AEON: Unifying Video and 3D World Models

Weiqi Zhang **Wenyuan Zhang** **Yu-Shen Liu**[†] **Junsheng Zhou**[†]

School of Software, Tsinghua University, Beijing, China

{zwq23, zhou-js24}@mails.tsinghua.edu.cn zhangweny98@gmail.com

liuyushen@tsinghua.edu.cn

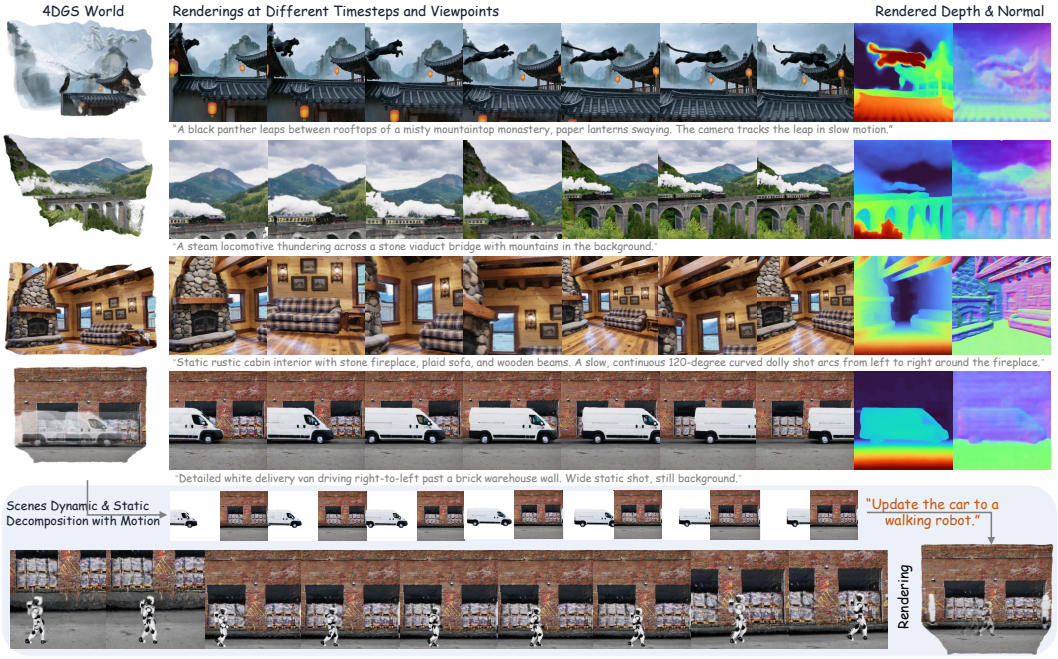


Figure 1: AEON generates 4DGS worlds that support real-time rendering at different timesteps and viewpoints. The 4DGS worlds are also ready for explicit editing and simulation (bottom).

Abstract

Current world models for explicit world simulation typically follow two distinct pathways: (1) video generative models (e.g., Sora), which simulate realistic appearances but fail to preserve 3D consistency; and (2) 3D generative models (e.g., Marble), which generate geometrically consistent scenes but often yield static, less realistic observations. We present AEON, a unified framework that bridges video-based and 3D-based world models and combines the benefits of both. The key insight of AEON lies in a generative model that operates within the text- or image-conditioned video latent space while producing renderable 4D graphics primitives—specifically, 4D Gaussian Splatting (4DGS)—as the explicit world representation. We train a 4D reconstruction model to predict both 4DGS and camera poses from monocular videos, serving as our initialized 4D decoder. Furthermore, we design a domain adapter that translates generative video latents into reconstructive 4D latents through a distillation process on the video VAE. AEON naturally supports video-based world modeling by rendering 4DGS at predicted camera

[†]The corresponding authors are Yu-Shen Liu and Junsheng Zhou.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

trajectories, and constructs a persistent 4D environment with scene dynamics for realistic world simulation. Consequently, AEON empowers a diverse range of world modeling applications and achieves state-of-the-art performance on standard benchmarks. *Project Page: https://weiqi-zhang.github.io/AEON.*

1 Introduction

World modeling has rapidly advanced to the forefront of AI research, driven by its immense potential in embodied AI, autonomous driving, and game design. Despite this progress, there remains a lack of consensus on the optimal representation for explicit world simulation, leaving the field divided between two dominant paradigms: video-based [27, 39] and 3D-based world models [46, 14]. As highlighted in Table 1, these approaches possess orthogonal strengths and suffer from fundamental, paradigm-specific bottlenecks. Consequently, a more powerful and unified representation is imperative to achieve comprehensive world modeling.

We begin by asking a fundamental question: What constitutes a good world model? To train robots to understand physical laws and perceive scenes, a world model should exhibit the following properties: (1) generating **realistic** appearances that align with real-world domains; (2) modeling **dynamic** scene evolution (e.g., fire, waves) and moving objects; (3) maintaining **3D consistent** environments free from geometric ambiguity; (4) ensuring scene **persistence** for continuous and repeatable simulation; and (5) supporting **real-time exploration** to facilitate efficient interaction. We argue that a representation unifying video- and 3D-based world models, thereby benefiting from their complementary strengths, is the key to achieving such an ideal world model.

We present AEON, a generative world model that introduces 4D Gaussian Splatting (4DGS) [42, 54] as its foundational representation, unifying video and 3D models to benefit from both paradigms. 4DGS represents scenes as a set of explicit graphics primitives localized in spatial geometry, and models scene dynamics through the velocity-driven movement and time-varying visibility of these primitives. Consequently, a learned 4DGS representation provides two independent yet crucial capabilities: it models world appearances through direct rendering along user-defined camera trajectories, and it establishes a persistent 4D environment that supports highly controllable 3D simulations. Additionally, it enables explicit modeling of world geometry and physical states, including depth, normals, motion, and optical flow.

Despite the comprehensive capabilities inherent to the 4DGS representation, most existing approaches restrict its application to reconstruction tasks, and generating a persistent 4DGS world directly from text or image conditions remains largely unexplored. In AEON, we propose a novel framework that integrates the generative latent space of a pretrained video Diffusion Transformer (DiT) [39, 30] with the reconstructive latent space of a 4D reconstruction model. Specifically, we train a 4D Gaussian Reconstruction Model (4DRM) to jointly predict 4DGS attributes and camera poses from monocular videos. By utilizing a large-scale annotated 4D video dataset and pretrained geometric priors from MapAnything [17], a scalable geometry model, the 4DRM develops robust reconstructive capabilities to recover world geometry and dynamics. A primary challenge, however, is the significant domain gap between the video and 4DRM latent spaces. We address this by introducing a domain adapter that seamlessly bridges the two. Finally, we employ a multi-stage training pipeline that optimizes the adapter and fine-tunes the reconstruction model through end-to-end distillation from the video model, supervised by video VAE-decoded outputs.

Consequently, AEON unlocks an extensive array of world modeling applications, including text/image/video-to-4DGS generation, 3D world generation, robotics simulation, physical state estimation (e.g., depth, normals, and optical flow), scene editing, and novel view and timestep synthesis.

Our main contributions can be summarized as follows:

- We introduce AEON, a unified world modeling framework that represents worlds as explicit 4D Gaussians, effectively bridging video-based and 3D-based models to benefit from their complementary strengths.
- We propose a novel methodology that integrates the generative latent space of a pretrained video DiT with the reconstructive latent space of a 4D Gaussian reconstruction model. To

Table 1: Comparison of different world modeling paradigms across fundamental characteristics.

Paradigm	Realistic	3D Consistent	Dynamic	Persistence	Real-time Exp.
Video WM (e.g. Sora)	✓	✗	✓	✗	✗
3D WM (e.g. Marble)	✗	✓	✗	✓	✓
AEON	✓	✓	✓	✓	✓

achieve this, we train a specialized domain adapter through an end-to-end distillation of the video model.

- AEON demonstrates comprehensive versatility, natively supporting a diverse spectrum of world modeling tasks, and establishes new state-of-the-art results on standard benchmarks.

2 Related Work

2.1 Video and 3D World Models

World modeling aims to generate explorable and realistic environments from conditional inputs (e.g., text or images), serving as a foundation for embodied AI and autonomous systems. Current paradigms primarily rely on video-based [27, 11, 33, 19, 53, 39] or 3D-based [46, 36, 52, 14, 34, 3, 10] approaches. Within the video-centric domain, JEPA-style methods [2, 1] excel at forecasting future latent states but lack explicit visual synthesis. Conversely, video world models [27, 39, 19] and their action-conditioned variants [5, 45, 37, 38] can synthesize photorealistic frames, yet they struggle with 3D consistency and incur heavy computational costs due to massive per-frame token generation.

Alternatively, 3D world models [46, 14] construct explicit environments for simulation. Recent works have explored panoramic generation [36, 52], progressive expansion [14], or leveraging video priors [3]. However, these methods typically employ static representations such as 3DGS [18] or meshes, precluding the modeling of world dynamics and limiting their ability to scale on dynamic internet video datasets. The same holds for 3D generation and reconstruction more broadly, which mostly build on static Gaussians [63, 60, 59, 57, 48] or distance fields [64, 65, 26, 25]. To bridge this gap, we propose AEON, a framework that generates renderable 4D Gaussian Splatting [54, 42] (4DGS) primitives as an explicit world representation. AEON effectively unifies the strengths of both paradigms, achieving dynamic modeling with 3D consistency while overcoming their respective limitations.

2.2 4D Gaussian Splatting from Videos

4D Gaussian Splatting [54, 42] advances the success of 3D Gaussian Splatting [18] for modeling scene dynamics. 4DGS learns from multi-view videos to predict Gaussian primitives with explicit temporal attributes such as lifespan, velocity, and time-varying opacity, allowing Gaussians to appear, move, and vanish independently. However, predicting 4DGS from monocular videos remains significantly more challenging due to limited observational data. While optimization-based methods [40, 20, 35] can achieve faithful visual appearances, they require lengthy fitting processes and struggle to generalize or scale. More recently, feed-forward frameworks [32, 49, 51, 22, 61] have been proposed for fast and scalable video-to-4DGS reconstruction, including bullet-time reconstruction at arbitrary timestamps [21] and online reconstruction from uncalibrated video streams [47]. Some of these works predict deformable 3DGS [22, 29] as their foundational representation, which often falls short when modeling complex and diverse dynamics. Related feed-forward models also recover dynamic geometry and camera poses from monocular videos [8]. All of these methods reconstruct scenes from observed videos. Native 4DGS generation has also been explored for animated objects [58]. Closer to our setting, Diff4Splat [29] and Lyra [3] build Gaussian scenes on top of video diffusion priors, but both require a camera trajectory as input and represent dynamics with deformable 3DGS and per-frame 3DGS, respectively. AEON differs in both respects: it generates a world from a text prompt or a single image without camera input, and it represents the world as one persistent set of 4D primitives, each with its own reference time, duration, and velocity, so that content can appear, move, and vanish independently and can be rendered at any continuous timestep. Concurrent with our work, Beyond Pixels [23] also bypasses RGB decoding, lifting video latents to point maps instead of 4DGS.

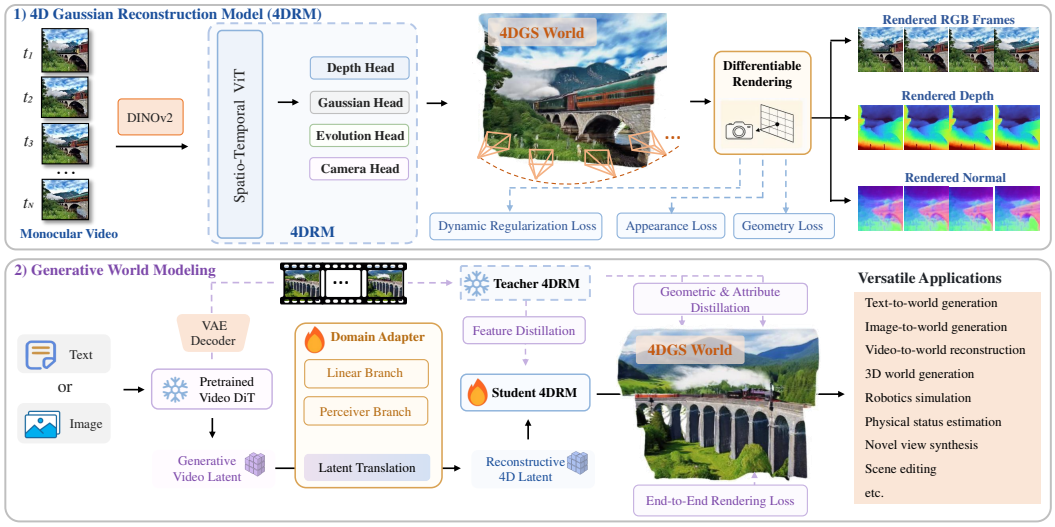


Figure 2: **Overview of AEON. Top (1).** The 4D Reconstruction Model (4DRM) decodes a monocular video into a 4DGS world through four task-specific heads that respectively predict depth, Gaussian appearance, dynamic evolution, and camera poses. The resulting primitives are rasterized into RGB, depth, and normal maps for end-to-end supervision. **Bottom (2).** We train AEON for generative world modeling by distilling the world generative capability of a frozen video diffusion model (e.g., Wan 2.2 [39]) and the 4D reconstructive capability of a teacher 4DRM via a Domain Adapter and a learnable student 4DRM. The generated explicit 4DGS world enables versatile world modeling applications.

3 Method

We present AEON, a generative world model that operates within the latent space of a pretrained video diffusion model while producing explicit 4D Gaussian Splatting (4DGS) as the world representation. Given a text prompt or conditioning image, our goal is to generate a set of 4D Gaussians encoding both appearance and dynamics, enabling photorealistic rendering and persistent 3D simulation. The overview of AEON is shown in Fig. 2. We begin by introducing the 4DGS representation in Sec. 3.1, which models dynamic scenes through velocity-driven motion and time-varying visibility. In Sec. 3.2, we present the 4D Gaussian Reconstruction Model, whose four task-specific heads jointly predict depth, Gaussian appearance, dynamic evolution, and camera poses from monocular videos. We then describe our domain adapter in Sec. 3.3, which bridges the generative video latent space and the reconstructive 4D latent space through distillation. Finally, Sec. 3.4 details our multi-stage training strategy that optimizes the full pipeline end-to-end.

3.1 4D Gaussian Splatting as the World Representation

An ideal world model should jointly capture realistic appearances, scene dynamics, and 3D-consistent geometry within a persistent and efficiently renderable representation. To this end, AEON adopts 4D Gaussian primitives [42, 54, 61] as its foundational scene representation, where each primitive simultaneously encodes geometry, appearance, and temporal evolution. Formally, a single primitive g_i is characterized by the following parameter set

$$g_i = \{\mathbf{x}_i, \mathbf{s}_i, \mathbf{r}_i, o_i, \mathbf{c}_i, t_i, l_i, \mathbf{v}_i, \boldsymbol{\omega}_i\}. \tag{1}$$

Specifically, $\mathbf{x}_i \in \mathbb{R}^3$ denotes the spatial center, $\mathbf{s}_i \in \mathbb{R}^2$ parameterizes anisotropic scaling along the principal axes, and $\mathbf{r}_i \in \mathbb{R}^4$ encodes the orientation as a unit quaternion. The base opacity $o_i \in [0, 1]$ together with the color $\mathbf{c}_i \in \mathbb{R}^3$ jointly determine the visual appearance of each primitive.

In addition to these spatial attributes, four temporal parameters govern the dynamic behavior of each primitive. The reference timestep $t_i \in [0, T]$ specifies the moment of peak activation, while the duration $l_i \in \mathbb{R}^+$ controls the temporal extent over which the primitive remains visible. The linear velocity $\mathbf{v}_i \in \mathbb{R}^3$ and the angular velocity $\boldsymbol{\omega}_i \in \mathbb{R}^3$ in axis-angle form capture translational and

rotational motion, respectively. This decomposition into explicit spatial and temporal components enables AEON to model both persistent scene structure and transient dynamic phenomena within a unified representation.

3.2 4D Gaussian Reconstruction Model

While 4DGS offers an ideal world representation, directly generating 4D Gaussians from diverse conditions is highly challenging without proper guidance and geometric initialization. To address this, we first train a 4D Gaussian Reconstruction Model (4DRM) as a foundation, focusing on the more tractable task of world recovery from monocular videos. In a single feed-forward pass, 4DRM uses four task-specific heads to jointly predict depth, camera poses, and 4D Gaussian attributes of appearance and evolution. The model is trained end-to-end via differentiable rendering. Once trained, 4DRM facilitates AEON's generative world modeling in two ways: (1) it provides a reconstructive foundation to initialize the student 4DRM that recovers the world from video latents, and (2) it acts as a teacher model, guiding the student to acquire 4D Gaussian modeling capabilities through distillation.

Encoding. Given an input video $V = \{I_t\}_{t=1}^{T}$ of T frames, we extract per-frame dense features $\mathbf{F}_t \in \mathbb{R}^{H/p \times W/p \times D}$ using a pretrained DINOv2 encoder [28]. To inject temporal identity, we add a sinusoidal positional embedding of the timestep index t to each spatial feature map. The resulting per-frame tokens are concatenated across all frames and processed by a transformer backbone of L layers that alternate between intra-frame and cross-frame self-attention, balancing computational efficiency with long-range temporal modeling. The output features $\mathbf{F}^S \in \mathbb{R}^{THW/p^2 \times D}$ encode holistic spatio-temporal context for subsequent decoding. We initialize 4DRM from the pretrained weights of MapAnything [17], a feed-forward metric 3D reconstruction model.

Decoding. We decode the shared spatio-temporal representation $\mathbf{F}_t^S$ through four task-specific heads, each responsible for a complementary component of the 4DGS world. The *Depth Head* predicts dense per-frame depth maps $\mathbf{D}_t \in \mathbb{R}^{H \times W}$, while the Camera Head regresses the intrinsics $\mathbf{K}_t$ together with per-frame rotations $\mathbf{R}_t$ (unit quaternions) and translations $\mathbf{t}_t$. The predicted depths are unprojected via $\mathbf{K}_t$ into the Gaussian centers $\mathbf{x}_i$, anchoring all downstream predictions in a metrically grounded 3D space. Conditioned on these centers, the *Gaussian Head* predicts the appearance attributes $\{\mathbf{s}_i, \mathbf{r}_i, o_i, \mathbf{c}_i\}$ (scale, orientation, opacity, and color), and the *Evolution Head* predicts the dynamic attributes $\{t_i, l_i, \mathbf{v}_i, \boldsymbol{\omega}_i\}$ (temporal reference, duration, and linear and angular velocities) that govern how each primitive moves and fades over time. This decomposition lets the Gaussian and Evolution Heads focus exclusively on appearance and dynamics, while the Camera Head's joint pose estimation enables 4DRM to operate directly on raw monocular videos, removing the dependency on pre-estimated poses or external Structure-from-Motion (SfM) pipelines required by prior work [49, 22]. As a result, 4DRM recovers a complete dynamic world state from a single monocular video and establishes a robust reconstructive latent space for the subsequent generative integration.

3.3 Generative World Modeling

Overview. AEON is designed to generate persistent and explorable 4D worlds from diverse conditions, such as text prompts and images. The complete pipeline is illustrated in the lower portion of Fig. 2. Given a condition (e.g., a text prompt), we leverage a pretrained video diffusion model (e.g., Wan 2.2) as the generative foundation to produce a video latent, z_{video}. Given that z_{video} encapsulates profound world-modeling priors capable of decoding into realistic, high-quality videos, we aim to unlock its full potential by recovering an explicit and persistent 4DGS world. We introduce a Domain Adapter that performs latent translation, mapping the generative video latent z_{video} to a reconstructive 4D latent z_{4D}. A learnable student 4DRM then utilizes z_{4D} to explicitly recover the 4DGS world. A simpler alternative would decode z_{video} into frames with the VAE and pass them to the Stage-1 4DRM. We avoid this cascade because the 4DRM is trained on real videos, whereas decoded generations carry VAE and generation artifacts outside that distribution, so decoding errors would compound with reconstruction errors.

A primary challenge in connecting the video generative model to the 4D reconstruction model is the significant domain gap between their latent representations. The video latent is optimized for

generative denoising, while the 4DRM consumes DINOv2 [28] tokens that are trained for geometric and semantic understanding. These two feature spaces therefore occupy fundamentally different distributions. Naively replacing the DINOv2 patch embedding with a randomly initialized projection would forfeit the geometric prior carried by the pretrained backbone and force the student 4DRM to relearn its feature distribution from scratch. To avoid this, we introduce a learnable domain adapter that bridges the two latent spaces while preserving the geometric capabilities of the pretrained 4DRM backbone layers.

Dual-Path Domain Adapter. We implement the domain adapter as a dual-path architecture that couples a linearly initialized convolutional branch with an expressive Perceiver Bridge. To establish a robust baseline, we first solve a closed-form ridge regression to map the video latents to the target DINOv2 activations at each candidate insertion depth ℓ. This linear mapping already recovers a coarse alignment, which makes it a good initialization, although on its own it falls well short of the full adapter (Sec. 4.4). Motivated by this, our dual-path design employs a linear branch to provide a stable foundation, while an expressive residual branch captures complex nonlinear transformations. Specifically, the linear branch $\mathbb{L}$ is instantiated as a strided 3D convolution initialized with the closed-form ridge solution. The residual branch $\mathbb{N}$ comprises a Perceiver Bridge module, detailed below. A learnable scalar gate γ fuses the two branches as:

$$Z = \mathbb{L}(z_{\text{video}}) + \gamma \cdot \mathbb{N}(z_{\text{video}}), \tag{2}$$

where γ is initialized to zero. Consequently, the adapter's optimization begins entirely from the stable linear solution and progressively incorporates the nonlinear residual branch as γ updates via gradient descent, ensuring stable optimization from the outset.

Perceiver Bridge. Implementing the residual branch with standard convolutions inherently dilutes fine-grained geometric information due to their restricted local receptive fields. Instead, we design a highly expressive Perceiver [15] Bridge module that leverages a learned token-to-token retrieval with a global receptive field. The video latent is first projected into the target feature dimension by a spatio-temporal Conv3D. An asymmetric temporal upsampler then restores the original temporal resolution by bypassing the independently encoded first frame and applying sub-pixel shuffle to the remaining compressed tokens so that the output length matches the original video exactly. The upsampled tokens are passed into a stack of Perceiver-style cross-attention blocks, where a set of learnable spatial queries attend to the projected video tokens as key-value pairs. Both queries and keys carry fixed 2D sinusoidal positional embeddings, enabling explicit spatial alignment without any resolution-specific learned parameters. Each frame is processed independently through the stack, confining the attention cost to the two-dimensional spatial grid while temporal modeling is handled by the preceding Conv3D and upsampler. A final LayerNorm aligns the output distribution with that of the DINOv2 tokens.

Token-Level Injection. The adapter output Z is integrated into the DINOv2 backbone at layer ℓ by bypassing the patch embedding and the initial ℓ transformer blocks, utilizing Z as the input tokens for the remaining blocks. Notably, these subsequent DINOv2 transformer blocks are fused into the 4DRM as part of the learnable spatio-temporal ViT. The selection of ℓ requires a careful trade-off: integrating the features too early imposes an excessive optimization burden on the adapter, whereas a late integration sacrifices essential geometric features relied upon by the downstream heads. Therefore, we set ℓ to the shallowest layer where the linear branch yields a reliable initialization. This strategy maximizes the number of retained upper blocks, effectively incorporating DINOv2's geometric knowledge into the 4DRM.

We train the student 4DRM and the domain adapter with both distillation losses and end-to-end rendering supervisions. The full training procedure is detailed in Sec. 3.4.

3.4 Training Strategy

We adopt a multi-stage pipeline that first builds a strong reconstructive backbone and then bridges the generative video latent space through generative distillation and end-to-end rendering refinement.

Stage 1: Reconstruction Model Training. The 4DRM learns to map monocular video to 4D Gaussian primitives through large-scale training on posed videos. From each training clip of T frames, we select K frames as network input and predict a unified set of 4D Gaussians. These Gaussians

are rendered back to all T frames via the differentiable rasterizer, yielding dense photometric supervision that extends well beyond the temporal coverage of the input. Our objective consists of three complementary terms that respectively enforce appearance fidelity, geometric grounding, and dynamic disentanglement.

Appearance fidelity. We supervise rendered frames $\tilde{\mathbf{I}}_t$ against ground-truth observations $\mathbf{I}_t$ with a composite loss that balances pixel accuracy and perceptual quality:

$$\mathcal{L}_{\text{app}} = \frac{1}{T}\sum_{t=1}^{T}\left(\|\mathbf{I}_t - \tilde{\mathbf{I}}_t\|_2^2 + \alpha \cdot \text{LPIPS}(\mathbf{I}_t, \tilde{\mathbf{I}}_t) + \beta \cdot \left(1 - \text{SSIM}(\mathbf{I}_t, \tilde{\mathbf{I}}_t)\right)\right). \tag{3}$$

The perceptual and structural terms prevent convergence to overly smooth solutions and encourage sharp, realistic textures in the rendered output.

Geometric grounding. Monocular video alone cannot resolve metric depth ambiguity. We therefore supervise predicted geometry directly with ground-truth depth and normal annotations from our training dataset. Specifically, we render per-frame depth $\tilde{\mathbf{d}}_t$ and normals $\tilde{\mathbf{n}}_t$ from the predicted 4DGS and minimize their discrepancy with the annotated counterparts $\mathbf{d}_t^*$ and $\mathbf{n}_t^*$:

$$\mathcal{L}_{\text{geo}} = \frac{1}{T}\sum_{t=1}^{T}\left(\mu\|\mathbf{d}_t^* - \tilde{\mathbf{d}}_t\|_2^2 + \nu\|\mathbf{n}_t^* - \tilde{\mathbf{n}}_t\|_2^2\right). \tag{4}$$

Unlike prior methods that rely on pseudo labels from monocular expert models, training against metric ground truth provides an unambiguous supervisory signal and anchors all predictions in a consistent world-scale coordinate frame.

Dynamic disentanglement. Without multi-view constraints, the network may collapse to a degenerate solution where every primitive is assigned a short lifespan, leading the representation to per-frame point clouds. To encourage proper separation of static structure from dynamic content, we penalize unnecessary motion and favor long-lived primitives:

$$\mathcal{L}_{\text{dyn}} = \frac{1}{M}\sum_{i=1}^{M}\left(\eta_1\|\mathbf{v}_i\|_1 + \eta_2\|\boldsymbol{\omega}_i\|_1 + \eta_3\, l_i^{-1}\right), \tag{5}$$

where M is the total number of predicted Gaussians. These penalties ensure that non-trivial velocity and short lifespans are assigned only to genuinely dynamic elements, enabling the model to learn meaningful motion decomposition without explicit motion labels.

The complete Stage 1 objective combines all three components:

$$\mathcal{L}_{\text{s1}} = \mathcal{L}_{\text{app}} + \mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{dyn}}. \tag{6}$$

Stage 2: Generative Distillation. In the second stage, we connect the generative video latent space to the reconstructive 4D space through end-to-end distillation. We freeze the pretrained 4DRM as a teacher and train the domain adapter alongside a student 4DRM. The teacher takes ground-truth RGB frames as input, while the domain adapter and student 4DRM operate solely on video latents. The training objective drives the student to reproduce the teacher's 4DGS world reconstruction in a generative manner from the video latents. The overall distillation loss combines three complementary terms:

$$\mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{4DGS}} + \mathcal{L}_{\text{geom}} + \lambda(t)\,\mathcal{L}_{\text{feat}}. \tag{7}$$

4D Gaussian attribute alignment. $\mathcal{L}_{\text{4DGS}}$ applies a weighted L1 loss between the per-primitive attributes predicted by the teacher and the student, covering spatial position, anisotropic scale, rotation, opacity, color, velocity, and lifespan. This term provides the most direct form of knowledge transfer by enforcing primitive-level correspondence in attribute space.

Dense geometric alignment. Attribute-space supervision alone is insufficient since distinct sets of Gaussian attributes can produce visually equivalent renderings, while small attribute deviations may cause large visual discrepancies. To address this ambiguity, $\mathcal{L}_{\text{geom}}$ supervises the per-view dense geometric predictions from the student 4DRM heads, combining a robust log-depth regression loss, a multi-scale depth gradient term for boundary sharpness, and an L1 penalty on predicted camera parameters. We apply a global scale normalization before computing this loss to eliminate the trivial scale ambiguity between the two models.

Feature alignment. $\mathcal{L}_{\text{feat}}$ regularizes the domain adapter by aligning the student's intermediate features at the bridge output with the corresponding teacher activations through an L1 loss. Its weight $\lambda(t)$ follows a cosine decay schedule that imposes strong regularization in early training to prevent catastrophic drift of the adapter and gradually relaxes to let the task-specific losses dominate.

Together, these objectives supervise the distillation process at complementary granularities. The attribute loss effectively instills the teacher's 4D reconstruction capabilities into the student, while the geometric and feature losses simultaneously drive domain adaptation by forcing the adapter to project the video latent space onto the reconstructive 4D latent space.

Stage 3: End-to-End Rendering Refinement. Distillation losses operate in attribute and feature space and do not directly reflect final rendering quality. We therefore fine-tune the domain adapter and student 4DRM with an end-to-end rendering loss, where pseudo ground-truth frames are obtained by decoding z_{video} through the frozen VAE decoder and compared against the rasterized 4DGS output using the same appearance objective as Stage 1. Gradients flow through the differentiable rasterizer into the full pipeline, so the student is optimized for rendering quality directly rather than only imitating the teacher. A linearly decayed distillation regularizer preserves the geometric structure acquired in Stage 2.

4 Experiments

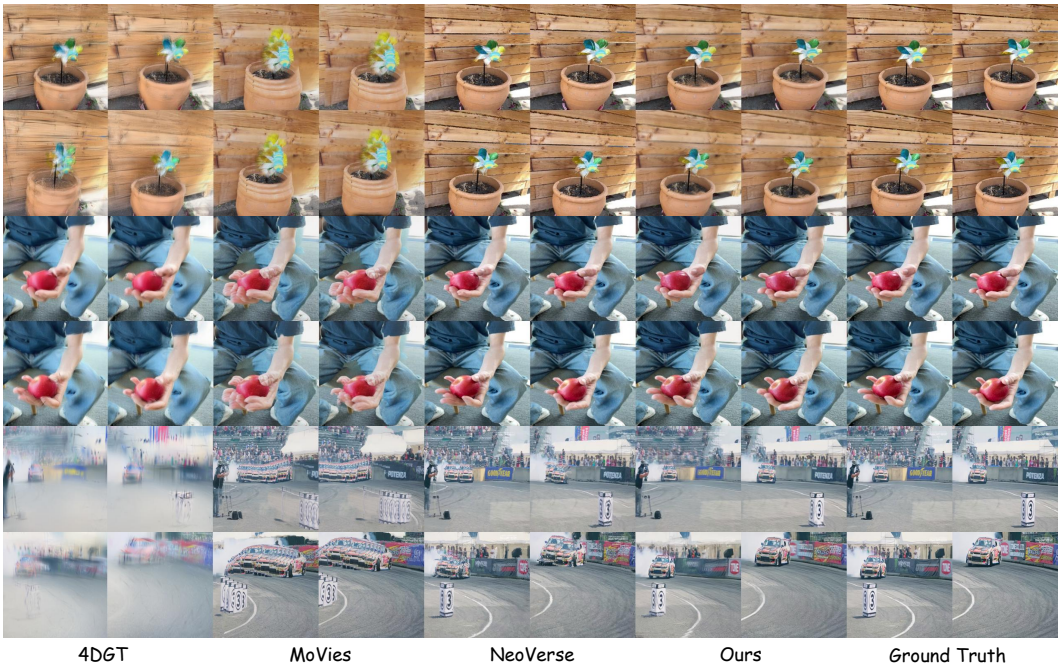


Figure 3: Qualitative reconstruction comparison on Dycheck and DaVis.

4.1 4D Generation

We evaluate AEON against 4DNex [7] and Lyra [3] for 4D generation. Unlike 4DNex (per-frame point clouds) and Lyra (per-frame 3DGS), AEON predicts a unified 4DGS representation for real-time exploration. We evaluate on 30 scenes, each specified by an image-text prompt. Evaluation videos, rendered along shared camera trajectories from identical image-text prompts, are assessed via VBench [13] metrics: I2V Subject, Subject Consistency, Motion Smoothness, and Aesthetic Quality. As shown in Table 2, AEON outperforms all baselines. While per-frame reconstruction lacks cross-frame coherence, AEON models scenes via persistent 4D Gaussians with explicit velocity and duration, naturally ensuring temporal consistency. Additionally, AEON generalizes across backbones (Wan 2.1 and Wan 2.2) without architectural modifications, yielding consistently strong results.

Table 2: Quantitative comparison on 4D generation. All outputs are rendered along shared preset camera trajectories and scored with VBench [13] metrics.

Method	**I2V_Subject ↑**	**Subject_Consistency ↑**	**Motion_Smoothness ↑**	**Aesthetic_Quality ↑**
4DNex [7]	0.796	0.790	0.968	0.342
Lyra [3]	0.865	0.864	0.974	0.459
AEON (Wan 2.1)	0.912	0.864	0.976	0.519
AEON (Wan 2.2)	**0.913**	**0.875**	**0.980**	**0.531**

4.2 Video-to-4D Reconstruction

We evaluate AEON for 4D reconstruction on Dycheck [9] and DaVis [31] against SOTA baselines: L4GM [32], 4DGT [49], MoVies [22], NeoVerse [51], and StreamSplat [47]. Unlike 4DGT and MoVies, which rely on external poses (from GT or π^3 [43]), AEON's integrated camera head jointly predicts camera parameters and 4DGS attributes from monocular RGB. StreamSplat additionally requires monocular depth from Depth Anything V2 [50], which AEON does not need. For Dycheck's novel view synthesis, we assess our video-to-4DGS pipeline by encoding input video into the VAE latent space and decoding it via the domain adapter and student reconstruction model.

As summarized in Table 3, AEON achieves superior PSNR, SSIM, and LPIPS across both benchmarks. Despite operating through the video VAE latent bottleneck, AEON's domain adapter effectively preserves the geometric details necessary for high-fidelity reconstruction, with especially pronounced gains on complex DaVis sequences. Furthermore, AEON needs neither camera poses nor depth at inference, and only Stage 1 uses annotated videos, so the later training stages can scale to raw videos. Qualitative comparisons in Fig. 3 further corroborate these quantitative trends.

Table 3: Video-to-4D reconstruction comparison.

Method	**Dycheck**			**DaVis**		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
L4GM [32]	8.673	0.354	0.652	20.589	0.667	0.311
4DGT [49]	12.267	0.413	0.542	20.362	0.667	0.313
MoVies [22]	11.840	0.341	0.545	20.736	0.688	0.307
NeoVerse [51]	14.269	0.363	0.501	23.929	0.735	0.283
StreamSplat [47]	11.876	0.375	0.532	23.748	0.712	0.273
AEON	**14.487**	**0.421**	**0.469**	**24.661**	**0.771**	**0.243**
Teacher 4DRM	*14.413*	*0.418*	*0.462*	*24.886*	*0.775*	*0.241*

4.3 Novel Trajectory Rendering

We collect 30 in-the-wild videos unseen during training and render each scene along several novel camera trajectories. We compare against TrajectoryCrafter [55], RecamMaster [4], and NeoVerse [51], reporting six VBench [13] metrics in Table 4. AEON achieves the best scores on four out of six metrics while remaining competitive on the other two. Since rendering a new trajectory from 4DGS is geometrically exact rasterization rather than learned hallucination, AEON naturally maintains strong consistency and visual quality across viewpoint changes.

Table 4: Novel trajectory rendering comparison under VBench [13] metrics.

Method	Subj. Consist. ↑	Back. Consist. ↑	Temp. Flick. ↑	Motion Smooth. ↑	Aesth. Quality ↑	Imag. Quality ↑
TrajectoryCrafter [55]	94.48	96.57	95.49	**98.79**	64.81	58.53
RecamMaster [4]	94.49	95.58	**97.64**	97.79	61.26	56.34
NeoVerse [51]	90.24	95.21	95.01	97.39	59.17	57.98
AEON	**94.64**	**96.77**	96.69	98.11	**69.02**	**58.65**

4.4 Ablation Studies

We ablate three key design choices of AEON under the Dycheck and DaVis benchmarks in Table 5. Conv3D projection replaces the full dual-path domain adapter with a single Conv3D layer, showing that the expressive Perceiver Bridge branch is

Table 5: Ablation study on Dycheck and DaVis.

Variant	Dycheck			DaVis		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Conv3D projection	12.086	0.314	0.606	12.813	0.276	0.701
W/o Distill	12.390	0.349	0.537	16.718	0.503	0.489
W/o Duration	12.500	0.372	0.535	22.057	0.718	0.328
Full	**14.487**	**0.421**	**0.469**	**24.661**	**0.771**	**0.243**

essential for capturing the complex mapping between the two latent spaces. *W/o Distill* removes the distillation losses and trains the adapter with rendering supervision alone, resulting in a significant performance drop, especially on DaVis, which highlights the importance of the teacher 4DRM as a geometric foundation. *W/o Duration* disables the temporal opacity evolution of 4DGS, forcing all primitives to remain visible throughout the sequence; the degradation shows that explicit duration modeling of 4DGS representation is critical for handling transient dynamics. The full model combines all components and achieves the best performance on every metric.

5 Conclusion

We present AEON, a unified world modeling framework that bridges video-based and 3D-based paradigms and combines the benefits of both. The core of AEON is a generative model that operates within the latent space of a pretrained video diffusion model while producing explicit 4D Gaussian Splatting as the world representation. We train a 4DRM for robust geometric guidance and design a domain adapter that translates generative video latents into the reconstructive 4D latents through end-to-end distillation. Experiments across 4D generation, dynamic-scene reconstruction, and novel trajectory rendering demonstrate that AEON achieves state-of-the-art performance while supporting a broad spectrum of world modeling applications within a single unified framework.

Limitations and future work. The temporal extent of generated 4D scenes is currently bounded by the video diffusion model's sequence length. Because AEON stores the world as a persistent 4DGS rather than as frames, it offers a natural path beyond this window: render conditioning frames from the current 4DGS at the latest camera pose, generate the next window, decode it into new primitives, and fuse them with the existing world where they overlap. Since each step is anchored to explicit geometry rather than extrapolated from previous frames, drift should accumulate more slowly than in video-only autoregression. In addition, regions never observed in the generated video cannot be recovered, so very large viewpoint changes may reveal holes that a generative completion module could fill.

Acknowledgments and Disclosure of Funding

This work was partially supported by Deep Earth Probe and Mineral Resources Exploration – National Science and Technology Major Project (2024ZD1003405) and the National Natural Science Foundation of China (62272263).

References

[1] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[3] Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B. Lindell, Zan Gojcic, Sanja Fidler, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra: Generative 3D scene reconstruction via video diffusion model self-distillation. In *International Conference on Learning Representations (ICLR)*, 2026.

[4] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[5] Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning (ICML)*, 2024.

[6] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey

Tulyakov. Panda-70M: Captioning 70M videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

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

[8] Juntong Fang, Zequn Chen, Weiqi Zhang, Donglin Di, Xuancheng Zhang, Chengmin Yang, and Yu-Shen Liu. MoRe: Motion-aware feed-forward 4D reconstruction transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[9] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[10] Hyojun Go, Dominik Narnhofer, Goutam Bhat, Prune Truong, Federico Tombari, and Konrad Schindler. Text-to-3D by stitching a multi-view reconstruction network to a video generator. In *International Conference on Learning Representations (ICLR)*, 2026.

[11] Google DeepMind. Veo 2. `https://deepmind.google/technologies/veo/veo-2/`, 2024.

[12] Yuan-Ting Hu, Jiahong Wang, Raymond A. Yeh, and Alexander G. Schwing. SAIL-VOS 3D: A synthetic dataset and baselines for object detection and 3D mesh reconstruction from video data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[13] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[14] Team HY-World, Chenjie Cao, Xuhui Zuo, Zhenwei Wang, Yisu Zhang, Junta Wu, Zhenyang Liu, Yuning Gong, Yang Liu, Bo Yuan, et al. HY-World 2.0: A multi-modal world model for reconstructing, generating, and simulating 3D worlds. *arXiv preprint arXiv:2604.14268*, 2026.

[15] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning (ICML)*, 2021.

[16] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. DynamicStereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[17] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision (3DV)*, 2026.

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4):139:1–139:14, 2023.

[19] Kuaishou. Kling AI. `https://klingai.kuaishou.com/`, 2024.

[20] Jiahui Lei, Yijia Weng, Adam W. Harley, Leonidas Guibas, and Kostas Daniilidis. MoSca: Dynamic Gaussian fusion from casual videos via 4D motion scaffolds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[21] Hanxue Liang, Jiawei Ren, Ashkan Mirzaei, Antonio Torralba, Ziwei Liu, Igor Gilitschenski, Sanja Fidler, Cengiz Oztireli, Huan Ling, Zan Gojcic, and Jiahui Huang. Feed-forward bullet-time reconstruction of dynamic scenes from monocular videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[22] Chenguo Lin, Yuchen Lin, Panwang Pan, Yifan Yu, Tao Hu, Honglei Yan, Katerina Fragkiadaki, and Yadong Mu. MoVieS: Motion-aware 4D dynamic view synthesis in one second. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[23] Zihao Liu, Xiaolong Shen, Zhenglin Zhou, Ruijie Quan, and Yi Yang. Beyond Pixels: From video priors to 4D worlds. *arXiv preprint arXiv:2608.10744*, 2026.

[24] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[25] Takeshi Noda, Chao Chen, Weiqi Zhang, Xinhai Liu, Yu-Shen Liu, and Zhizhong Han. MultiP-ull: Detailing signed distance functions by pulling multi-level queries at multi-step. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[26] Takeshi Noda, Chao Chen, Junsheng Zhou, Weiqi Zhang, Yu-Shen Liu, and Zhizhong Han. Learning bijective surface parameterization for inferring signed distance functions from sparse point clouds with grid deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[27] OpenAI. Video generation models as world simulators. `https://openai.com/index/video-generation-models-as-world-simulators/`, 2024.

[28] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud 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. *Transactions on Machine Learning Research (TMLR)*, 2024.

[29] Panwang Pan, Chenguo Lin, Jingjing Zhao, Chenxin Li, Yuchen Lin, Haopeng Li, Honglei Yan, Kairun Wen, Yunlong Lin, Yixuan Yuan, et al. Diff4Splat: Controllable 4D scene generation with latent dynamic reconstruction models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[30] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[31] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Hanxue Liang, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, and Huan Ling. L4GM: Large 4D Gaussian reconstruction model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[33] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[34] Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, et al. Lyra 2.0: Explorable generative 3D worlds. In *SIGGRAPH Asia*, 2026.

[35] Yang-Tian Sun, Yi-Hua Huang, Lin Ma, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Splatter a Video: Video Gaussian representation for versatile processing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[36] HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, et al. HunyuanWorld 1.0: Generating immersive, explorable, and interactive 3D worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025.

[37] InSpatio Team, Donghui Shen, Guofeng Zhang, Haomin Liu, Haoyu Ji, Hujun Bao, Hongjia Zhai, Jialin Liu, Jing Guo, Nan Wang, et al. INSPATIO-WORLD: A real-time 4D world simulator via spatiotemporal autoregressive modeling. *arXiv preprint arXiv:2604.07209*, 2026.

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

[39] 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.

[40] Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of Motion: 4D reconstruction from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[41] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A dataset to push the limits of visual SLAM. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[42] Yifan Wang, Peishan Yang, Zhen Xu, Jiaming Sun, Zhanhua Zhang, Yong Chen, Hujun Bao, Sida Peng, and Xiaowei Zhou. FreeTimeGS: Free Gaussian primitives at anytime anywhere for dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[43] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations (ICLR)*, 2026.

[44] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4):600–612, 2004.

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

[46] World Labs. Marble: A multimodal world model. `https://www.worldlabs.ai/blog/marble-world-model`, 2025.

[47] Zike Wu, Qi Yan, Xuanyu Yi, Lele Wang, and Renjie Liao. StreamSplat: Towards online dynamic 3D reconstruction from uncalibrated video streams. In *International Conference on Learning Representations (ICLR)*, 2026.

[48] Peng Xiang, Liang Han, Hui Zhang, Yu-Shen Liu, and Zhizhong Han. VGGS: VGGT-guided Gaussian splatting for efficient and faithful sparse-view surface reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

[49] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[50] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[51] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[52] Zhongqi Yang, Wenhang Ge, Yuqi Li, Jiaqi Chen, Haoyuan Li, Mengyin An, Fei Kang, Hua Xue, Baixin Xu, Yuyang Yin, et al. Matrix-3D: Omnidirectional explorable 3D world generation. *arXiv preprint arXiv:2508.08086*, 2025.

[53] 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. In *International Conference on Learning Representations (ICLR)*, 2025.

[54] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4D Gaussian splatting. In *International Conference on Learning Representations (ICLR)*, 2024.

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

[56] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[57] Wenyuan Zhang, Jimin Tang, Weiqi Zhang, Yi Fang, Yu-Shen Liu, and Zhizhong Han. MaterialRefGS: Reflective Gaussian splatting with multi-view consistent material inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[58] Weiqi Zhang, Junsheng Zhou, Juntong Fang, Kanle Shi, Shenkun Xu, and Yu-Shen Liu. TESLA: Native 4D Gaussian splatting generation with temporally structured latents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

[59] Weiqi Zhang, Junsheng Zhou, Haotian Geng, Kanle Shi, Shenkun Xu, Yi Fang, and Yu-Shen Liu. GaussianGrow: Geometry-aware Gaussian growing from 3D point clouds with text guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[60] Weiqi Zhang, Junsheng Zhou, Haotian Geng, Wenyuan Zhang, and Yu-Shen Liu. GAP: Gaussianize any point clouds with text guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[61] Weiqi Zhang, Junsheng Zhou, Xuancheng Zhang, Juntong Fang, Zequn Chen, Donglin Di, and Yu-Shen Liu. GAMMA: Scalable 4D Gaussian reconstruction model for novel view synthesis of monocular videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

[62] Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. PointOdyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[63] Junsheng Zhou, Weiqi Zhang, and Yu-Shen Liu. DiffGS: Functional Gaussian splatting diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[64] Junsheng Zhou, Weiqi Zhang, Baorui Ma, Kanle Shi, Yu-Shen Liu, and Zhizhong Han. UDiFF: Generating conditional unsigned distance fields with optimal wavelet diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[65] Junsheng Zhou, Weiqi Zhang, Baorui Ma, Kanle Shi, Yu-Shen Liu, and Zhizhong Han. UDFStudio: A unified framework of datasets, benchmarks and generative models for unsigned distance functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2026.

[66] Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, Mingyu Liu, Dingning Liu, Jiange Yang, Zhoujie Fu, Junyi Chen, Chunhua Shen, Jiangmiao Pang, Kaipeng Zhang, and Tong He. OmniWorld: A multi-domain and multi-modal dataset for 4D world modeling. In *International Conference on Learning Representations (ICLR)*, 2026.

A Implementation Details

The 4DRM encoder is built on DINOv2-Large [28] with 24 transformer blocks of dimension 1024. The 4DRM is initialized from the pretrained MapAnything weights [17]. The domain adapter is injected at layer ℓ=4, retaining the upper 20 pretrained transformer blocks as a geometric foundation for the 4DRM. The Linear branch projects the 16-channel Wan video latent to 1024 channels, initialized via closed-form ridge regression against the fourth-layer DINOv2 activations. The Perceiver Bridge employs 4 cross-attention blocks with 16 heads and 1024 learnable queries, and a temporal interleave of factor 4 inverts the video VAE's temporal compression. All three stages are optimized with AdamW using a weight decay of 10^{-4}, a cosine schedule with 2 000 linear warmup steps, a base learning rate of 5×10^{-6}, and gradient clipping at a global norm of 1.0.

Stage 1 trains the full 4DRM for 20 epochs on video clips with ground-truth camera poses and depth annotations. The appearance loss weights are α=2.0 and β=0.2, the geometry weights are μ=1.0 and ν=0.5, and the dynamic regularization coefficients are η_1=η_2=η_3=0.3. Stage 2 freezes the Stage-1 model as the teacher and trains the domain adapter and the student 4DRM for 30 additional epochs with a $10\times$ learning rate multiplier on the adapter parameters. Since the frozen teacher provides all supervisory signals through distillation, Stage 2 requires only raw RGB videos as input, removing the dependency on pose and depth annotations and enabling training on large-scale unannotated data. The feature alignment weight λ decays from 1.0 to 0 following a cosine schedule. All experiments are conducted on 16 NVIDIA H200 GPUs. The Stage-1 training corpus combines DynamicReplica [16], SAILVOS3D [12], Spring [24], TartanAir [41], PointOdyssey [62], and OmniWorld [66], each providing ground-truth camera poses and depth. Because these datasets differ in size by orders of magnitude, we resample the smaller ones by the factors listed in Table 6, so that every source contributes to geometric supervision. In Stage 2, we further incorporate Panda-70M [6] to expand the training distribution beyond annotated domains; it supplies real-world appearance diversity without geometric labels. Stage 3 continues to train the domain adapter and student 4DRM for 10 additional epochs with the end-to-end rendering loss, using the same optimizer configuration as Stage 2. The distillation regularization weight λ_{reg} is initialized to 0.5 and linearly decayed to zero over the course of training.

Table 6: Training datasets. Resample is the factor by which each dataset is repeated during training.

Dataset	Dynamic	Pose	Depth	Real	#Scenes	Resample
PointOdyssey [62]	✓	✓	✓	✗	0.1K	100×
TartanAir [41]	✓	✓	✓	✗	0.5K	50×
DynamicReplica [16]	✓	✓	✓	✗	0.5K	20×
OmniWorld-Game [66]	✓	✓	✓	✗	0.06K	100×
SAIL-VOS 3D [12]	✓	✓	✓	✗	0.16K	50×
Spring [24]	✓	✓	✓	✗	0.03K	1000×
Panda-70M [6] (subset)	✓	✗	✗	✓	2.4M	1×

B Applications

Text/Image-to-world generation. As illustrated in Fig. 1, AEON synthesizes high-quality 4D Gaussian Splatting (4DGS) environments from text or image prompts. This is achieved by first generating video latents using Wan-DiT conditioned on the input prompts. These latents are subsequently translated into 4D spatial-temporal latents and decoded to reconstruct the final 4DGS world.

Video-to-world reconstruction. For video-to-world reconstruction, input videos are encoded into video latents using Wan-VAE. These latents are then processed to generate 4DGS scenes, following the same pipeline as our text/image-to-world generation.

Novel view synthesis. The generated 4DGS representations inherently support real-time novel view synthesis across arbitrary viewpoints and timesteps.

Physical status estimation. The generated 4DGS encapsulates rich geometric and physical properties, which can be directly extracted to derive depth maps, normal maps, motion masks, and optical

flow. Specifically, motion masks are obtained by filtering the lifespans of the Gaussians, as dynamic elements typically correspond to Gaussians with shorter lifespans.

3D world generation. Beyond dynamic scene generation, AEON is also highly effective for static 3D world generation. An example of this capability is provided in the third row of Fig. 1.

Scene editing and simulation. Because AEON generates explicit 4D representations, it inherently facilitates seamless scene editing and structural decomposition for downstream simulations (shown in the bottom row of Fig. 1). By leveraging the motion masks derived from Gaussian lifespan filtering, we can cleanly separate a generated scene into its static and dynamic components. These isolated parts from various generated 4DGS worlds can then be freely recomposed to simulate novel scenarios. The same explicit structure is what embodied simulation needs. The metric geometry of the static component can supply occupancy and free space for an agent, the dynamic primitives move according to their own velocities, and real-time rasterization renders egocentric observations at every step. An agent can therefore be placed in a generated world, roll out its actions, and receive rendered observations in a closed loop, and because the world is persistent, the same scenario can be replayed for repeatable evaluation.

C Evolution of 4DGS Representation

Temporal opacity evolution. To faithfully represent the appearance and disappearance of scene elements over time, the opacity of each primitive is modulated by a Gaussian kernel centered at its reference timestep t_i. Given a query time t, the effective opacity is computed as

$$\sigma_i = \sqrt{\frac{(l_i/2)^2}{-2\log(\epsilon)}}, \qquad o_i(t) = o_i \cdot \exp\left(-\frac{(t-t_i)^2}{2\sigma_i^2}\right), \tag{8}$$

where $\epsilon = 0.05$ serves as the activation threshold at the duration boundary. This temporal gating mechanism ensures full visibility near the reference timestep while producing smooth fade-in and fade-out transitions. As a result, the representation can naturally accommodate transient phenomena such as fire, splashes, and object entrances without resorting to explicit per-object tracking.

Velocity-driven spatial evolution. The position and orientation of each primitive evolve over time according to a constant-velocity motion model. At query time t, the updated position and rotation are given by

$$\mathbf{x}_i(t) = \mathbf{x}_i + \mathbf{v}_i \cdot (t - t_i), \tag{9}$$

$$\mathbf{r}_i(t) = \mathbf{r}_i \otimes \mathrm{AxisAngle2Quat}(\boldsymbol{\omega}_i \cdot (t - t_i)), \tag{10}$$

where $\otimes$ denotes quaternion multiplication. The scaling $\mathbf{s}_i$ remains fixed across time, ensuring that each primitive preserves its intrinsic shape throughout the sequence. Factoring translational and rotational dynamics into independent velocity terms yields a compact yet expressive motion parameterization that supports rigid-body motion while incurring minimal per-primitive overhead. Combined with the temporal opacity mechanism above, the full 4DGS representation enables AEON to unify the photorealistic rendering capability of video models with the geometric consistency and persistence of 3D models.

D Evaluation Metrics

We describe all evaluation metrics used in the main paper. Our experiments rely on two families of metrics: per-frame reconstruction metrics that measure pixel-level fidelity and VBench metrics that assess perceptual video quality along multiple complementary dimensions.

Reconstruction metrics. We evaluate dynamic scene reconstruction with three standard metrics that operate at increasing levels of perceptual abstraction. Peak Signal-to-Noise Ratio (PSNR) measures per-pixel reconstruction accuracy in decibels:

$$\mathrm{PSNR} = 10 \cdot \log_{10} \frac{R^2}{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{I}_i - \tilde{\mathbf{I}}_i)^2}, \tag{11}$$

where R is the maximum pixel intensity, N is the total number of pixels, and $\mathbf{I}$, $\tilde{\mathbf{I}}$ denote the ground-truth and rendered frames respectively. Higher values indicate less distortion. Structural Similarity Index (SSIM) [44] moves beyond pixel-level comparison by jointly evaluating local luminance, contrast, and structural patterns:

$$\text{SSIM}(\mathbf{I}, \tilde{\mathbf{I}}) = \frac{(2\mu_{\mathbf{I}}\mu_{\tilde{\mathbf{I}}} + C_1)(2\sigma_{\mathbf{I}\tilde{\mathbf{I}}} + C_2)}{(\mu_{\mathbf{I}}^2 + \mu_{\tilde{\mathbf{I}}}^2 + C_1)(\sigma_{\mathbf{I}}^2 + \sigma_{\tilde{\mathbf{I}}}^2 + C_2)}, \tag{12}$$

where μ and σ denote the local mean and standard deviation computed over sliding windows, $\sigma_{\mathbf{I}\tilde{\mathbf{I}}}$ is the local cross-covariance, and C_1, C_2 are small stability constants. SSIM ranges from 0 to 1, with higher values indicating greater structural agreement. Learned Perceptual Image Patch Similarity (LPIPS) [56] further captures high-level perceptual differences by comparing deep features extracted from a pretrained network:

$$\text{LPIPS}(\mathbf{I}, \tilde{\mathbf{I}}) = \sum_l w_l \left\| \phi_l(\mathbf{I}) - \phi_l(\tilde{\mathbf{I}}) \right\|_2^2, \tag{13}$$

where ϕ_l denotes the normalized activations at layer l and w_l are learned linear weights. Lower LPIPS values indicate higher perceptual similarity.

VBench video quality metrics. For 4D generation and novel trajectory rendering we adopt six metrics from the VBench [13] benchmark, each targeting a distinct aspect of video quality. **I2V Subject** measures the visual fidelity between the conditioning image and the generated subject, quantifying how faithfully the output preserves the identity of the input. **Subject Consistency** evaluates the temporal coherence of the foreground subject across frames by computing pairwise feature similarity over the video. **Background Consistency** similarly measures the stability of background regions to detect unwanted scene drift. **Temporal Flickering** captures high-frequency luminance and color variations between consecutive frames, where higher scores correspond to smoother transitions. **Motion Smoothness** assesses whether object trajectories follow physically plausible paths without abrupt jumps. **Aesthetic Quality** and **Imaging Quality** respectively evaluate the overall visual appeal and the low-level image attributes such as sharpness and color fidelity. All VBench metrics are reported on a scale where higher values indicate better performance.

E Additional Ablation Studies

Table 7 reports additional ablations on Dycheck and DaVis, where each variant changes one component of the full model.

Table 7: Additional ablation studies on Dycheck and DaVis.

	Variant	Dycheck			DaVis		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Adapter	Conv3D projection	12.086	0.314	0.606	12.813	0.276	0.701
	Conv branch only	12.332	0.333	0.533	15.409	0.465	0.533
	Perceiver branch only	13.086	0.389	0.596	14.262	0.442	0.552
Training	W/o Stage 3	13.670	0.361	0.517	22.953	0.705	0.299
4DRM	W/o geometric prior	12.341	0.358	0.543	22.268	0.695	0.349
	W/o $\mathcal{L}_{\text{geo}}$	12.260	0.331	0.578	20.347	0.655	0.323
Full		**14.487**	**0.421**	**0.469**	**24.661**	**0.771**	**0.243**

Adapter. *Conv3D projection* replaces the dual-path adapter with a single Conv3D projection, while *Conv branch only* and *Perceiver branch only* use only a convolutional branch or only the Perceiver branch, respectively. All three variants lower DaVis PSNR to 12.8–15.4 dB, so no single path can bridge the two latent spaces on its own.

Training. *W/o Stage 3* stops after Stage 2 and keeps Stages 1 and 2 unchanged. Skipping Stage 3 reduces DaVis PSNR by 1.7 dB, demonstrating the benefit of end-to-end rendering refinement.

4DRM. *W/o geometric prior* trains 4DRM without the MapAnything initialization, and *W/o* $\mathcal{L}_{geo}$ removes the depth and normal supervision of Stage 1. Both variants perform worse than the full model on Dycheck and DaVis, so the quality of 4DRM matters, although performance degrades gracefully rather than collapsing.

F Additional Qualitative Results

We provide additional qualitative reconstruction comparisons on the DaVis benchmark in Fig. 4. These results further corroborate that AEON recovers sharper appearance details and more temporally coherent geometry than baselines, particularly in regions with fast motion and transient occlusions.

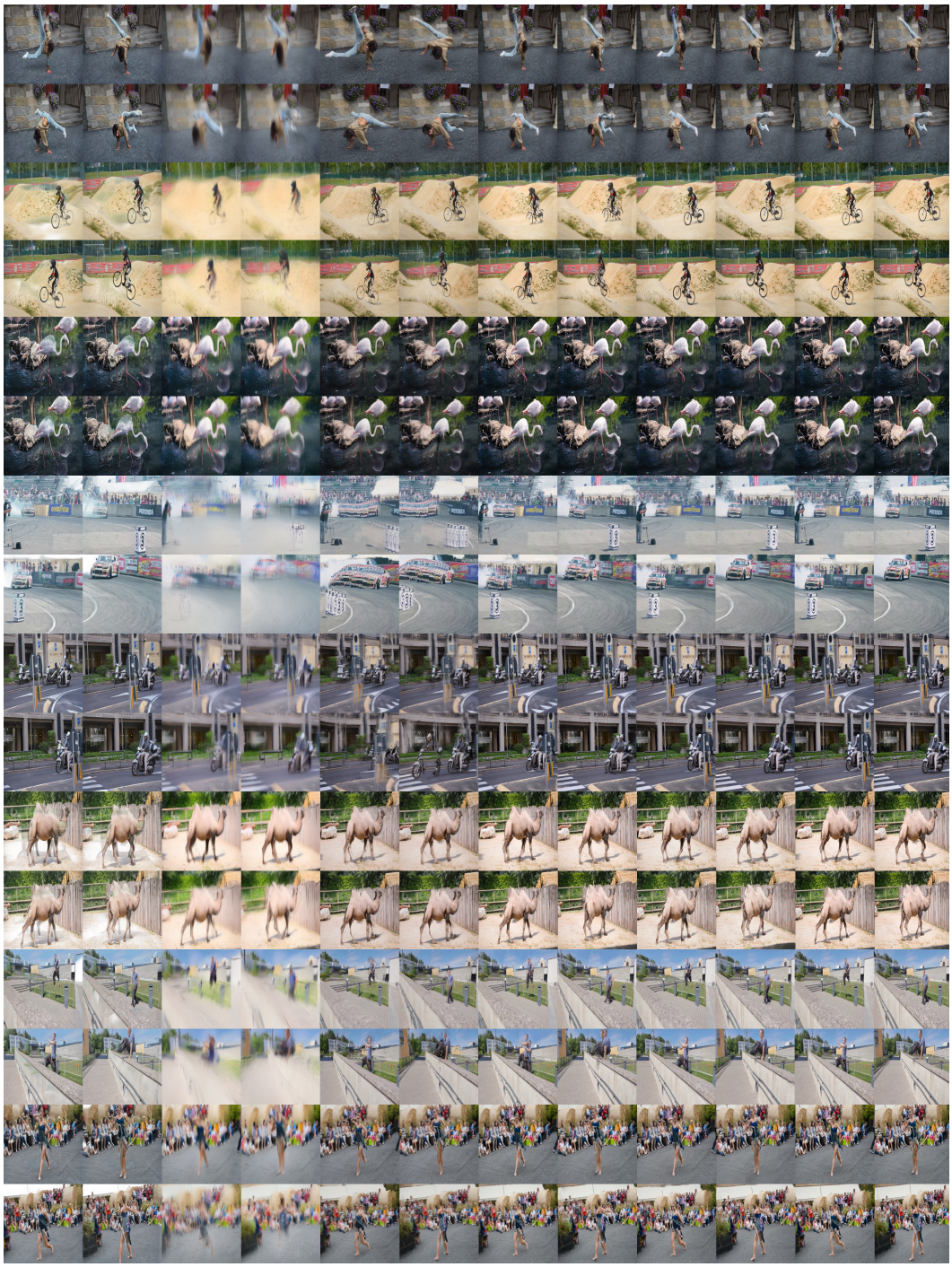


Figure 4: Additional qualitative reconstruction comparisons on DaVis. AEON produces more coherent geometry and finer details in dynamic regions.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three main contributions (unified 4DGS world representation, domain adapter with generative distillation, and state-of-the-art results), all of which are supported by experiments in Sec. 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss limitations in the Conclusion section (Sec. 5), noting that the temporal extent of generated 4D scenes is bounded by the video diffusion model's sequence length.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper does not include theoretical results. All formulas describe model architecture and loss functions rather than theorems requiring proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: Sec. 4 provides full implementation details including model architecture, optimizer, learning rate, loss weights, training epochs, hardware, and datasets for both stages.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Demonstration code is provided as supplementary material. The complete source code, data, and pretrained checkpoints will be publicly released upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 4 specifies the optimizer (AdamW), learning rate, weight decay, warmup schedule, all loss weights, training epochs, datasets, and evaluation benchmarks with their respective metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report the average performance as the experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: Sec. 4 reports that all experiments are conducted on 16 NVIDIA H200 GPUs, along with training epochs for both stages.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics and confirm that this research conforms with it in every respect.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the applications and potential impacts of our method in the introduction.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The proposed method focuses on 4D scene reconstruction and generation for research purposes and does not pose significant risks beyond those of existing video generation models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets (DynamicReplica, SAILVOS3D, Spring, TartanAir, PointOdyssey, OmniWorld, Panda-70M) and pretrained models (DINOv2) used in this work are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method does not involve LLMs as any component. LLMs were not used in the methodology, experiments, or evaluation of this work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.