PERSISTENT4D: TRAINING-FREE STREAMING WORLD GENERATION WITH GEOMETRIC MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive world generation requires a video model to follow camera actions, remember previously visited places, and continue producing frames with a small active context. We introduce Persistent4D, a training-free streaming system that separates persistent world evidence from the working context of a frozen video generator. A timestamped RGB-D memory converts camera actions into target-view color and visibility conditions for an existing video completion interface. Retrieval prioritizes the initial observation and selects complementary historical coverage, while provenance-aware evidence controls which generated surfaces can condition future views. Reliable observed regions are reinforced within the generator's native denoising steps, without an added action-conditioned neural branch. A resident four-step generator runs alongside asynchronous geometry updates and an output-only video refiner. On the complete 158-case WBench navigation suite, Persistent4D achieves 74.56 under a Doubao frame-based evaluation protocol, with a navigation-accuracy score of 88.04/100 on its 78-case pure first-person subset. The evaluated configuration produces refined frames at a median 5.78 FPS on four H200 GPUs. Controlled repeated-view experiments show that persistent retrieval and observation priority preserve scene evidence beyond the generator's context, establishing a practical route from pretrained video components to memory-based streaming world generation.

1 INTRODUCTION

An interactive visual world must remain recognizable after the camera turns away. A user should be able to explore a room, look down a street, and return to a previously observed view without asking the generator to recreate that place from a short recent clip. At the same time, navigation demands continuous responses to camera motion. The resulting challenge combines three distinct responsibilities: preserving world evidence, translating actions into visual constraints, and synthesizing plausible newly exposed content.

Video generation has made substantial progress on each component. Camera-conditioned models learn motion interfaces from pose-video supervision (He et al., 2025; Bai et al., 2025), memory-augmented models recover historical scene information (Xiao et al., 2025; Li et al., 2025a), and causal video backbones support low-step continuation (Huang et al., 2025; Yang et al., 2026). We organize these pretrained capabilities around explicit world memory to support persistent, camera-controlled generation without task-specific training. This design makes interactive sessions available directly from frozen components, removing the need to collect action-conditioned training sequences, fit another adapter, or optimize each scene.

Our starting point is that persistent evidence and neural working context serve different purposes. A generator needs a compact description of what is visible now, while a world memory must preserve what was seen before. Keeping both inside the same sliding video context ties revisitation to context length. Instead, geometry can turn a potentially long history into a small target-view condition. An action first changes the camera pose. That pose retrieves and projects relevant RGB-D evidence. A frozen video prior then completes the uncovered regions through its existing visual conditioning interface.

This decomposition creates an important question of authority. An observed image and a self-generated image are not interchangeable evidence. Coverage alone can favor recent generated views

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.

and progressively displace the initial observation, even when the camera returns to it. Persistent4D therefore preserves the provenance of every memory entry. A visible observed record receives a reserved historical slot, and the remaining budget is chosen for complementary target-view coverage. Generated surfaces accumulate support across sufficiently distinct views, and unreliable conflicts become completion regions. At the interface to generation, exact observed RGB is recovered where geometric correspondence is reliable and reinforces the native denoising prediction. Memory thus controls both *which evidence to recall* and *how strongly to present it*.

The resulting system uses a frozen four-step causal video generator, a frozen depth model, and a frozen video refiner. Geometry updates run asynchronously after raw frames are committed. Refinement improves the displayed output without feeding its appearance changes into the world archive. Resident model weights and bounded active buffers support chunk-wise continuation, while the timestamped RGB-D archive remains available beyond the neural context window. No task-specific training, test-time weight optimization, or newly learned action branch is introduced.

We evaluate this design along three axes: multi-turn navigation, retention under repeated revisitation, and measured streaming cost. Across all 158 WBench navigation cases (Ying et al., 2026), the system obtains a five-dimension aggregate of 74.56 and a consistency score of 82.00 under our stated Doubao evaluation protocol. Navigation accuracy reaches 88.04/100 on the nested 78-case pure first-person subset. Six-scene controlled revisitation experiments isolate the roles of historical access, observation priority, conditioning strength, and output refinement. The same configuration used for WBench provides 5.78 refined FPS per four-H200 stream.

Our contributions are threefold: (i) a training-free camera-to-visual-condition pipeline that joins persistent geometric memory with a frozen streaming generator; (ii) a provenance-aware read/write interface that retains observation authority under a small retrieval budget; and (iii) navigation, revisitation, and runtime evidence linking this design to effective camera control, scene retention, and efficient four-step generation. Together, these results demonstrate camera-controlled streaming generation with persistent scene evidence and frozen pretrained components.

2 Related Work

Camera control and training-free interfaces. CameraCtrl learns a camera-control module for video diffusion (He et al., 2025), while ReCamMaster studies camera-controlled generative rendering from an input video (Bai et al., 2025). VACE provides a general video creation and editing interface that accepts visual content and masks (Jiang et al., 2025). Recent work also establishes training-free control routes. Warp-as-History offers a zero-shot geometric-history interface as well as optional single-video adaptation (Wang & He, 2026). World in World (WiW) uses clean visual keys and values in native self-attention to control a frozen world model, including event-synchronous novel-view rendering of source videos (Song et al., 2026). Persistent4D supports image-initialized online navigation by selecting evidence from an external timestamped RGB-D archive and delivering target-view conditions through a frozen VACE RGB/mask interface to a resident generator.

Persistent memory and geometry-guided retrieval. WorldMem studies long-term consistency through memory (Xiao et al., 2025). VMem constructs a surfel-indexed view memory from off-the-shelf geometric estimates and retrieves relevant historical views (Li et al., 2025a). COVRAG selects references by their additional target-view coverage and maintains depth information for long video generation (Joo et al., 2026). Building on explicit memory and coverage-based retrieval, Persistent4D prioritizes evidence within a bounded streaming interface: observed records retain priority, generated evidence receives support-dependent trust, and uncertain boundaries are delegated to completion. Its persistent state consists of immutable posed RGB-D records and lightweight evolving evidence maintained independently of the generator's attention history.

Streaming generation, restoration, and evaluation. Self Forcing and LongLive develop causal generation for efficient video continuation (Huang et al., 2025; Yang et al., 2026). We use the released Krea Realtime 14B backbone through Scope (Millon, 2025; Daydream, 2026), retaining its native four-step schedule. Depth Anything 3 supplies geometric estimates (Lin et al., 2026), and SwiftVR supplies one-step generative video restoration (Yan et al., 2026). All weights remain frozen. WBench evaluates interactive video models across multi-turn behavior and visual criteria

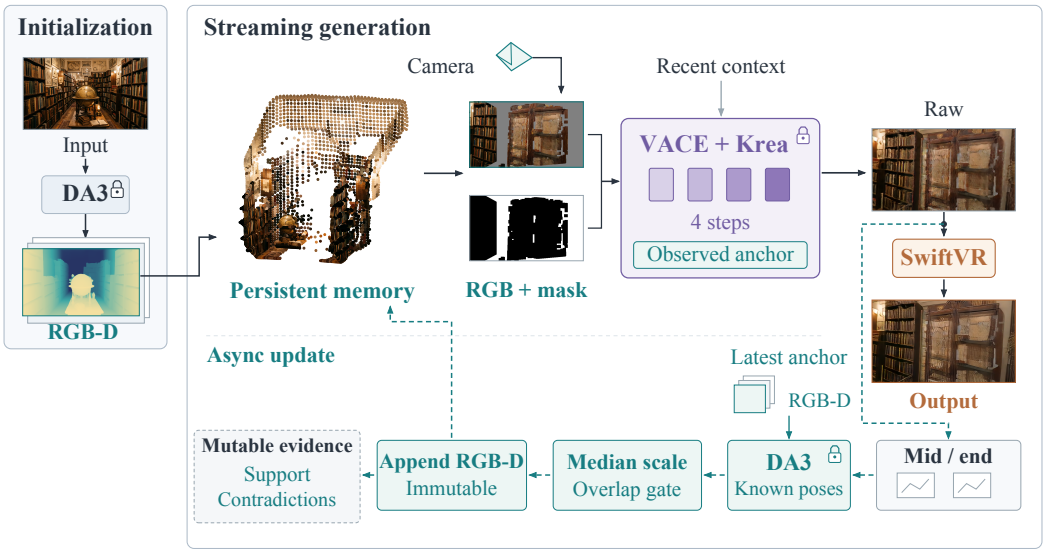


Figure 1: **Training-free streaming through persistent geometric memory.** Camera queries retrieve records with observation priority and form mask-aware view conditions for frozen four-step generation. The asynchronous write path reconstructs middle/end raw keyframes with the latest archived anchor. Reliable overlap aligns depth scale before immutable RGB-D append; cross-view support and contradictions update a separate evidence state for subsequent retrieval and rendering. SwiftVR refines the output. All networks are frozen. Conditions and outputs show bookstore case 53 at frame 393; the point cloud visualizes its initial RGB-D record.

(Ying et al., 2026). We use its complete navigation suite for absolute capability measurement and matched internal ablations for mechanism comparisons. MUSIQ (Ke et al., 2021) and SSIM (Wang et al., 2004) provide complementary measures of perceptual quality and repeated-view agreement.

3 PERSISTENT4D

3.1 OVERVIEW

Persistent4D couples persistent geometric memory with a frozen video generator (Figure 1). A session starts from an RGB image I_0, a camera T_0, and fixed intrinsics K. A deterministic integrator maps each camera command a_k to the requested poses for the next chunk. Memory retrieval and rendering convert these poses into visual evidence, and a mask-aware interface supplies it to the pretrained VACE conditioning path. At a chunk boundary,

$$\mathcal{C}_k = \mathrm{Render}(\mathrm{Retrieve}(\mathcal{M}_k, T_{k:k+B}), T_{k:k+B}), \quad Y_k = G_\theta(H_k, \mathcal{C}_k, p), \tag{1}$$

where B is the chunk length, H_k is the bounded recent generator context, p is the scene/perspective prompt, and $\mathcal{M}_k$ contains committed historical evidence. The condition $\mathcal{C}_k$ comprises target-view RGB, an unknown-region mask, and an observed-support mask. Committed raw frames Y_k feed asynchronous reconstruction and aligned RGB-D writes for later queries. Optional SwiftVR refinement operates on the output path; future generation and geometry continue to use raw frames. All learned components remain frozen, with camera control supplied by geometry rather than a newly trained action branch.

Sections 3.2–3.4 describe memory writing, retrieval, and the generation interface; Section 3.5 explains the acceleration strategies that support streaming execution.

3.2 MEMORY WRITING

A memory entry is

$$m_i = (I_i, D_i, Q_i, T_i, K_i, t_i, s_i), \qquad s_i \in \{\text{observed}, \text{generated}\}, \tag{2}$$

with image, depth, geometric quality, camera-to-world pose, intrinsics, timestamp, and provenance. Entries are immutable. A separate evidence state records support and contradictions. The term *4D memory* denotes this time-indexed archive of spatially posed observations and generated keyframes. It is accessed by unprojecting selected records on demand, without fitting a scene-specific radiance field.

Initialization and aligned updates. Depth Anything 3 initializes observed depth at a median of eight world units. Each reconstruction job uses a selected committed chunk's middle and final raw frames with a recent archived anchor under fixed known poses. Median depth ratios on reliable anchor overlap align new depth to the archive scale; insufficient overlap blocks publication. Confidence normalization by the 10th and 90th percentiles yields relative quality Q_i, rather than a calibrated probability.

Provenance and cross-view support. Observed provenance is permanent. Generated pixels gain support from distinct generation groups with adequate geometric quality, relative depth agreement, and compatible normals and color on a sparse evidence grid. Translations below 0.04 world units and rotations below 3° do not provide a new baseline; repeated evidence from one chunk adds no vote. Generated trust rises from 0.45 to 0.65 and then 0.75 after one and two supporting groups. Observed trust is one.

A surface substantially in front of reliable target depth can contradict its evidence; one behind it can simply be occluded. Two reliable contradictions reduce unsupported generated trust to 0.15. Unsupported evidence older than 96 frames becomes inactive but remains archived. This separates storage from eligibility and generated agreement from actual observation. Appendix A gives thresholds and publication details.

3.3 Memory Retrieval and Reliable View Conditions

Observation priority and complementary coverage. Every query includes the latest ready record and at most two historical records. Selection is causal and evaluated at the requested chunk's midpoint pose. A candidate's visible support is summarized on a 20×12 image grid. If an observed record covers at least six bins, it receives one reserved historical slot. The remaining slots maximize coverage not already explained by selected records:

$$j^* = \arg\max_{j \in \mathcal{J}} |\mathcal{B}_j \setminus \mathcal{U}|\, \bar{q}_j \bar{\tau}_j \left(0.35 + 0.65 \max(0, d_j^\top d_{\text{target}})\right), \tag{3}$$

where $\mathcal{J}$ is the candidate set, $\mathcal{B}_j$ is projected support, and $\mathcal{U}$ is its selected union. The weights $\bar{q}_j$ and $\bar{\tau}_j$ denote visible geometric quality and the support-based coverage surrogate in Appendix A; d is viewing direction. Historical selections are separated by at least 12 frames. Observation priority protects scene identity, while residual coverage retrieves complementary views within the remaining budget, without a learned retrieval model.

Visibility-aware RGB and unknown regions. A selected pixel u is unprojected as

$$X = R_i\left[D_i(u) K_i^{-1} \tilde{u}\right] + \mathbf{t}_i, \qquad (R_i, \mathbf{t}_i) = T_i, \tag{4}$$

where $\tilde{u}$ is the homogeneous pixel coordinate and $\mathbf{t}_i$ is camera translation. At each requested pose, points are depth-tested and splatted with a two-pixel radius. Quality, trust, and relative depth distinguish reliable support, tentative new-space support, and foreground conflicts. Unresolved regions remain unknown. The latest ready record is composed after history under the same evidence rules. Depth discontinuities and their neighborhoods are marked for completion, preventing uncertain splat boundaries from imposing rigid appearance constraints.

The system samples observed color from the immutable original RGB image. Under translation, resampling requires a reliable observed winner and depth-consistent bilinear correspondence. Under same-center rotation, projective mapping can also fill unknown regions wherever the correspondence is valid, while preserving accepted generated foreground. This separates the quality of observed color from the rasterization footprint of a sparse point cloud. Reprojection supplies the conditioning image, and the frozen generator synthesizes the raw video frames. The renderer outputs $(C, \mathbf{M}, A)$: visual content C, an unknown mask $\mathbf{M}$ with one indicating synthesis, and reliable observed support A.

3.4 A Mask-Aware Interface to Frozen Generation

Sparse renderings contain both reusable evidence and uncertain boundaries. Treating both with the same conditioning strength can preserve rasterization fragments; weakening every location also weakens established appearance. We therefore modulate the existing VACE condition using the renderer's support. After expanding unknown regions by eight pixels and aggregating them into the native conditioning tokens, the continuation weights are

$$w(v) = \begin{cases} 1, & v \text{ has clean known support,} \\ 0.35, & v \text{ intersects an unknown region or boundary.} \end{cases} \tag{5}$$

The bootstrap uses weight one. Unknown RGB is zero in the normalized conditioning representation. The weighting changes the strength of the pretrained visual hint and leaves the network weights unchanged.

Observed interiors receive a second, complementary constraint. We erode observed support by 16 pixels and retain a latent anchor only where all pixels and frames in its aggregation footprint are supported. Let z_{obs} be the observed inactive latent already computed by the VACE path, A_z its reliable support, and $\hat{x}_0^{(\ell)}$ the native clean prediction at denoising step ℓ. We apply

$$\hat{x}_0^{(\ell)} \leftarrow \hat{x}_0^{(\ell)} + \lambda A_z \odot \left(z_{\text{obs}} - \hat{x}_0^{(\ell)}\right), \qquad \lambda = 0.5. \tag{6}$$

This correction reuses existing conditioning latents at each of the four native steps. It introduces no extra generator pass, VAE encoding, or random draw. The decoded result and subsequent recent context both contain the correction. Mask-aware weighting decides how strongly a region conditions generation. Latent anchoring reinforces the subset supported by the original observation.

3.5 Acceleration Strategies

Krea-accelerated VACE completion. Scope combines the released distilled Krea Realtime 14B backbone with pretrained VACE-14B (Millon, 2025; Daydream, 2026). VACE uses Krea's four-step denoising schedule, with one conditional evaluation per step and no additional unconditional guidance pass. BF16 weights stay resident. RGB/mask inputs are encoded once per chunk; conditioning latents are reused across steps, while VACE hints are recomputed. Unchanged text embeddings and cross-attention keys and values are cached. An additional context forward refreshes self-attention keys and values from retained latent context. Two latent frames, nine RGB frames, and streaming TAE state carry context across chunks. A nine-frame bootstrap precedes 12-frame continuations.

Incremental DA3 reconstruction. A resident DA3 worker uses BF16 autocast on H200. Each job reconstructs the latest archived anchor with the middle and final raw frames of a selected committed chunk. This three-view window bounds inference input independently of archive length and avoids reconstructing every output frame. Known camera poses condition depth estimation, and the anchor links local estimates to archive scale (Section 3.2). Persistent geometry resides in external memory across these local inference jobs.

Asynchronous execution across stages. Generation, depth estimation, rendering, and refinement use separate GPUs. At a chunk boundary, the scheduler publishes completed geometry, retrieves ready past records, and renders the next condition without waiting for unfinished depth jobs. Committed raw output enters the geometry and refinement workers while generation can continue. Scale alignment and evidence publication remain boundary operations. The geometry queue holds at most one running and one pending job; under backpressure, newer work replaces pending reconstruction, preserving committed RGB. Rendering uses at most three records, and RGB-D and evidence caches each hold at most 12. The archive and its lightweight index grow while resident working tensors remain bounded.

Output-only refinement. SwiftVR uses 17 preceding raw frames with each continuation and discards the overlapping output prefix. Its weights remain resident, and bounded output queues apply backpressure to generation. Refinement overlaps subsequent computation without feeding appearance changes into future generation or geometry. All components remain frozen.

4 EXPERIMENTS

4.1 EVALUATION SETUP

Models and deployment. We use Krea Realtime 14B with pretrained Wan2.1 VACE-14B through Scope, DA3-GIANT-1.1, and SwiftVR, with no additional training. Videos are generated at 640×368. Each stream uses four H200 GPUs, assigned to generation, depth, rendering, and refinement. Six independent streams across 24 GPUs process the benchmark. Per-stream measurements are reported independently of this batch parallelism.

Navigation benchmark. We evaluate all 158 WBench navigation cases: 84 first-person and 74 third-person scenarios, containing 617 original turns. Each turn contributes 64 evaluation frames at an encoded 16 FPS, for 39,488 scored frames. The action adapter integrates first-person camera commands and follows the benchmark's follow/orbit camera rules for third-person cases. The task is camera-controlled navigation. Mixed non-navigation text turns hold the camera. No separate subject-action controller is introduced.

Scoring follows WBench's metric grouping and applicability rules. Navigation accuracy is computed from camera trajectories reconstructed from the generated videos by MegaSAM (Li et al., 2025b), so it evaluates the visual response to actions. We use Doubao Seed 2.0 Lite through a frame-based service adapter: video requests are represented by at most 12 uniformly sampled, timestamped images, while explicitly requested image lists are retained. Table 1 pairs our scores with published navigation results and labels the evaluation protocol for each source. Controlled ablations use matched protocols. Appendix B gives the complete scoring setup and metric breakdown.

Controlled revisitation. We additionally use six WBench initial scenes for a fixed repeated-view probe: three cycles of right rotation, left rotation, forward motion, backward motion, and hold. Each action lasts four seconds, giving 969 frames over 60.56 seconds of playback, with a maximum yaw of 72° and translation of 1.92 world units. Every variant uses the same scene-specific seed, trajectory, models, and chunk boundaries. Geometry publication is synchronized at boundaries and output is unpaced to give variants the same opportunity to access preceding history. Deployment throughput is measured separately using asynchronous WBench runs.

We measure *long-gap agreement* between frames requested at identical camera poses more than 16 seconds apart, excluding the initial pose, and *observation return* against the input image at scheduled returns to that pose. SSIM is evaluated on complete RGB frames (Wang et al., 2004). These complementary probes measure generated-view repeatability and retention of the actual input observation. MUSIQ measures perceptual quality on 32 uniformly sampled lossless continuation frames per clip (Ke et al., 2021). The fixed protocol averages scores within each scene and then across all six scenes.

4.2 NAVIGATION WITHOUT A NEW ACTION BRANCH

Table 1 establishes multi-turn navigation capability for the integrated training-free pipeline. The camera-derived visual interface yields 78.18 on navigation interaction, while the aggregate across quality, setting, interaction, consistency, and physical criteria is 74.56. Consistency reaches 82.00, including geometric consistency of 90.90 and segment continuity of 94.94. These dimensions directly address the intended use: a scene must respond to the requested motion and remain coherent across its generated turns.

The 78 pure first-person scenarios form a nested subset that directly exercises camera commands without third-person follow/orbit conventions. Here, navigation accuracy is 88.04/100, trajectory consistency is 79.12/100, and their official navigation combination is 83.58. These results show that action information can reach a frozen video generator through projected scene evidence: a newly trained action-to-feature branch is not required to obtain effective camera steering in this setting. The remaining evaluation categories and applicability counts are reported in Appendix B.

Visual navigation across scenes. Figure 2 shows selected rollouts through a museum gallery, a misty lakeshore, and a stylized bamboo forest. Turning reveals side galleries and shoreline structures, while forward motion advances along the forest path. Reversing the commands returns toward the

Table 1: **WBench Navi158.** Six nearest entries to ours by reported Overall from the official leaderboard (September 24, 2026). Scores range from 0 to 100; Overall equally averages the five dimensions.

Method	Overall↑	Quality↑	Setting↑	Interact.↑	Consist.↑	Physical↑
DreamX-World (5B AR)	75.0	77.5	80.8	78.6	74.9	63.3
ABot-World	74.7	76.8	71.4	84.0	79.5	61.7
Cosmos 2.5	74.6	72.9	83.3	63.1	86.5	67.4
Helios (Distilled)	74.3	73.8	75.4	64.4	81.5	76.2
Cosmos3-Nano	74.2	77.4	87.3	59.1	83.7	63.6
LTX 2.3	74.2	77.1	85.2	66.4	77.2	64.9
Persistent4D (ours)†	74.56	74.74	76.60	78.18	82.00	61.31

1st 2nd 3rd Highest displayed values in each column.
†Our SwiftVR outputs use Doubao Seed 2.0 Lite with up to 12 sampled frames per video request; published rows report the official native-video protocol. See Appendix B for protocol details.

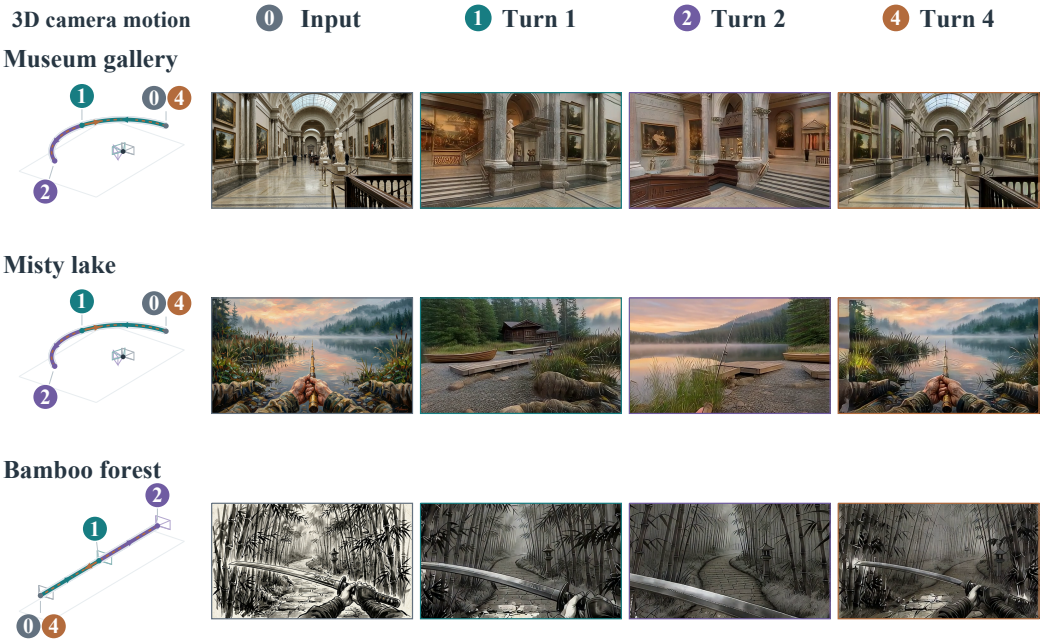


Figure 2: **Multi-turn navigation with a frozen pipeline.** Selected WBench cases 5, 78, and 82, showing the input and endpoints of turns 1, 2, and 4. Insets connect all commanded camera poses: arrows indicate progression, and dashed paths trace the return. Gallery and lake rotate at a fixed center, with arcs showing the viewing-direction sweep; bamboo translates along a line. Colors match the image panels, and poses 0 and 4 coincide. Insets are independently scaled. Generated panels use SwiftVR and retain the full field of view.

initial composition. The same frozen pipeline supports these different scene appearances through camera-derived visual conditions.

4.3 What Makes Persistent Memory Useful?

Persistent access contributes beyond recent context. The eight-second-window variant restricts the age of generated historical records while preserving the observed anchor, latest record, writing procedure, and retrieval budget. This isolates access to older generated evidence under a fixed conditioning budget. The full system improves mean long-gap SSIM from 0.3514 to 0.3553. On

Table 2: **Controlled 60-second revisitation across six scenes.** Numbers use refined output; larger is better. "Long-gap" compares identical requested poses more than 16 seconds apart; "return" compares scheduled returns with the input observation. The full system achieves the highest MUSIQ among these variants and improves both retention measures over restricted history and coverage selection without observation priority.

Condition	MUSIQ	Long-gap SSIM	Return SSIM
Full system	**74.90**	0.3553	0.4797
Generated-history window: 8 s	74.76	0.3514	0.4765
No observation-priority slot	74.65	0.3184	0.4099
Uniform low condition (0.35)	74.80	0.3497	0.4681
Uniform high condition (1.0)	74.53	**0.3732**	**0.4908**
Feathered observed anchor	74.85	0.3565	0.4815

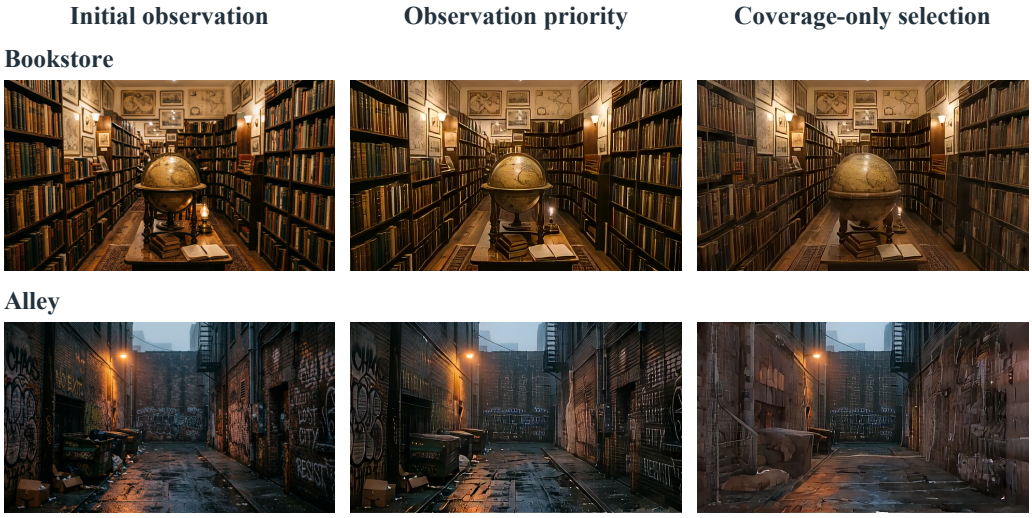


Figure 3: **Observation authority after repeated navigation.** Bookstore and alley scenes at the scheduled third return (frame 969). Both generated columns use SwiftVR and the same requested camera pose. Reserving an observed-history slot better retains the original shelf geometry, globe proportions, and alley wall appearance. The coverage-only variant can select self-generated history in place of that observation. Full-frame quantitative results use all six scenes.

repeated poses at least 45° from the initial orientation, agreement rises from 0.2694 to 0.2825, with improvements in all six scenes. Persistent retrieval is especially useful at these larger rotations, where historical records provide complementary views beyond the recent context. The full system retrieves generated history older than eight seconds in all six scenes, reaching an age of 52.5 seconds.

Observation priority prevents evidence displacement. Removing the reserved observed slot leaves the observation eligible for ordinary coverage selection. This isolates the value of provenance priority within the same small history budget. The full design increases refined long-gap SSIM from 0.3184 to 0.3553 in all six scenes, and observation-return SSIM from 0.4099 to 0.4797 on average. The corresponding raw return scores are 0.5061 and 0.6141, showing that the effect precedes refinement. A visible input observation provides an independent appearance reference that cannot be replaced by additional agreement among self-generated images. Figure 3 illustrates its role after the third return.

Spatially selective conditioning preserves visual quality. Mask-aware conditioning assigns full strength to reliable interiors and lower strength to uncertain boundaries. Compared with uniform low conditioning, the full design improves observation-return SSIM in every scene and increases mean

Table 3: **Measured quality and cost.** Quality uses the full configuration on the six-scene revisit probe; throughput uses the complete 158-case asynchronous WBench run with the same component configuration. FPS measures distinct refined frames becoming available on the server, excluding cold startup, export, and evaluation.

Evidence	Measurement	Value
Revisit quality	Raw MUSIQ	67.15
	Refined MUSIQ	74.90
	Mean improvement	+7.75
Per-stream deployment	GPUs	4× H200
	Native denoising steps / chunk	4
	Continuation frames / chunk	12
	Median continuation critical time	2.11 s
	Median distinct refined-frame availability	5.78 FPS
	Range across 158 cases	5.47–6.10 FPS

MUSIQ from 74.80 to 74.90. Compared with uniform high conditioning, it improves raw MUSIQ in all six scenes and refined MUSIQ by 0.37 on average. The full configuration achieves the highest refined MUSIQ among the tested variants while retaining observation-supported appearance. The feathered anchor produces similar aggregate results, showing that the observed-evidence interface supports both tested boundary profiles. Appendix C reports the raw and refined scores.

4.4 Refinement and Streaming Cost

SwiftVR improves the full system's MUSIQ from 67.15 to 74.90 (Table 3). Across all six ablation conditions, every one of the 36 clips and all 1,152 sampled frames improve in MUSIQ. Output-only refinement couples this perceptual gain with a stable memory interface: future views use the raw, geometrically associated stream, while the user receives restored frames. Appendix C reports both output domains, including their pixel-level agreement metrics.

Four-step Krea achieves a median 5.78 distinct refined FPS per four-H200 stream (5.47–6.10 across 158 cases), with a median continuation critical time of 2.11 seconds per 12 frames. These are server-side availability measurements; 16 FPS is the playback time base. Resident weights amortize initialization, and bounded neural context enables generation to overlap geometry updates. Camera control and memory require no new training. Archive storage and indexing grow with history; expensive resident tensors remain limited to a small active set.

5 Conclusion

Persistent4D turns frozen video components into a streaming, camera-controlled world through persistent geometric memory. Camera actions become visual conditions, observation priority preserves scene evidence, and asynchronous reconstruction extends history beyond the generator's active context. Navigation performance, repeated-view ablations, and four-step inference demonstrate a practical design: organize persistent evidence around a pretrained video prior to obtain world behavior without training another action branch.

AI Use Statement

Generative AI tools assisted with method and experiment design, implementation, experiment organization, analysis of recorded results, literature verification, manuscript drafting, figure preparation, and translation into Chinese for internal review. Doubao is also an evaluation component, as described in Section 4 and Appendix B. The authors are responsible for the manuscript's content, citations, and experimental claims.

Ethics Statement

The system is designed for interactive visualization and research on persistent generative environments. Experiments use WBench scene inputs and pretrained components, whose respective data and model terms govern reuse. Generated views are presented as synthetic model outputs.

Reproducibility Statement

The method specifies its frozen components, camera interface, memory representation, selection rules, and conditioning operations in Section 3. Appendix A records implementation constants and scheduling details; Appendix B specifies benchmark scope, VLM transport, and aggregation; Appendix C defines the controlled trajectories, metric sampling, seeds, and complete ablation results. Runtime measurements refer to the same component configuration used for the navigation scores.

References

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*, pp. 14834–14844, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Bai_ReCamMaster_Camera-Controlled_Generative_Rendering_from_A_Single_Video_ICCV_2025_paper.pdf.

Daydream. Scope: A tool for running and customizing real-time, interactive generative AI pipelines and models. Software, 2026. URL https://github.com/daydreamlive/scope. Accessed September 26, 2026.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Z4evOUYrk7.

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*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f4823f831af67a3ef15e41a85434422a-Paper-Conference.pdf.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Jiang_VACE_All-in-One_Video_Creation_and_Editing_ICCV_2025_paper.pdf.

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. URL https://arxiv.org/abs/2606.02479.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Ke_MUSIQ_Multi-Scale_Image_Quality_Transformer_ICCV_2021_paper.html.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. VMem: Consistent interactive video scene generation with surfel-indexed view memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25690–25699, 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/html/Li_VMem_Consistent_Interactive_Video_Scene_Generation_with_Surfel-Indexed_View_Memory_ICCV_2025_paper.html.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10486–10496, 2025b. doi: 10.1109/CVPR52734.

2025.00981. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_MegaSaM_Accurate_Fast_and_Robust_Structure_and_Motion_from_Casual_CVPR_2025_paper.html.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/pdf/9d69fc515ade062e6a76a0e4c7e6835ad3e21402.pdf.

Erwann Millon. Krea Realtime 14B: Real-time video generation. Software, 2025. URL https://github.com/krea-ai/realtime-video.

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

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

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://www.ece.uwaterloo.ca/~z70wang/research/ssim/.

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. URL https://arxiv.org/abs/2504.12369.

Jiaqi Yan, Xiangyu Chen, Xinlin Zhong, Haibin Huang, Chi Zhang, Jie Liu, Jiantao Zhou, and Xuelong Li. SwiftVR: Real-time one-step generative video restoration. *arXiv preprint arXiv:2606.09516*, 2026. URL https://arxiv.org/abs/2606.09516.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time interactive long video generation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=nCAODkpsPJ.

Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. WBench: A comprehensive multi-turn benchmark for interactive video world model evaluation. *arXiv preprint arXiv:2605.25874*, 2026. URL https://arxiv.org/abs/2605.25874.

A Implementation Details

A.1 Frozen Components and Camera Interface

The generator is Krea Realtime 14B with the pretrained Wan2.1 VACE-14B visual module, integrated through Scope. Generation uses BF16 and the native timestep list $[1000, 750, 500, 250]$. DA3-GIANT-1.1 estimates depth and relative quality with the supplied camera poses. SwiftVR is used as the output refiner. No component is trained, distilled, finetuned, or weight-optimized as part of Persistent4D; the released pretrained weights supply their learned priors.

The next camera chunk is determined before rendering. First-person actions update the camera trajectory directly. Third-person control uses the benchmark's follow/orbit camera conventions with a fixed character heading and description. Scene and perspective prompts remain fixed within the prescribed generation setup. Sixteen mixed non-navigation turns in Navi158 (ten event edits, five subject actions, and one perspective switch) are retained as camera holds; their text-adherence categories are outside the navigation Interaction metric.

The nine-frame bootstrap is followed by 12-frame continuation chunks, corresponding to three new native latent frames per continuation. At the evaluated resolution, Krea uses local attention size six and a three-frame KV setting, recomputing keys and values from its retained chunk context. It keeps

two context latent frames and nine decoded RGB frames, while temporal autoencoder state persists across continuations. The current positional-table implementation supports up to 4,089 native output frames in one session; our controlled long-revisit experiments generate 969.

A.2 Evidence and Rendering Constants

Table 4: Memory and visual-interface constants used by the full system.

Operation	Setting
Initial depth scale	Median depth = 8 world units
Quality normalization	Per-record 10th/90th percentiles
Evidence sampling	Stride 4 in image space
Support baseline	Translation ≥ 0.04 or rotation $\geq 3°$
Support quality / depth agreement	Both $Q > 0.4$; relative error ≤ 0.1
Support normal / color agreement	$\lvert n_i^\top n_j \rvert \geq 0.8$; mean RGB error ≤ 0.25
Observed / generated initial trust	1.0 / 0.45
Generated trust after 1 / 2 supports	0.65 / 0.75
Unsupported contradiction penalty	Trust 0.15 after 2 contradictions
Unsupported generated expiry	Age greater than 96 frames
Historical slots / spacing	2 slots / at least 12 frames
Coverage grid / observed reservation	20×12 bins / at least 6 visible bins
Splat radius	2 pixels
Reliable surface in arbitration	Quality ≥ 0.35, trust ≥ 0.60
Final known-mask quality floor	Observed $Q \geq 0.25$; generated $Q \geq 0.55$
Tentative new-space support	Quality ≥ 0.55, trust ≥ 0.35
Foreground conflict winner	Quality ≥ 0.65, trust ≥ 0.70
Depth-edge handling	Relative gradient > 0.2; dilation 2; closing 8
Unknown-boundary expansion	8 pixels before condition-token aggregation
Clean / boundary conditioning	1.0 / 0.35
Observed latent anchor	Erosion 16 pixels, strength 0.5

A support vote must originate from a distinct generation group and adequate baseline, and votes are capped at two for the trust schedule. A contradiction is recorded for a source depth $z_s < 0.8 z_t$, with target trust at least 0.6 and both qualities above 0.6. A source behind a target is not penalized simply for being occluded. Depth alignment uses reliable overlap with an anchor whose previous depth is already in archive coordinates. Rejected updates do not replace committed records.

For observed RGB reprojection under translation, the observed record must be the accepted unmasked winner, its quality must be at least 0.5, and all nonzero-weight bilinear contributors must meet the same quality threshold with relative depth error at most 0.1. Under same-center rotation, the projective correspondence can fill missing observed support while preserving already accepted generated foreground. Reprojected RGB supplies the conditioning image used by the frozen generator to synthesize its output.

Mask-aware VACE weights use 16×16 spatial token support and continuation temporal samples 2, 6, and 10 in zero-based indexing. The observed latent anchor uses an all-supported rule across its four-frame and 8×8 spatial footprint. The corresponding observed latent is taken from the existing inactive VACE representation. Equation 6 therefore adds elementwise arithmetic at native denoising steps, not another model invocation.

The coverage selector uses a record-level trust surrogate distinct from the renderer's pixelwise trust. Let s be the fraction of the coarse evidence grid with at least one support vote, r the fraction with at least two contradictions and no support, and a the record's age in frames. Then

$$\bar{\tau}_i = \begin{cases} 1, & \text{observed,} \\ 0.65s + 0.45\,\mathbf{1}[a \leq 96](1 - s - r), & \text{generated.} \end{cases} \tag{7}$$

The quality term in Equation 3 averages projected coarse-point quality inside the queried image. The latest ready record initializes the covered-bin union. Observation reservation precedes greedy selection; zero-score candidates are omitted.

A.3 Publication, Caches, and Refinement

Geometry is submitted after raw-frame commitment. An asynchronous worker can contain one running reconstruction and one pending reconstruction; pending jobs can be coalesced under back-pressure. At a chunk boundary the renderer sees published results from past committed frames, maintaining causal conditioning. Archive writing has a bounded active/pending queue of four. CPU RGB-D and evidence caches each retain at most 12 records; the conflict renderer materializes at most the latest record and two selected historical clouds. The driver retains 64 recent raw frames for its local operations.

SwiftVR processes the bootstrap and then a bounded window containing 17 previous raw frames and the new 12-frame chunk. It retains model weights while resetting restoration stream state for each window, discards prefix outputs, and sends new refined frames to the output stream. Both DA3 and future generation consume raw frames. Refinement changes presentation without creating an additional appearance-feedback loop in the memory.

The active buffers above have fixed configured limits. Immutable records and lightweight metadata are kept for the session, and historical candidate selection scans that index. Persistent storage and candidate-search work scale with accumulated history, while the neural context and resident RGB-D tensors remain bounded. Keeping the archive external preserves historical evidence across successive active contexts.

B WBench Evaluation Protocol

B.1 Scope, Scoring, and Visual-Judge Transport

Navi158 contains all 158 navigation cases, with 84 first-person and 74 third-person perspectives. It comprises 617 original turns and 39,488 evaluation frames at 64 frames per turn. Generation uses native nine-/12-frame chunks, and clips are assembled to the prescribed per-turn evaluation lengths. Scored videos use SwiftVR output. The nested 78-case pure first-person subset contains 309 turns and 19,776 evaluation frames.

The VLM judge is `doubao-seed-2.0-lite-260215-qdbbseed`. Video requests are adapted to at most 12 uniformly sampled images in chronological order, each paired with its actual timestamp. The maximum image edge is 1,024 pixels and JPEG quality is 85. When an evaluator requests an explicit list of images, all requested images are retained. Original WBench question templates and metric applicability rules are preserved. Published leaderboard scores use the standard native-video transport, with JPEG quality 50 for encoded image inputs.

Non-VLM quality metrics follow the evaluator's standard implementations. HPSv3 uses its evaluator normalization, with percentile anchors -5.21 and 8.61. For each metric, we average over applicable cases. The report rounds metric means to four decimal places, averages metrics within each of five dimensions, then equally averages the dimensions. There are 6 Quality, 2 Setting, 1 Interaction, 8 Consistency, and 2 Physical metrics. Thus, the aggregate is a mean over dimensions rather than a flat mean over 19 metrics. All 2,286 applicable per-case metric outputs are present.

On a 0–100 scale, navigation accuracy is 81.27 across 158 cases and navigation consistency is 75.09 across its 147 applicable cases. The evaluator combines their separately applicable means into the navigation trajectory score. In the pure first-person subset, accuracy is 88.04 over 78 cases and consistency is 79.12 over 77 applicable cases, giving the reported 83.58 navigation combination.

B.2 Navigation Trajectory Measurement

The evaluator reconstructs camera-to-world poses from the scored refined videos using MegaSAM (Li et al., 2025b). Its depth preprocessing and camera tracking are separate from our production DA3 memory. Evaluation therefore measures the trajectory expressed by the synthesized video. The

Table 5: Complete Navi158 metric breakdown on a 0–100 scale. Counts indicate applicable cases.

Dimension	Metric	Score	Cases
Quality	Aesthetic quality	57.47	158
	Imaging quality	72.91	158
	Temporal flickering	92.71	158
	Dynamic degree	74.68	158
	Motion smoothness	96.30	158
	HPSv3 quality	54.34	158
Setting	Scene adherence	68.61	158
	Subject adherence	84.59	74
Interaction	Navigation trajectory	78.18	158
Consistency	Background consistency	90.28	158
	Segment continuity	94.94	158
	Perspective consistency	45.99	103
	Subject consistency	85.21	103
	Geometric consistency	90.90	158
	Photometric consistency	82.73	158
	Spatial consistency	83.82	60
	Gated spatial consistency	82.13	60
Physical	Visual plausibility	55.14	158
	Causal fidelity	67.47	50

benchmark constructs an adaptive reference trajectory from each action and perspective, adapting displacement or orbit parameters to reconstructed motion. Each turn is resampled to 20 points by translational arc length, with uniform temporal sampling when translational motion is negligible.

Let E_t denote mean Euclidean camera-center error and E_r mean angular rotation error relative to that reference, L the reconstructed path length, and Θ cumulative rotation in degrees. WBench defines

$$e_t = \min\left(\frac{E_t}{\max(L, 0.5)}, 1\right), \quad e_r = \min\left(\frac{E_r}{\max(\Theta, 10^\circ)}, 1\right), \quad A = 1 - \frac{e_t + e_r}{2}. \tag{8}$$

Navigation accuracy continuously aggregates normalized translation and rotation errors. Navigation consistency compares repeated or symmetric action turns in normalized starting frames, averages similarly normalized translation and rotation errors over eligible pairs, and applies the same $1 - \frac{1}{2}(e_t + e_r)$ form. Cases without a qualifying pair have no consistency value. Accuracy and consistency are averaged over their respective applicable cases before combination.

B.3 Runtime Definitions

Runtime statistics cover all 158 cases with the same Krea/VACE, memory, DA3, and SwiftVR settings used for scoring. Four H200 GPUs support each independent stream, one each for the generator, depth worker, renderer, and refiner. Three eight-GPU nodes execute six independent streams for batch evaluation.

Distinct refined-frame availability is measured as $(N-1)/(t_N - t_1)$, where t_i is the server availability time of the ith distinct refined frame. The median per-case value is 5.7828 FPS, with range 5.4738–6.0963. Steady raw-frame commit throughput has median 5.6693 FPS, and active-compute throughput has median 5.6855 FPS. The latter steady measurements skip the first three chunk reports when available. The median of per-case mean continuation critical times is 2.1106 seconds per 12-frame chunk. Cold model loading, bootstrap, file export, and evaluation are excluded from steady throughput. Throughput counts distinct generated frames using server-availability timestamps under the prescribed benchmark camera plans.

C Controlled Ablations

C.1 Fixed Repeated-View Protocol

The six scenes are WBench cases 51, 52, 53, 69, 70, and 72. Each uses seed $2025 +$ case ID. The trajectory repeats right, left, forward, backward, and hold three times, with four seconds per action. This custom repeated-view trajectory uses WBench scene inputs to probe retention over successive visits. Every clip contains a nine-frame bootstrap and 960 continuation frames, totaling 969 frames and 81 chunks. All cases are generated at 640×368, with a 16 FPS playback timeline.

Long-gap pairs have matching requested camera-to-world matrices and intrinsics within 10^{-6}, a temporal gap strictly greater than 16 seconds, and a non-initial pose. There are 16 pairs in each of cycles two and three, giving 32 pairs per scene. The 45° subset contains three pairs per cycle, or six per scene. Observation-return frames are 266, 297, and 329 in the first cycle; 586, 617, and 649 in the second; and 906, 937, and 969 in the third. Each is compared with the actual input image. The revisit probe matches commanded poses, while WBench navigation metrics evaluate the motion expressed by the synthesized videos.

SSIM uses a 7×7 RGB window on complete frames, without alignment or selective cropping. Return SSIM measures full-frame appearance retention relative to the initial observation, including changes in dynamic content. Long-gap and observation-return scores are averaged within each cycle, across cycles, then across scenes. MUSIQ uses the KonIQ model on 32 uniformly sampled lossless continuation frames, excluding the bootstrap. All six scenes produce valid values for the 60-second protocol.

C.2 Variant Definitions and Complete Results

The full system corresponds to Section 3. Each variant changes a single memory/interface choice:

- **Eight-second generated-history window:** generated historical candidates older than 128 frames are ineligible. The initial observed record and latest ready record are exempt; writing and the two-slot historical budget are unchanged.
- **No observation priority:** the observed slot is not reserved. The observation remains an ordinary candidate for the same coverage selector.
- **Uniform low / high conditioning:** every continuation token uses strength 0.35 / 1.0, respectively. Bootstrap strength remains one.
- **Feathered observed anchor:** multiscale support at 0, 4, 8, and 16 pixels replaces the strict 16-pixel erosion profile; anchor strength remains 0.5.

Geometry publication is synchronized between chunks for these comparisons, and generation is unpaced. All six variants otherwise share the same frozen models and camera plan.

Table 6: Complete raw/refined results for the 60-second probe. Each cell reports raw / refined.

Condition	MUSIQ	Long-gap SSIM	Return SSIM
Full	67.153 / 74.901	0.3782 / 0.3553	0.6141 / 0.4797
8-s history	67.287 / 74.765	0.3735 / 0.3514	0.6104 / 0.4765
No observed priority	66.775 / 74.648	0.3446 / 0.3184	0.5061 / 0.4099
Uniform low	67.167 / 74.798	0.3800 / 0.3497	0.5994 / 0.4681
Uniform high	66.326 / 74.535	0.3990 / 0.3732	0.6230 / 0.4908
Feathered anchor	67.161 / 74.851	0.3797 / 0.3565	0.6194 / 0.4815

The full system improves long-gap SSIM over the no-priority variant in all six scenes for both raw and refined output. Against the eight-second window, the 45° subset improves in all six scenes in both domains. Scenes are the units of paired comparison, and each scene uses one shared seed across all conditions.

Generated history older than eight seconds is selected 112 times in the full runs, including 47 selections older than 20 seconds. All six scenes use evidence older than eight seconds and five use

Table 7: Repeated poses at least 45° from the initial orientation. All six scenes are included.

Condition	Raw SSIM	Refined SSIM
Full	0.3062	0.2825
8-s history	0.2850	0.2694
No observed priority	0.2876	0.2648
Uniform low	0.3303	0.2959
Uniform high	0.3428	0.3248
Feathered anchor	0.3070	0.2843

evidence older than 20 seconds. The oldest selected source is 52.5 seconds old. These counts are recorded at memory selection, before pixelwise visibility arbitration.

C.3 Complementary Short-Revisit Component Study

A separate 30-second probe isolates three additional mechanisms: access to generated historical records, coverage versus recent-history selection, and the observed RGB/latent interface. It uses the same six initial scenes, with 477 frames per clip. Its repeated-view metric uses a gap of at least four seconds and 16 pairs per valid scene; five scenes have valid pairs, while case 70 has none.

Table 8: Short-revisit component study. MUSIQ averages six scenes; repeated-view SSIM averages the five applicable scenes. Each cell reports raw / refined.

Condition	MUSIQ	Repeated-view SSIM
Full	67.158 / 74.166	0.3174 / 0.2960
No generated history	67.649 / 74.515	0.3103 / 0.2825
Recent-history selection	67.339 / 74.225	0.2889 / 0.2653
No observed RGB reprojection	67.671 / 74.289	0.2847 / 0.2581
No observed latent anchor	67.463 / 74.359	0.3017 / 0.2757
Uniform conditioning	67.294 / 74.092	0.3202 / 0.2861

The full observed interface has the highest refined repeated-view agreement in this component study. Observed RGB reprojection raises refined SSIM from 0.2581 to 0.2960, and native latent anchoring raises it from 0.2757 to 0.2960 under the corresponding single-component comparisons. Both operations contribute to repeated-view agreement by preserving the original observation in the generation interface.