# WorldReel: 4D Video Generation with Consistent Geometry and Motion Modeling

Shaoheng Fang[1] Hanwen Jiang[2] Yunpeng Bai[1] Niloy J. Mitra[2,3] Qixing Huang[1]
[1]The University of Texas at Austin [2]Adobe Research [3]University College London

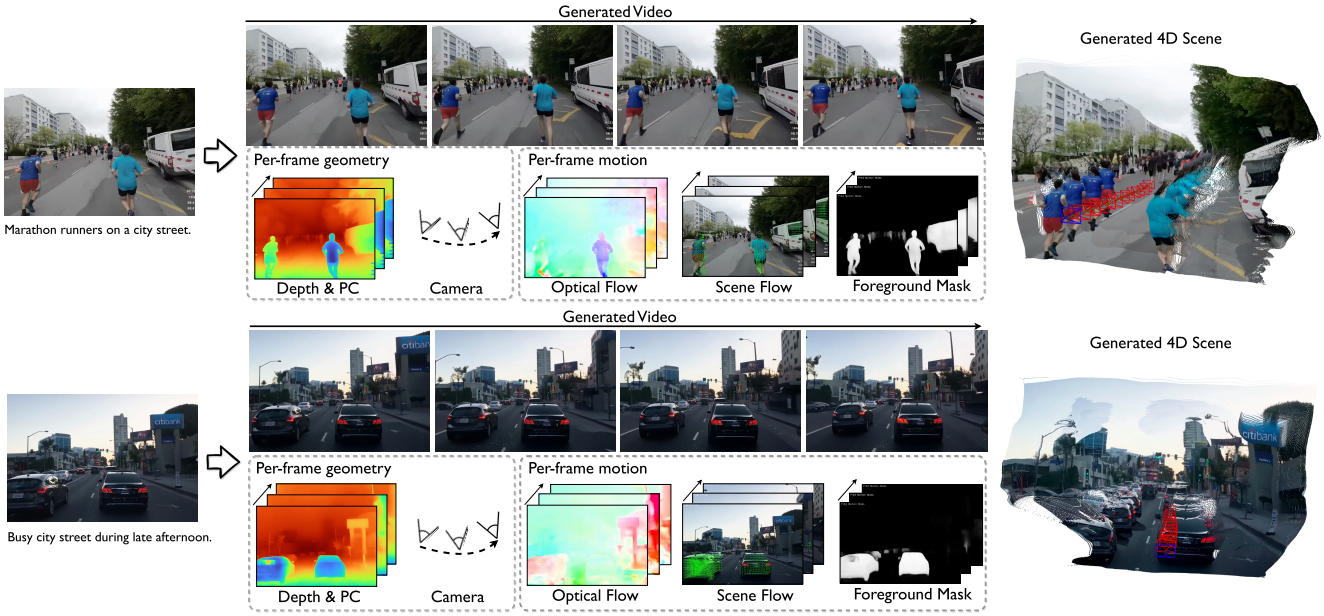


Figure 1. **End-to-end 4D generation.** Given a text prompt and a single input image (left), WorldReel generates a video (center) *together with* explicit 4D scene representations: per-frame geometry (depth + point cloud) with calibrated camera poses, and per-frame motion (optical flow, scene flow) with object masks (bottom panels). The rendered 4D scenes (right) exhibit consistent structure over time, even under non-rigid dynamics, illustrating spatiotemporal consistency and tight coupling of appearance, geometry, and motion. Project page: https://bshfang.github.io/worldreel/

## Abstract

*Recent video generators achieve striking photorealism, yet remain fundamentally inconsistent in 3D. We present WorldReel, a 4D video generator that is natively spatio-temporally consistent. WorldReel jointly produces RGB frames together with 4D scene representations, including pointmaps, camera trajectory, and dense flow mapping, enabling coherent geometry and appearance modeling over time. Our explicit 4D representation enforces a single underlying scene that persists across viewpoints and dynamic content, yielding videos that remain consistent even under large non-rigid motion and significant camera movement. We train WorldReel by carefully combining synthetic and real data: synthetic data providing precise 4D supervision (geometry, motion, and camera), while real videos contribute visual diversity and realism. This blend allows WorldReel to generalize to in-the-wild footage while preserving strong geometric fidelity. Extensive experiments demonstrate that WorldReel sets a new state-of-the-art for consistent video generation with dynamic scenes and moving cameras, improving metrics of geometric consistency, motion coherence, and reducing view-time artifacts over competing methods [3, 10, 26]. We believe that WorldReel brings video generation closer to 4D-consistent world modeling, where agents can render, interact, and reason about scenes through a single and stable spatiotemporal representation.*

## 1. Introduction

Recent video generators [32, 56, 72] have rapidly gained popularity, delivering impressive perceptual quality and realism

across diverse prompts and scenes. Despite their impressive success in image fidelity and temporal smoothness, these models do *not* maintain a single stable 3D scene that evolves over time. The resulting inconsistencies manifest as view-time drift, geometric flicker, and entangled camera/scene motion; these limitations become acute when extrapolating viewpoints or editing content as in emerging world model settings.

A natural next aspiration is 4D generation, in which a model maintains a coherent spatiotemporal scene representation. Prior efforts either optimize explicit dynamic scene representations [2, 38, 43] or post-hoc lift controllably generated 2D videos into 3D structure [45, 52, 76]. These approaches improve geometric alignment, but either are computationally heavy or fundamentally inherit the geometric inconsistency of 2D video priors, with limited capability to generalize to in-the-wild dynamics. None natively integrates a true 4D structure into a generative prior.

We propose *WorldReel*, a unified 4D generator that jointly produces RGB, per-frame geometry, and motion in both 2D and 3D, explicitly outputting pointmaps, calibrated camera trajectories, and scene flow. In contrast to 2D optical flow or keypoint tracking used in prior methods [6, 25], scene flow directly encodes 3D dynamics, (i) cleanly disentangling camera motion from object motion and (ii) operating in a physically-grounded evolving 3D frame. To prevent appearance from leaking into geometry/motion channels, we introduce an appearance-independent representation that stabilizes learning signals across viewpoints and lighting of both synthetic and real data.

At the core of WorldReel is an augmented geometry-motion latent for a video diffusion transformer. This latent explicitly carries geometry and motion information through the generative process, yielding two benefits: (i) stronger inductive bias for 4D consistency and (ii) appearance-agnostic conditioning that improves generalization. Crucially, we demonstrate how to leverage this appearance-independent representation for effectively using synthetic datasets that provide accurate 4D labels (geometry, motion, camera) without sacrificing realism when mixed with real videos.

We design an architecture that predicts 4D from the geo-motion latent via multi-task learning with a customized temporal DPT-style decoder. Specifically, a lightweight shared backbone captures correlations among geometric tasks, with multiple task-specific heads for pointmaps, camera, dynamic mask, and scene flow. We train these outputs jointly with regularization terms that explicitly decouple static and dynamic components of the scene. This enforces geometric consistency on static structure and motion consistency on non-rigid regions, resulting in 4D coherence under both object and camera motion.

We extensively compare WorldReel with state-of-the-art video and 4D generation models [3, 10, 52] in terms of both (i) geometry and motion consistency of generated videos, and (ii) geometric quality of generated 4D scenes, where WorldReel delivers consistent and substantial improvements. For video generation, WorldReel produces stronger camera and subject motion, achieving the best dynamic degree on both general and complex motion, while maintaining state-of-the-art photorealism. For generated 4D scenes, our geometry is significantly more accurate, reducing depth error from $0.353 \rightarrow 0.287$ and achieving the lowest camera pose errors compared to recent 4D/3D-aware baselines [3, 10].

In summary, our main contributions are: (i) WorldReel as a unified generative framework that outputs RGB, pointmaps, calibrated cameras, optical flow, and scene flow, enforcing a persistent dynamic 3D scene through time; (ii) an appearance-agnostic geo-motion latent that embeds explicit geometry and motion, improving generalization and enabling strong supervision from synthetic and real data; (iii) a shared, lightweight DPT backbone with multi-task heads and targeted regularizers that decouple camera motion from dynamic components for tight geometric and motion consistency. As a result, WorldReel sets a new SoTA in terms of 4D consistent output, especially for dynamic assets, taking a step towards 4D-consistent world modeling, where scenes can be rendered, edited, and reasoned about from a single stable spatiotemporal representation.

## 2. Related Works

### 2.1. Diffusion Models for Video Generation

The success of diffusion models in image synthesis [15, 47] has been extended to video generation, demonstrating promising results [13, 19, 50, 67]. Specifically, latent diffusion [5, 7, 16, 46, 55] is applied, where a variational autoencoder [31] first encodes the image or video into a compact latent representation, and the diffusion process is performed in the latent space. This variational approach forms the basis for current state-of-the-art video generation models, including cascaded diffusion systems [18, 64, 75] and large diffusion models [20, 32, 56, 72] built on DiT backbones [39]. However, a commonly recognized limitation of these models [5, 20] is their struggle with physical plausibility; they often fail to produce consistent 3D geometric structures and coherent temporal motions.

**Geometry-Aware Video Generation.** To address the lack of 3D consistency, many works have focused on integrating geometric priors into the generation process. Some methods jointly model video and geometry in the form of depth [3, 11, 23, 69] or point clouds [10, 11, 74], enabling the simultaneous generation of video and its underlying geometric structure; OmniVDiff [69] models appearance, depth, Canny edges, and semantic segmentation simultaneously, enabling multi-modal conditional control; GeoVideo [3] additionally introduces an explicit geometric regularization

loss, further improving 3D consistency in generated videos. Furthermore, [66] uses a 3D foundation model [57] to implicitly incorporate geometric priors, aligning the video latent space with the foundation model's features. However, these methods largely neglect scene dynamics and motion, focusing primarily on static scenes.

**Motion-Aware Video Generation.** In a parallel effort to improve temporal coherence, motion priors are incorporated into the generation process and have shown effective results. Track4Gen [25] adds an auxiliary point tracking task to provide explicit spatial supervision over diffusion latents. Additionally, optical flow, as a simple, generic, and easily obtainable motion representation, is frequently utilized. Videojam [6] jointly models video and optical flow during generation, and introduces an inference-time guidance mechanism using the motion prediction as a dynamic guidance signal. In addition, optical flow is also useful as an input condition to achieve coherent and smooth motion during generation [28, 48]. Alternatively [37, 62], optical flow is used to assess motion quality and intensity, guiding video generation models to generate smoother/dynamic motions.

### 2.2. Feed-Forward 4D Perception

Latest 4D feed-forward models aim to recover camera parameters and consistent geometric properties, such as depth and tracking, directly from image sequences. Following the success of Dust3R [61] on static scenes, several works have adapted the transformer-based framework to dynamic scenarios for image pairs [9, 14, 73] or longer sequences [27, 60, 81]. Alternative approaches leverage large-scale foundation models. For instance, Geo4D [26] and GeometryCrafter [71] repurpose video generation models [5, 70] for multi-modal geometry generation, leveraging the strong dynamic priors learned by pre-trained video diffusion models. Meanwhile, L4P [1] utilizes a pre-trained video encoder [59] to extract video features for various downstream 4D tasks. In our method, we generate various 4D properties together with the videos.

### 2.3. 4D Generation

Early methods [2, 38, 43, 51] optimize explicit 4D scene representations [41, 65] with score distillation sampling (SDS) [40], leveraging priors from video diffusion models [5, 50] and 3D-aware multi-view diffusion models [35, 49]. These optimization-based pipelines suffer from high computational cost and long generation times, and are typically constrained to single dynamic object generation. To improve scalability, recent methods have moved to feed-forward prediction that outputs a 4D scene representation in a single pass. L4GM [44] extends the LGM [53] to predict per-frame 3D Gaussian splats, but remains limited to single-object settings. Others [45, 52, 68, 76] leverage controllable video generation to synthesize dynamic videos, but require an additional reconstruction stage to create the 4D representation. TesserAct [77] proposes jointly predicting RGB-DN (RGB, Depth, and Normal) videos to build a 4D world model. However, it is specifically tailored for embodied robotics and focuses largely on manipulation scenarios. 4DNeX [10] jointly models videos and point cloud geometry, training on dynamic datasets to enable dynamic point cloud generation. Despite this, their approach does not explicitly model scene dynamics and produces nearly fixed camera motion.

We follow these works to build our method on a video generation model. Moreover, we generate all 2D and 3D geometry and motion simultaneously from an augmented video latent space and design regularization terms to enforce motion and geometry consistency across time.

## 3. WorldReel

Based on existing powerful video generation models, we improve the spatio-temporal consistency of generated videos by introducing 4D inductive bias. In the following part, we first review the basics of video generation via latent diffusion (Sec. 3.1). We then introduce two key designs of WorldReel, including a latent space augmented by additional 2.5D inputs (Sec. 3.2), and the unified 4D outputs that are directly supervised for improving consistent video and 4D geometry modeling (Sec. 3.3).

### 3.1. Preliminary: Video Latent Diffusion

Our method is built on the latest latent diffusion video generation methods with transformer-based models [32, 56, 72, 79], where the diffusion process occurs in a lower-dimensional latent space of a pre-trained 3D VAE [72]. Given a clean video clip $X^{1:N} \in \mathbb{R}^{N\times H\times W\times 3}$, the latent can be denoted as $\mathbf{z}_0 = \mathcal{E}(X^{1:N})$, where $\mathcal{E}$ is the encoder of the 3D VAE. The diffusion involves a forward process that adds a Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ to obtain the noisy latent $\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \sigma_t \epsilon$, where $t = 1, \cdots, T$ is the diffusion time step and $\alpha_t$ and $\sigma_t$ are noise scheduler parameters. The reverse process is to remove noise through a denoiser $f_\theta$ with condition $\mathbf{c}$, trained by minimizing:

$$\min_{\theta} \mathbb{E}_{t\sim U[1,T],\, \epsilon\sim\mathcal{N}(0,I)} \left[ \left\| f_\theta(\mathbf{z}_t, t, \mathbf{c}) - \epsilon \right\|_2^2 \right] \quad (1)$$

### 3.2. Geometry-Motion Augmented Latent

Our objective is to generate dynamic 4D scenes using a video diffusion model, which requires the ability to jointly model video appearance, 3D geometry, and complex motion. To achieve this, we propose to extend the latent space of video generation models by introducing explicit priors for geometry and motion. We use frame-aligned depth maps and optical flow to jointly model 2.5-D geometry and motion. Concretely, depth captures the scene structure under perspective, while optical flow summarizes the per-pixel displacement induced by both camera and object motion.

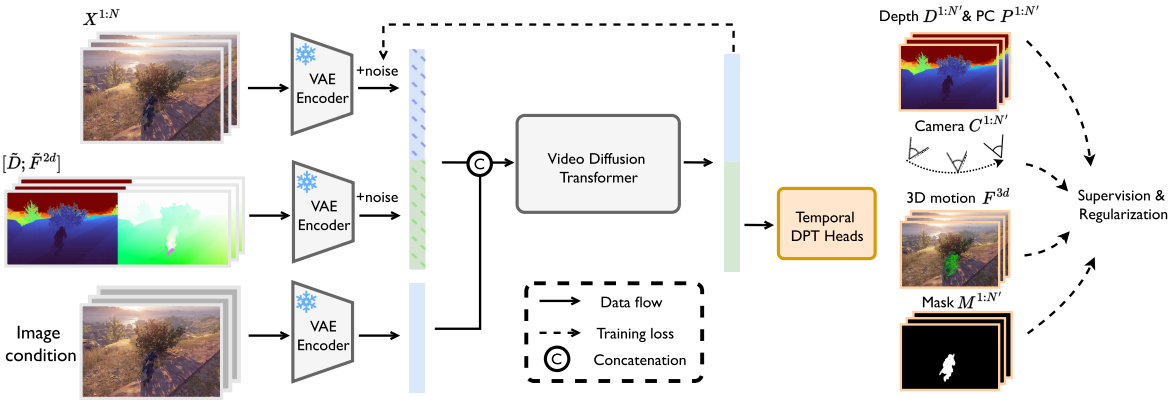


Figure 2. **Overview of WorldReel.** We augment a video diffusion transformer with a geo–motion latent (from RGB and 2.5D cues such as depth/optical flow) to inject a 4D inductive bias for spatio-temporal consistency. A temporal DPT decoder is trained with direct supervision and regularization to predict unified 4D outputs (depth/point cloud, calibrated camera, 3D scene flow, and masks).

Our choice of depth and optical flow as the preliminary geo-motion representation is based on several advantages: i) Both depth maps and optical flow fields possess a dense, image-like structure that is intrinsically aligned to RGB frames. This modality compatibility allows us to leverage powerful, pre-trained 3D VAEs to directly encode and decode them (as shown in [21, 30]) and is easy to integrate into existing video generation pipelines. ii) High-quality depth maps and optical flow are readily obtainable from powerful foundation models [21, 63], which enables our model training to be conducted at scale. iii) The geo-motion latent is 3D-focused, which factors out the appearance and texture of the generated scene, resulting in a smaller distribution gap between synthetic and real-world data. When training subsequent 4D generation tasks, this disentanglement enables better leverage of accurate ground-truth annotations from synthetic datasets, leading to stronger generalization performance.

To encode the additional inputs of depth $D_i \in \mathbb{R}^{H\times W\times 1}$, and forward optical flow $F_i^{2\text{d}} \in \mathbb{R}^{H\times W\times 2}$ with existing 3D VAE, we need to align them to the same value range with the RGB images. Thus, we get the normalized depth and optical flow as $\tilde{D}_i = 2\cdot\frac{D_i - d_{min}}{d_{max}-d_{min}} - 1$ and $\tilde{F}_i^{2\text{d}} = \frac{F_i^{2\text{d}}}{|F^{2\text{d}}|_{max}}$, where $d_{max}$ and $d_{min}$ are the maximum and minimum depth and $|F^{2d}|_{max}$ is the maximum displacement scale across all frames. The geo-motion latent can be encoded by the pre-trained 3D VAE, denoted as $\mathbf{z}_0^{\text{gm}} = \mathcal{E}([\tilde{D}; \tilde{F}^{2\text{d}}])$, where $[\cdot;\cdot]$ denotes concatenation along the channel dimension. With the original video latent $\mathbf{z}_0^{\text{rgb}} = \mathcal{E}(X)$, we can obtain the augmented latent for the diffusion model by channel-wise concatenation $\mathbf{z}_0 = [\mathbf{z}_0^{\text{rgb}}; \mathbf{z}_0^{\text{gm}}]$

**Adapt Existing Video Generation Models.** We then train the video diffusion DiT to model the joint distribution over appearance and geo-motion in the augmented latent space. To effectively leverage pre-trained weights, we adapt the original transformer architecture with minimal modifications. Specifically, only the input and output projection layers are modified for the doubled channel dimension of $\mathbf{z}_0$, while all intermediate blocks remain unchanged. To further stabilize training, we employ a zero-initialization scheme for the input projection layer: weights corresponding to the original video latent $\mathbf{z}^{rgb}$ are loaded from the pre-trained model, while new expanded parameters corresponding to the geo-motion latent $\mathbf{z}^{gm}$ are initialized to zero. This strategy ensures that, at the beginning of training, the model's behavior is identical to that of the original video diffusion model, improving training stability.

### 3.3. 4D Video Modeling

**Unified 4D Representations as Outputs.** While leveraging additional inputs of depth and optical flow provides strong priors for improving video generation in latent space, they are insufficient to recover the 3D scene structure, especially for disentangling camera and object motion which is ill-posed under the 2.5-D setting. Thus, we opt to predict finer-grained 4D representations of the generated video clip and apply explicit supervision. In detail, except for the video clip, the model also outputs the corresponding camera trajectory, point clouds, and object foreground masks. The supervision of 4D representations enables consistent spatial alignment across frames, by back-propagating geometry-related gradients to the latent space. This design choice also effectively helps the disentanglement between camera motion and object motion, encouraging the model to capture 3D dynamics in a better latent space.

Formally, for a downsampled set of latent frames $i = 1, \cdots N'$, we define the output unified 4D representation as $(D_i, P_i, C_i, F_i^{3\text{d}}, M_i)$. $C_i \in \mathbb{R}^9$ are camera intrinsics and extrinsics following the parameterization in [57]. $P_i \in \mathbb{R}^{H\times W\times 3}$ is the point cloud, $F_i^{3d} \in \mathbb{R}^{H\times W\times 3}$ is the forward

scene flow, and $M_i \in \mathbb{R}^{H\times W}$ is the dynamic (foreground) mask. Camera parameters $C_i$, point clouds $P_i$, and scene flow $F_i^{3d}$ are all represented in the first-frame canonical coordinate system for a consistent scene description.

**Model Design.** To predict the unified 4D representations, we design a customized temporal DPT [42] architecture. The model first extracts multi-scale dense features from the input latent, which are then processed and aggregated using a DPT-like fusion backbone incorporating temporal transformers (see supp.). Critically, our design is motivated by the high correlation between target geometric representations. We utilize a single, shared DPT-style decoder to process and generate unified dense features. Only at the final output layer do multiple lightweight, task-specific heads predict their respective tasks. This not only ensures significant parameter efficiency but also acts as a strong form of regularization, forcing the model to learn a unified and geometrically consistent representation for all tasks.

**Training.** Our model is trained using a two-stage strategy to ensure stability and optimize both generative quality and 4D accuracy. The first stage trains the video diffusion transformer and the temporal DPT heads in isolation. The second stage then jointly trains the full model end-to-end with additional regularization terms.

Specifically, we first finetune the video diffusion transformer to accommodate the new geo-motion augmented latent. We use a standard diffusion loss as defined in Eq. 1. The total loss combines losses on the appearance component and the geo-motion components: $\mathcal{L}_{\text{diff}} = \mathcal{L}_{\text{diff}}^{\text{rgb}} + \mathcal{L}_{\text{diff}}^{\text{gm}}$.

To pre-train the temporal DPT heads on 4D prediction, we use the clean geo-motion latent $\mathbf{z}_0^{\text{gm}}$ as input and optimize using a comprehensive multi-task loss $\mathcal{L}_{\text{dpt}}$.

$$\mathcal{L}_{\text{dpt}} = \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{pc}} + \mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{mask}} + \lambda_{\text{flow}}\mathcal{L}_{\text{flow}}. \quad (2)$$

We use masked L1 loss for $\mathcal{L}_{\text{depth}}$ and $\mathcal{L}_{\text{pc}}$ on pixels with valid depth values; Huber loss for $\mathcal{L}_{\text{cam}}$; BCE loss for $\mathcal{L}_{\text{mask}}$. In $\mathcal{L}_{\text{flow}}$, we reweight the loss pixel-wise to focus on the prediction of the foreground motion according to the dynamic mask $\hat{M}_i$ (see supp. for details).

Then, we jointly train the video transformer model and the temporal DPT heads end-to-end. Regularization terms are added for geometry and temporal consistency. To ensure geometry consistency in the static background and motion smoothness for dynamic objects, we apply different regularization terms. We use $\hat{M}_i^{\text{bg}}$ and $\hat{M}_i^{\text{fg}}$ to represent the background and foreground masks derived from $\hat{M}_i$ label.

$$\begin{aligned}
\mathcal{L}_{\text{reg}}^{\text{depth}} &= \sum_i \sum_j \left\| \hat{M}_i^{\text{bg}} \odot (D_j - \text{Proj}(D_i, T_{i\to j})) \right\|_2 \\
\mathcal{L}_{\text{reg}}^{\text{flow}} &= \sum_i \left( \left\| \hat{M}_i^{\text{fg}} \odot \nabla_x F_i^{3d} \right\|_2 + \left\| \hat{M}_i^{\text{fg}} \odot \nabla_y F_i^{3d} \right\|_2 \right) \\
\mathcal{L}_{\text{reg}} &= \mathcal{L}_{\text{reg}}^{\text{depth}} + \mathcal{L}_{\text{reg}}^{\text{flow}}
\end{aligned} \quad (3)$$

Here $T_{i\to j}$ is the transformation matrix derived from $C_i, C_j$. The joint training objective is

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{dpt}}\mathcal{L}_{\text{dpt}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \quad (4)$$

### 3.4. 4D Data Preparation

Training a model for comprehensive 4D video generation requires a diverse, high-quality video dataset with corresponding 4D labels, encompassing a wide range of dynamics and sufficient scene complexity. Therefore, we employ a mixed-data strategy [4, 29, 58, 78, 80], leveraging both synthetic datasets with ground-truth labels and real-world videos with pseudo-annotations.

Synthetic datasets include PointOdyssey [78], BEDLAM [4], Dynamic Replica [29], and Omniworld-Game [80]. Synthetic data alone is limited in scale and complexity, which can impede the model's generalization to real-world scenarios. We therefore supplement our training with real-world videos. While some recent works construct 4D datasets [10, 58] by filtering large-scale sources [8, 12] and using off-the-shelf annotation models [33, 73], the resulting labels are noisy and of insufficient quality. We start with high-quality raw videos filtered from Panda-70M [8] by SpatialVid [58] and re-annotate them using state-of-the-art vision foundation models to obtain superior pseudo-labels. Specifically, for per-frame depth, we use GeometryCrafter [71] to obtain temporally smooth depth sequences. For feed-forward 4D task labels corresponding to latent frames, we employ ViPE [22] to obtain camera parameters, depth maps, and foreground masks. We obtain the point cloud annotations by back-projecting the depth map with camera poses, and all camera poses and point clouds are in the first-frame canonical coordinates.

**2D/3D Motion Labels.** Accurate motion supervision is critical, yet ground-truth scene flow is rarely available. For all datasets, we estimate frame-to-frame optical flow using SEA-RAFT [63]. Then, inspired by zero-MSF [34], we compute large-scale, dense scene flow labels from optical flow and corresponding geometry labels. We derive dense 3D scene-flow pseudo-labels $\hat{F}_i^{3d} \in \mathbb{R}^{H\times W\times 3}$ in the first-camera canonical frame for each adjacent pair $(X_i, X_{i+1})$. We first get the forward and backward optical flow $F_{i\to i+1}^{2d}, F_{i+1\to i}^{2d} \in \mathbb{R}^{H\times W\times 2}$, and their per-pixel uncertainties $\sigma_{i\to i+1}^{2d}$, $\sigma_{i+1\to i}^{2d} \in \mathbb{R}^{H\times W}$ using SEA-RAFT [63]. For pixel $\mathbf{u} = (x, y)$ in frame $i$, define the forward-mapped pixel $\mathbf{q}(\mathbf{u}) = \mathbf{u} + F_{i\to i+1}^{2d}(\mathbf{u})$. Scene flow is then retrieved by

$$\hat{F}_i^{3d}(\mathbf{u}) = \begin{cases} P_{i+1}(\mathbf{q}(\mathbf{u})) - P_i(\mathbf{u}), & \text{if } \hat{M}_i(\mathbf{u}) = 1, \\ \mathbf{0}, & \text{otherwise.} \end{cases} \quad (5)$$

The scene flow labels generated by finding the correspondence in adjacent point clouds may contain large

Table 1. Quantitative comparison on image-to-video (I2V) generation under two splits: *General motion* and *Complex motion*. Metrics: dynamic degree (d.d.; ↑), motion smoothness (m.s.; ↑), I2V-subject/background (i2v-s./i2v-b.; ↑), subject consistency (s.c.; ↑), Fréchet Video Distance (FVD; ↓), and FID (↓). WorldReel achieves the best overall performance, with notably higher dynamic degree while maintaining strong s.c. and perceptual quality (lower FVD/FID). **Bold** indicates best; underline second-best. Gray rows denote methods that primarily focus on nearly-static scenes.

| | General motion | | | | | | | Complex motion | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| method | d.d. | m.s. | i2v-s. | i2v-b. | s.c. | FVD | FID | d.d | m.s. | i2v-s. | i2v-b. | s.c. | FVD | FID |
| Cogvideox-I2V [72] | 0.37 | 0.976 | 0.967 | 0.970 | 0.928 | 617.4 | 47.69 | 0.52 | 0.972 | 0.954 | 0.960 | 0.916 | 824.8 | 52.97 |
| 4DNeX [10] | 0.03 | **0.994** | **0.993** | **0.990** | **0.983** | 712.5 | 44.97 | 0.19 | **0.994** | **0.987** | **0.985** | **0.983** | 632.8 | 49.79 |
| DimensionX [52] | 0.47 | 0.987 | 0.974 | 0.979 | 0.943 | 470.7 | 42.28 | 0.93 | 0.980 | 0.963 | 0.970 | 0.910 | 605.3 | 46.97 |
| GeoVideo [3] | 0.54 | 0.987 | 0.979 | 0.980 | 0.932 | 371.3 | 46.78 | 0.79 | 0.985 | 0.971 | 0.974 | 0.914 | 409.9 | 49.92 |
| WorldReel (ours) | **0.73** | 0.990 | 0.986 | 0.986 | 0.953 | **336.1** | **36.58** | **1.00** | 0.987 | 0.980 | 0.980 | 0.927 | **394.2** | **44.95** |

noise. Therefore, we utilize various existing information to eliminate potentially incorrect scene flow labels. A per-pixel validity mask $M_i^{\text{flow}}$ only keeps pixels that pass foreground/instance, uncertainty, and forward-backward consistency checks (see supp.) During training, the motion validity mask is applied when calculating $\mathcal{L}_{\text{flow}}$ and $\mathcal{L}_{\text{reg}}^{\text{flow}}$. Also, to align the frame number for 2D/3D flow, we retrieve one more frame after the video clip during all data processing.

# 4. Experiments

## 4.1. Experimental Settings

**Implementation Details.** We utilize CogVideoX-5B-I2V [72] as our base video generation model, which generates videos at a resolution of $480 \times 720$ over 49 frames. Our predicted 4D dynamic scene representations are generated at the same resolution but for a downsampled set of 13 frames. All training is conducted on the same mixed dataset of synthetic and real videos described in Sec. 3.4, which also details the generation process for pseudo annotations. Following the two-stage strategy described in Sec. 3.3, we first finetune the geo-motion augmented DiT for 20K steps and separately train the temporal DPT heads from scratch for 100K steps. In the second stage, the entire model is trained end-to-end for an additional 10K steps. All training is performed on 8×H200 GPUs with a batch size of 8. We employ the AdamW optimizer [36] with a constant learning rate of 2e-5. For optimization objectives, we set the loss weight $\lambda_{\text{flow}} = 5.0$ in Eq. 2 and the final joint loss weights $\lambda_{\text{dpt}} = 0.1$ and $\lambda_{\text{reg}} = 0.5$ in Eq. 4.

**Evaluation.** To assess 4D generation for dynamic scenes, we construct two benchmarks from the SpatialVid [58] validation split. The (i) *general* motion set contains 500 randomly sampled videos, while the (ii) *complex* motion set consists of 500 videos with the highest 3D motion magnitude, representing scenes with the most complex dynamics, ensuring a rigorous test of dynamic scene modeling (see supp. for details). We compare WorldReel with the base model CogVideoX-I2V [72] and other recent 4D video generation methods: DimensionX [52], 4DNeX [10], and GeoVideo [3]. We use the released checkpoints for DimensionX [52] and 4DNeX [10], and train GeoVideo [3] on our dataset. For dynamic-scene video generation quality, we report five metrics from VBench [24]: (a) i2v-subject, (b) i2v-background, (c) subject consistency, (d) motion smoothness, and (e) dynamic degree, which jointly measure the quality and consistency of dynamics in a video. We also report FVD [54] and FID [17] for video visual quality.

For *4D geometry generation* assessment, we evaluate the geometry terms generated with the videos. We utilize ViPE [22], the state-of-the-art dynamic-scene reconstruction method, to obtain a pseudo ground truth aligned with the video. We normalize the scene's scale using the median depth. For generated depth, we report log-RMSE and $\delta$-accuracy. For generated camera trajectories, we report the camera motion magnitude using the trajectory length (sum of inter-frame translations) and the cumulative viewpoint change (sum of inter-frame rotation angles). In addition, we use standard error metrics to evaluate the camera poses generated: Absolute Translation Error (ATE), Relative Translation Error (RTE), and Relative Rotation Error (RRE).

## 4.2. Experimental Results

**Video Generation Comparisons.** Table 1 demonstrates the quantitative comparison with other baseline models. Across both the *general* and *complex* motion sets, our method shows superior performance. Our model substantially improves visual quality: relative to GeoVideo [3] that is trained on the same data, FVD drops from 371.3 to **336.1** (-9.5%) on *general* and from 409.9 to **394.2** (-3.8%) on *complex*. Most notably, our model shows a significant improvement in the dynamic degree metric, far exceeding all baselines and achieving a perfect 1.0 on the *complex* motion set. Furthermore, our method achieves leading scores on motion smoothness, i2v-subject, i2v-background, and subject consistency, indicating our ability to generate smoother and more consistent dynamics. These results demonstrate that our method significantly enhances the effective motion of both the camera and

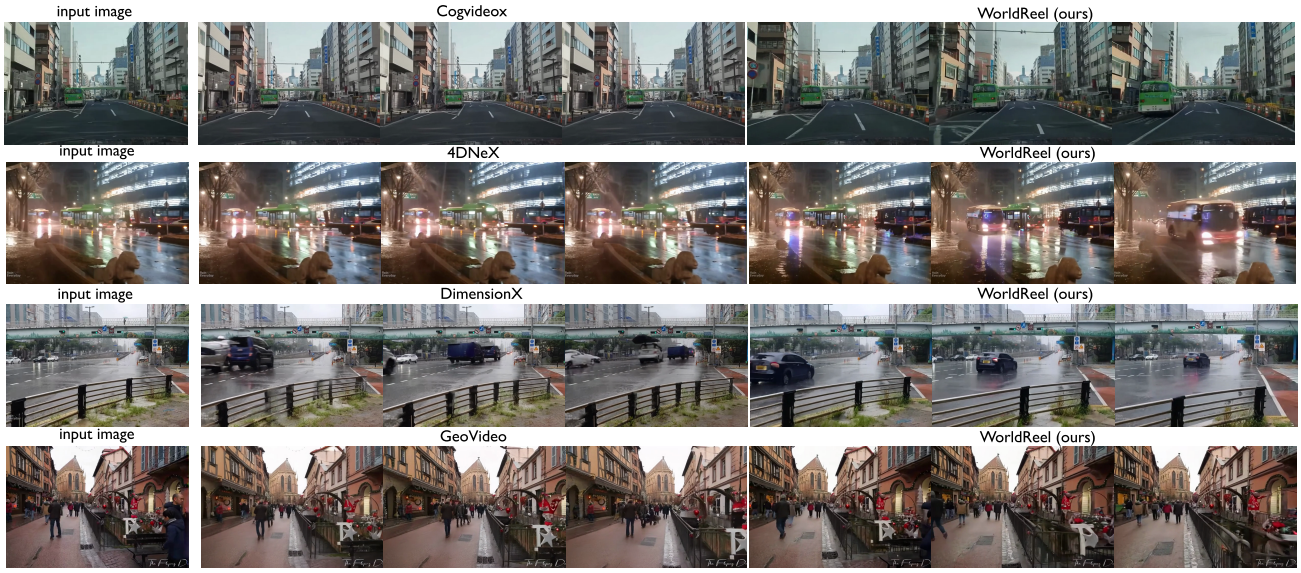


Figure 3. Qualitative image-to-video comparison on in-the-wild scenes. Given a single *input image* (left), we show sampled frames from videos generated by 4DNeX [10], DimensionX [52], GeoVideo [3], and *WorldReel* (ours). Prior methods often exhibit geometry drift and motion inconsistencies (e.g., warped facades, misaligned vehicles), while our results better preserve scene layout and maintain coherent camera and non-rigid dynamics. See the supplementary for prompts, full videos for all methods, and additional comparisons.

Table 2. Scene geometry evaluation on depth, camera pose, and camera trajectory. WorldReel achieves the best depth and camera accuracy across all pose metrics, with competitive trajectory estimates. Ablations (*w/o geomotion*, *w/o joint*) degrade depth or pose quality, confirming the importance of geo–motion latent and joint training. **Bold** = best, underline = second-best.

| | Depth | | Camera | | | Camera traj. | |
|---|---|---|---|---|---|---|---|
| method | log-rmse | $\delta_{1.25}$ | ATE | RTE | RRE | length | rotation |
| 4DNeX [10] | 0.479 | 39.9 | 0.006 | 0.017 | 0.378 | 0.034 | 0.55 |
| GeoVideo [3] | 0.353 | 63.4 | 0.011 | 0.012 | 0.443 | 0.331 | 4.61 |
| w/o geomotion | 0.352 | 67.2 | 0.010 | 0.013 | 0.372 | **0.379** | 5.34 |
| w/o joint | 0.399 | 57.6 | 0.006 | 0.014 | 0.410 | 0.294 | **5.86** |
| WorldReel | **0.287** | **71.1** | **0.005** | **0.007** | **0.317** | 0.358 | 3.83 |

the scene *without* sacrificing appearance quality or temporal consistency (see Figure 3).

While GeoVideo [3] introduces geometry modeling and regularization for consistency, our results suggest that the focus on static geometry penalizes the generation of dynamic content. In contrast, WorldReel, by jointly and explicitly modeling both geometry and motion, avoids this trade-off and prevents the model from preferring static content to maintain geometric consistency. Also, although 4DNeX [10] reports high framewise consistency (e.g., s.c.), its extremely low dynamic degree (0.03) and poor FVD (712.5) indicate a collapse toward near-static video generation.

**4D Scene Quality.** Table 2 assesses the fidelity of the generated 4D scene geometry. Because 4DNeX [10] outputs raw point clouds, we follow their pipeline to register frames and recover camera poses. Although 4DNeX attains a low camera ATE, its trajectory length and rotation are near-zero, indicating little actual camera motion, consistent with its tendency to collapse toward static scenes observed in our video metrics. In contrast, WorldReel delivers the best depth and pose accuracy across all measures, with camera parameters that faithfully match the generated videos.

Ablations further underscore our design choices: removing the joint training stage ("w/o joint") noticeably degrades both geometric and camera accuracy. This confirms that our joint optimization with targeted regularization – decoupling static structure from dynamic regions and supervising them separately – is crucial for high 4D consistency. The resulting high-quality geometry and poses highlight WorldReel's potential for realistic dynamic 4D scene generation.

**Ablation.** We present an ablation study on video generation results in Table 3. We show the effectiveness of two main designs in our method: (i) the geo-motion augmented latent for the DiT model; (ii) joint optimization of generation and 4D modeling with geometry and motion regularization. Removing the geo-motion augmented latent causes a clear drop in performance, especially on the *complex* set. Applying joint training with regularization on the RGB-only model ("w/o g.m.") leads to even worse FVD results than simply finetuning the base model ("base finetuned") on the complex set. This demonstrates that our geo-motion latents are critical for modeling complex dynamics. On the other hand, removing the joint training stage ("w/o joint") also hurts performance, which confirms that our regularization terms, applied via joint optimization, successfully align appearance and geometry, and improve motion quality. Additionally, we present the results following the GeoVideo [3] setup, where we freeze the temporal DPT heads ("freeze dpt") during joint

Table 3. Ablation study on image-to-video generation under *General* and *Complex* motion. Variants: *base finetuned*, *w/o g.m.* (without geo–motion latent), *w/o joint* (no joint multi-task decoding/regularizers), *freeze dpt* (freeze temporal DPT backbone in stage-2), and *full* (ours). Metrics: dynamic degree (d.d.; ↑), motion smoothness (m.s.; ↑), I2V-subject/background (i2v-s./i2v-b.; ↑), subject consistency (s.c.; ↑), FVD (↓), and FID (↓). Our *full* model delivers the best overall quality (lowest FID and highest d.d. on complex motion), while *freeze dpt* attains the lowest FVD; removing geo–motion latent or joint training degrades consistency. **Bold** = best, underline = second-best.

| | General motion | | | | | | | Complex motion | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| method | d.d | m.s. | i2v-s. | i2v-b. | s.c. | FVD | FID | d.d | m.s. | i2v-s. | i2v-b. | s.c. | FVD | FID |
| base finetuned | **0.90** | 0.986 | 0.974 | 0.979 | 0.921 | 383.4 | 42.68 | 0.98 | 0.985 | 0.971 | 0.976 | 0.913 | 437.0 | 52.01 |
| w/o g.m. | 0.85 | 0.989 | 0.982 | 0.983 | 0.943 | 359.2 | 42.07 | 0.93 | 0.987 | 0.975 | 0.977 | 0.918 | 452.8 | 48.02 |
| w/o joint | 0.73 | 0.989 | 0.983 | 0.984 | 0.946 | 354.5 | 40.42 | 0.96 | 0.988 | 0.978 | 0.979 | 0.926 | 411.8 | 47.2 |
| freeze dpt | 0.77 | **0.991** | 0.984 | 0.985 | **0.956** | **336.0** | 38.02 | 0.98 | **0.990** | **0.981** | **0.981** | **0.94** | **382.3** | 45.33 |
| full | 0.73 | 0.990 | **0.986** | **0.986** | 0.953 | 336.1 | **36.58** | **1.00** | 0.988 | 0.980 | **0.981** | 0.928 | 394.2 | **44.95** |

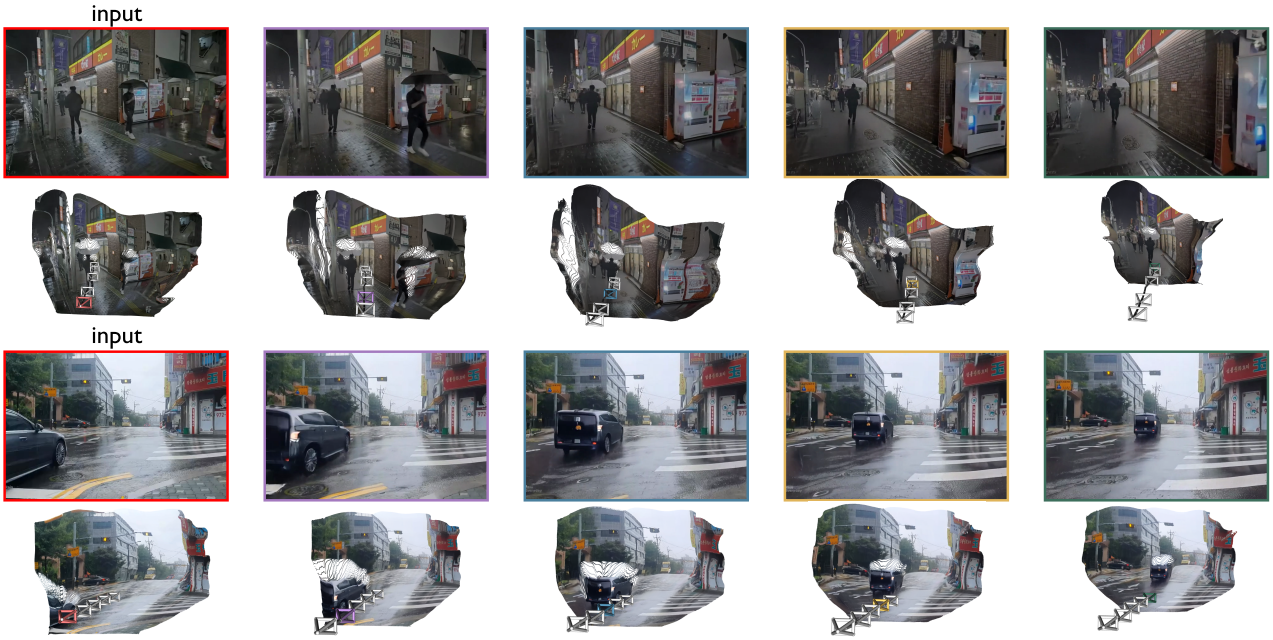


Figure 4. Qualitative 4D generation and geometry. For two in-the-wild inputs (left, red boxes), we show selected frames from our generated videos (top rows) alongside the corresponding *dynamic point clouds* rendered from our pointmaps and camera trajectories (bottom rows). The persistent structure and consistent camera/object motion illustrate a single, stable 3D scene across time, evidencing strong geometric consistency in the underlying world state. See supplementary for additional examples.

training and apply only regularization terms.

## 5. Conclusion

We have introduced WorldReel, a feed-forward 4D video generator that natively couples appearance with object geometry and dynamics. By jointly emitting RGB, pointmaps, camera trajectories, and dense (scene and optical) flow, WorldReel maintains a single persistent scene state, yielding consistent motion and stable geometry even for dynamic scenes. Trained in a mix of synthetic and real videos, WorldReel achieves state-of-the-art 4D consistency while requiring no extra input at inference time, moving video generation one step closer to editable and agent-ready world models.

**Limitations and Future Work.** WorldReel requires additional 4D supervision during training (e.g., camera, geometry, scene flow), which we currently obtain from synthetic data; Although WorldReel introduces strategies to mitigate domain gaps, domain gaps still remain that constrain generalization to unusual motion and dynamics. Also, since the temporal window is finite, failure modes appear under significant topology changes, heavy occlusions, and fast motions. We expect future work to reduce supervision by leveraging weak/self-supervised 4D signals from monocular videos, extend temporal context with streaming/causal diffusion and a persistent world state, and add controllable scene decomposition for more faithful long-horizon, interactive 4D generation.

## References


[1] Abhishek Badki, Hang Su, Bowen Wen, and Orazio Gallo. L4P: Towards unified Low-level 4D vision perception. 2025. 3

[2] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B. Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[3] Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, and Qixing Huang. Geovideo: Introducing geometric regularization into video generation model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1, 2, 6, 7

[4] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 5

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 3

[6] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025. 2, 3

[7] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2

[8] 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, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 5

[9] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3r: Estimating disentangled motion from dust3r without training. *arXiv preprint arXiv:2503.24391*, 2025. 3

[10] 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. 1, 2, 3, 5, 6, 7

[11] Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. *arXiv preprint arXiv:2509.21657*, 2025. 2

[12] Weichen Fan, Chenyang Si, Junhao Song, Zhenyu Yang, Yinan He, Long Zhuo, Ziqi Huang, Ziyue Dong, Jingwen He, Dongwei Pan, et al. Vchitect-2.0: Parallel transformer for scaling up video diffusion models. *arXiv preprint arXiv:2501.08453*, 2025. 5

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

[14] Jisang Han, Honggyu An, Jaewoo Jung, Takuya Narihira, Junyoung Seo, Kazumi Fukuda, Chaehyun Kim, Sunghwan Hong, Yuki Mitsufuji, and Seungryong Kim. Dˆ 2ust3r: Enhancing 3d reconstruction with 4d pointmaps for dynamic scenes. *arXiv preprint arXiv:2504.06264*, 2025. 3

[15] Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. Diffit: Diffusion vision transformers for image generation. In *European Conference on Computer Vision*, pages 37–55. Springer, 2024. 2

[16] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 2

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022. 2

[20] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2

[21] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2005–2015, 2025. 4

[22] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception. In *NVIDIA Research Whitepapers*, 2025. 5, 6

[23] Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson WH Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025. 2

[24] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive

benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 6

[25] Hyeonho Jeong, Chun-Hao P Huang, Jong Chul Ye, Niloy J Mitra, and Duygu Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7276–7287, 2025. 2, 3

[26] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4d: Leveraging video generators for geometric 4d scene reconstruction, 2025. 1, 3

[27] Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, and Aleksander Hołyński. Stereo4d: Learning how things move in 3d from internet stereo videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10497–10509, 2025. 3

[28] Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and Sunghyun Cho. Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2040–2049, 2025. 3

[29] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. *CVPR*, 2023. 5

[30] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9492–9502, 2024. 4

[31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[32] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 2, 3

[33] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10486–10496, 2025. 5

[34] Yiqing Liang, Abhishek Badki, Hang Su, James Tompkin, and Orazio Gallo. Zero-shot monocular scene flow estimation in the wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21031–21044, 2025. 5

[35] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 3

[36] Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5 (5):5, 2017. 6

[37] Hyelin Nam, Jaemin Kim, Dohun Lee, and Jong Chul Ye. Optical-flow guided prompt optimization for coherent video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7837–7846, 2025. 3

[38] Zijie Pan, Zeyu Yang, Xiatian Zhu, and Li Zhang. Efficient4d: Fast dynamic 3d object generation from a single-view video. *arXiv preprint arXiv 2401.08742*, 2024. 2, 3

[39] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2

[40] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 3

[41] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. *arXiv preprint arXiv:2011.13961*, 2020. 3

[42] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 5

[43] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 2, 3

[44] 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 *Proceedings of Neural Information Processing Systems(NeurIPS)*, 2024. 3

[45] 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*, 2025. 2, 3

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[48] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[49] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 3

[50] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2, 3

[51] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, and Yaniv Taigman. Text-to-4d dynamic scene generation. *arXiv:2301.11280*, 2023. 3

[52] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. In *International Conference on Computer Vision (ICCV)*, 2025. 2, 3, 6, 7

[53] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 3

[54] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

[55] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in neural information processing systems*, 34:11287–11302, 2021. 2

[56] 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. 1, 2, 3

[57] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 3, 4

[58] Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, Xiaoxiao Long, Hao Zhu, Zhaoxiang Zhang, Xun Cao, and Yao Yao. Spatialvid: A large-scale video dataset with spatial annotations, 2025. 5, 6

[59] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023. 3

[60] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state, 2025. 3

[61] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3

[62] Shijie Wang, Samaneh Azadi, Rohit Girdhar, Saketh Rambhatla, Chen Sun, and Xi Yin. Motif: Making text count in image animation with motion focal loss. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7773–7783, 2025. 3

[63] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, pages 36–54. Springer, 2024. 4, 5

[64] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078, 2025. 2

[65] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, 2024. 3

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

[67] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7623–7633, 2023. 2

[68] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26057–26068, 2025. 3

[69] Dianbing Xi, Jiepeng Wang, Yuanzhi Liang, Xi Qiu, Yuchi Huo, Rui Wang, Chi Zhang, and Xuelong Li. Omnivdiff: Omni controllable video diffusion for generation and understanding. *arXiv preprint arXiv:2504.10825*, 2025. 2

[70] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2024. 3

[71] Tian-Xing Xu, Xiangjun Gao, Wenbo Hu, Xiaoyu Li, Song-Hai Zhang, and Ying Shan. Geometrycrafter: Consistent geometry estimation for open-world videos with diffusion priors. *arXiv preprint arXiv:2504.01016*, 2025. 3, 5

[72] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 2, 3, 6

[73] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 3, 5

[74] Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista Martin, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21685–21695, 2025. 2

[75] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and

Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2

[76] Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan, Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim Hee Lee, and Lijuan Wang. Genxd: Generating any 3d and 4d scenes. In *ICLR*, 2025. 2, 3

[77] Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: Learning 4d embodied world models. 2025. 3

[78] 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*, pages 19855–19865, 2023. 5

[79] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 3

[80] 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, 2025. 5

[81] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025. 3