shujie Luo, Rui Wu, Congpei Qiu, Wei Ke, and Tong Zhang. Generating multimodal driving scenes via next-scene prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6844–6853, 2025.

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.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025a.

Xuemeng Yang, Licheng Wen, Tiantian Wei, Yukai Ma, Jianbiao Mei, Xin Li, Wenjie Lei, Daocheng Fu, Pinlong Cai, Min Dou, et al. Drivearena: A closed-loop generative simulation platform for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26933–26943, 2025b.

Ying Yang, Zhengyao Lv, Tianlin Pan, Haofan Wang, Binxin Yang, Hubery Yin, Chen Li, Ziwei Liu, and Chenyang Si. Stableworld: Towards stable and consistent long interactive video generation. *arXiv preprint arXiv:2601.15281*, 2026.

Deheng Ye, Fangyun Zhou, Jiacheng Lv, Jianqi Ma, Jun Zhang, Junyan Lv, Junyou Li, Minwen Deng, Mingyu Yang, Qiang Fu, et al. Yan: Foundational interactive video generation. *arXiv preprint arXiv:2508.08601*, 2025.

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

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

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

Junyi Zhang, Charles Herrmann, Junhwa Hur, Chen Sun, Ming-Hsuan Yang, Forrester Cole, Trevor Darrell, and Deqing Sun. Loger: Long-context geometric reconstruction with hybrid memory. *arXiv preprint arXiv:2603.03269*, 2026a.

Songchun Zhang, Yaowei Li, Junhao Zhuang, Weiyang Jin, Haoyu Wang, Xin Lu, Yilang Sun, Shiyi Zhang, Haoran Li, Xiaoxiao Ma, et al. Echowm: Open and enterable omnimodal world models. *arXiv preprint arXiv:2608.23189*, 2026b.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025.

Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video generation with updatable spatial memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4245–4257, 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. *arXiv preprint arXiv:2602.02214*, 2026b.

A Implementation Details

Memory Retrieval Configuration. During inference, we use the camera pose C_m of the middle frame in the current target trajectory as the retrieval reference. We retrieve $N_{pool} = 160$ historical frames with the closest camera poses, considering camera-center distances and relative rotations. From this candidate set, we select the $N_{temporal} = 10$ earliest frames, the $N_{spatial} = 10$ frames with camera poses closest to C_m, and the $N_{sim} = 40$ frames with the highest cosine similarity to the initial input image in DINOv2 feature space. The combined selections provide historical observations for joint reconstruction. The initial input image serves as a fixed similarity reference. We compute and cache similarity scores for historical frames during generation to avoid repeated computation in subsequent chunks.

Reference Jump. Using the camera pose C_{prev} of the last frame in the previous chunk, we retrieve $N_{jump} = 5$ historical frames with the closest camera poses and select the earliest among them. The selected frame and its pose replace the first-frame input of the current partial-view sequence. After generation, we discard the corresponding first output frame and retain the previous chunk's last frame as the first frame of the current chunk.

B Cost Analysis

We compare generation quality and computational cost across methods, reporting the overall score, parameter count, inference time, and inference GPU memory. The overall score is the mean of the six VBench metrics reported in Section 4.2. We generate 2,400-frame videos using the same test images and target camera trajectories, using a single NVIDIA A100 80 GB GPU and an output resolution of 832×480 whenever supported. For methods requiring multiple GPUs, we multiply the elapsed inference time by the number of GPUs.

The results show that PointWorld achieves competitive computational efficiency with a relatively small model, despite not undergoing few-step distillation, while outperforming the compared methods in the quantitative evaluation.

Table 3: Cost analysis.

Method	Overall Score↑	Params (B)	Inference Time (s) ↓	Inference GPU Memory (GB) ↓
Yume 1.5	0.8049	5.2	431	33.8
HY-World 1.5	0.8015	8.6	8632	37.2
Matrix-Game 3.0	0.7843	6.5	**305**	**30.4**
LingBot-World 2.0	0.7968	18.5	5160	71.9
PointWorld (ours)	**0.8302**	**2.0**	2624	32.8

C More Visual Comparisons

As shown in Figure 8 and Figure 9, we provide more visual comparisons with recent state-of-the-art methods (Yume1.5 (Mao et al., 2025a), HY-World1.5 (HunyuanWorld, 2025), Matrix-Game 3.0 (Wang et al., 2026b), and LingBot-World 2.0 (Gao et al., 2026)).

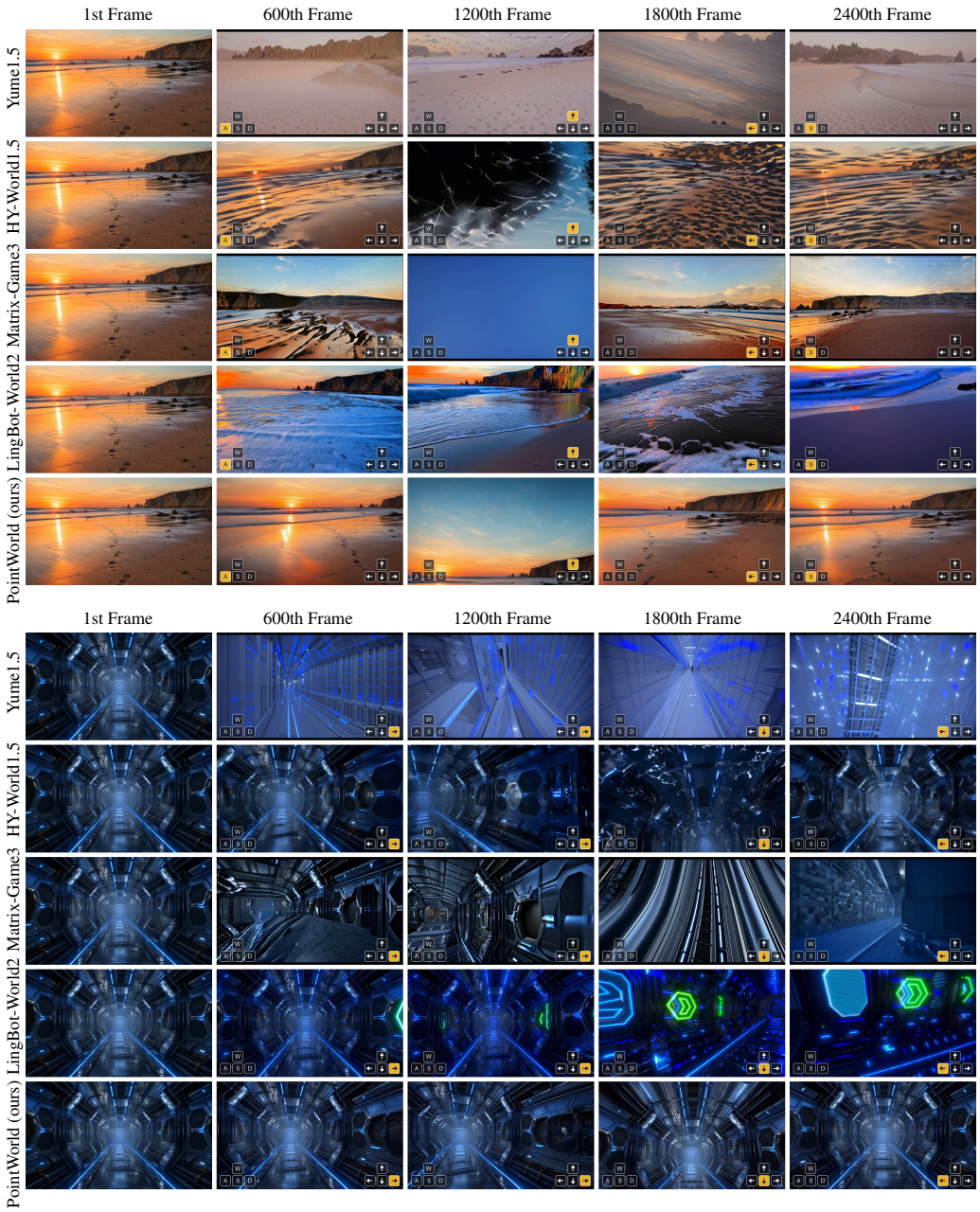

Figure 8: More visual comparisons. PointWorld can successfully extend consistent interactive generation, maintaining long-term coherence over 2,000 frames.

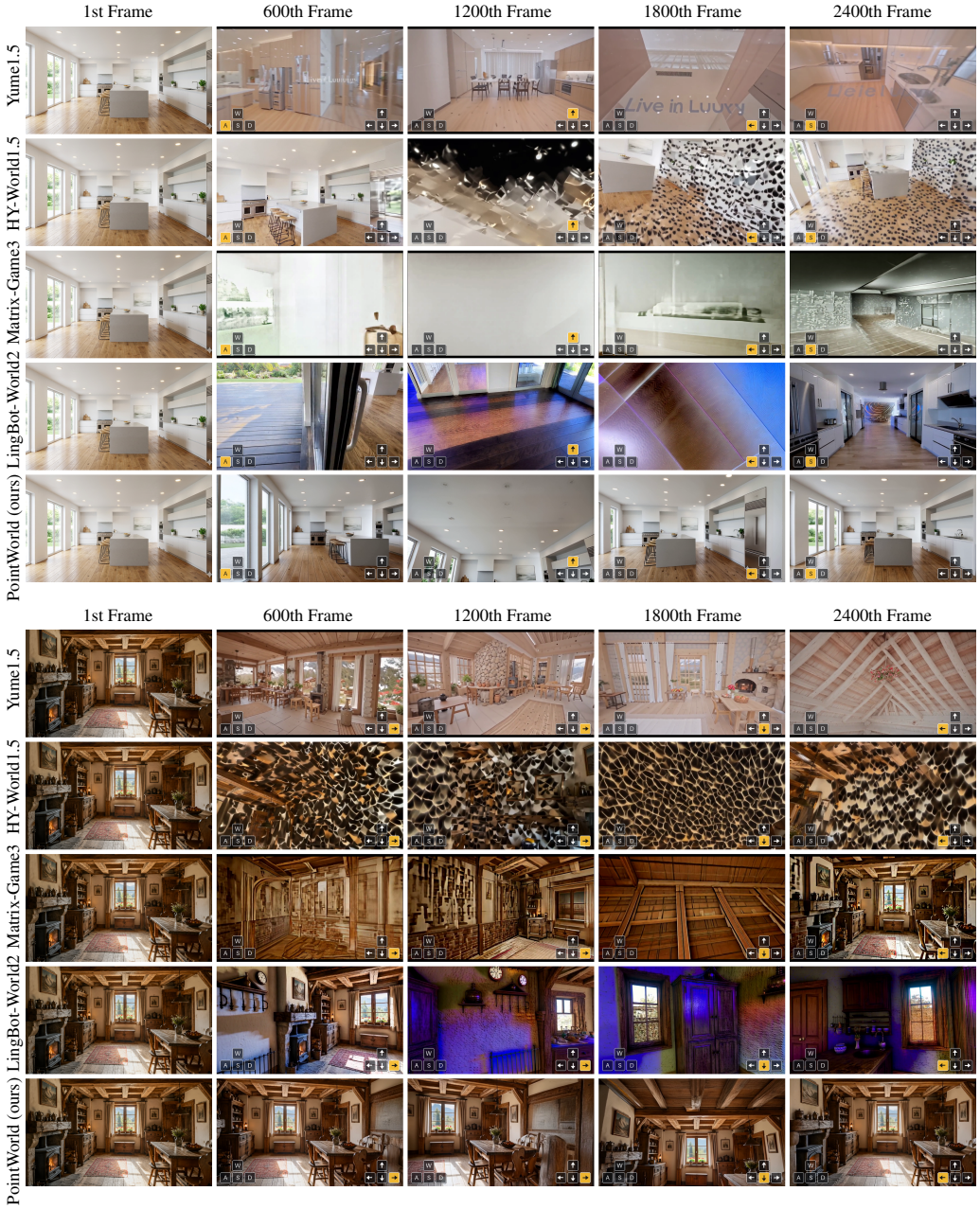

Figure 9: More visual comparisons. PointWorld can successfully extend consistent interactive generation, maintaining long-term coherence over 2,000 frames.