Beyond Pixels: From Video Priors to 4D Worlds

Zihao Liu[1], **Xiaolong Shen**[1], **Zhenglin Zhou**[1], **Ruijie Quan**[1], **Yi Yang**[1]*

[1]ReLER, CCAI, Zhejiang University

Project Page: https://hayd-zju.github.io/Beyond-Pixels/

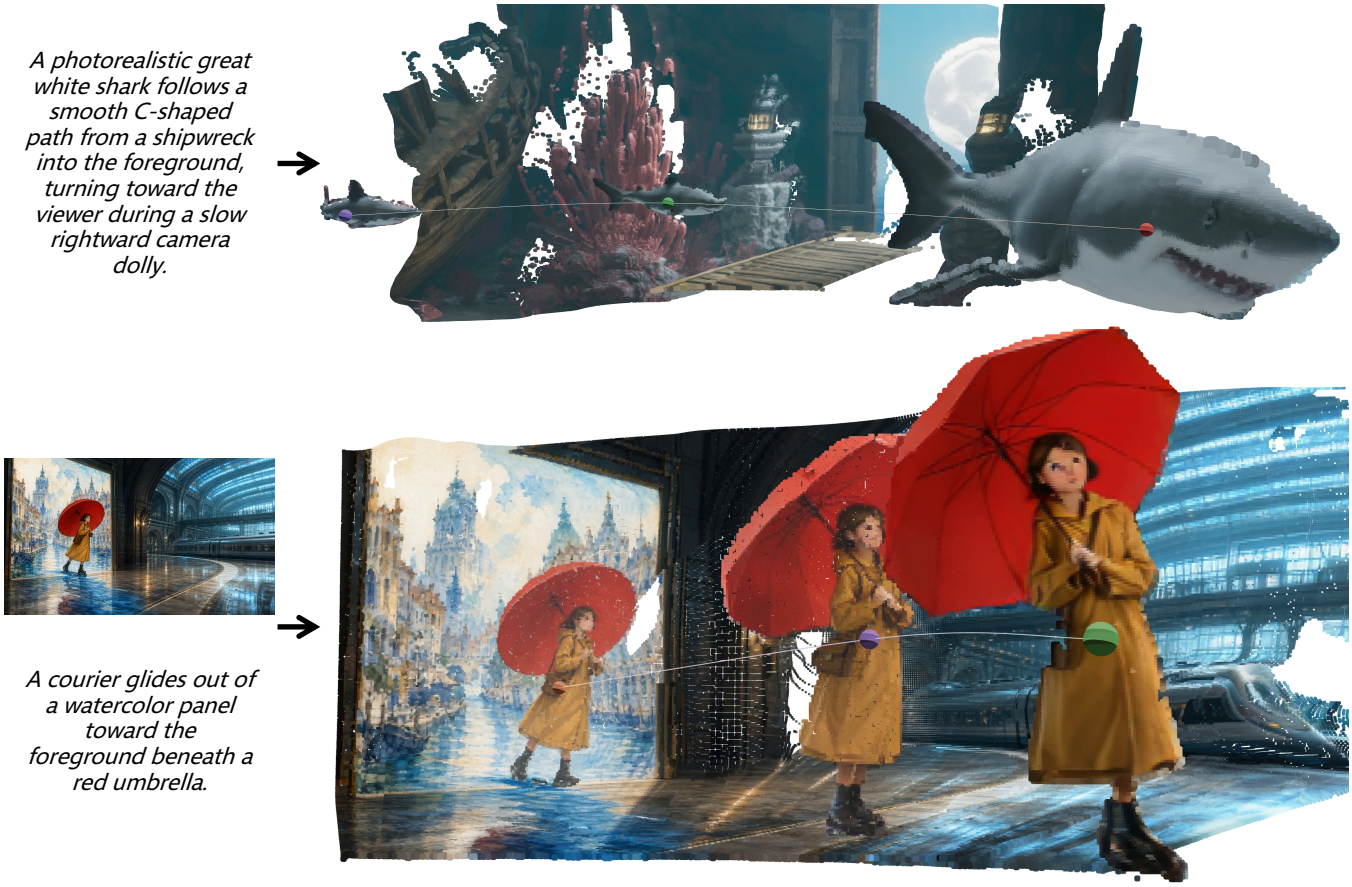

Figure 1: **Latent-to-4D** enables unified text- and image-conditioned 4D generation. It turns video-model latents directly into dynamic geometry and motion through a shared pathway, without task-specific adaptation.

Abstract

4D generation synthesizes dynamic 3D scenes from conditions such as text or images. Existing methods either reconstruct generated RGB videos with a separate 4D model or adapt a particular video generator to predict geometry directly. The former suffers from distribution mismatch and error propagation, whereas the latter ties 4D prediction to a specific generator and may require retraining when the generator or conditioning regime changes. We ask whether the final denoised latents of video models that share a variational autoencoder (VAE) can instead provide a reusable interface to explicit 4D prediction. Building on this insight, we introduce *direct latent-to-4D generation* and instantiate it as **Latent-to-4D**, which bypasses RGB by aligning a video latent with the token grid of a pretrained 4D decoder and refining it through frame-wise and global spatiotemporal attention. Trained on roughly 1K existing reconstruction clips, a single checkpoint transfers unchanged across multiple video diffusion transformers within the same VAE family. On Text4D-200 and I4D-200, Latent-to-4D surpasses matched same-latent Wan+4RC cascades in projection-based DINO-F1 by 2.88–3.45 and 5.81 points, respectively, while also being preferred by human raters for geometry, temporal stability, and overall quality.

Introduction

4D generation aims to synthesize dynamic 3D scenes with explicit geometry and motion from intuitive conditions such as text or images. Such outputs can be viewed and manipulated from novel viewpoints, making them valuable for virtual production, immersive VR/AR, simulation, navigation,

*Corresponding author.

and embodied intelligence. Recent advances in video generation and feed-forward 4D reconstruction have made this goal increasingly practical, motivating approaches that combine their complementary generative and geometric capabilities.

Recent 4D generation methods combine these capabilities through two broad paradigms. The first paradigm is **generate-then-reconstruct** (Liang et al. 2024; Zhang et al. 2024; Wu et al. 2024): They first synthesize multiview or temporal RGB observations, which are then converted into an explicit dynamic scene by a separate reconstruction model. The second is **integrated feed-forward generation** (Pan et al. 2026; Chen et al. 2025; Fang et al. 2025), which makes geometry a native output of the generative process through either direct prediction of explicit 4D representations or joint RGB–geometry video modeling. Collectively, these approaches demonstrate the value of video priors for efficient 4D generation.

Despite this progress, the two paradigms expose a fundamental trade-off in how video priors are transferred to 4D. Generate-then-reconstruct preserves generator modularity, but its RGB interface places a separately trained reconstructor, typically trained on much narrower reconstruction data, between the video prior and the 4D output. Open-domain content and temporal artifacts may therefore propagate into unstable geometry. Integrated feed-forward methods transfer video priors more directly, but typically bind 4D prediction to a particular generator or conditioning regime, so switching to another pretrained video model may require new geometry-supervised training. This trade-off is amplified by the scarcity of 4D supervision relative to large-scale video data. The key challenge is therefore to establish a reusable interface that bypasses generated RGB while allowing one geometry-supervised pathway to serve multiple compatible video generators.

A natural candidate for such an interface is the VAE latent space shared by compatible video generators. Although their DiT backbones and conditioning regimes may differ, models that share the same VAE checkpoint, latent normalization, layout, and compression convention produce final denoised latents in a common representation space. Because this representation lies upstream of RGB decoding, it provides a natural point at which a single geometry-supervised 4D pathway could access the appearance, motion, and conditioning information produced by different compatible DiTs (Fig. 1). This observation motivates our central question:

> *Can final denoised video latents serve as a reusable interface to explicit 4D prediction across compatible DiTs?*

To realize this interface, we propose **Latent-to-4D**, a direct 4D generation framework that connects a video model's final denoised latent to a 4D decoding hierarchy without passing through generated RGB (Fig. 2). Bridging these representations is non-trivial because their spatiotemporal grids and feature spaces are not aligned. Our Latent-to-4D Alignment and Refinement (**L4AR**) network aligns the video latent with the 4D token grid and refines it using frame-wise and

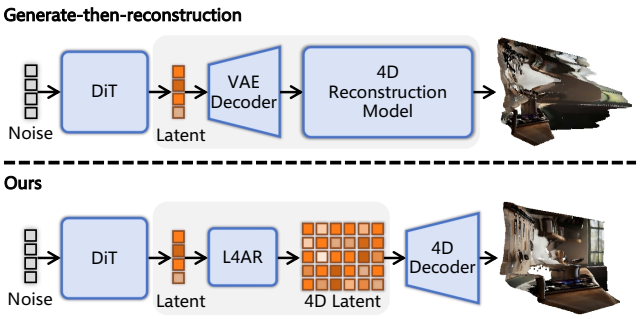

Figure 2: **Comparison of video-to-4D interfaces.** Previous methods decode video latents to RGB before reconstruction, whereas ours maps them directly to 4D through L4AR.

global spatiotemporal context, after which a decoder initialized from a pretrained reconstructor predicts cameras and dynamic world-space geometry. During training, a frozen VAE encodes roughly 1K existing reconstruction clips, and only the alignment module, lightweight refinement updates, and prediction heads are optimized from their 4D annotations. The video generators, VAE, and original Transformer weights remain frozen. At inference, the observed-video latent is replaced by the final denoised latent of a compatible DiT in the same VAE space, allowing the trained pathway to produce 4D outputs without further tuning.

We evaluate Latent-to-4D on **Text4D-200** and **I4D-200**, two generated-latent evaluation suites for text- and image-conditioned 4D generation. A single checkpoint operates unchanged across two text-to-video DiTs and one image-to-video DiT sharing the same VAE, while remaining compatible with their upstream controls. In controlled same-latent comparisons, Latent-to-4D surpasses matched Wan+4RC cascades in projection-based DINO-F1 by 2.88–3.45 points on Text4D-200 and 5.81 points on I4D-200, and is preferred in multi-view human evaluation for geometric plausibility, completeness, and temporal stability. These results support shared video latents as an effective and reusable interface within the evaluated common-VAE setting.

Our contributions are threefold:

- Conceptually, we formulate direct latent-to-4D generation by treating final denoised VAE latents as a reusable interface between compatible video generators and explicit 4D prediction. This formulation bypasses generated RGB and decouples 4D supervision from a particular generator or conditioning regime.
- Technically, we introduce **Latent-to-4D**, whose L4AR network aligns and refines video latents for structured 4D decoding. Trained on roughly 1K existing reconstruction clips, a single checkpoint transfers unchanged across multiple video generators sharing the same VAE.
- Empirically, we evaluate a single Latent-to-4D checkpoint across multiple video generators sharing the same VAE on Text4D-200 and I4D-200. Latent-to-4D achieves higher projection-based DINO-F1 than matched generate-then-reconstruct baselines, while human evaluation favors its outputs for geometric plausibility, completeness, and temporal stability.

Related Work

Video Generation. Modern video generators learn rich appearance and motion distributions and support text, image, camera, trajectory, and animation controls (Chen et al. 2024; Yang et al. 2025; Kong et al. 2025; Wan et al. 2025; Blattmann et al. 2023; Zhang et al. 2023; Xing et al. 2023; Wang et al. 2024; He et al. 2025; Bai et al. 2025; Chu et al. 2025; Cheng et al. 2025). Their representations have also been repurposed for depth, point maps, and dynamic geometry (Ke et al. 2024; Hu et al. 2024; Xu et al. 2025; Jiang et al. 2025; Mai et al. 2025; Zhu et al. 2026; Wang et al. 2026a), with VIST3A connecting a video VAE to a static 3D reconstructor (Go et al. 2026). These approaches typically convert a generator into a task-specific perception model or align one selected model pair. We instead keep compatible conditional DiTs and their shared VAE frozen and learn one interface from the final denoised VAE latent to a dynamic 4D decoder, preserving their generation and control capabilities.

4D Reconstruction. Feed-forward 4D reconstructors provide reusable geometric reasoning by predicting dynamic point maps, scene flow, trajectories, or dense motion from RGB videos (Feng et al. 2025; Sucar et al. 2025; Liu et al. 2025; Karhade et al. 2025; Sucar et al. 2026). Query-based models such as D4RT and 4RC recover cameras and geometry across viewpoint and time, while related models such as π^3 estimate cameras and dense point maps from image sets (Zhang et al. 2025; Luo et al. 2026; Wang et al. 2026b). Their pipelines nevertheless begin with RGB encoders trained on reconstruction data; when appended to a generator, these encoders must interpret potentially out-of-distribution generated frames. We retain the pretrained 4D reconstruction hierarchy as the decoder and replace this RGB interface with an aligned final denoised VAE latent, thereby using the video model itself as the encoder.

4D Generation. Existing 4D generation methods combine generative priors with dynamic scene representations through three main routes. Earlier text- and image-guided methods optimize an object-centric dynamic NeRF or Gaussian representation separately for each output (Singer et al. 2023; Zhao et al. 2024; Bahmani et al. 2024; Zheng et al. 2024; Ling et al. 2024; Xu et al. 2024; Chu, Ke, and Fragkiadaki 2024). Generate-then-reconstruct methods, including Diffusion4D, 4Diffusion, and CAT4D, synthesize multiview or temporal RGB observations before recovering an explicit dynamic scene (Liang et al. 2024; Zhang et al. 2024; Wu et al. 2024). Feed-forward alternatives such as 4DNeX and Diff4Splat adapt video generators to predict dynamic point clouds or Gaussians for selected image-conditioned settings (Chen et al. 2025; Pan et al. 2026); WorldReel jointly predicts RGB and geometry, while WorldForge adds inference-time camera control (Fang et al. 2025; Song et al. 2026). Latent-to-4D instead connects a frozen video model to a pretrained 4D decoder directly in latent space, using the former as the encoder: it neither optimizes each scene nor retrains each generator, and does not use generated RGB as the geometric input.

Method

Latent-to-4D instantiates direct latent-to-4D generation by replacing an RGB encoder with the video model's VAE-space representation. As in Fig. 3, a frozen VAE provides observed-video latents during training, whereas a compatible frozen DiT provides final denoised latents during generation. The same pathway aligns either source, refines it with pretrained 4D latent, and decodes cameras and dynamic geometry.

Problem Formulation

We formulate direct latent-to-4D generation as a cross-representation mapping from a video VAE latent space to the structured token space of a pretrained 4D reconstructor. Let $(E_{\mathrm{v}}, D_{\mathrm{v}})$ denote the spatiotemporal VAE of a video generator, and let $\mathcal{Z}_{\mathrm{v}}$ denote its latent space. A video latent $\mathbf{z}_{\mathrm{v}} \in \mathcal{Z}_{\mathrm{v}}$ is used to encode the appearance and motion information required for video decoding. By contrast, the pretrained 4D hierarchy operates on a structured token space $\mathcal{Z}_{\mathrm{4D}}$, whose elements $\mathbf{Q} \in \mathcal{Z}_{\mathrm{4D}}$ support camera and dynamic-geometry prediction. Although both representations are spatiotemporal, they differ in temporal resolution, spatial grid, and feature dimension and are therefore not directly interchangeable.

The central problem is to bridge these two representation spaces without passing through RGB. Specifically, we seek an alignment map

$$\mathcal{A}_\phi : \mathcal{Z}_{\mathrm{v}} \to \mathcal{Z}_{\mathrm{4D}}, \qquad \mathbf{Q}^{(0)} = \mathcal{A}_\phi(\mathbf{z}_{\mathrm{v}}), \tag{1}$$

such that the aligned tokens $\mathbf{Q}^{(0)}$ can be refined and decoded into a discrete 4D scene

$$\mathcal{Y} = \{(\mathbf{C}_t, \mathbf{P}_t)\}_{t=1}^{T}, \qquad \mathbf{P}_t \in \mathbb{R}^{H\times W\times 3}, \tag{2}$$

where $\mathbf{C}_t$ denotes the camera at time t and $\mathbf{P}_t$ is a dense point map expressed in a shared world coordinate system. This representation captures both time-varying geometry and the camera trajectory required to render the scene from novel viewpoints. A conventional generate-then-reconstruct pipeline predicts

$$\widehat{\mathcal{Y}}_{\mathrm{rgb}} = R(D_{\mathrm{v}}(\mathbf{z}_{\mathrm{v}}))\,, \tag{3}$$

where D_{v} first decodes the video latent into RGB frames and an independently trained reconstructor R subsequently lifts those frames into 4D. This pipeline introduces an RGB representation boundary through which generation and decoding artifacts can propagate to the geometry prediction. Our objective is instead to learn a direct mapping $\mathcal{Z}_{\mathrm{v}} \to \mathcal{Z}_{\mathrm{4D}} \to \mathcal{Y}$ that bypasses both RGB decoding and subsequent RGB re-encoding.

Shared-Latent Interface across Compatible DiTs

We next exploit the video VAE space as a shared input interface across compatible DiTs. Given an observed video tensor $\mathbf{V}$, the VAE encoder provides the posterior-mean latent $\mathbf{z}^{\mathrm{obs}} = \mu(E_{\mathrm{v}}(\mathbf{V})) \in \mathcal{Z}_{\mathrm{v}}$. Given a condition c and sampled noise $\boldsymbol{\epsilon}$, a compatible video DiT G_θ produces the final denoised latent $\mathbf{z}^{\mathrm{gen}} = G_\theta(c, \boldsymbol{\epsilon}) \in \mathcal{Z}_{\mathrm{v}}$ immediately before VAE decoding. Both latent sources follow the same VAE scaling, tensor layout, and compression convention, although their distributions need not coincide.

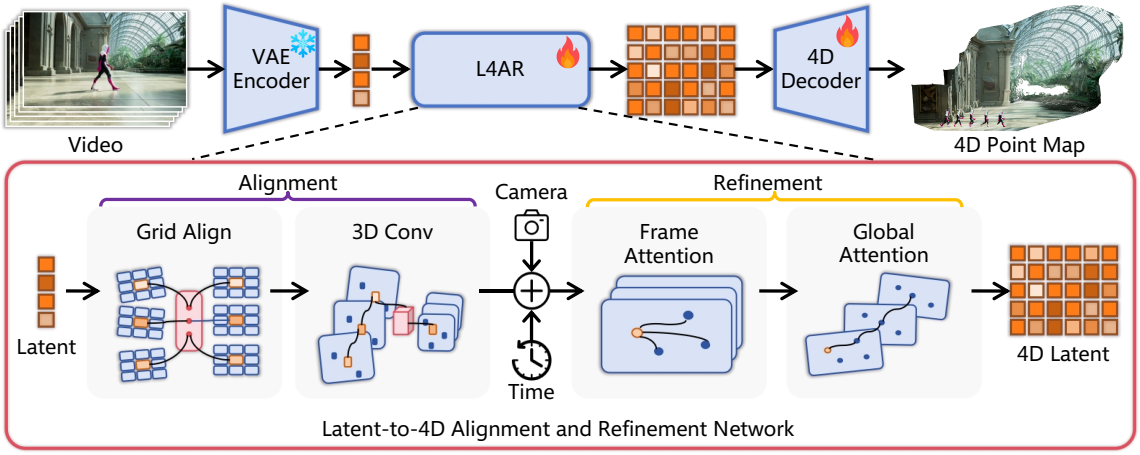

Figure 3: **Latent-to-4D training pipeline.** A frozen video VAE encodes an observed video into a VAE-space latent. L4AR aligns the latent grid through a learned 3D convolution, reuses frozen camera and time tokens, and refines the representation through alternating frame-wise and global attention. The 4D decoder predicts cameras and dynamic world-space geometry.

For either source $\mathbf{z} \in \{\mathbf{z}^{\text{obs}}, \mathbf{z}^{\text{gen}}\}$, Latent-to-4D factorizes the direct mapping into alignment, spatiotemporal refinement, and structured 4D decoding:

$$\widehat{\mathcal{Y}} = \mathcal{D}_\omega\left(\mathcal{H}_{\psi,\Delta\psi}\left(\mathcal{A}_\phi(\mathbf{z}); \mathbf{S}\right)\right). \tag{4}$$

The Alignment Module $\mathcal{A}_\phi$ maps the VAE latent tensor to the spatiotemporal token grid expected by the pretrained 4D reconstruction hierarchy. The Spatiotemporal Refinement Module $\mathcal{H}_{\psi,\Delta\psi}$ then transforms the aligned tokens into contextual 4D latent using frozen pretrained weights ψ and lightweight trainable updates $\Delta\psi$; $\mathbf{S}$ denotes the frozen camera and time tokens. Finally, the 4D Decoder $\mathcal{D}_\omega$ predicts per-frame cameras and dynamic world-space geometry.

During training, the pathway receives $\mathbf{z}^{\text{obs}}$ from videos with 4D annotations, and supervision updates the Alignment Module, the lightweight Refinement updates, and the geometry and camera heads. The video DiT is neither executed nor optimized. At inference, $\mathbf{z}^{\text{obs}}$ is replaced by $\mathbf{z}^{\text{gen}}$, while the complete downstream pathway remains unchanged. Because the condition has already been realized in the generated latent, no condition-specific 4D branch or explicit task identifier is required.

Compatibility requires the video DiTs to share the VAE checkpoint, latent normalization, tensor layout, compression convention, and a supported latent shape. Within this common-VAE boundary, their DiT architectures and conditioning regimes may differ. In the Experiments section, we evaluate a single Latent-to-4D checkpoint across two text-to-video DiTs and one image-to-video DiT, while the "Sensitivity to DiT-Derived Residuals" examines its sensitivity to a controlled DiT-derived residual component.

L4AR for Structured 4D Decoding

Alignment Module. The VAE latent cannot be consumed directly by the pretrained 4D hierarchy because the two representations differ in temporal resolution, spatial grid, and feature dimension. We first apply a fixed trilinear resampling operator $\mathcal{R}$ to match the required spatiotemporal resolution. A learned 3D convolution $\mathcal{S}_\phi$ then aggregates local spatiotemporal neighborhoods and projects the latent channels to the feature dimension expected by the 4D hierarchy. After flattening the spatial dimensions, the aligned tokens are

$$\mathbf{Q}^{(0)} = \mathcal{A}_\phi(\mathbf{z}) = \text{Flatten}(\mathcal{S}_\phi(\mathcal{R}(\mathbf{z}))) \in \mathbb{R}^{T\times M\times d}, \tag{5}$$

where T is the number of frames, M is the number of spatial tokens per frame, and d is the token dimension. The 3D convolution performs a shared local conversion across the latent grid, allowing neighboring appearance and motion evidence to be aggregated before global reasoning.

Spatiotemporal Refinement Module. Although alignment produces a compatible token grid, its local receptive field does not capture long-range correspondence, viewpoint changes, or motion across a sequence. We therefore introduce a hierarchical refinement module that combines frame-wise and global spatiotemporal self-attention. Given tokens of shape (B, T, M, d), frame-wise attention reshapes them to (BT, M, d) and consolidates spatial structure within each frame, whereas global attention operates on (B, TM, d) to exchange information across all spatial locations and time steps. The hierarchy first establishes per-frame structure through frame-wise attention and then alternates the two attention modes, allowing global blocks to propagate correspondence, viewpoint, and motion evidence while interleaved frame-wise blocks retain detailed spatial structure.

We collect intermediate representations at multiple depths of the hierarchy. At each selected depth, the most recent frame-wise and global features are concatenated to form a multi-level representation containing both spatial detail and sequence-level context. This representation provides the 4D Decoder with complementary intra-frame geometry and cross-frame motion cues.

4D Decoder. The 4D Decoder converts the refined feature hierarchy into explicit dynamic scene structure. For each frame t and pixel u, a multi-level geometry head predicts a depth value $\widehat{d}_t(u)$ and a world-space ray represented by origin $\widehat{\mathbf{o}}_t$ and unit direction $\widehat{\mathbf{r}}_t(u)$, together with their confi-

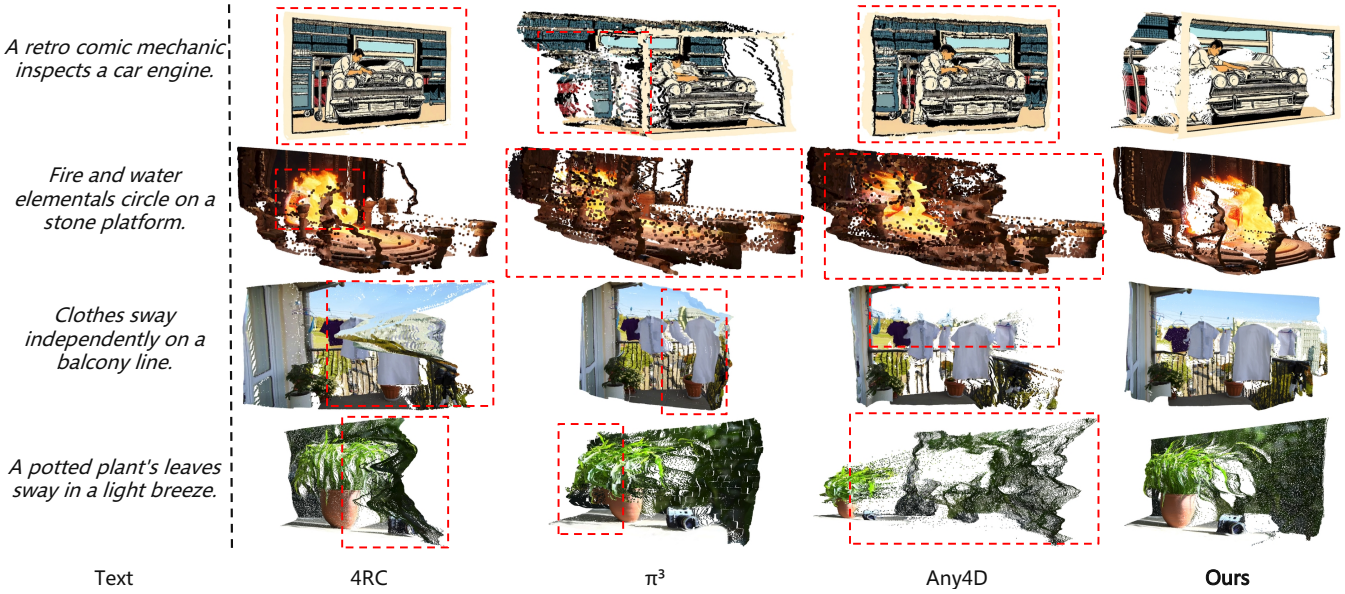

Figure 4: **Text-to-4D comparison.** Texts are abbreviated for display. RGB baselines reconstruct the decoded video, whereas Ours consumes its terminal latent directly.

dence values. In parallel, a camera head predicts $\widehat{\mathbf{C}}_t$ through a 9D pose–field-of-view parameterization. Following the decoder's ray parameterization, the corresponding world-space point is recovered as

$$\widehat{\mathbf{P}}_t(u) = \widehat{\mathbf{o}}_t + \widehat{d}_t(u)\widehat{\mathbf{r}}_t(u). \tag{6}$$

The predicted cameras and dynamic point maps form $\widehat{\mathcal{Y}}$, which can be reprojected across viewpoints and time rather than remaining restricted to the input video observations.

Training Objective. Given an annotated sequence $(\mathbf{V}, \mathcal{Y})$, the frozen VAE produces $\mathbf{z}^{\text{obs}}$, and the latent-to-4D pathway is supervised by

$$\mathcal{L} = \mathcal{L}_{\text{unc}} + \mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{geom}}. \tag{7}$$

The uncertainty-aware term $\mathcal{L}_{\text{unc}}$ contains confidence-weighted depth, depth-gradient, and world-ray losses. $\mathcal{L}_{\text{cam}}$ supervises camera translation, rotation, and field of view, while $\mathcal{L}_{\text{geom}}$ combines metric depth and ray supervision with mean and tail errors on world-space points and a surface-normal loss. Together, these objectives train the Alignment Module, the lightweight Refinement updates, and the geometry and camera heads from metric 4D annotations.

Throughout training, the video generators, VAE, original Transformer weights, camera and time tokens, motion decoder, and tracking head remain frozen. We progressively activate the trainable components to avoid destabilizing the pretrained geometric prior. Dataset composition, LoRA configuration, and stage-wise training budgets are provided in the Experimental Setup subsection and the Appendix.

Experiments

Experimental Setup

Implementation Details. We use the frozen Wan VAE and initialize the 31-block refinement hierarchy and prediction heads from 4RC (Oquab et al. 2024; Luo et al. 2026). We train the Alignment Module and prediction heads while adapting the refinement hierarchy with rank-16 LoRA, keeping the video models and pretrained base weights frozen. Starting from a 4RC-aligned latent adapter, the model undergoes multi-stage geometry-supervised training, with the final stage using 1,143 clips from six reconstruction datasets. All benchmarks are held out and use no test-time adaptation. Full architecture and training details are provided in the Appendix.

Method	Text CLIP ↑	RGB-ref. CLIP-I ↑	DINO global ↑	DINO match ↑	DINO F1 ↑
Text-to-4D					
CogVideoX-5B + 4RC	26.887	**75.33**	42.71	53.47	53.27
CogVideoX-5B + π^3	26.293	74.84	38.74	50.29	49.72
CogVideoX-5B + Any4D	25.414	72.66	29.46	45.97	45.12
Wan2.1-14B + 4RC	28.116	71.20	42.45	54.31	53.56
Wan2.1-14B + π^3	26.594	68.70	34.79	48.69	47.50
Wan2.1-14B + Any4D	26.261	67.56	29.97	46.80	45.55
Wan2.1-1.3B + 4RC	27.829	71.34	43.30	54.97	54.21
Wan2.1-1.3B + π^3	27.034	70.23	38.51	51.23	50.10
Wan2.1-1.3B + Any4D	26.326	68.19	31.00	47.26	46.06
Ours (Wan2.1-14B)	**28.544**	72.24	45.43	57.52	57.01
Ours (Wan2.1-1.3B)	28.434	72.32	**46.02**	**57.64**	**57.09**
Image-to-4D					
4DNeX	22.844	61.11	11.31	30.53	28.33
Wan2.2-I2V-A14B + 4RC	26.340	70.55	47.83	56.25	55.79
Wan2.2-I2V-A14B + π^3	24.678	66.65	33.50	45.08	43.82
Wan2.2-I2V-A14B + Any4D	24.362	65.30	27.21	43.54	42.17
Ours	**27.340**	**72.87**	**54.85**	**61.85**	**61.60**

Table 1: **4D generation on Text4D-200 and I4D-200.** Two-view projection scores (×100; higher is better); Ours and matched Wan cascades share the same generated latent.

Benchmarks and Metrics. Text4D-200 and I4D-200 are locked 200-case text- and image-conditioned benchmarks; every method is evaluated on every case. We render each predicted point sequence from two off-axis cameras and report Text CLIP, RGB-reference CLIP-I, DINO global similarity, valid-patch DINO matching, and DINO set F1. The corresponding generated RGB is used only for evaluation and col-

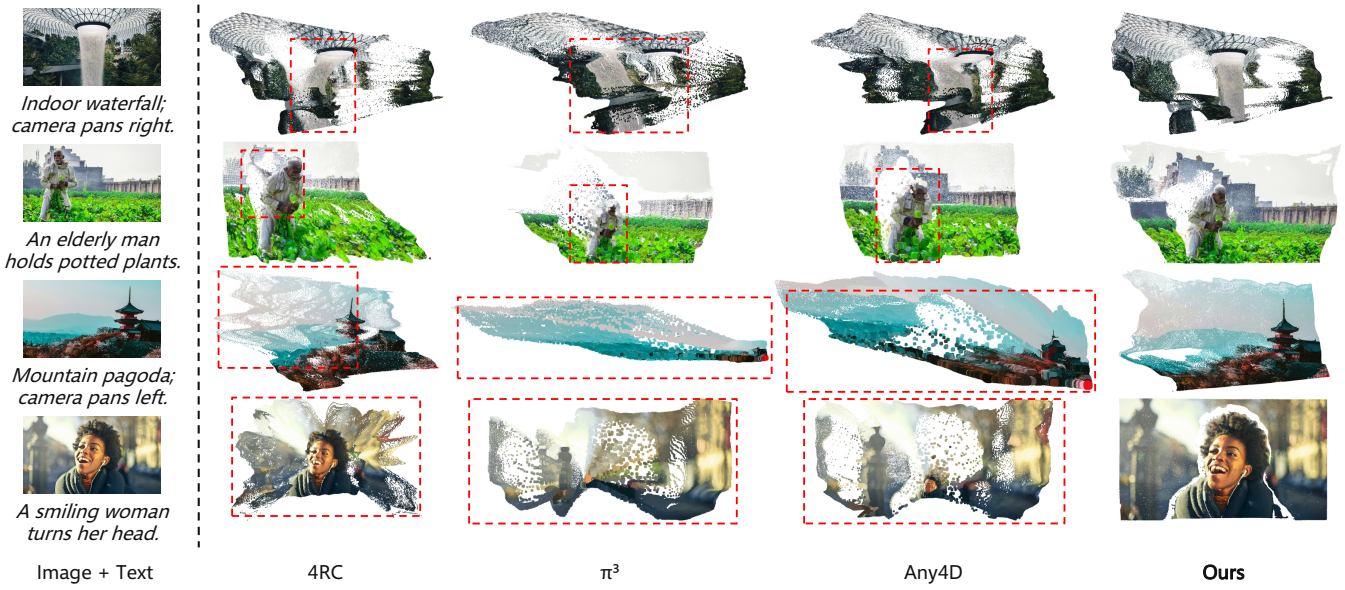

Figure 5: **Image-to-4D comparison.** Each row shows the input condition and 4D results. RGB baselines reconstruct the decoded video, whereas Ours consumes its terminal latent directly.

Task	Condition fidelity ↑	Geometry and completeness ↑	Temporal stability ↑	Overall quality ↑
Text-to-4D	59.2 [54.1–64.3]	66.8 [62.0–71.5]	63.5 [58.4–68.5]	65.7 [60.8–70.5]
Image-to-4D	66.4 [61.8–70.9]	72.1 [67.8–76.3]	68.3 [63.7–72.8]	70.6 [66.2–74.9]

Table 2: **User preference for Ours over baselines (%).**

	7-Scenes (18)			NRGBD (9)		
Variant	Acc↓	Comp↓	NC↑	Acc↓	Comp↓	NC↑
w/o Grid	3.783	6.844	0.608	5.823	9.686	0.726
w/o 3D Conv	6.944	15.806	0.554	12.439	26.511	0.594
w/o Frame	6.688	20.806	0.513	12.367	36.982	0.502
w/o Global	6.754	19.742	0.559	13.818	36.207	0.515
Full	**3.121**	**5.418**	**0.628**	**5.202**	**8.187**	**0.766**

Table 3: **Component ablation on 7-Scenes and NRGBD.** Acc/Comp are in cm; lower is better except for NC.

oring, not as geometry input. The off-axis DINO scores are appearance-dependent proxies for visible geometric coherence and completeness, not metric 4D accuracy; multi-view human judgments and ground-truth ablations provide complementary evidence. Benchmark construction and metric definitions are given in Appendix.

Baselines. One checkpoint serves Wan2.1-T2V-14B, Wan2.1-T2V-1.3B (Wan et al. 2025), and Wan2.2-I2V-A14B, which share the frozen Wan VAE. We compare with generate-then-reconstruct that decode the corresponding latent and apply 4RC (Luo et al. 2026), π^3 (Wang et al. 2026b), or Any4D (Karhade et al. 2025); CogVideoX-5B (Yang et al. 2025) provides an additional text-video generator, and official 4DNeX (Chen et al. 2025) provides a native Image-to-4D baseline. Ours versus the matched Wan cascades is the controlled same-latent comparison. These generated-latent benchmarks are our primary robustness test: L4AR is trained only on observed-video encodings but evaluated unchanged

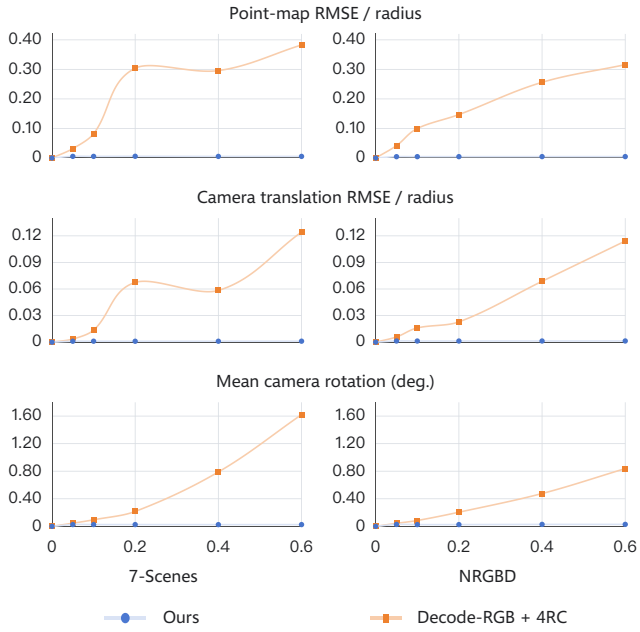

Figure 6: **Grid-Align-null DiT-residual sensitivity.** Geometry and camera drift on 7-Scenes and NRGBD.

on samples from three conditional DiTs. Further protocol details are in Appendix.

Quantitative 4D Generation Results

Text-to-4D. Direct latent lifting consistently outperforms the matched Wan+4RC cascades and the other tested reconstructors on structure-sensitive DINO metrics, with DINO-F1 gains of 2.88–3.45 points over matched 4RC. One checkpoint gives nearly identical performance with both text DiTs, supporting reuse without retraining. CogVideoX-5B+4RC retains the highest RGB-ref. CLIP-I, so we do not claim a uniform win on every metric.

Figure 7: **Additional controls.** Motion, appearance, pose, and trajectory inputs.

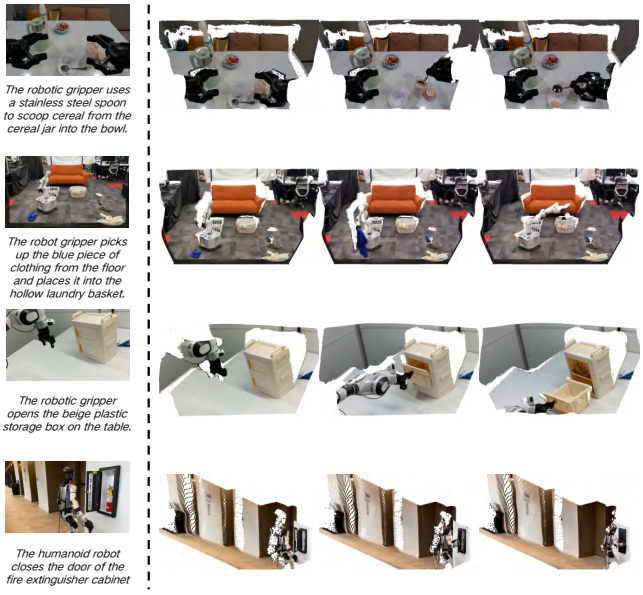

Figure 8: **Action-conditioned 4D.** Manipulation and navigation latents from a compatible backbone.

Image-to-4D. Ours ranks first on all I4D-200 metrics, including a 5.81-point DINO-F1 gain over the matched Wan2.2+4RC cascade, and exceeds the alternative reconstructors and native 4DNeX. Across both tasks, the off-axis DINO gains are consistent with more complete visible geometry, while not constituting metric geometry measurements.

Every interval exceeds 50%, with the strongest preferences for geometry and completeness. These judgments support more plausible and stable 4D outputs, but do not measure metric geometry.

Qualitative 4D Generation Results

Text-to-4D. Figure 4 covers articulated subjects, interacting elements, thin structures, translucency, and subtle motion. Ours generally preserves the principal subject and more surrounding scene support, whereas the RGB baselines exhibit holes, fragmented surfaces, or missing structures. The displayed point clouds show one off-axis frame and therefore assess geometric plausibility, completeness, and content preservation rather than temporal stability.

Image-to-4D. Figure 5 spans indoor and outdoor scenes, people, animals, and camera motion. Ours retains more foreground and source structure than the generate-then-reconstruct, consistent with Table 1.

Human Evaluation. Fifty participants evaluated 50 sampled cases per benchmark through randomized, anonymized pairwise comparisons. They could replay motion and inspect multiple viewpoints before judging condition fidelity, geometry and completeness, temporal stability, and overall quality. Each case received ten ratings; Table 2 reports case-averaged preferences with 95% bootstrap intervals. The full protocol is provided in Appendix.

Diagnostic Analyses

Sensitivity to DiT-Derived Residuals. We project the empirical near-terminal residual $\mathbf{z}_{45} - \mathbf{z}_{50}$ onto the Grid-Align width-null space and add it to an observed-video latent, applying the same perturbation to both pathways; Appendix provides the construction. All 30 comparisons favor Ours in Fig 6. At $\rho = 0.6$, point-map drift is $0.0053/0.0047$ for Ours versus $0.3827/0.3160$ for the baseline on 7-Scenes/NRGBD, with the same trend for camera estimation. This isolates a residual component rather than arbitrary DiT errors.

Benefits of Alignment and Refinement. Table 3 uses the matched protocol in Appendix. Every removal degrades both ground-truth benchmarks; the largest drops arise without the 3D convolution or either attention scope. This validates the design rather than claiming reconstruction OOD superiority.

Broader Applications

The same L4AR checkpoint inherits upstream motion, appearance, pose, trajectory, manipulation, and navigation controls (Figures 7–8). These examples show interface compatibility, not action success or physical correctness.

Conclusion

We introduce direct latent-to-4D generation, using a video model's final denoised VAE latent as a reusable interface to a 4D decoding hierarchy initialized from a pretrained reconstructor. Trained on roughly 1K reconstruction clips, Latent-to-4D transfers one checkpoint unchanged across three compatible text- and image-conditioned DiTs. Benchmarks yield higher projection-based DINO-F1 than RGB-decoding cascades, while multi-view human evaluation favors geometric plausibility, completeness, and temporal stability. A DiT-residual diagnostic and ground-truth ablations probe interface sensitivity and support the L4AR design. Qualitative results demonstrate compatibility with upstream controls without condition-specific training. Evidence remains limited to a shared VAE convention, and projection-based evaluation does not establish metric accuracy for generated scenes.

References

Bahmani, S.; Skorokhodov, I.; Rong, V.; Wetzstein, G.; Guibas, L.; Wonka, P.; Tulyakov, S.; Park, J. J.; Tagliasacchi, A.; and Lindell, D. B. 2024. 4D-fy: Text-to-4D Generation Using Hybrid Score Distillation Sampling. arXiv:2311.17984.

Bai, J.; Xia, M.; Fu, X.; Wang, X.; Mu, L.; Cao, J.; Liu, Z.; Hu, H.; Bai, X.; Wan, P.; and Zhang, D. 2025. ReCamMaster: Camera-Controlled Generative Rendering from A Single Video. arXiv:2503.11647.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; Jampani, V.; and Rombach, R. 2023. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. arXiv:2311.15127.

Chen, H.; Zhang, Y.; Cun, X.; Xia, M.; Wang, X.; Weng, C.; and Shan, Y. 2024. VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. arXiv:2401.09047.

Chen, Z.; Liu, T.; Zhuo, L.; Ren, J.; Tao, Z.; Zhu, H.; Hong, F.; Pan, L.; and Liu, Z. 2025. 4DNeX: Feed-Forward 4D Generative Modeling Made Easy. arXiv:2508.13154.

Cheng, G.; Gao, X.; Hu, L.; Hu, S.; Huang, M.; Ji, C.; Li, J.; Meng, D.; Qi, J.; Qiao, P.; Shen, Z.; Song, Y.; Sun, K.; Tian, L.; Wang, F.; Wang, G.; Wang, Q.; Wang, Z.; Xiao, J.; Xu, S.; Zhang, B.; Zhang, P.; Zhang, X.; Zhang, Z.; Zhou, J.; and Zhuo, L. 2025. Wan-Animate: Unified Character Animation and Replacement with Holistic Replication. arXiv:2509.14055.

Chu, R.; He, Y.; Chen, Z.; Zhang, S.; Xu, X.; Xia, B.; Wang, D.; Yi, H.; Liu, X.; Zhao, H.; Liu, Y.; Zhang, Y.; and Yang, Y. 2025. Wan-Move: Motion-controllable Video Generation via Latent Trajectory Guidance. arXiv:2512.08765.

Chu, W.-H.; Ke, L.; and Fragkiadaki, K. 2024. DreamScene4D: Dynamic Multi-Object Scene Generation from Monocular Videos. arXiv:2405.02280.

Fang, S.; Jiang, H.; Bai, Y.; Mitra, N. J.; and Huang, Q. 2025. WorldReel: 4D Video Generation with Consistent Geometry and Motion Modeling. arXiv:2512.07821.

Feng, H.; Zhang, J.; Wang, Q.; Ye, Y.; Yu, P.; Black, M. J.; Darrell, T.; and Kanazawa, A. 2025. St4RTrack: Simultaneous 4D Reconstruction and Tracking in the World. arXiv:2504.13152.

Go, H.; Narnhofer, D.; Bhat, G.; Truong, P.; Tombari, F.; and Schindler, K. 2026. Text-to-3D by Stitching a Multi-view Reconstruction Network to a Video Generator. arXiv:2510.13454.

He, H.; Xu, Y.; Guo, Y.; Wetzstein, G.; Dai, B.; Li, H.; and Yang, C. 2025. CameraCtrl: Enabling Camera Control for Text-to-Video Generation. arXiv:2404.02101.

Hu, W.; Gao, X.; Li, X.; Zhao, S.; Cun, X.; Zhang, Y.; Quan, L.; and Shan, Y. 2024. DepthCrafter: Generating Consistent Long Depth Sequences for Open-world Videos. arXiv:2409.02095.

Jiang, Z.; Zheng, C.; Laina, I.; Larlus, D.; and Vedaldi, A. 2025. Geo4D: Leveraging Video Generators for Geometric 4D Scene Reconstruction. arXiv:2504.07961.

Karhade, J.; Keetha, N.; Zhang, Y.; Gupta, T.; Sharma, A.; Scherer, S.; and Ramanan, D. 2025. Any4D: Unified Feed-Forward Metric 4D Reconstruction. arXiv:2512.10935.

Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R. C.; and Schindler, K. 2024. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. arXiv:2312.02145.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; Wu, K.; Lin, Q.; Yuan, J.; Long, Y.; Wang, A.; Wang, A.; Li, C.; Huang, D.; Yang, F.; Tan, H.; Wang, H.; Song, J.; Bai, J.; Wu, J.; Xue, J.; Wang, J.; Wang, K.; Liu, M.; Li, P.; Li, S.; Wang, W.; Yu, W.; Deng, X.; Li, Y.; Chen, Y.; Cui, Y.; Peng, Y.; Yu, Z.; He, Z.; Xu, Z.; Zhou, Z.; Xu, Z.; Tao, Y.; Lu, Q.; Liu, S.; Zhou, D.; Wang, H.; Yang, Y.; Wang, D.; Liu, Y.; Jiang, J.; and Zhong, C. 2025. HunyuanVideo: A Systematic Framework For Large Video Generative Models. arXiv:2412.03603.

Liang, H.; Yin, Y.; Xu, D.; Liang, H.; Wang, Z.; Plataniotis, K. N.; Zhao, Y.; and Wei, Y. 2024. Diffusion4D: Fast Spatial-temporal Consistent 4D Generation via Video Diffusion Models. arXiv:2405.16645.

Ling, H.; Kim, S. W.; Torralba, A.; Fidler, S.; and Kreis, K. 2024. Align Your Gaussians: Text-to-4D with Dynamic 3D Gaussians and Composed Diffusion Models. arXiv:2312.13763.

Liu, X.; Xiao, Y.; Chen, D. Y.; Feng, J.; Tai, Y.-W.; Tang, C.-K.; and Kang, B. 2025. Trace Anything: Representing Any Video in 4D via Trajectory Fields. arXiv:2510.13802.

Luo, Y.; Zhou, S.; Lan, Y.; Pan, X.; and Loy, C. C. 2026. 4RC: 4D Reconstruction via Conditional Querying Anytime and Anywhere. arXiv:2602.10094.

Mai, J.; Zhu, W.; Liu, H.; Li, B.; Zheng, C.; Schmidhuber, J.; and Ghanem, B. 2025. Can Video Diffusion Model Reconstruct 4D Geometry? arXiv:2503.21082.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193.

Pan, P.; Lin, C.; Zhao, J.; Li, C.; Lin, Y.; Li, H.; Yan, H.; Wen, K.; Lin, Y.; Yuan, Y.; and Mu, Y. 2026. Diff4Splat: Controllable 4D Scene Generation with Latent Dynamic Reconstruction Models. arXiv:2511.00503.

Singer, U.; Sheynin, S.; Polyak, A.; Ashual, O.; Makarov, I.; Kokkinos, F.; Goyal, N.; Vedaldi, A.; Parikh, D.; Johnson, J.; and Taigman, Y. 2023. Text-To-4D Dynamic Scene Generation. arXiv:2301.11280.

Song, C.; Yang, Y.; Zhao, T.; Li, R.; and Zhang, C. 2026. Taming Video Models for 3D and 4D Generation via Zero-Shot Camera Control. arXiv:2509.15130.

Sucar, E.; Insafutdinov, E.; Lai, Z.; and Vedaldi, A. 2026. V-DPM: 4D Video Reconstruction with Dynamic Point Maps. arXiv:2601.09499.

Sucar, E.; Lai, Z.; Insafutdinov, E.; and Vedaldi, A. 2025. Dynamic Point Maps: A Versatile Representation for Dynamic 3D Reconstruction. arXiv:2503.16318.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; Zeng, J.; Wang, J.; Zhang, J.; Zhou, J.; Wang, J.; Chen, J.; Zhu, K.; Zhao, K.; Yan, K.; Huang, L.; Feng, M.; Zhang, N.; Li, P.; Wu, P.; Chu, R.; Feng, R.; Zhang, S.; Sun, S.; Fang, T.; Wang, T.; Gui, T.; Weng, T.; Shen, T.; Lin, W.; Wang, W.; Wang, W.; Zhou, W.; Wang, W.; Shen, W.; Yu, W.; Shi, X.; Huang, X.; Xu, X.; Kou, Y.; Lv, Y.; Li, Y.; Liu, Y.; Wang, Y.; Zhang, Y.; Huang, Y.; Li, Y.; Wu, Y.; Liu, Y.; Pan, Y.; Zheng, Y.; Hong, Y.; Shi, Y.; Feng, Y.; Jiang, Z.; Han, Z.; Wu, Z.-F.; and Liu, Z. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. arXiv:2503.20314.

Wang, L.; Zhang, C.; Kabra, R.; Uijlings, J.; Waslander, S.; Zisserman, A.; Carreira, J.; He, K.; Andriluka, M.; Bazavan, E. G.; Zanfir, A.; and Sminchisescu, C. 2026a. Video Generation Models are General-Purpose Vision Learners. arXiv:2607.09024.

Wang, Y.; Zhou, J.; Zhu, H.; Chang, W.; Zhou, Y.; Li, Z.; Chen, J.; Pang, J.; Shen, C.; and He, T. 2026b. π^3: Permutation-Equivariant Visual Geometry Learning. arXiv:2507.13347.

Wang, Z.; Yuan, Z.; Wang, X.; Chen, T.; Xia, M.; Luo, P.; and Shan, Y. 2024. MotionCtrl: A Unified and Flexible Motion Controller for Video Generation. arXiv:2312.03641.

Wu, R.; Gao, R.; Poole, B.; Trevithick, A.; Zheng, C.; Barron, J. T.; and Holynski, A. 2024. CAT4D: Create Anything in 4D with Multi-View Video Diffusion Models. arXiv:2411.18613.

Xing, J.; Xia, M.; Zhang, Y.; Chen, H.; Yu, W.; Liu, H.; Wang, X.; Wong, T.-T.; and Shan, Y. 2023. DynamiCrafter: Animating Open-domain Images with Video Diffusion Priors. arXiv:2310.12190.

Xu, D.; Liang, H.; Bhatt, N. P.; Hu, H.; Liang, H.; Plataniotis, K. N.; and Wang, Z. 2024. Comp4D: LLM-Guided Compositional 4D Scene Generation. arXiv:2403.16993.

Xu, T.-X.; Gao, X.; Hu, W.; Li, X.; Zhang, S.-H.; and Shan, Y. 2025. GeometryCrafter: Consistent Geometry Estimation for Open-world Videos with Diffusion Priors. arXiv:2504.01016.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; Yin, D.; Zhang, Y.; Wang, W.; Cheng, Y.; Xu, B.; Gu, X.; Dong, Y.; and Tang, J. 2025. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. arXiv:2408.06072.

Zhang, C.; Moing, G. L.; Koppula, S.; Rocco, I.; Momeni, L.; Xie, J.; Sun, S.; Sukthankar, R.; Barral, J. K.; Hadsell, R.; Ghahramani, Z.; Zisserman, A.; Zhang, J.; and Sajjadi, M. S. M. 2025. Efficiently Reconstructing Dynamic Scenes One D4RT at a Time. arXiv:2512.08924.

Zhang, H.; Chen, X.; Wang, Y.; Liu, X.; Wang, Y.; and Qiao, Y. 2024. 4Diffusion: Multi-view Video Diffusion Model for 4D Generation. arXiv:2405.20674.

Zhang, S.; Wang, J.; Zhang, Y.; Zhao, K.; Yuan, H.; Qin, Z.; Wang, X.; Zhao, D.; and Zhou, J. 2023. I2VGen-XL: High-Quality Image-to-Video Synthesis via Cascaded Diffusion Models. arXiv:2311.04145.

Zhao, Y.; Yan, Z.; Xie, E.; Hong, L.; Li, Z.; and Lee, G. H. 2024. Animate124: Animating One Image to 4D Dynamic Scene. arXiv:2311.14603.

Zheng, Y.; Li, X.; Nagano, K.; Liu, S.; Kreis, K.; Hilliges, O.; and Mello, S. D. 2024. A Unified Approach for Text- and Image-guided 4D Scene Generation. arXiv:2311.16854.

Zhu, R.; Lu, J.; Hu, W.; Han, X.; Cai, J.; Shan, Y.; and Zheng, C. 2026. MotionCrafter: Dense Geometry and Motion Reconstruction with a 4D VAE. arXiv:2602.08961.