arXiv:2610.06687v2 [cs.CV] 6 Oct 2026

ChronoWorld: Camera-Controlled Consistent 4D World Generation via Spatiotemporal Cues and Geometric Reflections

Xiaoyu Zhou[1] **Dingwei Xian**[1] **Zhenyu Wang**[1]
Yajiao Xiong[1] **Yongtao Wang**[1,2]* **Ming-Hsuan Yang**[3]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]VGI Labs Co., Ltd. [3]University of California, Merced



Figure 1: ***Showcases of our ChronoWorld.*** Our method addresses the challenge of spatiotemporally consistent 4D scene generation by introducing a "Observation–State–Reflection" framework. This approach leverages spatiotemporal causal cues and reconstruction-driven geometric reflection to generate globally consistent, free-view 4D world.

Abstract

While existing camera-controllable video generation models can produce visually compelling sequences, preserving intrinsic 4D spatiotemporal coherence remains challenging. To address this limitation, we propose **ChronoWorld**, an "Observation–State–Reflection" framework that leverages spatiotemporal causal cues and reconstruction priors to generate globally consistent, free-view 4D scenes. Given a context video, we introduce a Spatiotemporal Epipolar Causal Attention mechanism that enforces multi-view epipolar constraints and temporal causality throughout the generation process. In addition, we develop a reconstruction-driven geometric reflection pipeline with a 4D retrieval strategy to enable dynamic self-assessment and correction of generated outputs, improving consistency and accuracy. Extensive experiments show that ChronoWorld achieves state-of-the-art performance in spatiotemporally consistent, cinematic-quality 4D scene generation, with strong generalization and high-fidelity geometry across diverse scenarios.

*Corresponding author.

Preprint.

1 Introduction

Transforming sparse observations into freely navigable 4D worlds has long been a central goal of the field. Driven by recent advances in video generation and world models, modern generative architectures can synthesize cinematic-quality sequences and enable real-time interactive rendering. Beyond enabling cost-effective construction of immersive VR/AR environments [1, 2], these capabilities provide scalable, high-fidelity simulations that are critical for embodied intelligence [3, 4] and autonomous driving [5, 6, 7].

Despite these advances, a "Tower of Babel" gap persists between 2D video generation and 4D scene reconstruction. While current video generation models [8, 9, 10, 11] can produce visually plausible, pixel-level sequences, they struggle to maintain intrinsic 4D spatiotemporal consistency in the underlying scene representation. As a result, generated videos often violate multi-view geometric constraints, undermining both physical validity and temporal coherence. Recent approaches [12, 13, 14, 15] attempt to mitigate this issue by incorporating projected 3D representations (e.g., point clouds) as multi-view guidance within iterative render–generate–update pipelines. However, constructing 3D representations from coarse reconstructions of generated outputs introduces a critical failure mode: errors in synthesized frames are back-projected and fused into the 3D prior, creating a self-reinforcing feedback loop that amplifies inconsistencies and can ultimately lead to collapse in subsequent generations.

Fundamentally, these challenges can be attributed to three key limitations: **1) 4D Spatiotemporal Inconsistency:** Current models generate videos that violate 3D epipolar geometry and 4D spatiotemporal constraints, preventing the formation of a consistent and persistent scene representation. **2) Error Accumulation and Exposure Bias:** Due to training–inference discrepancies and inherited artifacts, errors at inference time accumulate over time, leading to pronounced dynamic distortions and spatial drift. **3) Spatial Forgetting and Geometric Degradation:** Over temporal horizons and complex camera motions, existing methods fail to preserve global coherence and handle occlusions, resulting in "spatial forgetting" that can cause severe geometric degradation or collapse.

To address these challenges, we draw inspiration from human cognition and propose an "Observation–State–Reflection" framework, ChronoWorld, for camera-conditioned, globally consistent 4D scene generation. The framework extracts spatiotemporal cues from external inputs (**observations**), compresses them into a structured latent space via geometric invariants (**state**), and refines predictions through geometry-aware feedback (**reflection**), yielding a persistent and unified representation. At its core, we introduce a Spatiotemporal Epipolar Causal Attention mechanism that leverages contextual views to preserve intra- and inter-frame dependencies, while enforcing multi-view epipolar constraints and temporal causality. We further propose a Spatial Context Augmentation strategy integrated with rectified flow, where perturbed memory cues encourage robustness and reduce error accumulation and exposure bias during spatial transformations. In addition, we augment the video diffusion model with a reconstructive multi-head decoder that maps latent features to 4D Gaussian attributes and motion fields, supported by auxiliary representation supervision to enhance spatiotemporal consistency. To mitigate cumulative drift and address spatial forgetting and geometric degradation, we design a reconstruction-driven geometric reflection pipeline at inference time. Specifically, we build a unified 4D memory from feed-forward 4D Gaussians and global camera poses. An explicit 4D retrieval mechanism then evaluates generated outputs, filters inconsistent frames, and constructs a geometric consistency graph. Finally, we synchronize updates between the explicit memory and the implicit history cache to maintain accuracy and consistency over time. The main contributions of our work are summarized as follows:

- We propose **ChronoWorld**, an "Observation–State–Reflection" framework that bridges video generation and 4D reconstruction, enabling the synthesis of spatiotemporally consistent, free-view 4D scenes from input videos.
- We introduce a Spatiotemporal Epipolar Causal Attention mechanism that exploits spatiotemporal cues to enforce multi-view epipolar constraints and temporal causality during generation.
- We develop a reconstruction-driven geometric reflection pipeline with a 4D retrieval strategy to enable dynamic self-assessment and correction of generated outputs, improving accuracy and consistency of the 4D representation.
- Extensive experiments show that ChronoWorld achieves state-of-the-art performance in spatiotemporally consistent, cinematic-quality 4D scene generation, with strong generalization and high fidelity across diverse scenes and scales.

2 Related Work

Camera-Controlled Video Generation. Recent progress in video generation has been driven by large-scale diffusion models [16, 17, 18, 19] and autoregressive transformer architectures [20, 21]. While these approaches achieve strong visual fidelity, conditioning on a single text or image input provides limited control over camera motion and restricts free-view exploration. Camera-conditioned methods [8, 9, 22] incorporate trajectory signals to guide video synthesis along prescribed camera paths. However, preserving geometric fidelity and scene consistency under large viewpoint changes remains challenging. Warping-and-repainting approaches project images into point clouds and warp them along trajectories, which improves spatial alignment but often introduces artifacts in unseen regions. Joint RGB-depth generation enables point cloud reconstruction for novel-view rendering, yet several limitations persist in 4D inconsistency.

To address these issues, we integrate diffusion-based video generation with both implicit and explicit geometric priors. In addition, we introduce a spatiotemporal self-reflection mechanism that enables the model to preserve consistency throughout the generation process.

3D and 4D Scene Generation. Existing 3D/4D generation methods can be broadly categorized into: (1) compositional scene generation [23, 24, 25], which assembles scenes from discrete 3D assets using layouts or scene graphs, and (2) end-to-end volumetric generation, which directly models scenes using representations such as occupancy fields, point clouds, voxels, or meshes. Compositional approaches improve object diversity and controllability but often struggle with realistic backgrounds and object–environment interactions. In contrast, end-to-end methods better preserve global semantic consistency, yet are constrained by limited 3D data, high annotation costs, and lower visual fidelity, and they do not fully benefit from large-scale 2D pretraining.

Our approach bridges this gap by integrating video generation with 4D reconstruction through spatiotemporal self-reflection and explicit 4D representation prediction, unifying generative modeling with geometric reasoning to produce spatially coherent 4D scenes.

Generalizable Reconstruction. Traditional optimization-based approaches [26, 27, 28, 29] rely on dense multi-view observations and per-scene optimization, which limits scalability. Feed-forward methods [30, 31, 32] incorporate geometric priors to handle sparse inputs, but they often fail to recover occluded or unseen regions. Recent work combines generative models with reconstruction to improve multi-view consistency; however, performance remains dependent on accurate volumetric consistency, which are frequently violated in practice. In contrast, our method tightly couples controllable video generation with reconstruction-driven cue guidance, improving both geometric and semantic consistency and enabling photorealistic 4D reconstructive generation.

3 Method

This section presents ChronoWorld (Fig. 2-3), which unifies video generation and reconstruction via a hybrid training–inference paradigm under an "Observation–State–Reflection" framework. During training, spatiotemporal cues and a reconstructive multi-head decoder enforce geometric consistency in the latent space. During inference, a reconstruction-driven geometric reflection mechanism refines outputs and maintains a consistent 4D memory.

3.1 Overview Architecture

Our architecture builds on a latent video diffusion framework, comprising a causal 3D VAE encoder, a Diffusion Transformer with Spatiotemporal Cues (STC-DiT), and a reconstructive multi-head decoder (Fig. 2). During training, we sample triplets of context videos, target videos, and target camera trajectories, where context and target sequences are temporally aligned but captured from different viewpoints. The causal 3D VAE encodes the inputs into latent representations $Z_c, Z_t \in \mathbb{R}^{B\times C\times T\times H\times W}$, where B, C, T, H, and W denote batch size, channels, timesteps, height, and width, respectively. We encode the camera trajectory $\{(T_i, R_i)\}_{i=0}^{T-1}$ into dense positional features using Plücker embeddings, followed by a two-layer MLP modulation network that injects the features into the attention keys K and values V.

To simulate inference-time degradation, we apply Temporal Tube Masking and Spatial Context Augmentation to the context latents Z_c, modeling error accumulation and information loss. The

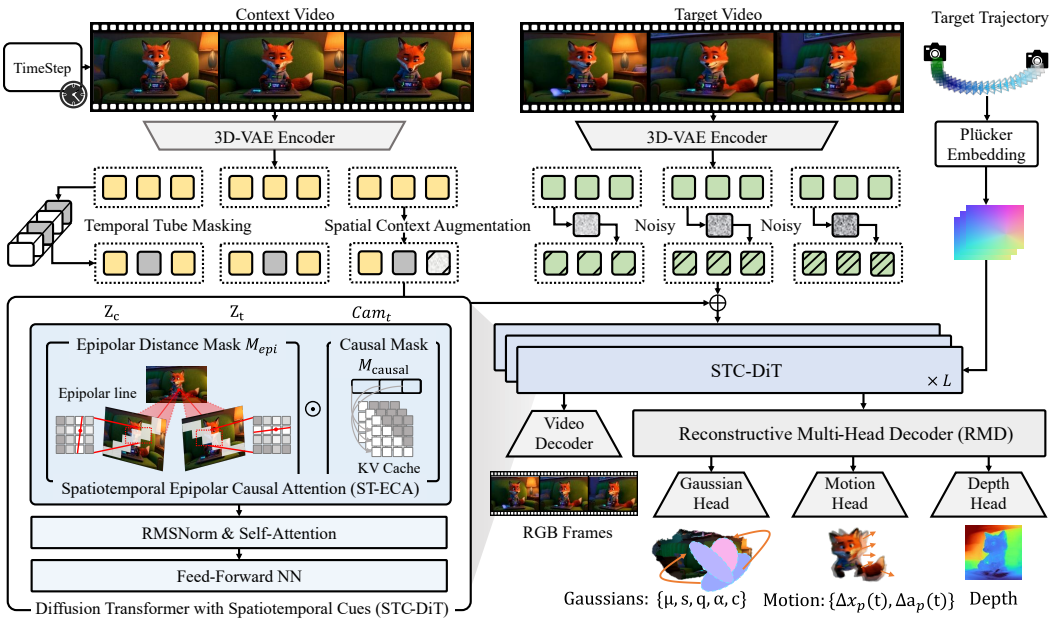


Figure 2: **Framework of ChronoWorld.** Our method is trained to denoise target video conditioned on the camera trajectories with a Diffusion Transformer with Spatiotemporal Cues (STC-DiT), and a reconstructive multi-head decoder for predicting 4D attributes.

model is trained to denoise Z_t conditioned on Z_c and the camera trajectory, enabling temporally coherent generation and globally consistent 4D geometry across the sequence. Finally, the multi-head decoder converts the block-wise, asynchronously denoised latents into explicit outputs, including RGB videos (appearance), Gaussian attributes (geometry), motion maps (dynamics), and depth maps.

3.2 Diffusion Transformer with Spatiotemporal Cues

Existing diffusion models typically capture intra-frame spatial and inter-frame temporal dependencies via full-sequence cross- or causal attention, largely overlooking cross-frame spatiotemporal cues (i.e., latent correspondences along epipolar trajectories). This oversight introduces computational redundancy and degrades geometric consistency across overlapping regions.

To address this issue, we propose a diffusion framework with Spatiotemporal Epipolar Causal Attention. By coupling multi-view geometric constraints with spatiotemporal augmentation, the model enforces coherent cross-view correspondence, enabling the generation of consistent 4D scenes.

Spatiotemporal Epipolar Causal Attention. To reduce the computational redundancy and geometric ambiguity of unconstrained full-sequence attention, we propose spatiotemporal epipolar causal attention (ST-ECA). By integrating multi-view 3D priors with strict temporal causality, ST-ECA confines cross-frame interactions to geometrically valid, causally ordered regions.

Given query $\mathbf{Q} \in \mathbb{R}^{N\times d}$, key $\mathbf{K} \in \mathbb{R}^{L\times d}$, and value $\mathbf{V} \in \mathbb{R}^{L\times d}$ projected from latent features, we first construct an epipolar distance mask $\mathbf{M}_{\text{epi}}$ to enforce multi-view geometric consistency at timestep T. For a query token at 2D coordinate $\mathbf{p}_i$ in the target view and a key token at $\mathbf{p}_j$ in the context view, valid correspondences lie on the epipolar line $\mathbf{l}_i = \mathbf{F}\mathbf{p}_i$, where $\mathbf{F}$ is the fundamental matrix induced by relative camera poses. We define the epipolar mask as:

$$\mathbf{M}_{\text{epi}}(i,j) = -\gamma \frac{|\mathbf{p}_j^T \mathbf{F}\mathbf{p}_i|}{\|\mathbf{F}\mathbf{p}_i\|_2}, \tag{1}$$

where γ is a learnable scaling factor controlling the sharpness of the geometric constraint.

To enforce temporal causality and prevent leakage from future states, we introduce a binary temporal mask $\mathbf{M}^{\text{causal}}_{i,j} \in \{0, -\infty\}$, defined as 0 if $\tau(i) \geq \tau(j)$ and $-\infty$ otherwise, where $\tau(\cdot)$ denotes the frame timestamp. Each token can thus attend only to the current and past frames.

We integrate both constraints by injecting the geometric and causal biases into the attention:

$$\text{ST-ECA}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \sum_j \text{Softmax}\left(\frac{\mathbf{q}_i \mathbf{k}_j^T}{\sqrt{d}} + \mathbf{M}_{\text{epi}}(i,j) + \mathbf{M}_{\text{causal}}(i,j)\right)\mathbf{v}_j, \tag{2}$$

where d is the feature dimension. The epipolar mask acts as a continuous geometric prior, restricting correspondence search to 1D epipolar trajectories rather than the full 2D plane, while the causal mask preserves temporal consistency and reduces the cost of global attention.

Temporal Tube Masking. To improve temporal reasoning and motion prediction, naive approaches apply independent spatial masks per frame, which allows trivial reconstruction by copying from adjacent frames. To prevent this shortcut, we introduce Temporal Tube Masking (TTM), which occludes consistent spatial regions across the entire temporal sequence.

We first sample a 2D binary spatial mask $\mathbf{M}_s \in \{0,1\}^{H\times W}$ from a Bernoulli distribution, $\mathbf{M}_s(x,y) \sim \text{Bernoulli}(p_{\text{drop}})$, where (x,y) denotes coordinates and p_{drop} is the masking ratio. This mask is then extended along the temporal dimension to form a 4D tube mask $\mathbf{M}_{\text{tube}}(t,x,y) = \mathbf{M}_s(x,y)$ for all $t \in \{1,\dots,T\}$. The corrupted context latent is obtained via element-wise masking:

$$\hat{\mathbf{Z}}_c = \mathbf{Z}_c \odot (\mathbf{1} - \mathbf{M}_{\text{tube}}), \tag{3}$$

where $\odot$ denotes the Hadamard product. By enforcing consistent occlusions over time, TTM prevents degenerate temporal copy-paste behavior and instead encourages reconstruction from long-range geometric cues and motion dynamics.

Spatial Context Augmentation. To mitigate exposure bias and spatial drift (error accumulation) in video generation, we introduce Spatial Context Augmentation (SCA), a lightweight self-augmentation strategy in the flow-matching latent space. Given ground-truth context and target frames, the causal 3D VAE encodes them into clean latents $\mathbf{Z}_c^0$ and $\mathbf{Z}_t^0$. To simulate inference-time degradation, we corrupt the context latent by sampling a flow timestep $\tau \sim \mathcal{U}(0, 0.5)$ and injecting noise via the rectified flow schedule:

$$\mathbf{Z}_c^\tau = (1-\tau)\mathbf{Z}_c^0 + \tau\epsilon, \tag{4}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and τ controls the corruption level. To avoid training collapse, τ is constrained to be lower than the noise level applied to the target latent.

The flow matching objective then supervises recovery of the target $\mathbf{Z}_t^0$ from corrupted context. This structured perturbation improves robustness to cumulative drift by explicitly training the model to correct errors induced by degraded spatial context.

Efficient Training Strategy. To reduce the computational cost and stabilize optimization in video diffusion, we introduce an efficient training strategy that leverages spatiotemporal structure for accelerated generation. We assign monotonically increasing noise levels τ_i along the temporal axis to enforce causal progression.

We then partition spatial latents into non-overlapping blocks $\{\mathcal{B}_k\}$ and perform block-wise asynchronous denoising with locally cached context:

$$\mathbf{Z}_{\tau-\Delta\tau}^{\mathcal{B}_k} = \mathbf{Z}_\tau^{\mathcal{B}_k} - \Delta\tau \cdot v_\theta(\mathbf{Z}_\tau^{\mathcal{B}_k}, \tau, \mathbf{C}_{\neg k}), \tag{5}$$

where $\mathbf{C}_{\neg k}$ denotes cached key-value features from neighboring blocks and $v_\theta(\cdot)$ is the velocity field. This design reduces redundant computation and decouples gradients across blocks.

Finally, we combine this scheme with rectified flow for few-step generation. We apply velocity-matching distillation, where the student v_θ learns a target velocity pointing to the teacher's multi-step prediction, reducing the denoising process from $\sim$50 steps to 5 steps.

3.3 Reconstructive Multi-Head Decoder

Directly mapping high-dimensional denoised latents to dynamic scenes is highly under-constrained. We address this by employing a dual-stage training with four parallel decoders that disentangle the output space into complementary attributes: RGB frames (appearance), Gaussian attributes (geometry), motion maps (dynamics), and depth maps.

Appearance decoder. We use the pre-trained 3D VAE decoder to project latents back to pixel space, producing $\mathbf{I} \in \mathbb{R}^{T\times 3\times H\times W}$. To preserve the learned latent manifold, this decoder is kept frozen while training the remaining heads.

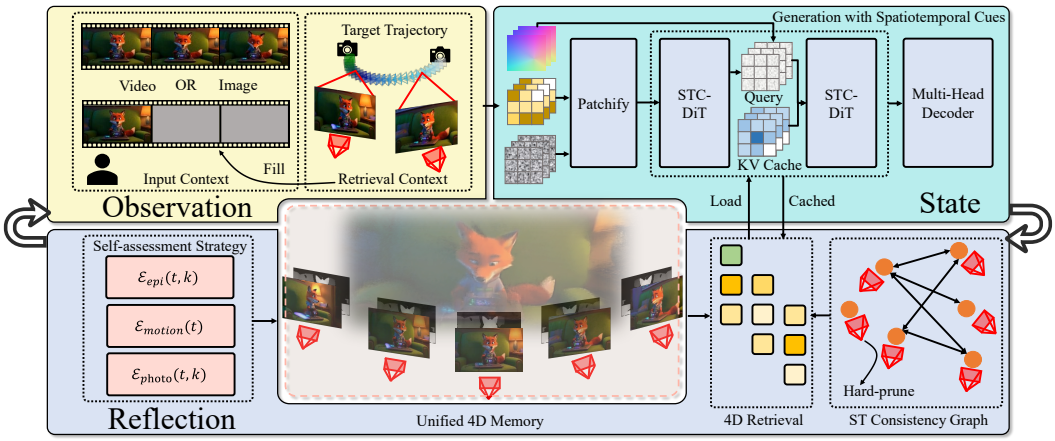


Figure 3: **"Observation-State-Reflection" inference strategy.** Leveraging a unified 4D memory, ChronoWorld enables dynamic self-assessment and correction of generated outputs, thereby enhancing the spatiotemporal consistency of 4D scenes.

Geometry, motion, and depth decoders. We adopt a Dense Prediction Transformer (DPT) architecture [30, 33]. Given denoised tokens $\mathbf{Z} \in \mathbb{R}^{N\times d}$, we reassemble them into multi-scale feature maps $\mathcal{F}_l \in \mathbb{R}^{H_l\times W_l\times C_l}$ across transformer layers l. A fusion module with residual convolutions and progressive upsampling aggregates these features into a dense representation $\mathbf{F}_{\text{dense}}$. Following [34], we apply geometry forcing to align intermediate diffusion features with geometry-aware supervision from 4D reconstruction models [35].

The Gaussian decoder maps $\mathbf{F}_{\text{dense}}$ to per-pixel 3D Gaussian attributes in canonical space. For each splat g, it predicts $A_g = \{\mu, \mathbf{s}, \mathbf{q}, \alpha, \mathbf{c}\}$, where $\mu \in \mathbb{R}^3$ is position, $\mathbf{s} \in \mathbb{R}^3$ anisotropic scale, $\mathbf{q} \in \mathbb{R}^4$ rotation quaternion, $\alpha \in [0, 1]$ opacity, and $\mathbf{c}$ spherical harmonics coefficients.

To model dynamics, the motion decoder predicts per-pixel displacements $m_p(t) = \{\Delta\mathbf{x}_p(t), \Delta\mathbf{a}_p(t)\}$ at query time t, where $\Delta\mathbf{x}_p(t) \in \mathbb{R}^3$ is the 3D translation relative to the canonical position μ_p, and $\Delta\mathbf{a}_p(t)$ captures temporal variations in other attributes (e.g., rotation or scale). This decoupling of static geometry and dynamic motion enables explicit temporal tracking of each Gaussian.

An auxiliary depth decoder predicts a dense depth map $\mathbf{D} \in \mathbb{R}^{H\times W}$, where d_p provides geometric regularization to improve spatial consistency and resolve scale ambiguity. Overall, the DPT-based design translates tokenized latents into pixel-aligned Gaussian primitives with coherent dynamics.

To supervise geometry, motion, and depth, we derive pseudo-labels from pre-trained reconstruction models [35, 36], caching their intermediate features as geometry-aware signals to align with diffusion hidden states via the reconstructive multi-head decoder.

3.4 Reconstruction-driven Geometric Reflection

A key limitation of existing video and 4D scene generation methods is that they lack reliable mechanisms to assess the quality of their own predictions, violating spatiotemporal consistency under zero-shot conditions. While recent 4D memory-augmented and post-reconstruction approaches attempt to enhance coherence, their uncorrected fusion of flawed predictions exacerbates error accumulation, inducing semantic drift, spatial forgetting, and geometric inconsistency.

To address this, we propose an *Observation-State-Reflection* inference strategy with Reconstruction-driven Geometric Reflection. The framework performs iterative retrieval, selection, and self-refinement, breaking the accumulation of single-pass errors and establishing a feedback loop for correction. This enables robust, globally consistent 4D scene generation across time.

Unified 4D Memory Construction.

While existing 3D-enhanced video generation methods passively unproject generated outputs into local spatial representations (e.g., depth maps, point clouds), their lack of global 4D modeling restricts them to static scenes and exacerbates error accumulation. To address this, we propose a unified 4D

memory construction strategy that aligns and fuses predicted representations within a Canonical Gaussian Space.

Specifically, we fix the first context view as the global reference frame $\mathcal{C}_{\text{ref}}$, and transform target camera poses $\mathbf{T}_{\text{tgt}}$ into this frame via $\mathbf{T}_{\text{rel}} = \mathbf{T}_{\text{ref}}^{-1}\mathbf{T}_{\text{tgt}}$. This ensures all Gaussians are generated in a consistent canonical space, mitigating global pose drift. For each splat p, the dynamic 4D Gaussian state at time t is constructed by augmenting the canonical base with predicted motion:

$$\mathbf{x}_p(t) = \mu_p + \Delta\mathbf{x}_p(t), \quad \mathcal{A}_p(t) = \mathcal{A}_p \oplus \Delta\mathbf{a}_p(t), \tag{6}$$

where $\mathbf{x}_p(t) \in \mathbb{R}^3$ is the Gaussian center at time t, $\mu_p \in \mathbb{R}^3$ is the canonical position anchored in $\mathcal{C}_{\text{ref}}$, and $\Delta\mathbf{x}_p(t)$ is the predicted motion offset. $\mathcal{A}_p(t)$ denotes time-varying attributes (e.g., rotation, scale), updated via $\Delta\mathbf{a}_p(t)$ using attribute-specific operators (e.g., quaternion composition).

The predicted depth map $\mathbf{D}$ further serves as a scale regularizer, grounding the 4D representation by normalizing Gaussian coordinates into a consistent metric range (e.g., $[-1, 1]^3$).

Self-Assessing for Reflection. To ensure fidelity of the unified 4D representation and suppress hallucinated structures, we introduce a hybrid self-assessment strategy that combines epipolar geometry, motion consistency, and dense reprojection errors. For a generated block at timestep t, the model constructs canonical dynamic Gaussians $\mathcal{G}_t$. We evaluate geometric reliability by differentiably reprojecting (rasterizing) $\mathcal{G}_t$ onto camera poses $\mathbf{T}_k$ of a retrieved set of verified anchor frames $\mathcal{M}_{\text{ref}}$.

We first compute the epipolar error using the symmetric Sampson distance:

$$\mathcal{E}_{\text{epi}}(t,k) = \frac{1}{\|\mathbf{M}_{t\to k}\|_0} \sum_{\mathbf{u}_t \leftrightarrow \mathbf{u}_k} \mathbf{M}_{t\to k}(\mathbf{u}_t) \frac{(\mathbf{u}_t^T \mathbf{F}_{t\to k}\mathbf{u}_k)^2}{\|\mathbf{F}_{t\to k}\mathbf{u}_k\|_2^2 + \|\mathbf{F}_{t\to k}^T\mathbf{u}_t\|_2^2}, \tag{7}$$

where $\mathbf{u}_t, \mathbf{u}_k \in \mathbb{R}^3$ are corresponding homogeneous pixels, $\mathbf{F}_{t\to k}$ is the fundamental matrix between $\mathbf{T}_t$ and $\mathbf{T}_k$, and $\mathbf{M}_{t\to k}$ is a depth-aware co-visibility mask. We then measure motion consistency via:

$$\mathcal{E}_{\text{motion}}(t) = \mathbb{E}_{p,q}\left[\|\Delta\mathbf{x}_p - \Delta\mathbf{x}_q\|_2^2\right] + \lambda\mathbb{E}_p\left[\|\partial_t\Delta\mathbf{a}_p\|_2^2\right], \tag{8}$$

where $\Delta\mathbf{x}_p$ and $\Delta\mathbf{x}_q$ denote motion vectors of temporally adjacent Gaussians, and $\partial_t\Delta\mathbf{a}_p$ penalizes high-frequency attribute variation. Finally, we evaluate photometric and perceptual consistency by comparing rendered projections $\hat{\mathbf{I}}_{t\to k}$ with observations $\mathbf{I}_k$:

$$\mathcal{E}_{\text{photo}}(t,k) = \|\mathbf{M} \odot (\hat{\mathbf{I}}_{t\to k} - \mathbf{I}_k)\|_1 + \lambda\mathcal{L}_{\text{LPIPS}}, \tag{9}$$

where $\mathbf{M}$ is a co-visibility mask and $\mathcal{L}_{\text{LPIPS}}$ is the perceptual loss [37]. These metrics are fused into a scale-invariant consistency graph spanning the current frame and historical context, enabling robust identification of unreliable predictions.

Cycle Refinement via 4D Retrieval Selection. Given an observation $\mathcal{O}$, our STC-DiT and reconstructive multi-head decoder output complementary representations (RGB, Gaussian parameters, motions, and depth) to construct a Unified 4D Memory. To prevent flawed frames from corrupting generations via passive caching, we formulate this memory as a dynamically evaluated topological graph. At inference, we execute a cyclic refinement loop: the model self-assesses this graph, hard-prunes geometrically inconsistent nodes, and retrieves the top-K reliable frames to dynamically refresh the KV cache. Conditioned on this curated context, corrupted regions are iteratively regenerated, thereby halting error propagation and reducing computational costs.

Active 4D Retrieval and KV Cache Refresh. Unlike conventional memory-augmented video generation that retrieves historical RGB frames or latent tokens based on appearance similarity, temporal proximity, or view overlap, our method retrieves structured 4D states in a unified Canonical Gaussian Space. Each state jointly represents persistent geometry, time-varying attributes, explicit motion, and depth. For the p-th Gaussian primitive, its position at time t is represented as $\mathbf{x}p(t) = \boldsymbol{\mu}p + \Delta\mathbf{x}p(t)$, where $\boldsymbol{\mu}p$ is the canonical position and $\Delta\mathbf{x}p(t)$ is its temporal motion. Its time-varying attributes are similarly defined as $\mathcal{A}p(t) = \mathcal{A}p \oplus \Delta\mathbf{a}p(t)$. This formulation explicitly associates the same Gaussian primitive across time, turning historical observations into a queryable 4D memory rather than independent frame-level context.

At each inference step, we first assess the reliability of historical states using the reconstruction-driven consistency graph, which jointly considers epipolar geometry, motion consistency, and reprojection fidelity. Denoting the aggregated inconsistency of memory node m_i by Φ_i, we retain only states

Table 1: **Quantitative evaluation** of our method against state-of-the-art models in camera-controlled video generation, 3D-enhanced scene generation, 4D dynamic scene generation, measuring visual generation quality, and 4D visual synchronization. All tests are conducted under fair comparison settings, using identical video inputs and target camera trajectories.

Method	Type	Visual Quality		4D Visual Synchronization			Camera Accuracy	
		CLIP-V↑	FID↓	FVD-F↓	FVD-V↓	FVD-4D↓	RPE-R↓	RPE-T↓
MotionCtrl [8]	2D	0.46	123.68	535.96	421.58	442.72	4.77	8.93
CameraCtrl [9]	2D	0.49	114.27	510.59	378.05	390.16	4.02	7.65
RecamMaster [38]	2D	0.64	101.58	403.18	214.33	198.52	3.01	6.24
TrajectoryCrafter [39]	3D	0.64	98.46	487.02	293.56	286.50	1.26	3.76
Vmem [40]	3D	0.66	113.41	389.64	252.08	264.73	2.59	6.42
GEN3C [41]	3D	0.58	87.60	474.86	274.71	283.48	2.76	5.21
DeepVerse [42]	4D	0.67	72.51	376.05	237.10	239.77	1.82	3.20
Free4D [13]	4D	0.71	88.32	347.56	224.95	216.67	1.73	3.28
Neoverse [43]	4D	0.72	90.16	353.48	246.71	223.10	1.68	3.02
Ours	4D	**0.80**	**64.58**	**314.28**	**201.46**	**177.41**	**0.95**	**2.34**

satisfying $\Phi_i \leq \tau_{\text{prune}}$, forming the clean memory $\mathcal{M}t^{\text{clean}}$. This hard pruning is performed before retrieval, preventing geometrically inconsistent states from contaminating subsequent generation. Given the current 4D state as query q_t, we then perform active Top-K retrieval according to the joint spatiotemporal ranking function $S\text{ST}(m_i, q_t)$, yielding $\mathcal{H}t = \text{TopK}\, m_i \in \mathcal{M}t^{\text{clean}} S\text{ST}(m_i, q_t)$. Here, relevance is determined by the joint geometric validity and spatiotemporal utility of each state rather than appearance similarity alone, allowing reliable long-range observations to be actively recalled when they provide useful constraints for the current generation.

The retrieved states are encoded into the generator's key-value context as $(\mathbf{K}_t, \mathbf{V}t) = \text{Encode}\,\text{KV}(\mathcal{H}_t)$, which replaces the previously accumulated historical context. Thus, the active KV cache is bounded by the retrieval budget K, rather than the total history length, reducing the context size from $O(N_t L)$ to $O(KL)$ when each state contributes L tokens. Importantly, this is selective compression rather than temporal truncation, since states are retained according to both reliability and query relevance.

The retrieval is performed recurrently throughout inference. Each generation step produces new RGB, Gaussian geometry, motion, and depth, which extend the Unified 4D Memory and are subsequently re-evaluated to update the consistency graph. The refined memory then undergoes hard pruning, Top-K retrieval, and KV refresh for the next generation cycle, forming a closed-loop "generate $\rightarrow$ assess $\rightarrow$ prune $\rightarrow$ retrieve $\rightarrow$ refresh $\rightarrow$ regenerate" process. In this way, 4D RAG preserves scene information by actively recalling reliable states beyond the current temporal window, maintains a bounded inference context by restricting the active KV cache to K retrieved states, and suppresses cumulative spatial and temporal drift by preventing inconsistent predictions from being repeatedly propagated through memory. This active, reliability-aware memory update is therefore essential for maintaining stable 4D reconstruction and generation over extended sequences.

4 Experiments

4.1 Experimental Settings

Implementation Details. Our framework builds on Wan2.1 I2V [19], augmented with a Diffusion Transformer with Spatiotemporal Cues (STC-DiT), Reconstructive Multi-Head Decoder (RMD), and camera pose embeddings with spatial context augmentation. We train on a mixture of public video generation and reconstruction datasets [38, 44], combining real and synthetic dynamic scenes. We perform rectified flow distillation for STC-DiT with learning rate 1×10^{-3}, batch size 10, and 20K steps, using AdamW (weight decay 0.05) and a cosine schedule with linear warm-up. Subsequently, the RMD is trained incorporating geometric forcing, with a learning rate of 1×10^{-6} and a batch size of 10. DPM++ is used for efficient epipolar sampling. Training runs on 8$\times$A100 (80GB) GPUs, while inference uses a single A100 GPU. Additional details are provided in the Supplementary Material.

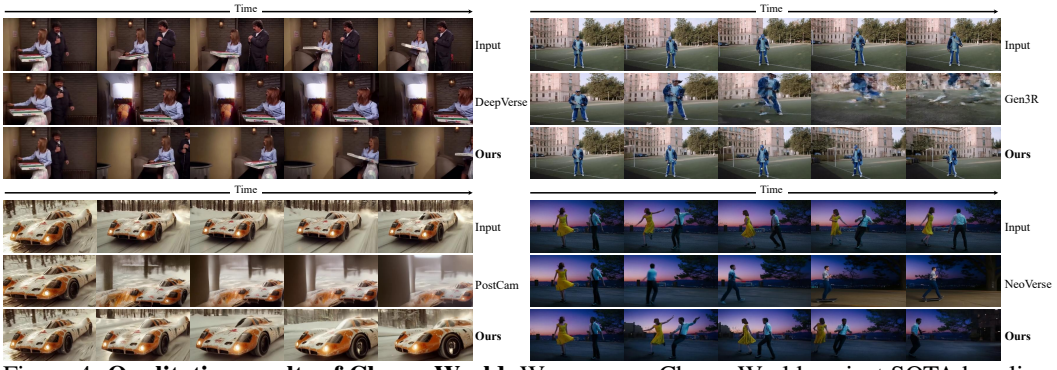


Figure 4: **Qualitative results of ChronoWorld.** We compare ChronoWorld against SOTA baselines across diverse camera trajectories to validate its effectiveness (zoom in for details).

Table 2: **Quantitative comparisons on spatiotemporal consistency** of our method against state-of-the-art models in generation and reconstruction of 4D scenes.

Method	Representation Consistency			Spatiotemporal Consistency					
	PSNR↑	SSIM↑	LPIPS↓	Aesthetic Quality ↑	Imaging Quality ↑	Temporal Flickering ↑	Motion Smoothness ↑	Subject Consistency ↑	Background Consistency ↑
Free4D [13]	18.03	0.64	0.48	0.51	0.54	0.96	0.98	0.92	0.94
More4D [51]	18.32	0.67	0.49	0.48	0.56	0.97	0.98	0.91	0.95
DeepVerse [13]	16.37	0.56	0.61	0.50	0.51	0.92	0.95	0.93	0.91
Lyra [49]	17.01	0.58	0.72	0.41	0.43	0.88	0.81	0.87	0.90
Gen3R [52]	17.28	0.64	0.68	0.44	0.46	0.49	0.88	0.90	0.92
Ours	**18.87**	**0.73**	**0.45**	**0.52**	**0.58**	**0.99**	**0.99**	**0.95**	**0.96**

Evaluation. We curate a comprehensive testing set of real and synthetic videos, which are sourced from diverse datasets, encompassing movies, games, animations, real-world captures, indoor scenes, in-the-wild driving scenarios, aerial drone footage, and web-sourced videos. We assess the model using multi-dimensional metrics: visual quality (CLIP-V, FID), 4D visual fidelity (FVD-F, FVD-V, FVD-4D), camera pose accuracy (RPE-R, RPE-T), reconstruction quality (PSNR, SSIM, LPIPS). VBench [45, 46] is also employed to measure spatiotemporal consistency.

4.2 Comparisons with State-of-the-Art Methods

We compare our proposed ChronoWorld with state-of-the-art approaches in camera-controlled video generation, 3D-enhanced scene generation, and 4D dynamic scene generation.

Quantitative Comparison. As shown in Tab. 1, our method consistently surpasses all baselines across all metrics, achieving stronger visual quality alongside significantly improved 4D consistency and trajectory accuracy. Among the baselines, camera-controlled video generation methods [38, 8, 9, 47] generate visually plausible videos but suffer from shape drift and trajectory misalignment without 4D constraints and explicit reflection. Current 3D-enhanced scene generation methods [40, 48] rely on static surfel- or point-based memory, struggling to synthesize high-fidelity, temporally coherent 4D dynamic scenes. Despite utilizing depth-warped conditioning, GEN3C [41] and TrajectoryCrafter [39] introduce pronounced textural artifacts stemming from unnatural geometric priors and flawed projections.

While existing 4D dynamic scene generation methods [13, 42, 49, 50, 43] utilize 4D memory, their naive "generate-and-reconstruct" relies on expensive post-reconstruction and per-frame querying, inevitably amplifying error accumulation and inconsistencies. In contrast, our model achieves globally consistent conditional generation by aligning the implicit latent and 4D spaces via spatiotemporal constraints, and rectifying discrepancies through explicit geometric reflection.

Qualitative Comparison. As shown in Fig. 4, we visualize the 4D scenes generated from a source video along arbitrary camera trajectories. Compared with baselines, our approach generates high-fidelity, spatiotemporally coherent 4D scenes, exhibiting realistic appearance, precise pose accuracy, and stable geometric structures. More results and videos are provided in the Supplementary Material.

4.3 Comparisons on Spatiotemporal Consistency

In this work, we aim to enhance the spatiotemporal consistency of scene generation, jointly synthesizing cinematic-quality videos and geometrically coherent 4D representations. Thus, we validate the unified 4D memory by rendering target novel views and quantifying their photometric and spatiotemporal fidelity. Tab. 2 shows that our approach consistently outperforms our main competitors in preserving multi-view geometry and spatiotemporal consistency for 4D scenes. Notably, our Spatiotemporal Epipolar Causal Attention enforces latent geometric alignment, while Reconstruction-driven Geometric Reflection iteratively rectifies textural artifacts and spatial distortions.

4.4 Ablation Studies

We perform comprehensive ablation studies to validate the key components of our framework, as summarized in Tab. 3. Further detailed ablations are deferred to the Supplementary Material.

Spatiotemporal Epipolar Causal Attention (ST-ECA) and Plücker ray embeddings jointly enforce multi-view geometric coherence and trajectory alignment. The introduced temporal tube masking, alongside spatial context augmentation, facilitates the mitigation of exposure bias and the rectification of spatial distortions. Reconstruction-driven Geometric Reflection proves effective in mitigating spatial forgetting and geometric degradation by enabling the generative model to iteratively rectify errors through self-reflection. Our proposed dual-stage training, coupled with geometry forcing, enables the model to decode 4D scene representations from latent features, thereby significantly enhancing the spatial consistency of predictions while accelerating the cyclic inference by $10\times$.

Table 3: **The effect of core components in our method.**

Method	PSNR↑	CLIP-V↑	FVD↓	RPE_{avg}↓
w/o ST-ECA	18.03	0.71	184.27	1.75
w/o Plücker Emb	18.71	0.79	181.42	1.79
w/o RMD	18.59	0.77	212.79	1.68
w/o ST Reflection	17.65	0.75	196.53	1.72
Ours-Full	18.87	0.80	177.41	1.64

5 Conclusion

In this paper, we presented ChronoWorld, an "Observation-State-Reflection" framework for camera-controllable video generation that explicitly enforces 4D spatiotemporal coherence. By grounding the model in Spatiotemporal Epipolar Causal Attention, our approach incorporates multi-view geometric constraints and temporal reasoning directly into the generation. Combined with reconstruction-driven geometric reflection and an active 4D retrieval strategy, the framework enables dynamic self-assessment and correction, effectively mitigating error accumulation. Extensive experiments demonstrate that ChronoWorld achieves state-of-the-art performance in synthesizing high-quality, free-view 4D scenes with strong geometric fidelity, temporal consistency, and trajectory accuracy.

References

[1] Xiaoyu Li, Qi Zhang, Di Kang, Weihao Cheng, Yiming Gao, Jingbo Zhang, Zhihao Liang, Jing Liao, Yan-Pei Cao, and Ying Shan. Advances in 3d generation: A survey. *arXiv preprint arXiv:2401.17807*, 2024. 2

[2] Rafael Pages, Konstantinos Amplianitis, David Monaghan, Jan Ondrej, and Aljosa Smolic. Affordable content creation for free-viewpoint video and vr/ar applications. *Journal of Visual Communication and Image Representation*, 53:192–201, 2018. 2

[3] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, et al. Holodeck: Language guided generation of 3d embodied ai environments. In *CVPR*, pages 16227–16237, 2024. 2

[4] Xinjie Wang, Liu Liu, Yu Cao, Ruiqi Wu, Wenkang Qin, Dehui Wang, Wei Sui, and Zhizhong Su. Embodiedgen: Towards a generative 3d world engine for embodied intelligence. *arXiv preprint arXiv:2506.10600*, 2025. 2

[5] Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. *arXiv preprint arXiv:2405.14475*, 2024. 2

[6] Yifan Lu, Xuanchi Ren, Jiawei Yang, Tianchang Shen, Zhangjie Wu, Jun Gao, Yue Wang, Siheng Chen, Mike Chen, Sanja Fidler, et al. Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models. In *ICCV*, pages 27272–27283, 2025. 2

[7] Zehuan Wu, Jingcheng Ni, Xiaodong Wang, Yuxin Guo, Rui Chen, Lewei Lu, Jifeng Dai, and Yuwen Xiong. Holodrive: Holistic 2d-3d multi-modal street scene generation for autonomous driving. *arXiv preprint arXiv:2412.01407*, 2024. 2

[8] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH*, pages 1–11, 2024. 2, 3, 8, 9

[9] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 2, 3, 8, 9

[10] Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In *ICCV*, pages 28785–28796, 2025. 2

[11] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion prompting: Controlling video generation with motion trajectories. In *CVPR*, pages 1–12, 2025. 2

[12] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *CVPR*, pages 5916–5926, 2025. 2

[13] Tianqi Liu, Zihao Huang, Zhaoxi Chen, Guangcong Wang, Shoukang Hu, Liao Shen, Huiqiang Sun, Zhiguo Cao, Wei Li, and Ziwei Liu. Free4d: Tuning-free 4d scene generation with spatial-temporal consistency. *arXiv preprint arXiv:2503.20785*, 2025. 2, 8, 9

[14] Chenxi Song, Yanming Yang, Tong Zhao, Ruibo Li, and Chi Zhang. Worldforge: Unlocking emergent 3d/4d generation in video diffusion model via training-free guidance. *arXiv preprint arXiv:2509.15130*, 2025. 2

[15] Stanislaw Szymanowicz, Jason Y Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T Barron, and Philipp Henzler. Bolt3d: Generating 3d scenes in seconds. In *ICCV*, pages 24846–24857, 2025. 2

[16] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3

[17] Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, et al. Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts. *arXiv preprint arXiv:2507.20939*, 2025. 3

[18] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025. 3

[19] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 8

[20] Vinay Joseph and Gustavo de Veciana. Nova: Qoe-driven optimization of dash-based video delivery in networks. In *INFOCOM*, pages 82–90. IEEE, 2014. 3

[21] Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, et al. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025. 3

[22] Songchun Zhang, Huiyao Xu, Sitong Guo, Zhongwei Xie, Hujun Bao, Weiwei Xu, and Changqing Zou. Spatialcrafter: Unleashing the imagination of video diffusion models for scene reconstruction from limited observations. In *ICCV*, pages 27794–27805, 2025. 3

[23] Xiaoyu Zhou, Xingjian Ran, Yajiao Xiong, Jinlin He, Zhiwei Lin, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Gala3d: Towards text-to-3d complex scene generation via layout-guided generative gaussian splatting. *arXiv preprint arXiv:2402.07207*, 2024. 3

[24] Ryan Po and Gordon Wetzstein. Compositional 3d scene generation using locally conditioned diffusion. In *3DV*, pages 651–663, 2024. 3

[25] Gege Gao, Weiyang Liu, Anpei Chen, Andreas Geiger, and Bernhard Scholkopf. Graphdreamer: Compositional 3d scene synthesis from scene graphs. In *CVPR*, pages 21295–21304, 2024. 3

[26] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[27] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuhler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 3

[28] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *CVPR*, pages 19447–19456, 2024. 3

[29] Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. In *ECCV*, pages 341–359. Springer, 2024. 3

[30] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306, 2025. 3, 6

[31] Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *arXiv preprint arXiv:2505.23716*, 2025. 3

[32] Nikhil Keetha, Norman Muller, Johannes Schonberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 3

[33] Haonan Wang, Hanyu Zhou, Haoyue Liu, and Luxin Yan. 4d-vggt: A general foundation model with spatiotemporal awareness for dynamic scene geometry estimation. *arXiv preprint arXiv:2511.18416*, 2025. 6

[34] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025. 6

[35] Chenguo Lin, Yuchen Lin, Panwang Pan, Yifan Yu, Honglei Yan, Katerina Fragkiadaki, and Yadong Mu. Movies: Motion-aware 4d dynamic view synthesis in one second. *arXiv preprint arXiv:2507.10065*, 2025. 6

[36] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 6

[37] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *NeurIPS*, 29, 2016. 7

[38] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 8, 9

[39] Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638*, 2025. 8, 9

[40] Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. *arXiv preprint arXiv:2506.18903*, 2025. 8, 9

[41] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Muller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *CVPR*, pages 6121–6132, 2025. 8, 9

[42] Junyi Chen, Haoyi Zhu, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, et al. Deepverse: 4d autoregressive video generation as a world model. *arXiv preprint arXiv:2506.01103*, 2025. 8, 9

[43] Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026. 8, 9

[44] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, pages 22160–22169, 2024. 8

[45] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, pages 21807–21818, 2024. 9

[46] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 9

[47] Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025. 9

[48] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 9

[49] Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B Lindell, Zan Gojcic, Sanja Fidler, et al. Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. *arXiv preprint arXiv:2509.19296*, 2025. 9

[50] Chun-Han Yao, Yiming Xie, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4d generation. *arXiv preprint arXiv:2503.16396*, 2025. 9

[51] Yanran Zhang, Ziyi Wang, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Joint 3d geometry reconstruction and motion generation for 4d synthesis from a single image, 2025. 9

[52] Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3r: 3d scene generation meets feed-forward reconstruction. *arXiv preprint arXiv:2601.04090*, 2026. 9