4DNeX: Feed-Forward 4D Generative Modeling Made Easy

Anonymous ICCV submission

Paper ID 5

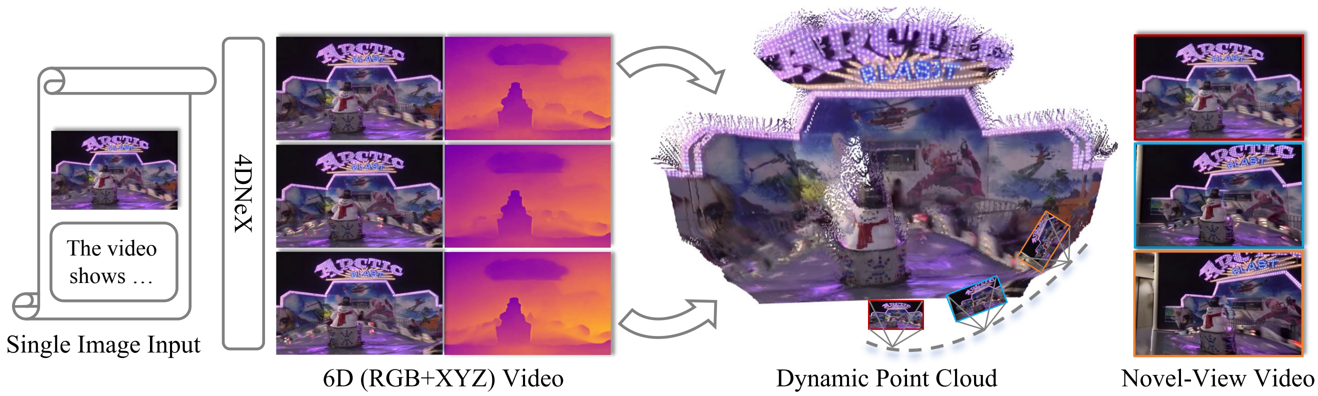

Figure 1. **4DNeX can generate dynamic point clouds from a single image.** By modeling dynamic 3D point clouds as learned representations of RGB and XYZ, 4DNeX effectively leverages priors from existing video generation models to achieve high-quality results. The resulting dynamic point clouds support downstream applications such as novel-view video synthesis.

Abstract

*We present **4DNeX**, the first feed-forward framework for generating 4D (i.e., dynamic 3D) scene representations from a single image. In contrast to existing methods that rely on computationally intensive optimization or require multi-frame video inputs, 4DNeX enables efficient, end-to-end image-to-4D generation by fine-tuning a pretrained video diffusion model. Specifically, **1)** to alleviate the scarcity of 4D data, we construct 4DNeX-10M, a large-scale dataset with high-quality 4D annotations generated using advanced reconstruction approaches. **2)** we introduce a unified 6D video representation that jointly models RGB and XYZ sequences, facilitating structured learning of both appearance and geometry. **3)** we propose a set of simple yet effective adaptation strategies to repurpose pretrained video diffusion models for 4D modeling. 4DNeX produces high-quality dynamic point clouds that enable novel-view video synthesis. Extensive experiments demonstrate that 4DNeX outperforms existing 4D generation methods in efficiency and generalizability, offering a scalable solution for image-to-4D modeling and laying the foundation for generative 4D world models simulating dynamic scene evolution.*

1. Introduction

The images we capture are 2D projections of the 4D (*i.e.*, dynamic 3D) physical world. Creating a 4D scene from such 2D observations, particularly from a single image, is a highly challenging yet compelling task. As a core capability in generative modeling, image-to-4D generation lays the foundation for building 4D world models that can predict and simulate dynamic scene evolution, enabling applications in AR/VR, film production, and content creation.

Existing approaches for 4D scene modeling can be broadly classified into two categories. The first comprises 4D generation methods, which typically adopt representations such as Neural Radiance Fields (NeRF) [17] or 3D Gaussian Splatting (3DGS) [10]. These methods can be further divided into feed-forward [22, 26, 35, 46] and optimization-based variants [1, 16, 21, 38, 45, 48]. However, they either require video input or rely on object-centric, computationally intensive optimization procedures. The second category includes dynamic Structure-from-Motion (SfM) approaches [9, 12, 31, 36, 42], which estimate dynamic 3D structures such as time-varying point clouds from video sequences. However, these methods cannot generate 4D representations from a single image.

To this end, we aim to develop a feed-forward framework for 4D scene generation from a single image. A straightforward solution is to fine-tune a pretrained video diffusion model. However, this approach faces two core challenges: **1)** how to mitigate the scarcity of 4D data, and **2)** how to adapt the pretrained model in a simple and efficient way.

For the **first** challenge, we curate 4DNeX-10M, a large-scale dataset comprising both static and dynamic scenes, with high-quality 4D annotations generated from monocular videos using state-of-the-art reconstruction methods [12, 30, 32, 33, 42]. To ensure geometric accuracy and scene diversity, we apply careful data selection, pseudo-annotation generation, and multi-stage filtering. To address the **second** challenge, we first introduce a unified 6D video representation that models RGB and XYZ sequences jointly, enabling the structured modeling of both appearance and geometry. We then systematically investigate different fusion strategies between the two modalities and show that width-wise fusion achieves the most effective cross-modal alignment. Moreover, we incorporate a set of carefully designed techniques, including XYZ initialization, XYZ normalization, mask design, and modality-aware token encoding, to adapt pretrained video diffusion models in a simple manner while preserving their generative priors.

To summarize, we present 4DNeX, the first feed-forward framework for image-to-4D generation (Fig. 1). We qualitatively demonstrate the plausibility of the generated dynamic point clouds. Furthermore, to validate their utility, we leverage TrajectoryCrafter [39] to transform the generated 4D point clouds into novel-view videos, achieving comparable results to existing 4D generation methods. In addition, we perform comprehensive ablation studies to validate the effectiveness of our proposed fine-tuning strategies.

Our main contributions can be summarized as follows:

- We propose 4DNeX, the first feed-forward framework for image-to-4D generation, capable of producing dynamic point clouds from a single image.
- We construct 4DNeX-10M, a large-scale dataset with high-quality 4D annotations.
- We introduce simple yet effective strategies to adapt pretrained video diffusion models for 4D generation.

2. Related Work

2.1. Optimization-based 4D Generation

Recent methods leverage pre-trained diffusion models to optimize 3D and 4D representations [10, 17, 34] via score distillation sampling [19] or multi-view synthesis. A major challenge is maintaining spatio-temporal consistency. Some approaches [1, 5, 8, 13, 20, 38, 41, 45, 48] start from static 3D representations and incorporate motion using video diffusion priors. Others [16, 18, 21, 23, 25, 37] begin with generated videos and enforce multi-view consistency to reconstruct 4D content. Beyond consistency issues,

Figure 2. **Visualization of 4DNeX-10M Dataset.** Our dataset spans a wide range of dynamic scenarios, including indoor, outdoor, close-range, far-range, static, high-speed, and human-centric scenes. The word cloud summarizes common visual concepts captured in the dataset, while the 4D point clouds and camera trajectories demonstrate the spatial precision of our pseudo-annotations.

these optimization-based techniques are typically computationally expensive, slow, and unstable due to multi-stage training. In this work, we propose a feed-forward framework to efficiently generate 4D representations, improving scalability and speed.

2.2. Feed-forward 4D Generation

Feed-forward 4D generation methods directly produce 4D representations in a single pass, offering efficiency and stability over optimization-based approaches. Existing works either generate multi-view videos with implicit geometry [2, 11, 26, 39, 40, 46] or output explicit 4D representations [22, 35] that struggle to generalize beyond specific data sources. Specialized methods like TesserAct [47] target robotic applications, while dynamic SfM techniques [6, 9, 12, 31, 36, 42] reconstruct geometry from videos rather than generating from images. In contrast, we propose a general-purpose framework that efficiently generates full 4D representations from a single image.

3. 4DNeX-10M

To address the data scarcity in 4D generative modeling, we introduce 4DNeX-10M, a large-scale hybrid dataset tailored for training feed-forward 4D generative models. It aggregates videos from public sources and internal pipelines, encompassing both static and dynamic scenes. All data undergoes rigorous filtering, pseudo-annotation, and quality assessment to ensure geometric consistency, motion diversity, and visual realism. As shown in Figure 2, our proposed dataset encompasses a highly diverse range of scenes, including indoor and outdoor environments, distant landscapes, close-range settings, high-speed scenarios, static scenes, and human-inclusive situations. Furthermore, 4DNeX-10M encompasses a wide variety of lighting conditions and a profusion of human activities. Meanwhile, we provide precise 4D pointmaps and camera trajectories of these corresponding scenes. In total, 4DNeX-10M contains over 9.2 million video frames with pseudo annotations. For data curation, as illustrated in Figure 3, we curate this data using an automated acquisition and filtering pipeline comprising several stages: 1) data cleaning, 2) data captioning, and 3) 3D/4D annotation.

3.1. Data Preprocessing

Data Sources. We collect monocular videos from several sources. DL3DV-10K (DL3DV) [14] and RealEstate10K (RE10K) [49] offer static indoor and outdoor videos with diverse camera trajectories. The Pexels dataset provides a large pool of human-centric stock videos with auxiliary metadata such as movements, OCR, and optical flow. The Vimeo Dataset, selected from [4], contributes in-the-wild dynamic scenes. Synthetic data sourced from [7] contains dynamic sequences using video diffusion models (VDM).

Initial Filtering. For large-scale sources like Pexels, we apply metadata filtering, including optical flow, motion, and OCR, to eliminate non-compliant videos, such as those exhibiting excessive motion blur or text-saturated videos. Across all data sources, brightness filtering is applied based on average luminance ($0.299R + 0.587G + 0.114B$) to discard videos with extreme illumination conditions.

Video Captioning. For datasets without textual annotations (*e.g.*, DL3DV-10K and RE-10K), we use LLaVA-Next-Video [44] to generate captions. We sample 32 frames uniformly per video (or clip) and feed them to the LLaVA-NeXT-Video-7B-Qwen2 model with the prompt: *"Please provide a concise description of the video, focusing on the main subjects and the background scenes."* For scenes with consistent content (*e.g.*, DL3DV-10K, Dynamic Replica), we generate one caption per video. For RealEstate10K, we split each video into clips and caption them separately.

3.2. Static Data Processing

To learn strong geometric priors, we curate static monocular videos from DL3DV-10K [14] and RE-10K [49]. These cover a wide range of environments including homes, streets, stores, and landmarks, with varied camera trajectories providing rich multi-view coverage.

Pseudo 3D Annotation. As these datasets lack 3D ground-truth, we employ DUSt3R [32], a stereo reconstruction model, to generate pseudo point maps. For each video, DUSt3R is applied exhaustively over view pairs to form a view graph, followed by global fusion (per the original paper) to recover a consistent scene-level 3D structure.

Quality Filtering. To ensure high-quality annotations, we define two metrics using the confidence maps from DUSt3R: 1) the *Mean Confidence Value (MCV)*, averaging pixel-wise confidence scores over all frames, and 2) the *High-Confidence Pixel Ratio (HCPR)*, representing the proportion of pixels exceeding a threshold τ. We select the top-$r\%$ of clips for each metric and retain over 100K high-quality 28-frame clips with reliable pseudo point map annotations for static training.

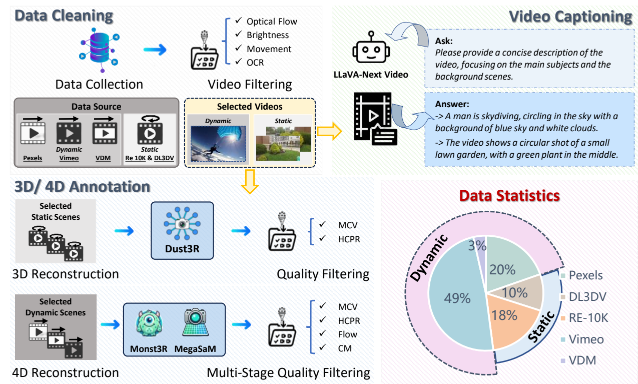

Figure 3. **Data Curation Pipeline.** The video data is collected from various sources and then selected by video filtering during Data Cleaning. The selected data is captioned via LLaVA-Next-Video model in Video Captioning. The selected data is processed and finally filtered out the video with high-quality annotation during 3D/4D Annotation. Data statistics is provided in bottom right.

3.3. Dynamic Data Processing

To enrich 4DNeX-10M with dynamic content, we collect monocular videos from Pexels, VDM, and Vimeo. These datasets contain diverse real-world scenes with motion and depth variation but lack ground-truth geometry.

Pseudo 4D Annotation. We employ MonST3R [42] and MegaSaM [12], two advanced dynamic reconstruction models, to generate pseudo 4D annotations. Each model recovers temporally coherent 3D point clouds and globally aligned camera poses from monocular videos, enabling the construction of time-varying scene representations.

Multi-Stage Filtering. To select high-quality clips, we apply three sequential filtering strategies. First, we use the final alignment loss in the global fusion stage, which reflects multi-view consistency and flow agreement with RAFT [28], to filter out low-quality reconstructions. Second, we assess camera smoothness (CS) by computing frame-wise velocity and acceleration from camera translations, and estimate local trajectory curvature as:

$$\kappa_i = \frac{\|\mathbf{v}_{i+1} - \mathbf{v}_i\|}{\|\mathbf{v}_{i+1}\|^2 + \|\mathbf{v}_i\|^2 + \epsilon}, \quad \epsilon > 0. \tag{1}$$

Clips with low average velocity, acceleration, and curvature are retained. Third, we apply the same *Mean Confidence Value (MCV)* and *High-Confidence Pixel Ratio (HCPR)* used in the static pipeline. After filtering, we retain approximately 32K clips from the MonST3R-processed set,

ICCV #5

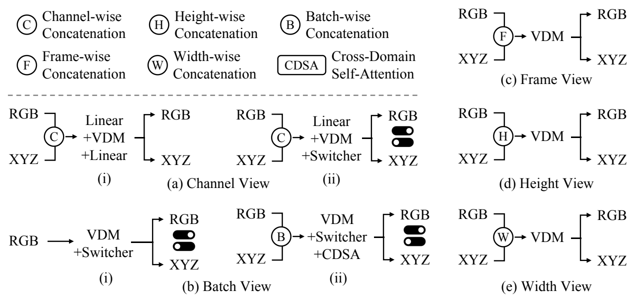

Figure 4. **Comparison of fusion strategies for joint RGB and XYZ modeling.** We explore five fusion strategies and analyze their impact on model compatibility and cross-modal alignment.

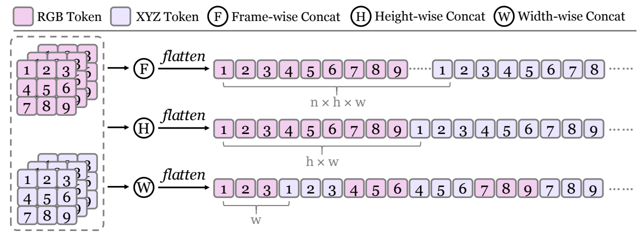

Figure 5. **Comparison of spatial fusion strategies.** We compare frame-, height-, and width-wise fusion in terms of the interaction distance between RGB and XYZ tokens.

5K clips from VDM, and 27K from Pexels, and over 80k clips from MegaSaM-processed set. Together, these yield a total over 110K high-quality clips with pseudo 4D annotations, enabling robust modeling of dynamic 3D scenes across a wide range of motions and appearances.

4. 4DNeX

4.1. Problem Formulation

Given a single image $I_0 \in \mathbb{R}^{H \times W \times 3}$, we aim to construct a 4D (*i.e.*, dynamic 3D) representation of the underlying scene geometry. This task can be formulated as learning a conditional distribution over dynamic point clouds:

$$p\left(\{P_t\}_{t=0}^{T-1} \mid I_0\right), \tag{2}$$

where $\{P_t\}_{t=0}^{T-1}$ denotes the sequence of dynamic point clouds. However, directly modeling point clouds is challenging due to their highly unstructured nature. To address this, inspired by [43], we use a pixel-aligned point map representation, *XYZ*, where each frame $X_t^{\text{XYZ}} \in \mathbb{R}^{H \times W \times 3}$ encodes the 3D coordinates of each pixel in the global coordinates. This format provides a structured and learnable structure, making it compatible with existing generative models. Instead of directly modeling $\{P_t\}$, we reformulate the problem as predicting paired RGB and XYZ image sequences:

$$p\left(\{X_t^{RGB},\ X_t^{XYZ}\}_{t=0}^{T-1} \mid I_0\right). \tag{3}$$

Accordingly, the joint distribution can be also factorized as:

$$p\left(\{X_t^{RGB}\}_{t=0}^{T-1},\ \{X_t^{XYZ}\}_{t=0}^{T-1} \mid I_0\right). \tag{4}$$

Therefore, a 4D scene can be effectively represented using a 6D video composed of paired RGB and XYZ sequences. This simple and unified representation offers two key advantages: it enables explicit 3D consistency supervision through pixel-aligned XYZ maps, and eliminates the need for camera, facilitating scalable and robust 4D generation.

To model this distribution, we adopt Wan2.1 [29], a video diffusion model trained under the flow matching [15] framework. We extend its image-to-video capability to generate 6D videos as $V = \{X_t^{RGB},\ X_t^{XYZ}\}_{t=0}^{T-1}$. V is first encoded into a latent space via a VAE encoder $\mathcal{E}$: $x_1 = \mathcal{E}(V)$, and interpolating with a noise latent $x_0 \sim \mathcal{N}(0, I)$:

$$x_t = (1-t)x_0 + t x_1, \quad t \sim \mathcal{U}(0,1). \tag{5}$$

And a velocity predictor u is trained to regress the velocity between endpoints:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\left[\left\| u(x_t, c_{\text{img}}, c_{\text{txt}}, t) - (x_1 - x_0)\right\|^2\right], \tag{6}$$

where c_{img} and c_{txt} denote image and text condition embeddings. This formulation enables efficient learning of temporally coherent and geometrically consistent 6D videos.

4.2. Fusion Strategies

To finetune the video diffusion model for joint RGB and XYZ generation, a key challenge is designing an effective fusion strategy that enables the model to leverage both modalities. Our goal is to exploit the strong priors of pretrained models through simple yet effective fusion designs. Latent concatenation is a widely adopted technique for joint modeling. We systematically explore fusion strategies across different dimensions, as illustrated in Fig. 4.

Channel-wise Fusion. A straightforward approach is to concatenate RGB and XYZ along the channel dimension, and insert a linear layer (*a.i*) or a modality switcher (*a.ii*) to adapt the input and output formats. However, this strategy disrupts the input and output distributions expected by the pretrained model, which undermines the benefits of pretraining. It requires large-scale data and substantial computational resources to achieve satisfactory performance.

Batch-wise Fusion. To maintain pretrained distributions, this strategy treats RGB and XYZ as separate samples and uses a switcher to control the output modality (*b.i*). While it preserves unimodal performance, it fails to establish cross-modal alignment. Even with additional cross-domain attention layers (*b.ii*), the modalities remain poorly correlated.

Frame-/Height-/Width-wise Fusion. These strategies concatenate RGB and XYZ along the frame (*c*), height (*d*),

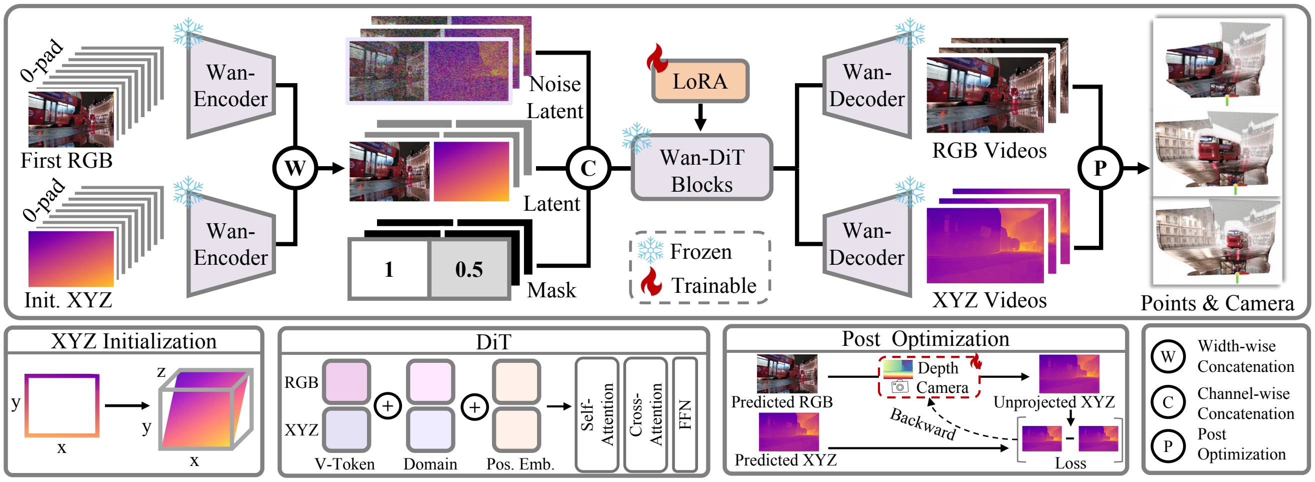

Figure 6. **Overview of 4DNeX.** Given a single image and an initialized XYZ map, 4DNeX encodes both with a VAE encoder and fuses them via width-wise concatenation. The fused latent, combined with a noise latent and a guided mask, is processed by a LoRA-tuned Wan-DiT model to jointly generate RGB and XYZ videos. A lightweight post-optimization recovers camera and depths from the outputs.

or width (e) dimensions, preserving the distributions of the pretrained model while enabling cross-modal interaction within a single sample. We analyze them from the perspective of token interaction distance. Intuitively, shorter interaction distance between corresponding tokens makes it easier for the model to learn cross-modal alignment. As shown in Fig. 5, width-wise fusion yields the shortest interaction distance, leading to more effective alignment and higher generation quality, as confirmed in Sec. 5.3.

4.3. Network Architecture

As illustrated in Fig. 6, our framework takes a single image $I_0 \in \mathbb{R}^{H\times W\times 3}$ and an initialized XYZ map $X^{init} \in \mathbb{R}^{H\times W\times 3}$ as conditions. Both are encoded by a frozen VAE encoder and concatenated along the width dimension. This fused condition is then combined with a noise latent x_t and a binary mask M along the channel dimension, and fed into a pretrained DiT with LoRA tuning. The output latent is decoded by a VAE decoder to generate paired RGB and XYZ video sequences. A lightweight post-optimization step further recovers camera and depths from the predicted outputs.

XYZ Initialization. We initialize the first-frame XYZ map X^{init} using a sloped depth plane. Specifically, we define a normalized 2D coordinate grid over the range $[-1,1]^2$ and compute the initial XYZ values as:

$$X^{init}_{i,j} = \left(\frac{2j}{W-1}-1,\ \frac{2i}{H-1}-1,\ \frac{2i}{H-1}-1\right). \quad (7)$$

This results in a sloped plane where depth values gradually increase from the bottom to the top of the image, reflecting common depth priors in natural scenes (*e.g.*, sky regions appearing farther away). Such initialization provides a stable starting point for geometry learning.

XYZ Normalization. Since the VAE is pretrained on RGB images, directly encoding XYZ inputs with different distributions can cause instability and suboptimal performance. To mitigate this issue, inspired by [3] , we apply a modality-aware normalization strategy to adapt the XYZ latent to the pretrained VAE's distributional priors. Specifically, we compute the mean μ and standard deviation σ of XYZ latent across the training dataset, and normalize the encoded representation as:

$$\hat{x} = \frac{x-\mu}{\sigma}, \quad (8)$$

where x denotes the XYZ latent. Before passing into the VAE decoder, we perform de-normalization to recover the original scale:

$$x = \hat{x}\cdot\sigma + \mu. \quad (9)$$

Mask Design. Following [29], we introduce a guided mask $M \in [0,1]^{T\times H\times W}$, where $M_{t,i,j} = 1$ indicates a known pixel and $M_{t,i,j} = 0$ indicates a pixel to be generated. Since we use an approximate initialization for the first-frame XYZ map, we assign a soft mask:

$$M^{XYZ}_{0,i,j} = 0.5, \quad \forall i,j, \quad (10)$$

which encourages the model to refine the initial geometry.

Modality-Aware Token Encoding. To preserve pixel-wise alignment across modalities during joint modeling, we adopt a shared rotary positional encoding (RoPE) [24] for RGB and XYZ tokens. To further distinguish their semantic differences, we introduce a learnable domain embedding. Given RGB and XYZ token sequences $x^{\text{RGB}}, x^{\text{XYZ}} \in \mathbb{R}^{L\times D}$, we apply the following encoding:

$$\begin{aligned} x^{RGB} &\leftarrow \mathrm{RoPE}(x^{RGB}) + e_{RGB}, \\ x^{XYZ} &\leftarrow \mathrm{RoPE}(x^{XYZ}) + e_{XYZ}, \end{aligned} \quad (11)$$

Table 1. **4D Generation Results on VBench [7].** We report the consistency, dynamics, and aesthetics of the generated videos, together with the inference time of each method.

Method	Consistency	Dynamic	Aesthetic	Time
4Real [38]	95.7%	32.3%	50.9%	90min
Free4D [16]	96.0%	47.4%	64.7%	60min
Ours	96.4%	58.0%	59.5%	15min
Animate124 [45]	90.7%	45.4%	42.3%	\
Free4D [16]	96.9%	40.1%	60.5%	60min
Ours	97.2%	58.3%	53.0%	15min
GenXD [46]	89.8%	98.3%	38.0%	\
Free4D [16]	96.8%	100.0%	57.9%	60min
Ours	96.8%	100.0%	52.4%	15min

where $\mathrm{RoPE}(\cdot)$ denotes the shared rotary positional encoding, and $e_{RGB}, e_{XYZ} \in \mathbb{R}^{1\times D}$ are learnable domain embeddings broadcasted across the sequence.

Post-Optimization. Since our method produces XYZ videos that represent dense 3D points in global coordinates, we can recover the corresponding camera parameters $C = (R, t, K)$ and depth maps d for the generated RGB frames via a lightweight post-optimization step. Specifically, we minimize the reprojection error between the generated and back-projected 3D coordinates:

$$\min_{R,t,K,d} \sum_{i,j} \left\| \tilde{q}^{XYZ}_{i,j} - \hat{q}^{XYZ}_{i,j} \right\|_2^2, \tag{12}$$

where $\hat{q}^{XYZ}_{i,j}$ denotes the generated 3D coordinate, and $\tilde{q}^{XYZ}_{i,j}$ is computed by back-projecting depth into 3D space:

$$\tilde{q}^{XYZ}_{i,j} = [R \mid t]^{-1} K^{-1} \left(d_{i,j} \cdot [i, j, 1]^\top \right). \tag{13}$$

This optimization is efficient and can be parallelized across views, producing physically plausible and geometrically consistent estimates of camera poses and depth maps.

5. Experiments

5.1. Setting

Baselines. Following [16], we compare our method with existing 4D generation approaches, which can be grouped into two categories: text-to-4D and image-to-4D methods. For text-to-4D, we compare against 4Real [38], a state-of-the-art method in this category. For image-to-4D, we benchmark against the state-of-the-art Free4D [16], the feed-forward method GenXD [46], and the object-levle approach Animate124 [45]. For text-to-4D methods, we first generate an image from the input text prompt and then convert it into the image-to-4D setting. To ensure fairness, we use the same single-image or text prompt during evaluation.

Table 2. **User study results.** Percentages indicate user preference.

Comparison	Consistency	Dynamic	Aesthetic
Ours / Free4D [16]	56% / 44%	59% / 41%	53% / 47%
Ours / 4Real [38]	79% / 21%	85% / 15%	93% / 7%
Ours / Ani124 [45]	75% / 25%	56% / 44%	100% / 0%
Ours / GenXD [46]	90% / 10%	85% / 15%	100% / 0%

Datasets and Metrics. We conduct evaluations on a collection of images and texts sourced from the official project pages of the compared methods. To assess the quality of generated novel-view videos, We report standard VBench metrics [7], including Consistency (averaged over subject and background), Dynamic Degree, and Aesthetic Score. Given the lack of a well-established benchmark for 4D generation, we further conduct a user study involving 23 evaluators to enhance the reliability of our evaluation.

Implementation Details. We opt for the vanilla Wan2.1 [29] image-to-video model as our final base model with a total of 14B parameters. For the modality-aware normalization, we trace the statistics (mean and standard deviation) of XYZ domain in the latent space over 5K random samples from the training dataset. It results in $\mu = -0.13$ and $\sigma = 1.70$, which serves as the constant normalization term for XYZ latent during training and inference. To fully transfer the capability of original image-to-video generation from the base model to the target image-to-4D task, we train a LoRA with a rank of 64 for the sake of parameter and data efficiency instead of full-parameter supervised finetuning. The Lora finetuning is run with a batch size of 32 using an AdamW optimizer. The learning rate is set to 1×10^{-4} with a cosine learning rate warmup. The training is distributed on 32 NVIDIA A100 GPUs with 5k iterations at a spatial resolution of 480×720 for each modality. To generate novel-view videos, we first produce a 4D point cloud representation of the scene using our feed-forward model, and then render the results using [39].

5.2. Main Results

4D Geometry Generation. As illustrated in Fig. 7, we visualize the paired RGB and XYZ video generated from a single image. The results demonstrate that our method can simultaneously infer plausible scene motion and the corresponding 4D geometry from a single image. This high-quality geometric representation of dynamic scenes is essential for consistent and photorealistic novel view synthesis in the subsequent rendering stage.

Novel-View Video Generation. Quantitative results on VBench [7] are presented in Table 1. Our method achieves performance comparable to state-of-the-art approaches, and notably outperforms others in terms of Dynamic Degree. Free4D [16] benefits from the proprietary Kling [27] model

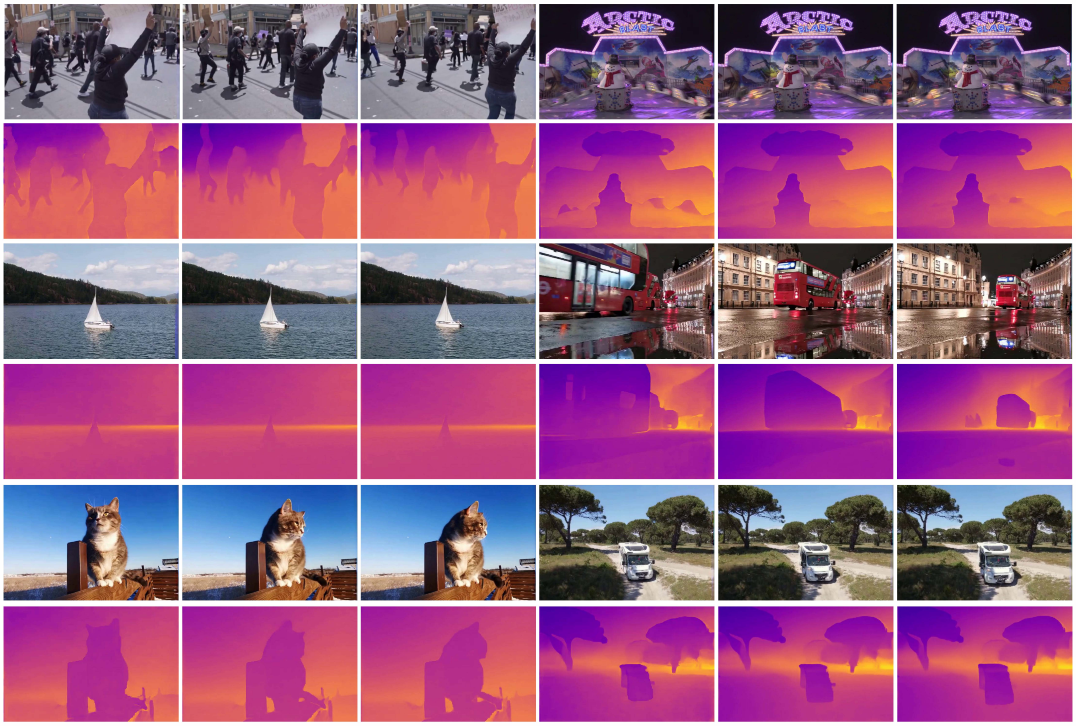

Figure 7. **Generated RGB and XYZ sequences from single-image input.** Each pair shows RGB and XYZ video sequence.

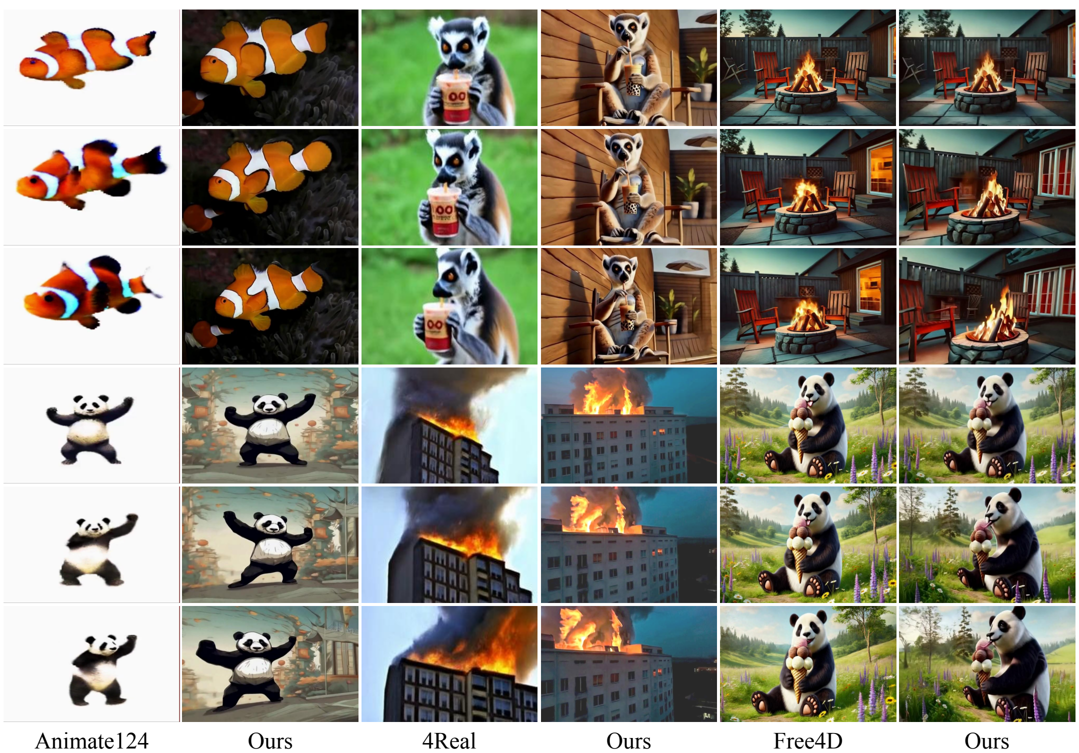

Figure 8. **Qualitative comparison.** Our method generates results with higher consistency, better aesthetics, and notably larger motion.

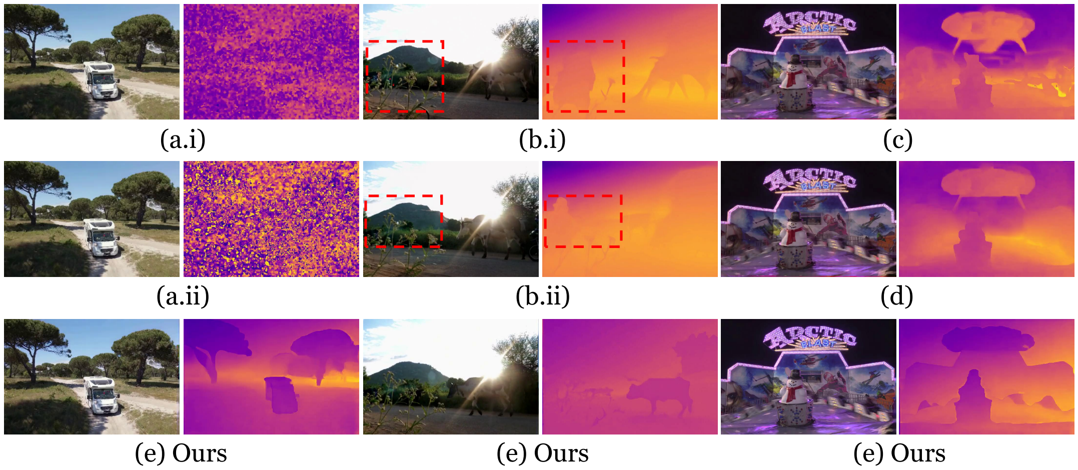

Figure 9. **Ablation study on fusion strategies.** We compare channel-wise (a), batch-wise (b), frame-wise (c), height-wise (d), and our width-wise fusion (e) for RGB and XYZ inputs.

for image animation, which contributes to its higher aesthetic scores. Qualitative comparisons are shown in Fig. 8, where our results demonstrate more significant and coherent scene dynamics, especially under camera motion. Furthermore, user study results (Table 2) show that our method is consistently preferred over most baselines in terms of consistency, dynamics, and aesthetics. Although the results are comparable to Free4D, it is important to note that the evaluation was conducted on the Free4D test set, which predominantly features object-centric scenes. In contrast, our method generalizes effectively to diverse, in-the-wild scenarios, as shownw in the supplementary material. In addition, our method is feed-forward and highly efficient, capable of generating a 4D scene within 15 minutes. By comparison, Free4D relies on a time-consuming pipeline, typically requiring over one hour to produce results.

5.3. Ablations and Analysis

To validate the effectiveness of our used width-wise fusion strategy and support the analysis presented in Sec. 4.2, we conduct an ablation study comparing five different fusion designs, as illustrated in Fig. 9. Among these, channel-wise fusion introduces a severe distribution mismatch with the pretrained prior, often leading to noisy or failed predictions (*a.i-a.ii*). Batch-wise fusion preserves unimodal quality but fails to capture cross-modal alignment, yielding inconsistent RGB-XYZ correlation (*b.i-b.ii*). Frame-wise (*c*) and height-wise (*d)* strategies provide moderate improvements, yet still suffer from suboptimal alignment and visual quality. In contrast, our width-wise fusion brings corresponding RGB and XYZ tokens closer in the sequence, shortening the cross-modal interaction distance. This facilitates more effective alignment and yields sharper, more consistent geometry and appearance across frames, as shown in Fig. 9 (e).

6. Conclusion

We propose 4DNeX, the first feed-forward model for single-image 4D scene generation. Our approach fine-tunes a pretrained video diffusion model to enable efficient image-to-4D generation. To tackle data scarcity, we introduce 4DNeX-10M, a large-scale dataset with pseudo-4D labels. We also design a unified 6D video representation that jointly encodes appearance and geometry, alongside effective adaptation strategies to repurpose video diffusion models for 4D tasks. Experiments show that 4DNeX generates high-quality point clouds, serving as a strong geometric basis for novel-view videos synthesis. It achieves competitive results with higher efficiency and better generalization, advancing scalable 4D world modeling from single images.

Limitations and Future Work While 4DNeX demonstrates promising results in single-image 4D generation, several limitations remain. First, our method relies on pseudo-4D annotations for supervision, which may introduce noise or inconsistencies, particularly in fine-grained geometry or long-term temporal coherence. Introducing high-quality real-world or synthetic dataset would be fruitful for general 4D modeling. Second, although the image-driven generated results are 4D-grounded, controllabilities over lighting, fine-grained motion and physical property are still lacking. Third, the unified 6D representation assumes relatively clean input images and may degrade under occlusions, extreme lighting conditions, or cluttered backgrounds. Future work includes improving temporal modeling with explicit world priors, incorporating real-world 4D ground-truth data when available, and extending our framework to handle multi-object or interactive scenes. Additionally, integrating multi-modal inputs like text or audio could further enhance controllability and scene diversity.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas J. 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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 7996–8006. IEEE, 2024. 1, 2

[2] Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. *CoRR*, abs/2412.07760, 2024. 2

[3] Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, et al. 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26576–26586, 2025. 5

[4] 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. 3

[5] Quankai Gao, Qiangeng Xu, Zhe Cao, Ben Mildenhall, Wenchao Ma, Le Chen, Danhang Tang, and Ulrich Neumann. Gaussianflow: Splatting gaussian dynamics for 4d content creation. *arXiv preprint arXiv:2403.12365*, 2024. 2

[6] 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. 2

[7] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 21807–21818. IEEE, 2024. 3, 6

[8] Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848*, 2023. 2

[9] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4d: Leveraging video generators for geometric 4d scene reconstruction. *arXiv preprint arXiv:2504.07961*, 2025. 1, 2

[10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. 1, 2

[11] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J. Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 2

[12] 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. *arXiv preprint arXiv:2412.04463*, 2024. 1, 2, 3

[13] Zizhang Li, Hong-Xing Yu, Wei Liu, Yin Yang, Charles Herrmann, Gordon Wetzstein, and Jiajun Wu. Wonderplay: Dynamic 3d scene generation from a single image and actions. *arXiv preprint arXiv:2505.18151*, 2025. 2

[14] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 3

[15] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 4

[16] Tianqi Liu, Zihao Huang, Zhaoxi Chen, Guangcong Wang, Shoukang Hu, Liao Shen, Huiqiang Sun, Zhiguo Cao, Wei Li, and Ziwei Liu. Free4d: Tuning-free 4d scene generation with spatial-temporal consistency. *arXiv preprint arXiv:2503.20785*, 2025. 1, 2, 6

[17] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, 2022. 1, 2

[18] 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

[19] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[20] Ohad Rahamim, Ori Malca, Dvir Samuel, and Gal Chechik. Bringing objects to life: 4d generation from 3d objects. *CoRR*, abs/2412.20422, 2024. 2

[21] 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. 1, 2

[22] Jiawei Ren, Cheng Xie, Ashkan Mirzaei, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, Huan Ling, et al. L4gm: Large 4d gaussian reconstruction model. *Advances in Neural Information Processing Systems*, 37:56828–56858, 2025. 1, 2

[23] 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. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 31915–31929. PMLR, 2023. 2

[24] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5

[25] Qi Sun, Zhiyang Guo, Ziyu Wan, Jing Nathan Yan, Shengming Yin, Wengang Zhou, Jing Liao, and Houqiang Li. EG4D: explicit generation of 4d object without score distillation. *CoRR*, abs/2405.18132, 2024. 2

[26] 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. *CoRR*, abs/2411.04928, 2024. 1, 2

[27] KLING AI Team. Kling image-to-video model, 2024. 6

[28] Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow (extended abstract). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4839–4843. ijcai.org, 2021. 3

[29] 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. 4, 5, 6

[30] 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. 2

[31] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025. 1, 2

[32] 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. 2, 3

[33] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning, 2025. 2

[34] 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*, pages 20310–20320, 2024. 2

[35] 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. *arXiv preprint arXiv:2411.18613*, 2024. 1, 2

[36] 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. 1, 2

[37] Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *CoRR*, abs/2312.17225, 2023. 2

[38] Heng Yu, Chaoyang Wang, Peiye Zhuang, Willi Menapace, Aliaksandr Siarohin, Junli Cao, László A. Jeni, Sergey Tulyakov, and Hsin-Ying Lee. 4real: Towards photorealistic 4d scene generation via video diffusion models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 1, 2, 6

[39] Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638*, 2025. 2, 6

[40] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 2

[41] Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. Stag4d: Spatial-temporal anchored generative 4d gaussians. In *European Conference on Computer Vision*, pages 163–179. Springer, 2024. 2

[42] 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. 1, 2, 3

[43] 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. 4

[44] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 3

[45] Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Animate124: Animating one image to 4d dynamic scene. *CoRR*, abs/2311.14603, 2023. 1, 2, 6

[46] 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. *CoRR*, abs/2411.02319, 2024. 1, 2, 6

[47] Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025. 2

[48] Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text- and image-guided 4d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7300–7309, 2024. 1, 2

[49] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view syn-

thesis using multiplane images. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 37, 2018. 3