4LC: 4D LATENT CACHE FOR EFFICIENT CONDITIONING IN DYNAMIC WORLD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

For video generation in dynamic world modeling, observed static and dynamic content should be efficiently reused to provide conditions for repeated generation at different times and viewpoints. We introduce 4LC, a 4D latent cache with composite visual representations and a flexible spatiotemporal access interface. As a latent cache, 4LC directly reuses visual representations of static and dynamic content to provide generation conditions, avoiding per-request RGB rendering and re-encoding. As a composite cache, 4LC binds visual representations from multiple visual encoders (e.g., VAE, DINO, V-JEPA, VGGT) to shared geometry and motion, allowing the composite representations to describe the dynamic world with rich information. Besides, we design a flexible spatiotemporal interface, which directly reads the cache at independently specified times and viewpoints and combines readouts across observations, providing a common conditioning format for diverse downstream tasks. Building on 4LC, we develop 4LC-Video to address multiple video generation tasks through different configurations of source observations, spatiotemporal requests, and historical reuse, including camera control, temporal control, novel-view synthesis, scene revisits, and long-term generation. Our experiments show more efficient condition acquisition, effective spatiotemporal controllability, and scene consistency across tasks using a single 4LC-Video model without task-specific fine-tuning.

1 INTRODUCTION

In dynamic world modeling, observed scene information provides conditions for video generation at specified times and viewpoints. These conditions should convey persistent static structure together with dynamic content at the requested scene time. Geometry and motion help relate the observed static and dynamic content across viewpoints and scene times (Yao et al., 2025; Qiao et al., 2026), supporting consistency in the generated video (Fang et al., 2026). Because the same observations serve different requests and subsequent generation segments, reusing previously extracted information can reduce the recurring cost of condition acquisition.

Existing conditioning pipelines retain scene information in explicit 3D reconstructions or spatial memories and render it into target views (Ren et al., 2025; Zhao et al., 2026; Wu et al., 2025). Static and dynamic content can be maintained in separate components (Duan et al., 2026; Zheng et al., 2026) or represented together through shared 4D geometry, as in 4D Gaussian Splatting (4DGS) (Wu et al., 2024). This shared geometry allows both to be rendered at requested times and viewpoints (Yang et al., 2026). However, these geometric conditioning pipelines still require RGB rendering and re-encoding for each new generation request (Fig. 1b). This repeated processing can consume a substantial share of total generation time. Latent reuse already reduces repeated computation in video world models, both in history processing (Yi et al., 2026) and in denoising (Feng et al., 2026). For conditioning, Mirage caches latent features with scene geometry and warps them directly to target views (Wang et al., 2026b). These features lie in the latent space of a specific generator, and dynamic content is excluded from the cache. These limitations motivate us to investigate a cache mechanism that extends latent reuse to dynamic content, enabling efficient conditioning in dynamic world modeling.

To this end, we introduce 4LC, a 4D latent cache that jointly stores visual representations of static and dynamic content extracted by multiple visual encoders (Fig. 1a). Motion determines the 3D po-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

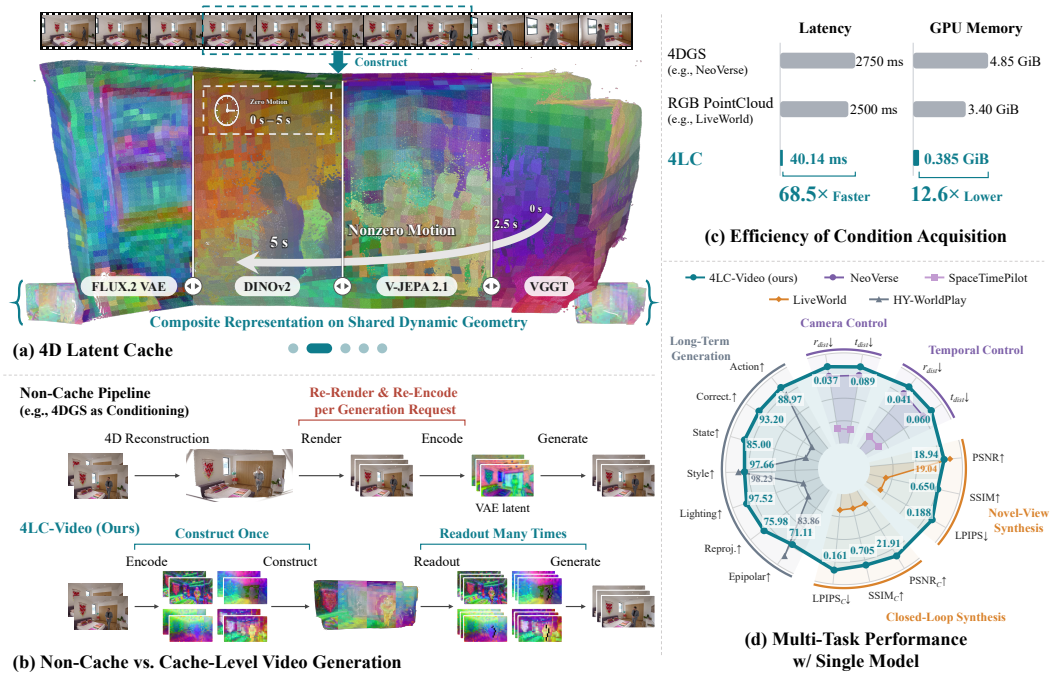


Figure 1: **Overview of 4LC.** (a) In 4LC, static and dynamic scene content is retained in multiple visual representations for reuse across time and viewpoint requests. (b) Non-cache pipelines render and re-encode a reconstructed scene for each generation request; 4LC encodes source observations once and reads out the cache for subsequent requests. (c) With the same Wan2.1 VAE, 4LC is superior to 4DGS and RGB point clouds in condition-acquisition latency and peak GPU memory. (d) 4LC-Video shows superior performance across five video generation tasks.

sitions associated with cached representations at the requested scene time, with static content held at its reference position. These visual representations are bound to the same geometry and motion while retaining their own formats, allowing their complementary information to be used together for conditioning. 4LC provides a spatiotemporal interface that directly reads these representations at independently specified times and viewpoints, combining readouts across observations to supplement uncovered regions. Once observations are cached, subsequent requests read the stored representations directly instead of rendering and re-encoding them (Fig. 1b), reducing condition-acquisition latency and peak GPU memory (Fig. 1c).

We build 4LC-Video to use these cache readouts as conditions for a shared video generator. Visual representations of generated observations are appended to the cache as new entries for subsequent segments. The same model can be configured for camera control, temporal control, novel-view synthesis, scene revisits, and long-term generation by specifying source observations, spatiotemporal requests, and historical reuse. Across these tasks, a single checkpoint achieves effective spatiotemporal controllability and scene consistency without task-specific fine-tuning (Fig. 1d).

Our main contributions are as follows:

- We introduce 4LC, a 4D latent cache with rich visual representations from composite encoders. Its spatiotemporal interface directly reads the cache at independently requested times and viewpoints, without per-request RGB rendering and re-encoding, thereby forming an efficient dynamic cache mechanism for 4D scenes.
- We develop 4LC-Video, which includes a spatiotemporal condition adapter to align the rich visual representations of 4LC with a pretrained video diffusion model for downstream video generation.
- Without task-specific fine-tuning, a single 4LC-Video model achieves state-of-the-art control accuracy and scene consistency across five generation tasks spanning spatiotemporal control, view synthesis, and long-term generation. 4LC accelerates condition acquisition by up to $68.5\times$ with $12.6\times$ lower peak GPU memory and reduces its share of end-to-end generation time from 14.9% to 0.25%, paving a new route for efficient conditioning and multi-task video generation.

2 Related Work

Dynamic World Modeling. Dynamic world modeling aims at controllable video generation within a persistent scene that evolves over time (Zhu et al., 2025; Chen et al., 2025b; InSpatio Team et al., 2026; Huang et al., 2025a). BulletTime (Wang et al., 2026c) and SpaceTimePilot (Huang et al., 2026b) control viewpoint and scene time as separate inputs, including frozen and reversed motion. Spatia (Zhao et al., 2026) maintains consistency across segments by updating a point-cloud memory. WorldMem (Xiao et al., 2025) and VMem (Li et al., 2025) retrieve past views by field-of-view overlap and surfel indexing, respectively. Interactive world models support long-term generation over successive segments (Wu et al., 2026; Sun et al., 2026; Robbyant Team et al., 2026; Huang et al., 2026a; Mao et al., 2026). WorldKV (Yi et al., 2026) retrieves relevant cached history and GIM-World (Wei et al., 2026) compresses past observations into compact memory. When scene time advances between visits, dynamic content must also remain coherent (Yu et al., 2026; Chen et al., 2026). LiveWorld (Duan et al., 2026) simulates out-of-sight dynamics, and VerseCrafter (Zheng et al., 2026) separates a static point cloud from per-object trajectories. To avoid the complexity of maintaining separate components, 4LC binds representations of static and dynamic content to shared geometry and motion, reading them flexibly at independently specified times and viewpoints.

Reconstruction-based Video Generation. Reconstruction-based video generation guides video models with renderings of reconstructed geometry. Several methods condition generation on point-cloud or 3D-cache renderings to complete unseen or poorly reconstructed regions (Ren et al., 2025; Chen et al., 2025a). For dynamic scenes, 4DGS (Wu et al., 2024) and feed-forward 4D reconstruction (Zhang et al., 2026a; Sucar et al., 2026; Luo et al., 2026; Xu et al., 2025b) model geometry and motion, and NeoVerse (Yang et al., 2026) renders them at requested times and viewpoints as conditions. TrajectoryCrafter (Yu et al., 2025b) and Vista4D (Lin et al., 2026) redirect the camera of monocular videos through point-cloud renderings that preserve source motion. These pipelines re-encode rendered RGB views for every generation request. Feature-space approaches instead condition on features carried by reconstructed geometry (Aliev et al., 2020; Wewer et al., 2024; Zhou et al., 2025). MVSplat360 (Chen et al., 2024b) and Splatent (Hirschorn et al., 2026) render Gaussian features into diffusion latents for novel-view synthesis. Mirage (Wang et al., 2026b) warps cached latent features to target views. MVSplat360, Splatent, and Mirage avoid target RGB encoding, yet their features lie in a specific generator's latent space and cover only static content. For efficient conditioning, 4LC reads out cached representations directly without per-request re-encoding, and its multi-encoder representations add complementary information.

3 Method

The core of 4LC is a cache $\mathcal{C} = \{\mathcal{C}_q\}_{q=0}^{Q-1}$ and a spatiotemporal interface equipped with three operations. Each entry $\mathcal{C}_q$ stores the visual representations of its source observations, bound to their 3D geometry and motion, and $\mathcal{C}$ organizes these entries by insertion order q. *Construction* encodes source observations with multiple visual encoders and stores their visual representations in a new cache entry (Fig. 2a). *Readout* propagates an entry's geometry to requested times and projects it through independently specified target cameras, reusing stored representations to produce readout features and coverage masks for static and dynamic content (Fig. 2b). *Composition* uses these masks to combine readouts across selected entries, filling uncovered regions while retaining existing readouts (Fig. 2c). 4LC-Video conditions a pretrained video diffusion model on these readouts (Sec. 3.4). Appendices A and B detail the implementation of 4LC and 4LC-Video, respectively.

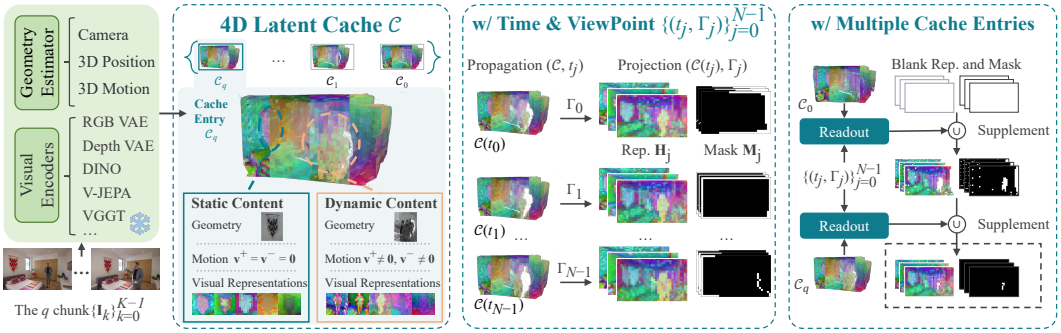
3.1 4LC Construction

Construction builds an entry $\mathcal{C}_q = \{u_i\}_{i=0}^{P-1}$ of P cache units from K timestamped source images (keyframes) $\{\mathbf{I}_k\}_{k=0}^{K-1}$ in shared world coordinates and appends it to $\mathcal{C}$. Each unit is defined as

$$u_i = \big(\mathbf{x}_i, t_i, \mathbf{v}_i^+, \mathbf{v}_i^-, b_i, \{\mathbf{f}_i^{(r)}\}_r\big).$$

Geometry. Using depth and cameras from a pretrained estimator, we back-project source pixels to reference positions $\mathbf{x}_i \in \mathbb{R}^3$ in shared world coordinates. Reference time t_i is the source timestamp.

Motion. Motion estimates from the same estimator toward adjacent observations give local forward and backward world velocities $\mathbf{v}_i^+, \mathbf{v}_i^- \in \mathbb{R}^3$. The motion mask $b_i \in \{0, 1\}$ retains the



(a) 4LC Construction **(b) 4LC Readout** **(c) 4LC Composition**

Figure 2: **Architecture of 4LC. (a) Construction** binds visual representations from multiple encoders to shared geometry and motion in one cache entry. **(b) Readout** propagates the entry to a requested time and projects it through a target camera, yielding readout features and coverage masks. **(c) Composition** fills uncovered regions with readouts from further entries.

estimated motion when $b_i = 1$ and holds the reference position when $b_i = 0$. A single source image ($K = 1$) provides no motion estimates between observations, so we set $b_i = 0$.

Visual representations. Visual encoders produce visual representations as feature maps at different spatial and temporal resolutions. For each visual representation r, we associate source frame k with step $\alpha_r(k)$ of its encoded feature sequence $\mathbf{Z}^{(r)}$. For unit u_i, its source frame $k_i \in \{0, \dots, K-1\}$ and source-pixel coordinates $\mathbf{p}_i \in \mathbb{R}^2$ locate the bound feature vector:

$$\mathbf{F}_k^{(r)} = \mathbf{Z}_{\alpha_r(k)}^{(r)}, \qquad \mathbf{f}_i^{(r)} = \mathbf{F}_{k_i}^{(r)} \left[\left\lfloor \frac{\mathbf{p}_i}{d_r} \right\rfloor \right] \in \mathbb{R}^{D_r}. \tag{1}$$

Here, d_r is the spatial stride and D_r the feature dimension. Units sharing a source feature cell reuse the same vector while retaining their own geometry and motion.

3.2 4LC Readout

The target sequence $\{(t_j, \Gamma_j)\}_{j=0}^{N-1}$ specifies N output frames, with scene time t_j and camera configuration Γ_j (intrinsics and world pose) at index j. For a cache entry $\mathcal{C}_q$ and a pair (t, Γ), the readout features and coverage for representation r are

$$\big(\mathbf{H}^{(r)}, \mathbf{M}^{(r)}\big) = \mathrm{Projection}\big(\mathrm{Propagation}(\mathcal{C}_q, t), \Gamma; r\big). \tag{2}$$

Temporal propagation. The local support $\mathcal{T}_i$ contains t_i and the open intervals to adjacent source times with available motion estimates. Under a locally linear approximation, the target position is

$$\mathbf{x}_i(t) = \begin{cases} \mathbf{x}_i, & b_i = 0 \text{ or } t = t_i, \\ \mathbf{x}_i + (t - t_i)\mathbf{v}_i^+, & b_i = 1,\ t \in \mathcal{T}_i,\ t > t_i, \\ \mathbf{x}_i + (t_i - t)\mathbf{v}_i^-, & b_i = 1,\ t \in \mathcal{T}_i,\ t < t_i. \end{cases} \tag{3}$$

The backward velocity points toward the past, hence the positive interval $t_i - t$. The propagated entry $\mathcal{C}_q(t) = \mathrm{Propagation}(\mathcal{C}_q, t)$ retains units with defined target positions and preserves their feature bindings. The source entry $\mathcal{C}_q$ remains unchanged and available for subsequent requests. For $b_i = 0$, the reference position remains defined beyond $\mathcal{T}_i$. To suppress spurious motion on static structure, we set b_i by thresholding projected displacement in the fixed source camera. The threshold ϵ is measured in source-image pixels per video frame.

Camera projection. For each visual representation r, we project $\mathcal{C}_q(t)$ through Γ onto its target grid, obtaining candidate units at each cell $\mathbf{g} \in \mathbb{Z}^2$. Depth alone would let closer units from other times hide moving content, so we take moving candidates from the nearest source time first and fill remaining cells with static ones. The selected unit index $i^\star(\mathbf{g})$ determines the readout features:

$$\mathbf{H}^{(r)}[\mathbf{g}] = \mathbf{f}_{i^\star(\mathbf{g})}^{(r)}. \tag{4}$$

Coverage $\mathbf{M}^{(r)}[\mathbf{g}] = 1$ when a unit is selected; otherwise, both are zero, and the coverage mask distinguishes missing content from valid zero-valued features.

An entry constructed from K keyframes can provide conditions for an N-frame target sequence, including $N > K$, by reading each representation r at its sampled target indices $\mathcal{J}_r \subseteq \{0, \dots, N-1\}$ to obtain readouts $(\mathbf{H}_j^{(r)}, \mathbf{M}_j^{(r)})$. Different visual representations therefore share the same access procedure while retaining their own feature grids, dimensions, and temporal sampling.

3.3 4LC Composition

Composition applies Readout to selected cache entries in the specified order and combines their readouts to supplement coverage for the same target time and camera sequences (Fig. 2c). For an entry $\mathcal{C}_q$, let $(\mathbf{H}_{q,j}^{(r)}, \mathbf{M}_{q,j}^{(r)})$ denote its readout at target index j for representation r. The subscript q retains the entry's insertion index in $\mathcal{C}$, independently of the order of access. For each representation r and $j \in \mathcal{J}_r$, we initialize the accumulated outputs as $\mathbf{H}_j^{(r)} = \mathbf{0}$ and $\mathbf{M}_j^{(r)} = \mathbf{0}$ on that representation's grid. Each selected entry contributes through the updates

$$\mathbf{H}_j^{(r)} \leftarrow \mathbf{H}_j^{(r)} + (\mathbf{1} - \mathbf{M}_j^{(r)}) \odot \mathbf{H}_{q,j}^{(r)}, \qquad \mathbf{M}_j^{(r)} \leftarrow \mathbf{M}_j^{(r)} \vee \mathbf{M}_{q,j}^{(r)}. \tag{5}$$

We apply the feature update before the coverage update, using the previous coverage to retain the first valid readout at overlapping locations. The mask is broadcast across channels in the elementwise product $\odot$; $\vee$ denotes elementwise logical OR. For each representation, the coverage ratio is the fraction of covered grid cells across its sampled target sequence. Composition returns when every representation reaches the prescribed coverage-ratio threshold or no selected entries remain.

3.4 4LC-Video

Representation adaptation. Visual representations differ in spatial resolution, temporal sampling, and feature dimension. A condition adapter aligns the outputs of Readout or Composition to the generator's spatiotemporal latent grid, using coverage masks to exclude missing values (Fig. 3). Spatial alignment interpolates masked features and normalizes by interpolated coverage, keeping uncovered locations at zero. Temporal alignment maps the initial readout to the first latent step and aggregates subsequent readouts within each latent time interval, averaging valid contributions and recording their coverage. Separate linear projections map the aligned representations, merged coverage mask, and target-camera rays before concatenation and fusion into $\mathbf{h}_{\text{cond}}$.

Conditional generation. A control branch injects this condition into the diffusion model to guide generation of target observations and complete uncovered regions. Construction converts generated

Table 1: **Task configurations for 4LC-Video.** Requests use one entry $\mathcal{C}_q$ or selected entries from $\mathcal{C}$. $\bar{t}_j$ preserves the source or default single-image timeline, and $\bar{\Gamma}_j$ the source camera trajectory. Scene revisits use one readout along a camera path that returns to its start.

Task	Initial frames	Target request (t_j, Γ_j)	Condition source
Camera control	$\{\mathbf{I}_k\}_{k=0}^{K-1}$	$(\bar{t}_j, \Gamma_j)$	$\mathcal{C}_q$
Temporal control	$\{\mathbf{I}_k\}_{k=0}^{K-1}$	$(t_j, \bar{\Gamma}_j)$	$\mathcal{C}_q$
Novel-view synthesis	$\mathbf{I}_0$	$(\bar{t}_j, \Gamma_j)$	$\mathcal{C}_q$
Scene revisits	$\mathbf{I}_0$	$(\bar{t}_j, \Gamma_j)$	$\mathcal{C}_q$
Long-term generation	$\mathbf{I}_0$	$(\bar{t}_j, \Gamma_j)$	$\mathcal{C}$

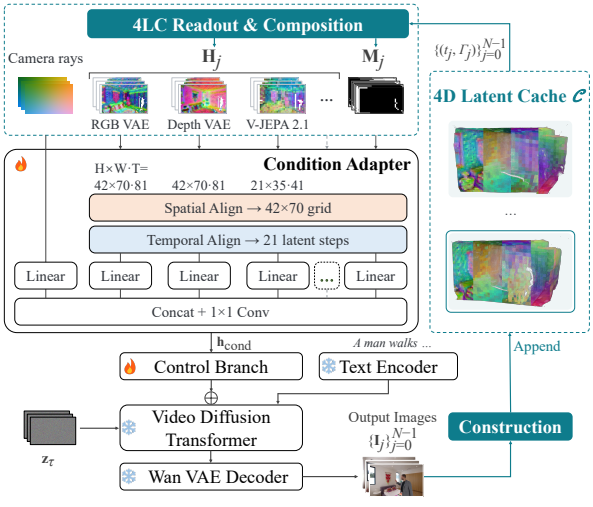


Figure 3: **4LC-Video framework.** The condition adapter aligns readouts to the Wan2.1 VAE latent grid and fuses them with masks and camera rays for the control branch. Generated observations extend the cache.

observations and their prescribed cameras and timestamps into new entries in $\mathcal{C}$ for reuse. Because Readout takes times and cameras independently and Composition combines entries, the tasks in Table 1 share one model and differ only in initial frames, target requests, and condition sources.

Training. We train the condition adapter and control branch with a weighted flow-matching objective (Lipman et al., 2023), keeping the visual encoders and video diffusion backbone frozen. Monocular training videos supply target sequences and sampled source observations, whose cache entries are queried at the original frame times and cameras.

4 Experiments

4.1 Experimental Setup

Implementation details. 4LC-Video uses Wan2.1-T2V-14B (Team Wan et al., 2025) with a VACE control branch (Jiang et al., 2025) and the pretrained NeoVerse reconstructor (Yang et al., 2026) for geometry and motion estimation. We select FLUX.2 VAE (Black Forest Labs, 2025) and V-JEPA 2.1 (Mur-Labadia et al., 2026) as the default encoders for extracting visual representations. Training uses 200K monocular clips from SpatialVID (Wang et al., 2026a), RealCam-Vid (Zheng et al., 2025), and OpenVid (Nan et al., 2025) without paired multiview supervision. Inference generates 81-frame 560×336 segments with the LightX2V 4-step LoRA (LightX2V Contributors, 2025).

Benchmarks and baselines. Camera control compares with SpaceTimePilot (Huang et al., 2026b), TrajectoryCrafter (Yu et al., 2025b), GEN3C (Ren et al., 2025), and NeoVerse (Yang et al., 2026), and temporal control with SpaceTimePilot and NeoVerse. Both tasks report camera errors and six VBench (Huang et al., 2024) scores. Novel-view synthesis on RealEstate10K (Zhou et al., 2018) and scene revisits on WorldScore (Duan et al., 2025) compare with ViewCrafter (Yu et al., 2025c), FlexWorld (Chen et al., 2025a), Voyager (Huang et al., 2025b), Spatia (Zhao et al., 2026), Mirage (Wang et al., 2026b), and LiveWorld (Duan et al., 2026) with PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). Long-term generation on MBench-A compares with published results (Zhang et al., 2026b) of HY-WorldPlay (Sun et al., 2026), Infinite-World (Wu et al., 2026), Yume-1.5 (Mao et al., 2026), and LingBot-World (Robbyant Team et al., 2026). Appendix C and Appendix D provide evaluation protocols and additional results.

4.2 Main Results on Multiple Video Generation Tasks

We evaluate a single 4LC-Video checkpoint on the five tasks in Table 1 without task-specific fine-tuning. Camera and temporal control read 4LC built from a source video with static and dynamic content, novel-view synthesis and scene revisits read an entry built from a single image, and long-term generation reads entries accumulated across generated segments.

Camera control. 4LC-Video attains the lowest camera errors together with the highest subject and background consistency (Table 2), whereas the baselines trade one for the other: NeoVerse follows the cameras with lower consistency, GEN3C and SpaceTimePilot stay consistent but close to static views with several-fold camera errors, and TrajectoryCrafter trails on both. 4LC-Video avoids this trade-off because 4LC keeps static structure persistent and dynamic content accurate. GEN3C, TrajectoryCrafter, and NeoVerse instead render conditions from reconstructed point clouds or 4DGS, whose artifacts and holes appear on moving subjects and static structures in NeoVerse outputs, while 4LC-Video stays faithful to the source observations (Fig. 4).

Temporal control. 4LC-Video keeps this advantage under temporal control (Table 2), while the baselines show the same trade-off. For time and camera pairs never observed together, 4LC reads each moving subject from its own observations and moves it by its estimated motion, while static structure stays in place. SpaceTimePilot generates the retimed motion without explicit conditions, and NeoVerse renders it from 4DGS with the artifacts noted above. In the frozen-time examples of Fig. 4, 4LC-Video holds the subject at frame 0 and keeps later-observed background intact, whereas NeoVerse deforms the subject inconsistently.

Table 2: **Quantitative comparison for camera and temporal control.** We evaluate 100 randomly sampled unseen videos under four camera trajectories (400 cases) and three temporal trajectories (300 cases). r_{dist} measures rotation error in radians and t_{dist} scale-normalized translation error; higher VBench scores are better. Best and second-best results per task are bold and underlined.

Method	$r_{\text{dist}}\downarrow$	$t_{\text{dist}}\downarrow$	Subj. Consist.	Back. Consist.	Temp. Flick.	Motion Smooth.	Aesth. Quality	Imag. Quality
Camera control								
SpaceTimePilot	0.179	0.321	89.79	93.43	95.26	**98.69**	47.60	57.23
TrajectoryCrafter	0.244	0.395	86.37	92.19	92.69	96.17	45.16	60.97
GEN3C	0.203	0.285	90.45	93.79	**95.66**	97.55	45.62	56.39
NeoVerse	0.059	0.120	88.63	92.77	94.43	98.17	**50.38**	**67.84**
4LC-Video (ours)	**0.037**	**0.089**	**90.71**	**94.14**	95.47	98.06	48.16	65.49
Temporal control								
SpaceTimePilot	0.163	0.192	88.42	92.65	**94.64**	**98.37**	47.44	56.40
NeoVerse	0.057	0.080	87.55	92.11	93.38	97.88	**49.65**	**66.85**
4LC-Video (ours)	**0.041**	**0.060**	**88.58**	**92.85**	93.78	97.41	48.75	65.14

Novel-view synthesis and Scene revisits. 4LC-Video obtains the best novel-view LPIPS with PSNR within 0.1 dB of LiveWorld, and it exceeds ViewCrafter, FlexWorld, Voyager, and Spatia on all three metrics (Table 3). For static scenes, 4LC reuses the cached visual representations of the image without RGB rendering and re-encoding, so its conditions avoid the information loss of these steps. Mirage attains higher SSIM at a higher output resolution (1280×704 versus 560×336), while its PSNR and LPIPS are worse than those of 4LC-Video. Scene revisits expose accumulated drift. 4LC-Video leads LiveWorld by 1.84 dB in PSNR_C and attains the best LPIPS_C. The baselines update their point clouds or spatial memories with generated views along the path, allowing intermediate errors to re-enter the conditions of return views. Reusing one source-image entry, 4LC gives each revisited camera the same condition regardless of intermediate outputs.

Table 3: **Quantitative comparison for novel-view synthesis and scene revisits.** We use the protocols and baseline results reported by Spatia (Zhao et al., 2026) and Mirage (Wang et al., 2026b), and evaluate LiveWorld under the same protocols.

Method	Novel-view synthesis			Closed-loop consistency		
	PSNR↑	SSIM↑	LPIPS↓	PSNR_C↑	SSIM_C↑	LPIPS_C↓
ViewCrafter	15.78	0.580	0.396	14.79	0.481	0.365
FlexWorld	16.25	0.593	0.370	12.20	0.428	0.598
Voyager	17.79	0.636	0.297	17.66	0.540	0.380
Spatia	18.58	0.646	0.254	19.38	0.579	0.213
Mirage	18.38	**0.779**	0.250	20.05	**0.825**	0.228
LiveWorld	**19.04**	0.645	0.191	20.07	0.653	0.186
4LC-Video (ours)	18.94	0.650	**0.188**	**21.91**	0.705	**0.161**

Long-term generation. Reuse across generated segments separates the MBench-A baselines, which either preserve the environment or evolve the scene (Table 4). HY-WorldPlay and InfiniteWorld score high on spatial and rendering metrics but low on state evolution, as their scenes remain overly static (Zhang et al., 2026b). Yume-1.5 and LingBot-World evolve scene states but lose spatial consistency. 4LC-Video bridges the two with the highest Environment and Causal scores. Each generated segment is appended to 4LC as a new entry, and later segments reuse only the static units of earlier entries as history through Composition. Revisited structure is thus constrained, while past motion is not replayed and scene states can keep evolving. Return views reproduce the outbound content, whereas HY-WorldPlay shifts the layout of revisited regions (Fig. 6 in Appendix D).

Table 4: **Quantitative comparison for long-term generation on MBench-A (Zhang et al., 2026b).** We compare Environment and Causal scores with published leaderboard results.

Method	Env. score	Causal score	Spatial		Rendering		Self-evolution		Interaction
			Epipolar	Reproj.	Lighting	Style	State	Correct.	Action
HY-WorldPlay	83.23	65.86	**83.86**	68.17	82.67	**98.23**	49.50	62.40	85.69
Infinite-World	73.76	70.92	74.04	61.51	62.63	96.87	48.00	78.40	86.37
Yume-1.5	54.92	85.03	51.86	24.55	51.21	92.05	**97.90**	**95.00**	62.20
LingBot-World	38.90	82.91	22.12	7.57	40.06	85.87	96.00	89.40	63.32
4LC-Video (ours)	**85.56**	**89.06**	71.11	**75.98**	**97.52**	97.66	85.00	93.20	**88.97**

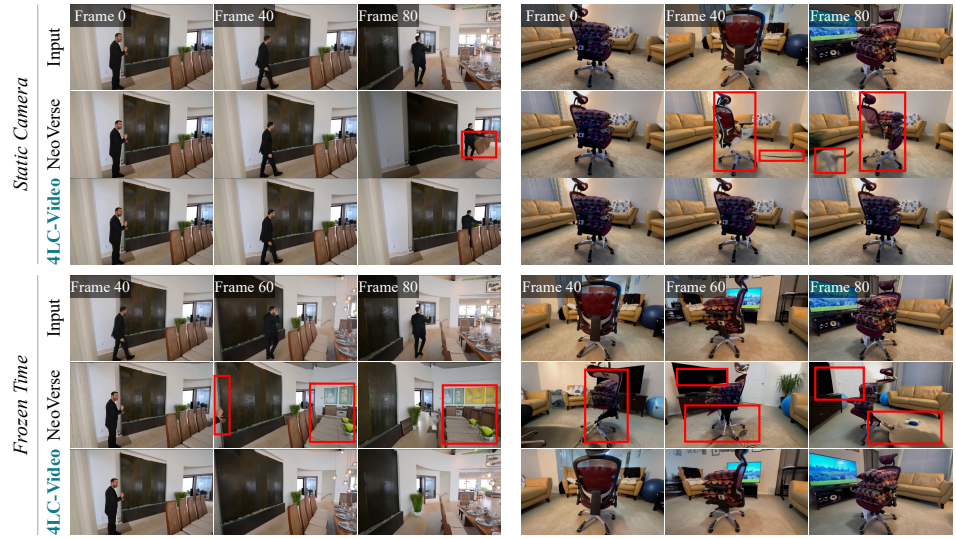


Figure 4: **Qualitative comparison for camera and temporal control.** Left: dynamic scene; right: static scene. **Top: Static camera.** The camera is fixed while scene time follows the source timeline. **Bottom: Frozen time.** Scene time is fixed at source frame 0 while the camera follows the source trajectory. Input frame labels identify each column; red boxes highlight distortions.

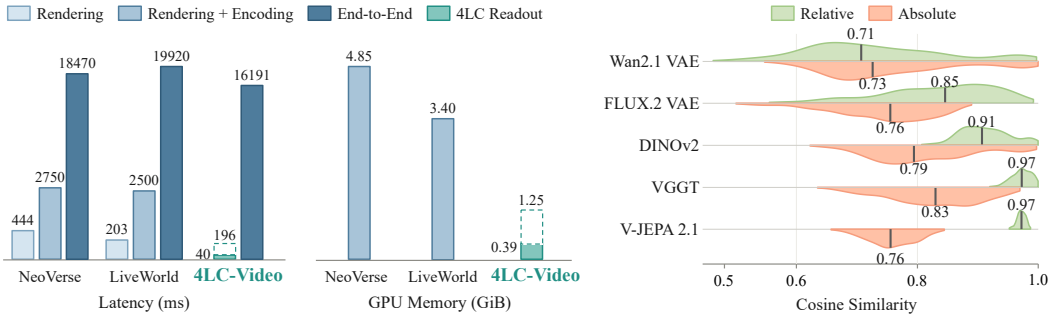


Figure 5: **Efficiency and feature similarity across condition-acquisition paths. Left: Latency and peak GPU memory** of 4LC readout and of RGB rendering and VAE encoding from NeoVerse's 4DGS and LiveWorld's RGB point clouds, with the end-to-end latency of each framework. Rendering + Encoding includes rendering time. Solid and dashed 4LC Readout bars use Wan2.1 VAE alone and with DINOv2, VGGT, and V-JEPA 2.1. Latency uses a square-root scale. **Right: Feature similarity** between direct readouts and re-encoded features, as per-case mean cosine similarity in the original (Absolute) and relative-representation (Relative) spaces; black ticks mark medians.

4.3 Conditioning Efficiency and Quality

Conditioning efficiency. Condition acquisition accounts for 14.9% of the 18.47 s end-to-end generation time in NeoVerse, compared with 0.25% of 16.19 s for 4LC-Video under the same generation settings, and this overhead recurs across segments in long-term generation. In Fig. 5, NeoVerse, LiveWorld, and 4LC-Video all use the Wan2.1 backbone and VAE encoder for 81-frame requests at 560 × 336 (Appendix C). The conditioning paths in NeoVerse (Yang et al., 2026) and LiveWorld (Duan et al., 2026) render image-resolution RGB views from 4DGS and RGB point clouds, respectively, followed by VAE encoding (Fig. 5, left). 4LC readout directly projects cached representations onto lower-resolution feature grids without repeated encoding. Compared with 4DGS rendering followed by VAE encoding in NeoVerse, 4LC readout is 68.5× faster, with 12.6× lower peak GPU memory. Adding DINOv2, VGGT, and V-JEPA 2.1 to Wan2.1 VAE raises readout latency from 40.14 to 196.14 ms and peak GPU memory from 0.385 to 1.245 GiB. The default FLUX.2 VAE with V-JEPA 2.1 takes 162.03 ms and 0.763 GiB, and adding DINOv2 and VGGT raises them to 255.20 ms and 1.605 GiB. All remain below the cost of rendering and encoding one representation.

Conditioning quality. To test whether direct readout preserves the information of re-encoding, we compare it with re-encoding views rendered from cached source pixels under the same Readout and encoders. All features retain comparable absolute similarity (Fig. 5, right), while semantic and geometric features also preserve stable relative structure (Moschella et al., 2023), with VGGT (Wang et al., 2025) and V-JEPA 2.1 concentrating near one in relative space and DINOv2 (Oquab et al., 2024) showing intermediate similarity. VAE features show lower relative similarity (medians 0.85 for FLUX.2 VAE and 0.71 for Wan2.1 VAE), as projection, resampling, and re-encoding alter their pixel-level detail. Sec. 4.4 examines the effect of condition-acquisition paths on generation.

4.4 Ablation Studies

Condition-acquisition path. For each VAE, we compare direct readout with RGB rendering and re-encoding (Table 5). Re-encoding updates features in the rendered target context, while direct readout retains the source encoding as content moves and visibility changes. With FLUX.2 VAE, direct readout leaves subject and background consistency nearly unchanged, with a small difference in camera errors. For Wan2.1 VAE, direct readout lowers both consistency scores and raises camera errors more. This matches FLUX.2 VAE's higher readout similarity in Sec. 4.3 and supports its selection.

Table 5: **Ablations of condition-acquisition paths and composite visual representations.** All variants share a frozen Wan2.1-T2V-1.3B backbone and the same 100 unseen videos with all trajectories. ALL denotes the other three encoders.

Visual representations	$r_{\text{dist}}\downarrow$	$t_{\text{dist}}\downarrow$	Subj. Consist.	Back. Consist.
RGB rendering and re-encoding				
Wan2.1 VAE	0.046	0.106	89.86	93.69
FLUX.2 VAE	**0.044**	0.106	90.00	93.80
Direct readout				
Wan2.1 VAE	0.056	0.128	89.03	93.03
FLUX.2 VAE	0.049	0.114	90.06	93.81
Direct readout with composite representations				
FLUX.2 VAE + DINOv2	0.051	0.116	89.89	93.69
FLUX.2 VAE + VGGT	0.048	0.120	89.89	93.70
FLUX.2 VAE + V-JEPA 2.1	0.046	0.112	90.05	93.85
FLUX.2 VAE + ALL	**0.044**	**0.101**	**90.15**	**93.90**

Composite visual representations. A single visual representation cannot meet the representational demands of dynamic world modeling, which span appearance, semantics, geometry, and dynamics, whereas existing encoders each specialize in one of them: FLUX.2 VAE reconstructs appearance, DINOv2 learns semantic features, VGGT infers multi-view geometry, and V-JEPA 2.1 models video dynamics. Since 4LC binds these representations to shared geometry and motion, we add DINOv2, VGGT, and V-JEPA 2.1 to FLUX.2 VAE under direct readout, individually and together (Table 5). Individually, DINOv2 and VGGT keep camera errors and consistency close to FLUX.2 VAE, indicating that their semantic and geometric cues are not fully exploited in pairwise combinations. V-JEPA 2.1 lowers both rotation and translation errors while leaving consistency nearly unchanged: unlike the image and multi-view encoders, it learns from video, and its temporally consistent features (Mur-Labadia et al., 2026) keep high similarity after readout, bringing temporal context into the condition. Combining all four lets appearance, semantic, geometric, and dynamic cues complement one another, making it the strongest configuration: it reduces the translation error of FLUX.2 VAE combined with V-JEPA 2.1 from 0.112 to 0.101, and all four metrics match or exceed FLUX.2 VAE re-encoding. Because Readout reuses cached representations, it adds no efficiency bottleneck (Fig. 5, Sec. 4.3), so 4LC is both composite and efficient, yielding rich, high-quality 4D conditions.

5 Conclusion

We presented 4LC, a 4D latent cache that stores composite visual representations of static and dynamic content and reads them directly at independently requested times and viewpoints. Conditioned on these readouts, a single 4LC-Video checkpoint achieved state-of-the-art control accuracy and scene consistency across five generation tasks. By avoiding per-request RGB rendering and re-encoding, 4LC turns condition acquisition from a recurring cost into a one-time encoding of observations. Its composite representations add complementary information from multiple encoders, giving dynamic world modeling conditions that are both efficient and rich.

Limitations and future work. 4LC relies on a feed-forward 4D reconstruction model to estimate depth, camera parameters, and motion, and its estimation errors can affect readout quality despite training augmentation. While 4LC supports both static and dynamic content, 4LC-Video reuses only static historical content for long-term generation. We leave dynamic reuse across generation segments to future work, as a further step toward efficient conditioning for multi-task video generation.

AI USE STATEMENT

We used generative AI tools to improve the grammar, wording, and readability of the manuscript, and to assist retrieval and discovery of related work. The authors reviewed and revised the AI-assisted text, verified all retrieved references, and take full responsibility for the accuracy and integrity of the final manuscript.

REFERENCES

Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *ECCV*, 2020.

Black Forest Labs. FLUX.2: Frontier visual intelligence. Official model release, 2025. URL https://bfl.ai/blog/flux-2.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, 2024a.

Kaijin Chen, Dingkang Liang, Xin Zhou, Yikang Ding, Xiaoqiang Liu, Pengfei Wan, and Xiang Bai. Out of sight but not out of mind: Hybrid memory for dynamic video world models. *arXiv preprint arXiv:2603.25716*, 2026.

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 flexible-view exploration. In *NeurIPS*, 2025a.

Yabo Chen, Yuanzhi Liang, Jiepeng Wang, Tingxi Chen, Junfei Cheng, Zixiao Gu, Yuyang Huang, Zicheng Jiang, Wei Li, Tian Li, Weichen Li, Zuoxin Li, Guangce Liu, Jialun Liu, Junqi Liu, Haoyuan Wang, Qizhen Weng, Xuan'er Wu, Xunzhi Xiang, Xiaoyan Yang, Xin Zhang, Shiwen Zhang, Junyu Zhou, Chengcheng Zhou, Haibin Huang, Chi Zhang, and Xuelong Li. TeleWorld: Towards dynamic multimodal synthesis with a 4D world model. *arXiv preprint arXiv:2601.00051*, 2025b.

Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. MVSplat360: Feed-forward 360 scene synthesis from sparse views. In *NeurIPS*, 2024b.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A unified evaluation benchmark for world generation. In *ICCV*, 2025.

Zicheng Duan, Jiatong Xia, Zeyu Zhang, Wenbo Zhang, Gengze Zhou, Chenhui Gou, Yefei He, Feng Chen, Xinyu Zhang, and Lingqiao Liu. LiveWorld: Simulating out-of-sight dynamics in generative video world models. In *ECCV*, 2026.

Paul Engstler, Andrea Vedaldi, Iro Laina, and Christian Rupprecht. Invisible Stitch: Generating smooth 3D scenes with depth inpainting. In *3DV*, 2025.

Weichen Fan, Chenyang Si, Junhao Song, Zhenyu Yang, Yinan He, Long Zhuo, Ziqi Huang, Ziyue Dong, Jingwen He, Dongwei Pan, Yi Wang, Yuming Jiang, Yaohui Wang, Peng Gao, Xinyuan Chen, Hengjie Li, Dahua Lin, Yu Qiao, and Ziwei Liu. Vchitect-2.0: Parallel transformer for scaling up video diffusion models. *arXiv preprint arXiv:2501.08453*, 2025.

Shaoheng Fang, Hanwen Jiang, Yunpeng Bai, Niloy J. Mitra, and Qixing Huang. WorldReel: 4D video generation with consistent geometry and motion modeling. In *CVPR*, 2026.

Weilun Feng, Guoxin Fan, Haotong Qin, Mingqiang Wu, Yuqi Li, Xiangqi Li, Zhulin An, Libo Huang, Dingrui Wang, Longlong Liao, Michele Magno, Yongjun Xu, and Chuanguang Yang. WorldCache: Accelerating world models for free via heterogeneous token caching. In *ICML*, 2026.

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

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo DAI, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *ICLR*, 2025.

Or Hirschorn, Omer Sela, Inbar Huberman-Spiegelglas, Netalee Efrat, Eli Alshan, Ianir Ideses, Frederic Devernay, Yochai Zvik, and Lior Fritz. Splatent: Splatting diffusion latents for novel view synthesis. In *CVPR*, 2026.

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

Junchao Huang, Ziyang Ye, Xinting Hu, Tianyu He, Guiyu Zhang, Shaoshuai Shi, Jiang Bian, and Li Jiang. LIVE: Long-horizon interactive video world modeling. In *ICML*, 2026a.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-range and world-consistent video diffusion for explorable 3D scene generation. *ACM TOG*, 44(6):1–15, 2025b.

Zhening Huang, Hyeonho Jeong, Xuelin Chen, Yulia Gryaditskaya, Tuanfeng Y. Wang, Joan Lasenby, and Chun-Hao Huang. SpaceTimePilot: Generative rendering of dynamic scenes across space and time. In *CVPR*, 2026b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

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

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *ICCV*, 2025.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. VMem: Consistent interactive video scene generation with surfel-indexed view memory. In *ICCV*, 2025.

LightX2V Contributors. LightX2V: Light video generation inference framework. GitHub repository, 2025. URL https://github.com/ModelTC/lightx2v.

Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan Burgert, Yuancheng Xu, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, Paul Debevec, and Ning Yu. Vista4D: Video reshooting with 4D point clouds. In *CVPR*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Yihang Luo, Shangchen Zhou, Yushi Lan, Xingang Pan, and Chen Change Loy. 4RC: 4D reconstruction via conditional querying anytime and anywhere. In *ICML*, 2026.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, and Kaipeng Zhang. Yume1.5: A text-controlled interactive world generation model. In *CVPR*, 2026.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *ICLR*, 2023.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. OpenVid-1M: A large-scale high-quality dataset for text-to-video generation. In *ICLR*, 2025.

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

Feng Qiao, Zhaochong An, Zhexiao Xiong, Serge Belongie, and Nathan Jacobs. Track2View: 4D-Consistent camera-controlled video generation via paired 3D point tracks. *arXiv preprint arXiv:2606.15534*, 2026.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-Informed world-consistent video generation with precise camera control. In *CVPR*, 2025.

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

Edgar Sucar, Eldar Insafutdinov, Zihang Lai, and Andrea Vedaldi. V-DPM: 4D video reconstruction with dynamic point maps. In *CVPR*, 2026.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. WorldPlay: Towards long-term geometric consistency for real-time interactive world modeling. In *ICML*, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, Hao Zhu, Zhaoxiang Zhang, Xun Cao, and Yao Yao. SpatialVID: A large-scale video dataset with spatial annotations. In *CVPR*, 2026a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025.

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*, 2026b.

Yiming Wang, Qihang Zhang, Shengqu Cai, Tong Wu, Jan Ackermann, Zhengfei Kuang, Yang Zheng, Frano Rajič, Siyu Tang, and Gordon Wetzstein. BulletTime: Decoupled control of time and camera pose for video generation. In *CVPR*, 2026c.

Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004.

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

Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentSplat: Autoencoding variational Gaussians for fast generalizable 3D reconstruction. In *ECCV*, 2024.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4D Gaussian splatting for real-time dynamic scene rendering. In *CVPR*, 2024.

Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, and Ming-Ming Cheng. Infinite-World: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. In *ICML*, 2026.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. In *NeurIPS*, 2025.

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

Jiaqi Xu, Kunzhe Huang, Xinyi Zou, Yunkuo Chen, Bo Liu, Mengli Cheng, Jun Huang, and Xing Shi. EasyAnimate: High-performance video generation framework with hybrid windows attention and reward backpropagation. In *ACM MM*, 2025a.

Zhen Xu, Zhengqin Li, Zhao Dong, Xiaowei Zhou, Richard Newcombe, and Zhaoyang Lv. 4DGT: Learning a 4D Gaussian transformer using real-world monocular videos. In *NeurIPS*, 2025b.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. NeoVerse: Enhancing 4D world model with in-the-wild monocular videos. In *CVPR*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025.

David Yifan Yao, Albert J. Zhai, and Shenlong Wang. Uni4D: Unifying visual foundation models for 4D modeling from a single video. In *CVPR*, 2025.

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

Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T. Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, and Charles Herrmann. WonderJourney: Going from anywhere to everywhere. In *CVPR*, 2024.

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T. Freeman, and Jiajun Wu. WonderWorld: Interactive 3D scene generation from a single image. In *CVPR*, 2025a.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *ICCV*, 2025b.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE TPAMI*, pp. 1–18, 2025c.

Wei Yu, Runjia Qian, Yumeng Li, Liquan Wang, Songheng Yin, Sri Siddarth Chakaravarthy P, Dennis Anthony, Yang Ye, Yidi Li, Weiwei Wan, and Animesh Garg. MosaicMem: Hybrid spatial memory for controllable video world models. *arXiv preprint arXiv:2603.17117*, 2026.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K. Barral, Raia Hadsell, Zoubin Ghahramani, Andrew Zisserman, Junlin Zhang, and Mehdi S. M. Sajjadi. Efficiently reconstructing dynamic scenes one D4RT at a time. In *CVPR*, 2026a.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Shengjun Zhang, Zhang Zhang, Simin Huang, Zhenyu Tang, Hanyang Wang, Chensheng Dai, Min Chen, Yifan Li, Yuxin Li, Yingjie Chen, Hao Liu, Chen Li, Jing Lyu, and Yueqi Duan. MBench: A comprehensive benchmark on memory capability for video world models. *arXiv preprint arXiv:2606.00793*, 2026b.

Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video generation with updatable spatial memory. In *CVPR*, 2026.

Guangcong Zheng, Teng Li, Xianpan Zhou, and Xi Li. RealCam-Vid: High-resolution video dataset with dynamic scenes and metric-scale camera movements. *arXiv preprint arXiv:2504.08212*, 2025.

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

Chaoyi Zhou, Xi Liu, Feng Luo, and Siyu Huang. Latent radiance fields with 3D-aware 2D representations. In *ICLR*, 2025.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM TOG*, 37(4):1–12, 2018.

Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. Allegro: Open the black box of commercial-level video generation model. *arXiv preprint arXiv:2410.15458*, 2024.

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *ICCV*, 2025.

A 4LC DETAILS

Visual encoders. Source observations are resized and center-cropped to 560×336 in the coordinates of the geometry, and each encoder follows the settings in Table 6. FLUX.2 VAE (Black Forest Labs, 2025), DINOv2 (Oquab et al., 2024), and VGGT (Wang et al., 2025) produce one feature map per keyframe, so $\alpha_r(k) = k$ in Equation 1. V-JEPA 2.1 (Mur-Labadia et al., 2026) produces one feature step per two frames, so we pair each keyframe with its predecessor, repeating the first keyframe, to obtain one step per keyframe. Wan2.1 VAE (Team Wan et al., 2025) encodes the first frame separately and compresses later frames by four; we fill each frame with the most recent keyframe and associate keyframe k at clip index n_k with step $\lceil n_k/4 \rceil$.

Table 6: **Visual encoder settings.** d_r is the spatial stride in pixels, D_r the feature dimension, and the target interval the number of frames between readouts.

Encoder	Features	d_r	D_r	Target interval
FLUX.2 VAE	Posterior mean of RGB and depth	8	32	1
DINOv2-L	Patch tokens	14	1024	1
VGGT-1B	Aggregator layer-11 tokens	14	2048	1
V-JEPA 2.1 ViT-L	Patch tokens	16	1024	2
Wan2.1 VAE	Latents of RGB and depth	8	16	4

Motion filtering. The motion mask b_i in Sec. 3.2 marks a unit as dynamic when its estimated motion displaces its projection in the source camera Γ_{k_i} by more than ϵ pixels per frame toward either adjacent observation:

$$b_i = \mathbf{1}\left[\max_{\pm} \frac{\left\| \pi_{\Gamma_{k_i}}\left(\mathbf{x}_i + \Delta t_i^{\pm}\mathbf{v}_i^{\pm}\right) - \mathbf{p}_i \right\|_2}{\Delta n_i^{\pm}} > \epsilon\right], \tag{6}$$

where π_Γ denotes pixel projection through camera Γ, and $\Delta t_i^{\pm}$ and $\Delta n_i^{\pm}$ are the time and frame gaps to the adjacent observations. When motion is estimated from multiple observations, units with motion toward neither neighbor are excluded from readout.

Camera projection. Each unit is assigned to the feature cell nearest to its projection in the target camera. At each cell, dynamic units from the source time closest to the requested time take priority, and static units fill the remaining cells (Sec. 3.2). Among the remaining candidates, we prefer the source camera closest to the target camera, then the frontmost surface, and finally the unit whose projections lie closest to the source and target cell centers.

B 4LC-VIDEO DETAILS

Condition adapter. Wan2.1 latents encode the first frame alone and each later group of four frames, so readouts with target intervals of one, two, and four frames contribute four, two, and one readouts to each later latent step; for 81 frames, they give 81, 41, and 21 readouts over 21 latent steps. Target-camera rays use Plücker coordinates (He et al., 2025).

Training. The 200K training clips comprise 100K from SpatialVID (Wang et al., 2026a), 40K from RealCam-Vid (Zheng et al., 2025), and 60K from OpenVid (Nan et al., 2025). We train the condition adapter and control branch for 20K steps on 16 NVIDIA A100 GPUs with a global batch size of 16, using AdamW (Loshchilov & Hutter, 2019) at a learning rate of 10^{-5}.

Data augmentation. Inspired by NeoVerse (Yang et al., 2026), which simulates degraded conditions from monocular videos, training augmentations vary the available conditions while preserving the target video, times, and cameras.

Each training clip provides K uniformly spaced keyframes, with K drawn from 11 to 21. To expose the generator to incomplete coverage, we use two geometric augmentations. Source-frame dropout excludes selected frames from the readout candidates. Visibility culling retains source units visible from a virtual camera near each source camera, which translates laterally while looking at the scene, yaws, pitches, zooms in, or zooms out.

To account for uncertainty in motion estimation, we perturb the motion threshold ϵ and construct training conditions with different motion masks. First-frame sampling restricts the observations to the initial image and constructs geometric conditions from its static state.

Representation and text dropout vary the visual and textual information available to the conditioning path. Representation dropout removes selected representations and recomputes coverage from the retained ones, and text dropout replaces the text condition with the empty-prompt embedding. Table 7 lists the probabilities and ranges.

Table 7: **Training augmentation parameters.** Geometric augmentation applies outside first-frame sampling. Lateral translation is in world units; focal-length factors are relative to the source camera.

Type	Parameter	Value
Geometric	Number of keyframes K	11–21
	Augmentation probability	0.6
	Maximum source-frame drop fraction	0.3
	Lateral translation	0.1–0.3
	Yaw and pitch	5°–25°
	Zoom-in / zoom-out focal-length factor	1.1–1.3 / 0.7–0.9
Motion	First-frame sampling probability	0.1
	Motion threshold ϵ (pixels/frame)	0.05–0.2
Representation and text	Dropout probability, all representations	0.1
	Dropout probability, per representation	0.1
	Text-dropout probability	0.1

C Experimental Details

Spatiotemporal control. As in NeoVerse (Yang et al., 2026), test videos use $K = 21$ uniformly spaced keyframes, and its reconstructor estimates camera poses of generated videos. Each generated and target trajectory is expressed relative to its first frame. Camera trajectories are pan left, reversed, static, and dolly out at the original scene times; temporal trajectories are reversal, freezing, and pause-and-resume under the source cameras. Translation error excludes static-camera cases.

Long-term generation. Each MBench-A (Zhang et al., 2026b) case generates 401 frames. Historical conditions use up to eight uniformly sampled cache entries and their static units ($b_i = 0$). Composition visits them from earliest to latest for Environment cases and from latest to earliest for Causal cases, with a coverage-ratio threshold of 0.95.

Conditioning efficiency. Acquisition latency and peak GPU memory are measured on an NVIDIA A100, averaging 50 runs after three warmup runs. Scene construction differs widely across frameworks and cannot be fully aligned: the NeoVerse reconstructor used by NeoVerse and 4LC takes 1.35 s, whereas LiveWorld takes 20.43 s. For a fair comparison, acquisition and end-to-end timing exclude scene construction, including source reconstruction and encoding, for all frameworks; memory includes weights used for condition acquisition.

Feature readout analysis. For each video, request, and encoder, we uniformly sample 2,048 tokens from the region covered by both acquisition paths. Relative representations (Moschella et al., 2023) use 128 corresponding anchor features with ten anchor seeds, keeping the statistical tokens fixed across seeds. We average relative cosine similarities over seeds within each case and give all cases equal weight in Fig. 5 (right). The five requests preserve the source time and camera sequence, reverse scene time, freeze scene time, hold the camera fixed, or reverse the camera trajectory.

Ablations. Each ablation configuration, including the RGB variants, trains its condition adapter and control branch independently on 10K clips for 4K steps on four GPUs with one sample per GPU, and samples with 20 steps.

D Additional Results

Long-term generation. Fig. 6 complements the MBench-A results in Sec. 4.2 with two cases whose camera trajectory reverses midway and returns to the initial viewpoint. HY-WorldPlay (Sun



Figure 6: **Qualitative comparison for long-term generation on MBench-A.** Each method occupies two rows: the top row follows the outbound trajectory from left to right, and the bottom row follows the return trajectory from right to left, so vertically aligned frames share a camera pose. Red boxes mark structures that change between the first and final frames.

et al., 2026) regenerates returned regions with shifted layouts and missing structures, and its final frame departs from the first. 4LC-Video adds each generated segment to the cache, and Composition reads out the static units of these entries for the return requests, so return views reproduce the outbound content. This consistency agrees with its highest Reprojection and Environment scores in Table 4.

Generalization to WorldScore. We evaluate the main model on WorldScore without fine-tuning, using the official pipeline, and compare it with published results (Duan et al., 2025) of static scene generators (Yu et al., 2024; Engstler et al., 2025; Yu et al., 2025a; Huang et al., 2025b) and foundation video models (Chen et al., 2024a; Xu et al., 2025a; Zhou et al., 2024; Yang et al., 2025; Fan et al., 2025; HaCohen et al., 2024; Team Wan et al., 2025) (Table 8). 4LC-Video obtains the highest Dynamic score and the best content alignment and style consistency among these methods. Compared with Wan2.1, it raises camera controllability from 23.53 to 86.87 and 3D consistency from 78.74 to 87.97, and its Static score approaches those of static scene generators.

Motion threshold. Using the main model on camera trajectories, we compare disabled motion filtering (Equation 6) with filtering at different thresholds (Table 9). Without filtering, estimated motion on static structure enters the conditions as spurious displacement, yielding the largest camera errors and the lowest subject consistency, background consistency, temporal flickering, and motion smoothness. Training samples the threshold from 0.05 to 0.2 to account for motion-estimation uncertainty (Table 7), and performance remains stable from 0.05 to 0.4, including values beyond this range, indicating robustness to the condition changes induced by threshold selection.

Table 8: **Quantitative comparison on WorldScore (Duan et al., 2025).** We compare against its published leaderboard results and, following Spatia (Zhao et al., 2026), report Avg. as the mean of the Static and Dynamic scores; missing motion metrics of static scene generators count as zero.

Method	WorldScore			Controllability			Quality				Dynamics		
	Avg.	Static	Dynamic	Camera	Object	Content	3D	Photo.	Style	Subj.	Acc.	Mag.	Smooth.
Static scene generation													
WonderJourney	54.19	63.75	44.63	84.60	37.10	35.54	80.60	79.03	62.82	66.56	–	–	–
InvisibleStitch	51.95	61.12	42.78	**93.20**	36.51	29.53	**88.51**	**89.19**	32.37	58.50	–	–	–
WonderWorld	61.79	72.69	50.88	92.98	51.76	71.25	86.87	85.56	70.57	49.81	–	–	–
Voyager	66.08	**77.62**	54.53	85.95	**66.92**	68.92	81.56	85.99	84.89	**71.09**	–	–	–
Foundation video generation													
VideoCrafter2	50.03	52.57	47.49	28.92	39.07	72.46	65.14	61.85	43.79	56.74	47.12	30.40	29.39
EasyAnimate	52.25	52.85	51.65	26.72	54.50	50.76	67.29	47.35	73.05	50.31	75.00	31.16	40.32
Allegro	53.64	55.31	51.97	24.84	57.47	51.48	70.50	69.89	65.60	47.41	54.39	**40.28**	37.81
CogVideoX-I2V	60.64	62.15	59.12	38.27	40.07	36.73	86.21	88.12	83.22	62.44	69.56	26.42	60.15
Vchitect-2.0	40.38	42.28	38.47	26.55	49.54	65.75	41.53	42.30	25.69	44.58	33.59	33.81	21.31
LTX-Video	55.99	55.44	56.54	25.06	53.41	39.73	78.41	88.92	53.50	49.08	**76.22**	29.95	**71.09**
Wan2.1	55.21	57.56	52.85	23.53	40.32	45.44	78.74	78.36	77.18	59.38	54.27	33.26	38.05
4LC-Video (ours)	**68.46**	72.18	**64.73**	86.87	41.04	**76.84**	87.97	68.02	**86.41**	58.10	55.66	35.78	50.65

Table 9: **Motion filtering and threshold robustness.** The main model is evaluated on the same 100 unseen videos with camera trajectories. ϵ is in pixels/frame; $\epsilon = 0$ disables filtering, and $\epsilon = 0.1$ is the default. Bold and red values mark the best and worst results, respectively.

Threshold ϵ	$r_{\text{dist}}\downarrow$	$t_{\text{dist}}\downarrow$	Subj. Consist.	Back. Consist.	Temp. Flick.	Motion Smooth.	Aesth. Quality	Imag. Quality
0	0.065	0.138	89.91	93.53	94.44	97.88	**49.08**	**67.07**
0.05	**0.035**	0.092	90.61	94.06	94.81	98.03	48.50	66.88
0.1	0.037	0.089	90.71	94.14	**95.47**	**98.06**	48.16	65.49
0.2	**0.035**	**0.083**	**90.76**	**94.24**	94.88	98.04	48.15	66.61
0.4	0.036	**0.083**	90.74	94.23	94.88	98.04	48.13	66.61