DimensionX: Create Any 3D and 4D Scenes from a Single Image with Decoupled Video Diffusion

Supplementary Material

Contents

A. Limitations and Future Work

Despite the impressive achievements, the performance of our DimensionX depends on the video diffusion backbone. Although current video diffusion models are capable of synthesizing vivid results, they still struggle with understanding and generating subtle details, which restricts the quality of the synthetic 3D and 4D scenes. Additionally, the prolonged inference procedure of video diffusion models hampers the efficiency of our generation process. In the future, it is worthy to investigate how diffusion models can be integrated to build a generalizable and efficient end-to-end 3D and 4D generation framework. We believe that our research provides a promising direction to create a dynamic and interactive environment with video diffusion models.

B. More Related Work

3D Generation with Diffusion Priors. Leveraging 2D diffusion priors for generating 3D content has revolutionized the field of 3D generation. Score Distillation Sampling (SDS) [13, 22, 32] distills 2D diffusion priors to produce high-fidelity 3D meshes from text inputs. To further enhance the 3D consistency, several works have explored the object-level generation through the multi-view diffusion [17, 19, 27, 33, 37]. Similar techniques have been further applied in the scene-level generation [25]. More recent approaches leverage video diffusion models to generate novel views from a single image, achieving impressive results at both the object-level [3, 7, 29] and scene-level [3, 7, 16]. Additionally, ReconX [15] addresses the challenge of sparse-view inputs by employing video interpolation techniques, showcasing the potential of video diffusion models for 3D scene generation. ViewCrafter [39] employs DUSt3R [30] to build a coarse 3D scene point cloud and resolves the gaps in areas not covered by the initial point cloud with a video diffusion model. In this work, we unleash the power of video diffusion models in a novel way to generate 3D scenes from a single image.

4D Generation with Diffusion Priors. Similar to 3D generation, 4D generation has seen significant advancements with the pre-trained diffusion models, including image and video diffusion. Early works [1, 23, 31, 40] adopt the SDS technique to per-scene optimize the 4D representation from a text or image input. However, these methods tend to cost hours to generate a 4D asset with obvious inconsistency. To improve the consistency and generation efficiency, subsequent works [12, 35] filter out high-quality dynamic meshes from the large-scale Objaverse dataset [5, 6] and render the multi-view videos to train a multi-view video diffusion model. Although these models can generate high-quality 4D multi-view videos, they mainly focus on the object-centric setting rather than the complex scene. For the generation of 4D scenes, the lack of sufficient data poses significant challenges in producing multi-view videos that are utilized to reconstruct the whole scene. More recently, 4Real [38] firstly proposes distilling the pre-trained video diffusion prior with the SDS loss to produce a photorealistic dynamic scene. Unlike the aforementioned works, our approach emphasizes generating temporally and spatially decomposed videos, which are subsequently merged to create

multi-view videos for high-quality 4D scene reconstruction.

C. Preliminary

Video Diffusion Models. Diffusion models [24, 28] represent a class of generative models that gradually transform Gaussian noise into structured data through a series of denoising steps. For a given data distribution $x_0 \sim q(x)$, the forward diffusion process produces a sequence $\{x_t\}_{t=0}^{T}$, where each x_t incorporates progressive Gaussian noise:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I),$$

with β_t controlling the noise level at each step.

Current powerful video diffusion models [36] utilize a 3D Variational Autoencoder (3D-VAE) to compress spatial and temporal information into a latent representation z, followed by Unet or Transformer architecture, such as the Diffusion Transformer (DiT) [21], which processes image tokens within this latent space.

Gaussian Splatting. 3D Gaussian Splatting (3DGS) [11] is an explicit scene representation that uses a set of 3D Gaussian spheres to present a scene. A Gaussian sphere is parameterized by its center $\mathbf{c} \in \mathbb{R}^3$, covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{3\times3}$, and an associated color $\mathbf{v} \in \mathbb{R}^3$ or other attributes such as opacity $\alpha \in [0, 1]$. Each Gaussian represents a localized density distribution in space, defined as:

$$G(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{c})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{c})\right),$$

where $\mathbf{x} \in \mathbb{R}^3$ is a point in space, and $\mathbf{\Sigma}$ controls the shape and orientation of the Gaussian.

The scene is represented as a collection of N Gaussians $\{(\mathbf{c}_i, \mathbf{\Sigma}_i, \mathbf{v}_i, \alpha_i)\}_{i=1}^{N}$, where N is the number of Gaussians. These Gaussians are distributed in 3D space to approximate the geometry and appearance of the scene.

Rendering is performed by projecting the Gaussians onto the image plane. For a given camera pose, the Gaussians are transformed using a projection matrix $\mathbf{P}$ derived from the camera's intrinsic and extrinsic parameters. The contribution of each Gaussian to the final pixel color is computed using a weighted sum:

$$\mathbf{I}(u, v) = \sum_{i=1}^{N} w_i(u, v) \cdot \mathbf{v}_i,$$

where $w_i(u, v)$ is the weight of the i-th Gaussian at pixel (u, v), calculated as:

$$w_i(u, v) = \alpha_i \cdot G(\mathbf{P}^{-1}(u, v) - \mathbf{c}_i).$$

The weights are normalized across all Gaussians contributing to a pixel to ensure proper blending. This process is differentiable, making it suitable for optimization tasks in 3D scene reconstruction.

Low-Rank Adaptation (LoRA). LoRA [10] introduces a low-rank adaptation mechanism to the transformer's weight matrices. For a given weight matrix $W \in \mathbb{R}^{d\times k}$, instead of updating W directly, LoRA constrains the update to a low-rank matrix ΔW represented as:

$$W = W' + \Delta W = W + AB^\top,$$

where $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{k\times r}$, with $r \ll \min(d, k)$. This factorization ensures that the adaptation of W' is restricted to a low-rank space, significantly reducing the number of parameters that need to be updated, while still enabling the model to capture essential task-specific information efficiently.

D. Additional Illustrations on ST-Director

To decompose spatial and temporal variations, we define equivalence relations that capture the behavior of points in the 4D space under different conditions.

S-Equivalence Relation($\sim_S$) For any two spacetime points (x_1, y_1, z_1, t_1) and $(x_2, y_2, z_2, t_2) \in \mathbb{R}^4$, we define $(x_1, y_1, z_1, t_1) \sim_S (x_2, y_2, z_2, t_2)$ if:

$$\begin{cases} \mathcal{S}(t_1) = \mathcal{S}(t_2) = \mathcal{S}(t_0) & \text{(Temporal dimension fixed)} \\ C(t_1) \neq C(t_2) & \text{(Camera viewpoint varies)} \end{cases} \tag{1}$$

where $\mathcal{S}(t_0)$ represents a fixed spatial slice in time t_0, such that $(x_1, y_1, z_1), (x_2, y_2, z_2) \in \mathcal{S}(t_0)$. This relation indicates that while the temporal component remains constant, implying no actual progression in physical time, the spatial observations differ exclusively due to variations in camera parameters. Accordingly, the S-equivalence class $[p]_S$ encapsulates the set of spatial point projections across different camera perspectives at the same temporal instance. This class thereby abstracts the notion of viewing geometry at a singular point in time, reflecting all spatial appearances that are permissible by adjusting the camera's intrinsic or extrinsic parameters, without temporal evolution.

T-Equivalence Relation ($\sim_T$): For any two spacetime points (x_1, y_1, z_1, t_1) and $(x_2, y_2, z_2, t_2) \in \mathbb{R}^4$, we define $(x_1, y_1, z_1, t_1) \sim_T (x_2, y_2, z_2, t_2)$ if there exists a trajectory function $f_o : \mathbb{R} \to \mathbb{R}^3$ for dynamic objects $\bigcup_i O^i(t)$ satisfying the following conditions:

$$\begin{cases} C(t_1) = C(t_2) = C_0, & \\ (x_1, y_1, z_1) = (x_2, y_2, z_2), & (x, y, z) \in B, \\ (x_j, y_j, z_j) = f_o(t_j), & (x_j, y_j, z_j) \in \bigcup_i O^i(t), \end{cases} \tag{2}$$

Where C_0 is a constant and $j = 1, 2$, the condition $C(t_1) = C(t_2) = C_0$ indicates that the camera viewpoint remains fixed, with $(x, y, z) \in B$ representing static background

points and $(x_j, y_j, z_j) = f_o(t_j)$ describing the positions of dynamic objects. Consequently, the T-equivalence class $[p]_T$ includes all points observed from the fixed camera position C_0 that either belong to the static background or lie along the trajectory of a moving object as it progresses over time.

E. Implementation Details

Algorithm 1 Multi-view video generation with DimensionX

Input: Reference image I, Prompt P, S-Director θ_s, T-Director θ_T, number of frames N, number of views K, noise level T, refine noise level T_{refine}

Output: Multi-view video $\left\{\{I_j^i\}_{i=1}^N\right\}_{j=1}^K$. $\{I_j^i\}_{i=1}^N$

Static camera video $V_0 = \theta_T(I)$
Select one frame I_{ref} as the reference frame
Generates reference multi-view using S-Director $V_{\text{ref}} = \theta_S(I_{\text{ref}})$
Get noised reference video $V_{\text{ref}}^* = \text{Add\_noise}(V_{\text{ref}}, T)$ ▷ Using Eq. 5
$V_{t-s} \leftarrow [\ ]$ ▷ Each item stands for one frame
for $i \in [N]$ **do**
    $V_{t-s}[i] = \theta_S(V_0[i] \mid V^*, T)$
end for
$V_{s-t} \leftarrow [\ ]$ ▷ Each item stands for one view
$V_{s-t} = \text{SwitchTempSpatial}(V_{t-s})$ ▷ Reshape into multi-view video
for $i \in [K]$ **do**
    $V_{s-t}^*[i] = \text{Add\_noise}(V_{s-t}[i], T_{\text{refine}})$ ▷ Add noise
    $V_{s-t}[i] = \theta_T(V_{\text{ref}}[i] \mid V_{s-t}^*[i], T_{\text{refine}})$ ▷ Refine by re-denoising
end for

E.1. Training Details

We choose the CogVideoX-I2V (5B) [36] as our backbone video diffusion model, which contains 42 DiT blocks. For the ST-director training, we freeze the backbone and only train the LoRA layers for 3000 steps, which takes around 1 day with 8 NVIDIA A800 GPUs. For the frame extension and frame interpolation, we fine-tune the model for 2000 steps, which takes around 10 hours with 8 NVIDIA A800 GPUs. The S-Director based on the video interpolation model is trained for 1000 steps. Specifically, we use DL3DV-10K [14] and OpenVid-1M [20] to train our ST-director. Meanwhile, OpenVid-1M and RealEstate-10K [41] are combined to train the interpolation and frame extension diffusion model.

OpenVid-1M [20] is a curated high-quality open-sourced video dataset, including 1 million video clips with diverse motion dynamics and camera controls. DL3DV-10K [14] is a widely-collected 3D scene dataset with high-resolution multi-view images, including diverse indoor and outdoor scenes. RealEstate-10K [41] is a dataset from youtube, mianly including the captures of indoor scenes. Applying our designed data collection framework, we build the dimension-variant dataset from DL3DV-10K and OpenVid-1M. We select 100 high-quality temporal-variant videos from OpenVid to train T-Director. For each S-Director type, 100 videos are rendered according to the specific camera trajectory to train the corresponding LoRA. During the inference stage, we adopt the DDIM sampler [28] with classifier-free guidance [9] and set the sampling step to 50.

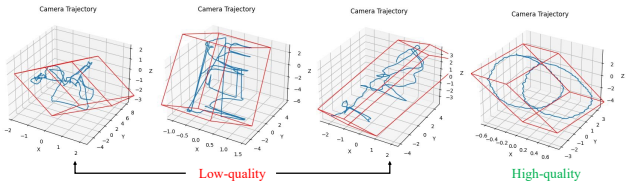

Figure 1. **Camera trajectory in DL3DV.**

E.2. Building Dimension-variant Dataset

We provide more details about how we build the dimension-variant dataset from the open-source datasets.

Trajectory planning for spatial-variant data. As shown in Fig. 1, though DL3DV [14] includes a vast collection of indoor and outdoor multi-view images, the camera paths are highly complex and diverse. On one hand, this makes it challenging for video diffusion models to learn the desired camera motion paradigms from such spatial-variant data. On the other hand, these complex camera motion patterns may degrade the generation quality of video diffusion models, as they deviate significantly from the pre-training dataset distribution of video diffusion.

Instead of directly using the 3D dataset, we propose reconstructing photorealistic 3D scenes and rendering videos consistent with our spatial variations. To select scenes that align with our target paths, we need to compute the coverage range of the cameras throughout the entire scene. Given N cameras in a scene, we first compute the center C and principal axes A along the direction x, y, and z using the Principal Component Analysis (PCA) technique:

$$C = \frac{\sum_{i=1}^{N} \boldsymbol{p}_i}{N}, \quad A = \text{SVD}(\mathcal{P} - C), \tag{3}$$

where p_i denotes the position of camera i, $\mathcal{P} = \{\boldsymbol{p}_i, 1 \le i \le N\} \in \mathbb{R}^{N\times 3}$ represents the position set of N cameras, and SVD is the Singular Value Decomposition operation. Next, we need to calculate the lengths L of each

axis from the projection distance D:

$$D = (\mathcal{P} - C) \cdot A \tag{4}$$

$$L = \max(D) - \min(D). \tag{5}$$

Built on the above calculation, we have already figured out the distribution of the camera throughout the entire scene. To cope with various scenes, we establish the following rules to filter out the qualifying data: **(1) Camera Distribution**: We calculate the center of the scene and judge how cameras capture around the scene, such as circle or semi-circle captures. **(2) Bounding Box Aspect Ratio**: The aspect ratio of the bounding box should meet the requirement for various S-Directors. For instance, the aspect ratio of x and y axis should not vary too greatly, which helps in selecting appropriate 360-degree surrounding videos. **(3) Camera-to-Bounding Box Distance**: We calculate the distance from each camera position to the closest plane of the bounding box and prioritize data with smaller total distances to ensure better camera placement. With the filtered dataset based on the above rules, it is necessary to compute the occupancy field within the scene to help us plan the feasible region for the rendering cameras. After reconstructing the entire scene's 3DGS from multi-view images, we render multi-view images and their corresponding depth maps, and then use TSDF [4] to extract the scene's mesh from the RGB-D data.

With the camera bounding box and TSDF of the scene, we can acquire spatial-variant videos consistent with our target path. For instance, if our goal is to generate a circular video, we can select scenes that meet our requirements based on the camera bounding box, as shown in the fourth column of Fig. 1. Then, we extract its TSDF, choose a starting camera, and utilize the scene's occupancy field to avoid collisions between the camera and objects in the scene during rendering.

In the 3D world, camera movements are defined by 6 degrees of freedom (DoF), with each DoF allowing movement in both positive and negative directions for translation and rotation, resulting in 12 distinct motion patterns. Additionally, we also train orbital motion category S-Director, where the camera follows a smooth, circular path around the subject, capturing a unique perspective beyond the standard DoF-based movements. Specifically, we provide the visualization of our designed S-Director in Fig. 9.

Flow guidance for temporal-variant data. We provide the optical flow maps of static and dynamic camera videos in Fig. 3.

E.3. 3DGS and 4DGS Optimization Details

We provide more details about the 3D and 4D scene generation. In the 3D scene generation experiments of our main paper, we adopt the orbit right S-Director for the single-view setting. For the sparse-view setting, we compare the estimated rotation and translation of input images with our designed S-Director and select the suitable S-Director. Due to our early-stopping strategy, the input images can flexibly correspond to S-Director, allowing the camera trajectory of generated videos to adaptively adjust based on the input image. In the 3DGS optimization stage, we select the first and end frames of generated videos to produce the initialization point cloud and 49 frames from videos to optimize the scene. Following the confidence-aware 3DGS in ReconX [15], we leverage the confidence map from DUSt3R [30] and set the optimization step to 3000 steps. The hyper-parameters λ_1, λ_{ssim}, and λ_{lpips} are 0.8, 0.2, and 0.3, respectively. Generating a 49-frame video with S-Director takes approximately 200 seconds on a single NVIDIA A800 GPU with 26.38GB memory. The corresponding 3DGS optimization also takes around 200 seconds and requires 14.36GB memory.

In the 4D scene generation experiments of our main paper, we adopt the orbit right and left S-Directors. In the multi-view video generation process, we set the strength of reference video to 0.9 and the appearance refinement strength to 0.7. During the deformable 3DGS phrase, we use the generated multi-view videos as the training multi-view videos, and apply the original deformable 3DGS to reconstruct the 4D scene. In particular, the per-view video frame is 49 and the view number is 22.

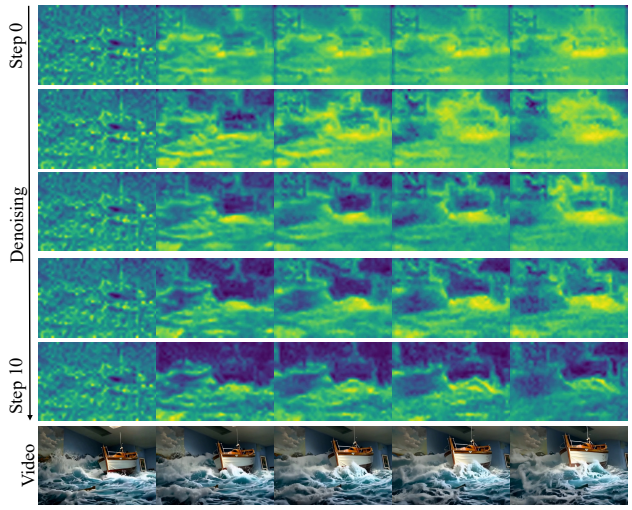

Figure 2. **Visualization of attention map of base model.** Starting from step 0, the early denoising steps (before step 10 of total denoising step 50) have determined the outline and layouts of output videos.

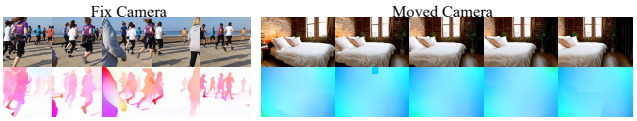

Figure 3. Optical flow of static and dynamic camera videos.

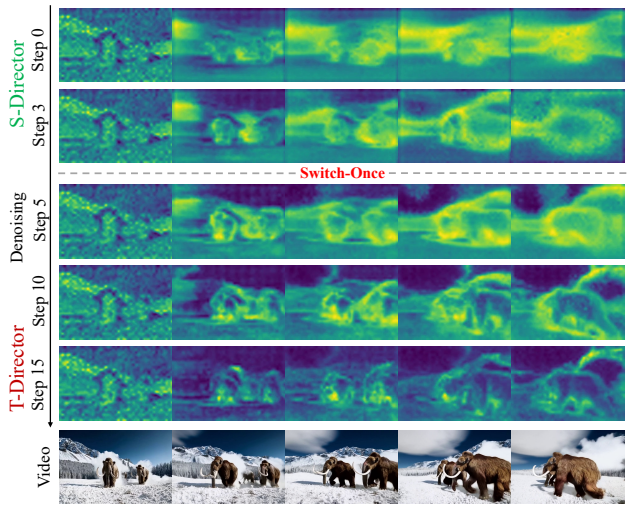

Figure 4. **Visualization of attention map with Switch-Once LoRA fusion.** At the start of the denoising loop, S-Director distributes high attention values across the scene, reflecting noticeable camera motion. After switching to T-Director at step 4, T-Director refocuses attention on scene objects.

E.4. Attention Map Visualization

We visualize the attention map of the 30th DiT block output from the base model during the denoising loop, as shown in Fig. 2. Compared to the attention maps generated with S-Director and T-Director, the base model's attention map reveals a blend of spatial and temporal information recovery throughout the denoising process. Additionally, we present the attention maps obtained using the Switch-Once method in Fig. 4. At the beginning of the denoising loop, when S-Director is applied, the attention maps show high attention values distributed across the scene, highlighting noticeable camera motion. After switching to T-Director at step 4 of the denoising loop, the high attention values shift to focus on objects within the scene. Over subsequent steps, the objects exhibit motion while preserving camera movement, resulting in coordinated motion of both objects and the camera in the generated video.

E.5. 4D Scene Generation Evaluation Metrics

Following SynCamMaster [18], we evaluate visual quality using CLIP-T and CLIP-F, while evaluating view synchronization through Mat. Pix. (K) and CLIP-V. Specifically, CLIP-T measures the average CLIP similarity between each frame and its corresponding text prompt, and CLIP-F quantifies the average CLIP similarity between adjacent frames. To further evaluate view synchronization, we utilize the image matching method GIM [26] to compute the number of matched pixels with confidence exceeding a specified threshold, denoted as Mat. Pix. (K). Additionally, CLIP-V captures the average CLIP similarity among multi-view frames captured simultaneously.

F. More Experiments

F.1. More Ablation Studies

Design of LoRA fusion mechnism. We ablate our proposed Switch-Once lora composition approach for hybrid-dimension control. We compare our approach with the naive fusion, which merges multiple LoRA weights together linearly [10]. As shown in Fig. 11, the LoRA merge approach cannot handle the hybrid-dimension control, while our Switch-Once achieves effective dimension-aware LoRA fusion, presenting impressive controllability along the spatial and temporal axis.

ST-Director for controllable video generation. We ablate the influence of our proposed ST-Director for dimension-variant video generation. Specifically, we compare our DimensionX with the base model, CogVideoX [36], to demenstrate the effectiveness of our approach. As presented in Fig. 12, even provided with detailed spatial and temporal prompt control descriptions, CogVideoX struggles to achieve the desired control, whereas our DimensionX enables controllable generation across multiple dimensions.

Frame extension for single-view 3D scene generation. We ablate the design of frame extension for single-view 3D scene generation. We compare our extended video diffusion (145 frames) with the original model (49 frames) on the 360-degree 3D scene generation setting. As shown in Fig. 13, with our proposed frame extension, our S-Director is able to generate a complex scene from a single image.

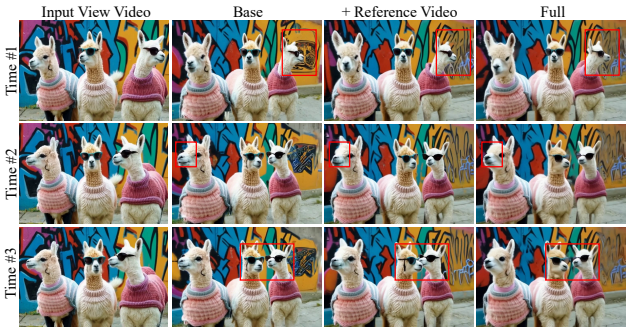

Figure 5. **Ablation study on 4D video generation.** We compare the novel views with different conditions.

Identity-preserving denoising for 4D generation. As illustrated in Fig. 5, we conduct ablation studies on the effectiveness of reference video latent sharing and appearance refinement. Specifically, we can observe that directly combing per-frame spatial-variant videos causes severe inconsistency, including the background and subject shape. Through the reference video latent sharing, the global background and appearance exhibit a high consistency across different frames. Building on reference video latent sharing, appearance refinement enhances the coherence of appearance details.

Method	RotErr↓	TransErr↓
MotionCtrl	1.23	10.21
CameraCtrl	1.07	8.27
Ours	**0.88**	**7.29**

Table 1. **Camera control Accuracy.**

F.2. More Results

Quantitative comparison of few-view 3D reconstruction Following CAT3D [7], we evaluate our method on the MipNeRF 360 dataset as shown in Tab. 2. Our approach outperforms existing methods across all metrics in the few-view 3D reconstruction setting.

Qualitative comparison in dimension-aware video generation Compared to Dream Machine 1.6, one of the state-of-the-art closed source video generation models, our method effectively decomposes the spatial and temporal factors of the video diffusion model, as demonstrated in Fig. 10.

Camera control accuracy To evaluate the accuracy of the camera control, following CameraCtrl [8], we provide a quantitative assessment of the accuracy of the camera in the Tab. 1. Our method achieves the lowest camera error in this setting compared to other camera control methods.

Figure 6. **Camera Fusion.** By flexibly integrating S-Director, our method enables diverse and versatile camera control.

Controllable video generation We provide more results on the controllable video generation in Fig. 15. Additionally, we present the estimated camera trajectory of these generated videos, demonstrating the impressive controllability of our approach. Furthermore, we present additional 360-degree orbit video generations in Fig. 14. In addition, Fig. 6 shows that the flexible combination of LoRAs enables diverse camera control.

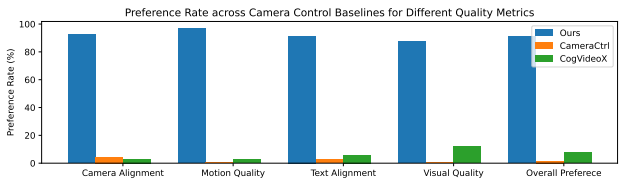

Figure 7. **User study for camera control.**

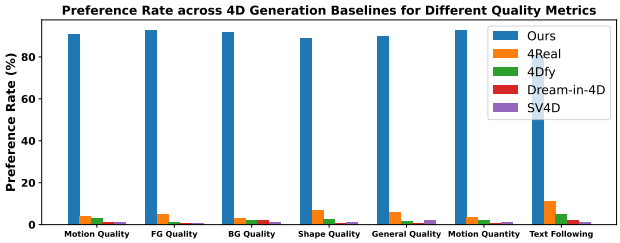

Figure 8. **User study for 4D generation.**

User Study As shown in Fig. 7, we conduct a user study with experienced evaluators to assess camera controllability. Evaluators were presented with anonymized video pairs and asked to select their preference based on the specified criteria. Our approach significantly outperforms previous baselines in terms of controllability. Additionally, we perform a separate user study illustrated in Fig. 8, focusing on evaluating the quality of generated 4D multi-view videos. Evaluators similarly assessed anonymized video pairs according to predefined criteria. Results indicate that our method achieves superior video quality compared to existing baselines.

3D scene generation We provide more 3D scene generation results, including the comparison with baselines in Fig. 16 and additional generated scenes of our approach in Fig. 17. These results demonstrate that our DimensionX achieves impressive 3D scene generation ability in real-world scenes.

4D scene generation We provide more 4D scene generation results of our DimensionX in Fig. 18.

References

[1] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7996–8006, 2024. 1

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields, 2023. 7

[3] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024. 1

[4] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 4

[5] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 1

[6] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte,

	3-view			6-view			9-view		
	PSNR↑	SSIM ↑	LPIPS ↓	PSNR↑	SSIM ↑	LPIPS ↓	PSNR↑	SSIM ↑	LPIPS ↓
Zip-NeRF [2]	12.77	0.271	0.705	13.61	0.284	0.663	14.30	0.312	0.633
ZeroNVS [25]	14.44	0.316	0.680	15.51	0.337	0.663	15.99	0.350	0.655
ReconFusion [34]	15.50	0.358	0.585	16.93	0.401	0.544	18.19	0.432	0.511
CAT3D [7]	16.62	0.377	0.515	17.72	0.425	0.482	18.67	0.460	0.460
DimensionX(Ours)	**17.59**	**0.478**	**0.378**	**18.32**	**0.484**	**0.329**	**20.41**	**0.657**	**0.243**

Table 2. **Quantitative comparison of few-view 3D reconstruction.**

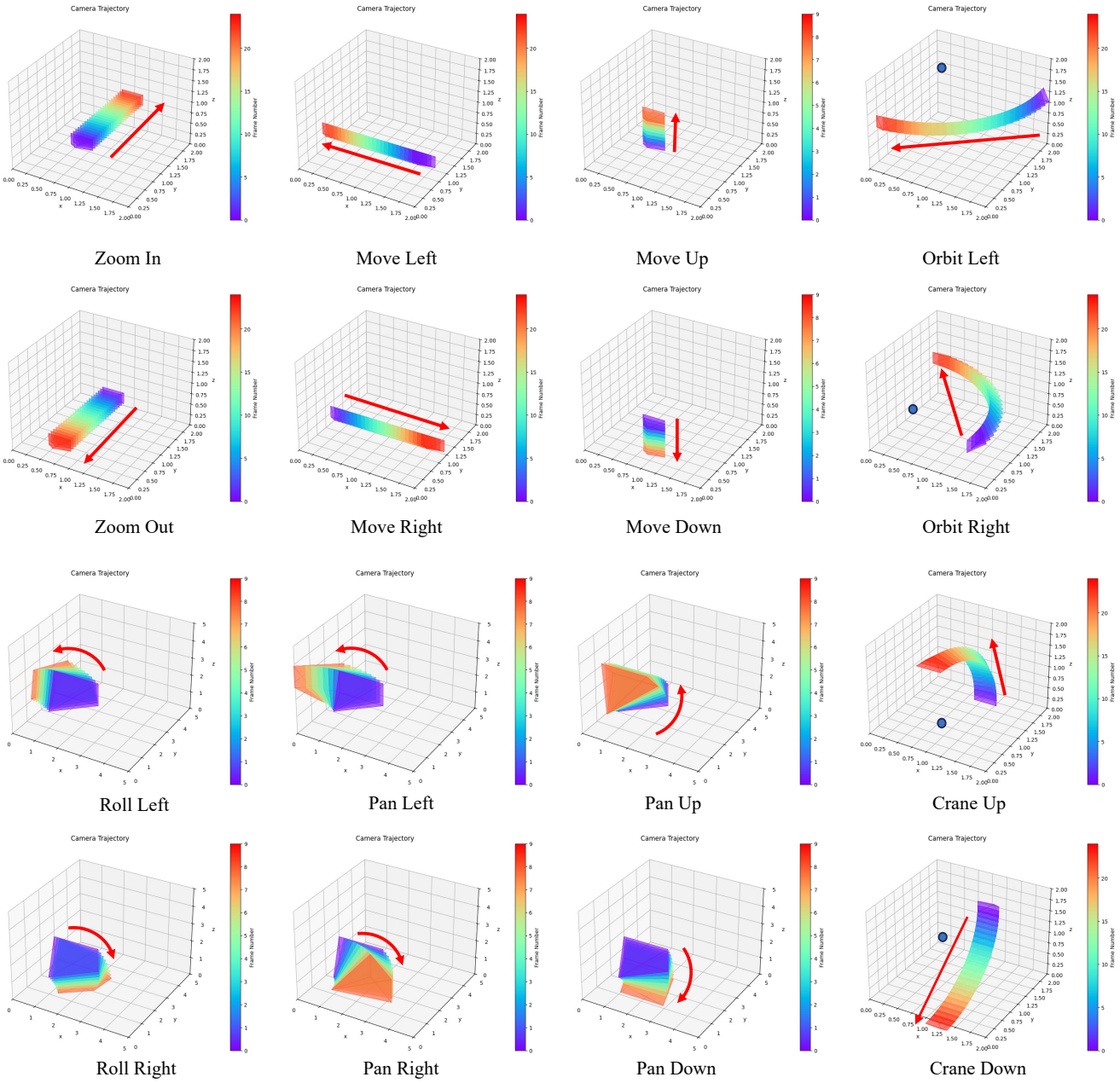

Figure 9. **Visualization of various S-Directors.**

Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[7] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur

Figure 10. **Qualitative comparison in dimension-aware video generation.** Given the same image and text prompt, the first row is the temporal-variant video generation (camera static), the second row is the spatial-variant video generation (object motion static while camera zoom out), and the third row is the spatial- and temporal-variant video generation (camera orbit right).

Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 1, 6, 7

[8] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 6

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 5

[11] 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, 2023. 2

[12] Hanwen Liang, Yuyang Yin, Dejia Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models. *arXiv preprint arXiv:2405.16645*, 2024. 1

[13] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 1

[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] Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024. 1, 4

[16] Fangfu Liu, Hanyang Wang, Shunyu Yao, Shengjun Zhang, Jie Zhou, and Yueqi Duan. Physics3d: Learning physical properties of 3d gaussians via video diffusion. *arXiv preprint arXiv:2406.04338*, 2024. 1

[17] Fangfu Liu, Diankun Wu, Yi Wei, Yongming Rao, and Yueqi Duan. Sherpa3d: Boosting high-fidelity text-to-3d generation via coarse 3d prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20763–20774, 2024. 1

[18] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 5

[19] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceed-*

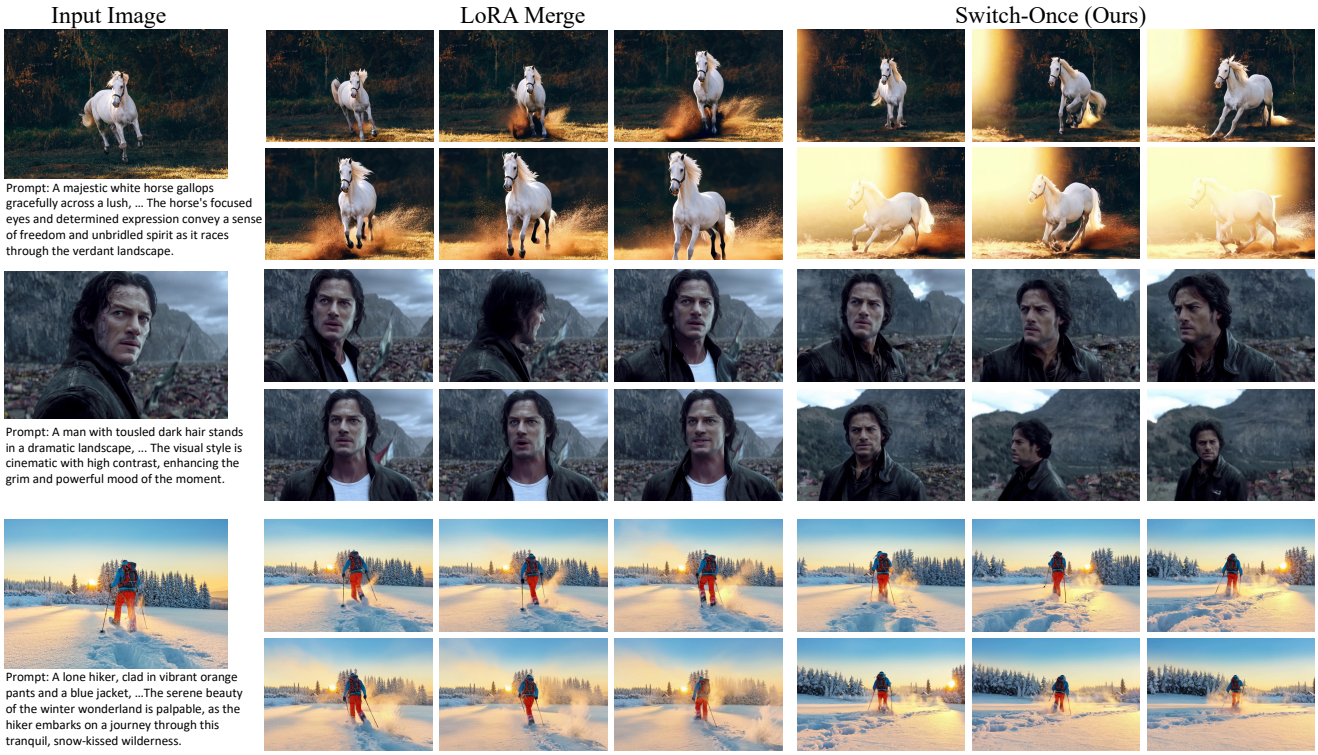

Figure 11. **Ablation on Swtich-Once.** We ablate our designed dimension-aware LoRAs fusion approach with the LoRA merge method. The desired camera control is orbit right.

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9970–9980, 2024. 1

[20] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. 3

[21] 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

[22] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

[23] 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

[24] 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

[25] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023. 1, 7

[26] Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, and Cheng Wang. Gim: Learning generalizable image matcher from internet videos. In *The Twelfth International Conference on Learning Representations*, 2024. 5

[27] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 1

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[29] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2025. 1

[30] 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. 1, 4

[31] Xinzhou Wang, Yikai Wang, Junliang Ye, Zhengyi Wang, Fuchun Sun, Pengkun Liu, Ling Wang, Kai Sun, Xintong Wang, and Bin He. Animatabledreamer: Text-guided non-

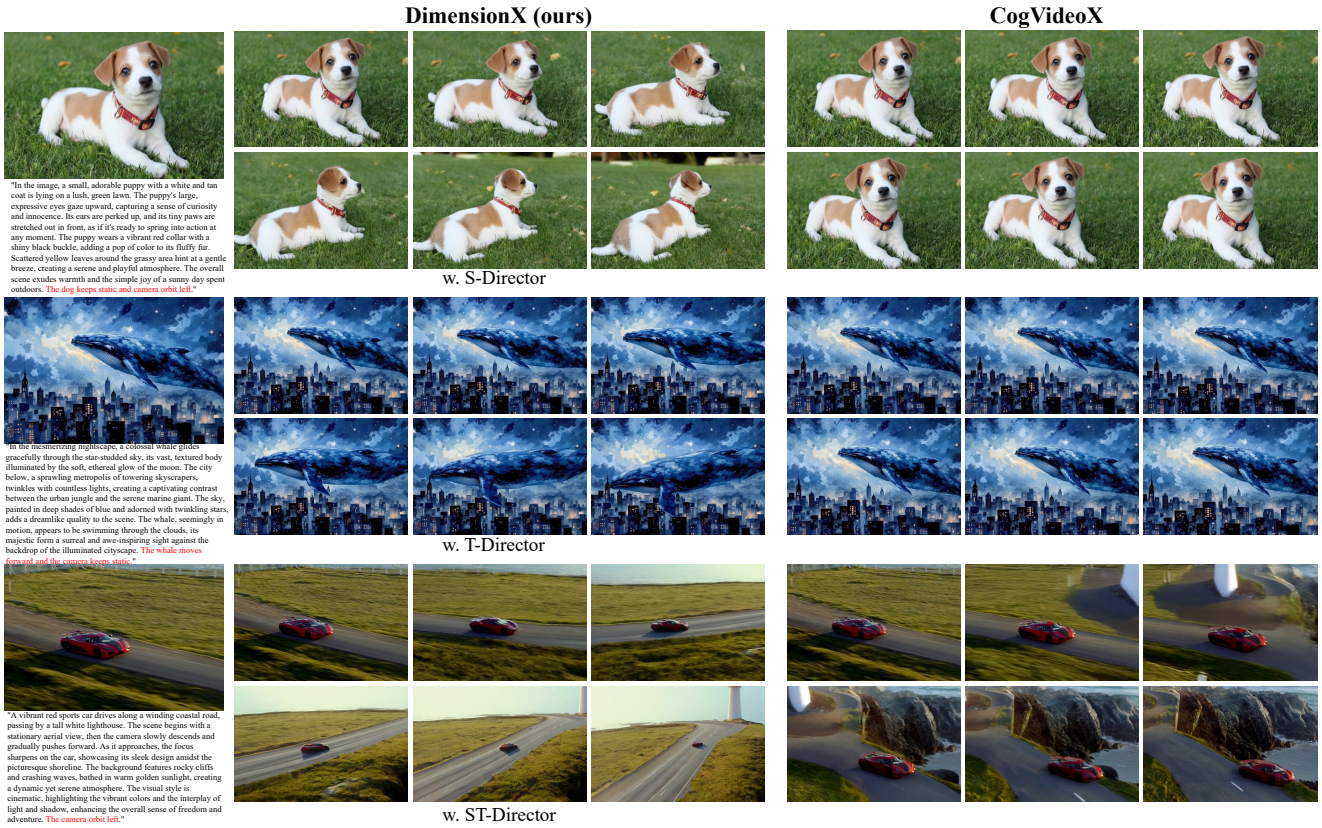

Figure 12. **Ablation on ST-Director for controllable video generation.** Given the same image and text prompt, the first row is the spatial-variant video generation (object static and camera orbit left), the second row is the temporal-variant video generation (camera static and object moves), and the third row is the spatial- and temporal-variant video generation (camera orbit left).

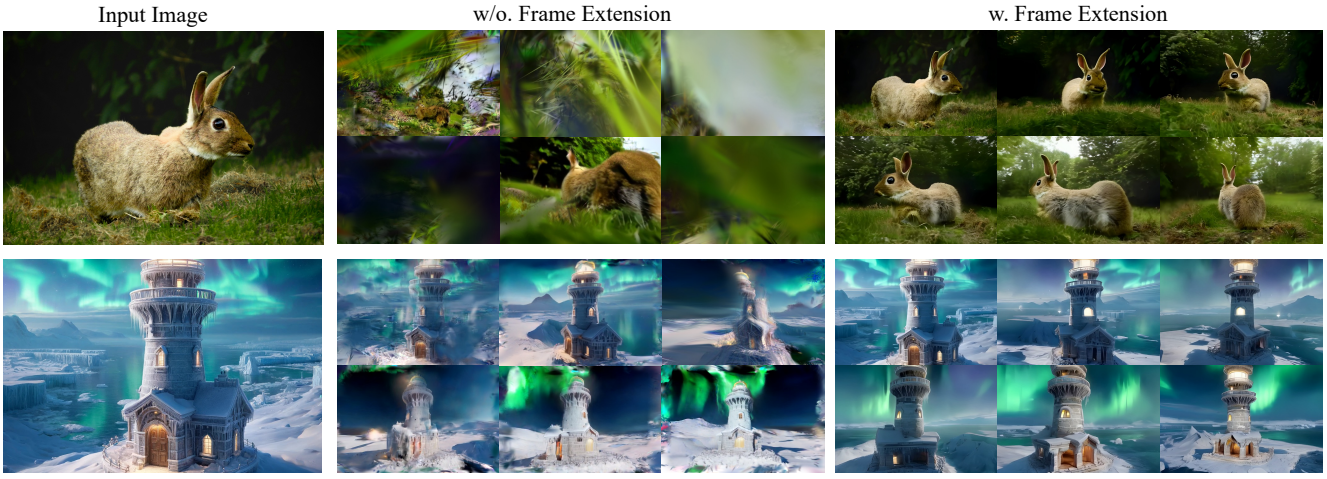

Figure 13. **Ablation on frame extension for circular 3D scene generation.** Given a single image, we compare the circular 3D scene generation performance w. and w/o. frame extension. With the frame extension, our DimensionX can generate consistent 3D scenes with high-quality renderings.

rigid 3d model generation and reconstruction with canonical score distillation. *arXiv preprint arXiv:2312.03795*, 2023. 1

[32] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and

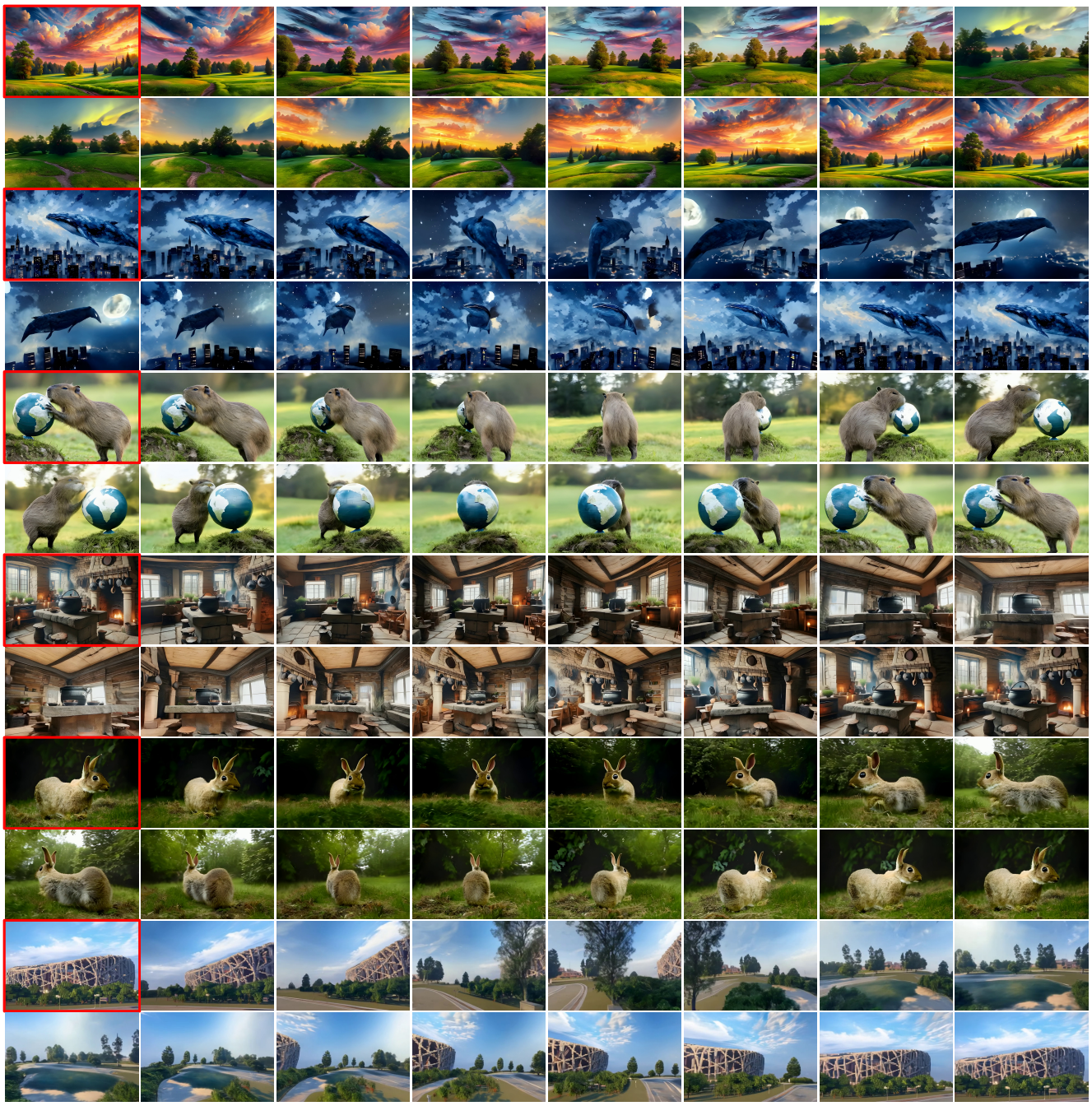

Figure 14. **More 360 Degree Orbit Video Results.**

diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[33] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv preprint arXiv:2405.20343*, 2024. 1

[34] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21551–21561, 2024. 7

[35] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*, 2024. 1

[36] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu

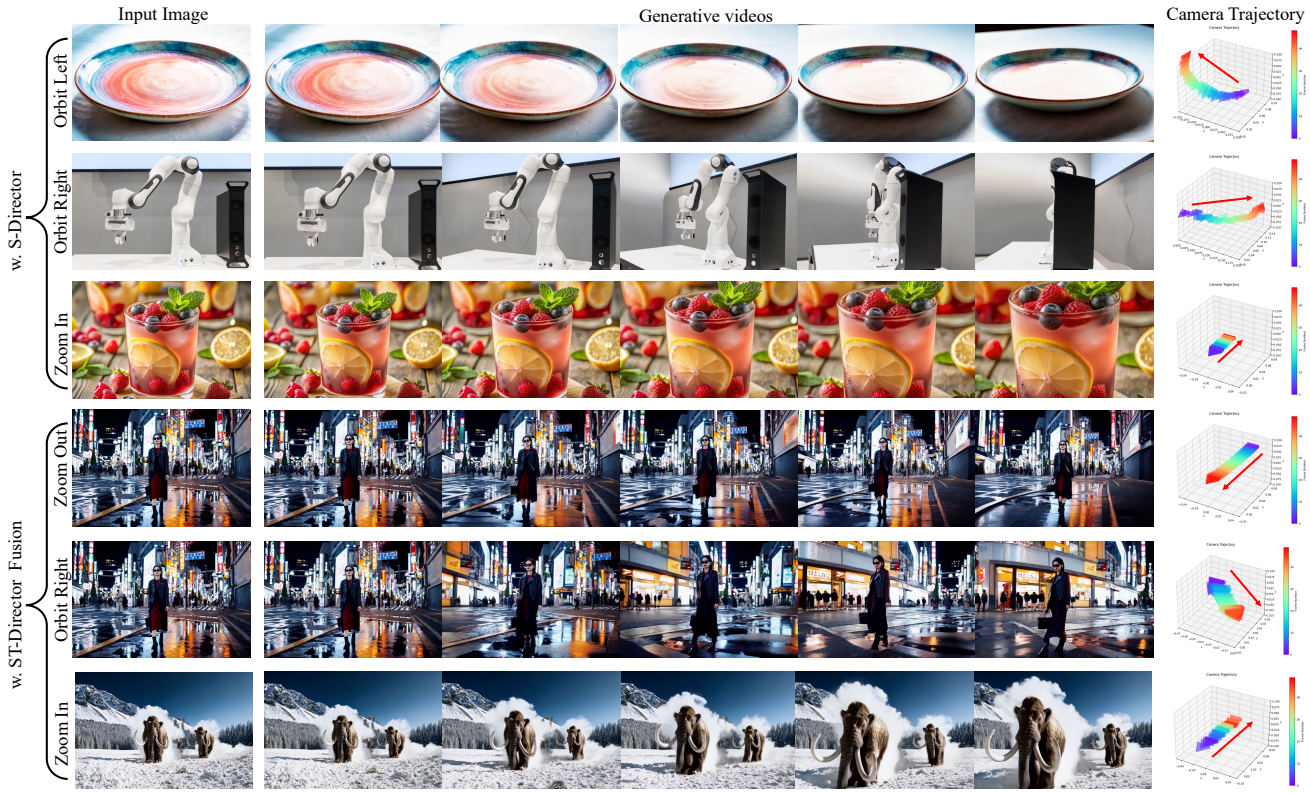

Figure 15. **Camera trajectory visualization of generated videos.**

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. 2, 3, 5

[37] Junliang Ye, Fangfu Liu, Qixiu Li, Zhengyi Wang, Yikai Wang, Xinzhou Wang, Yueqi Duan, and Jun Zhu. Dreamreward: Text-to-3d generation with human preference. *arXiv preprint arXiv:2403.14613*, 2024. 1

[38] Heng Yu, Chaoyang Wang, Peiye Zhuang, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Laszlo A Jeni, Sergey Tulyakov, and Hsin-Ying Lee. 4real: Towards photorealistic 4d scene generation via video diffusion models. *arXiv preprint arXiv:2406.07472*, 2024. 1

[39] 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. 1

[40] Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023. 1

[41] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018. 3

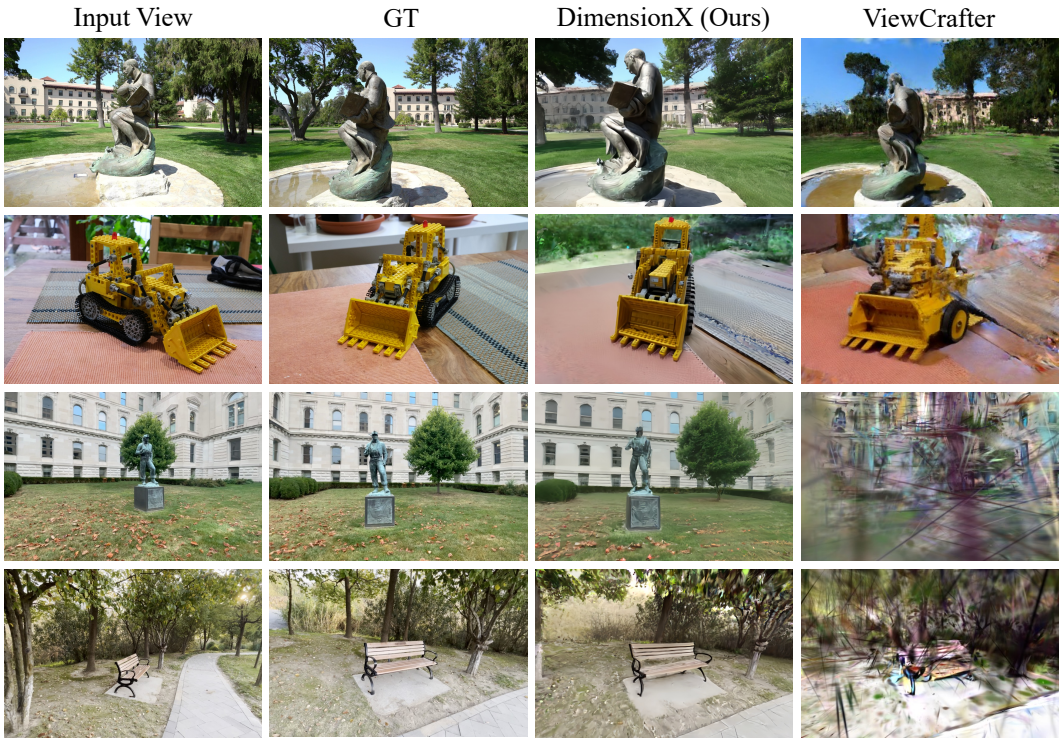

Figure 16. **Qualitative comparison in single-view 3D generation.** Given a single image, our approach can create much better 3D scenes compared with other baselines.

Figure 17. **More results of 3D scene generation.** We present more 3D scene generation results in both single-view and sparse-view settings.

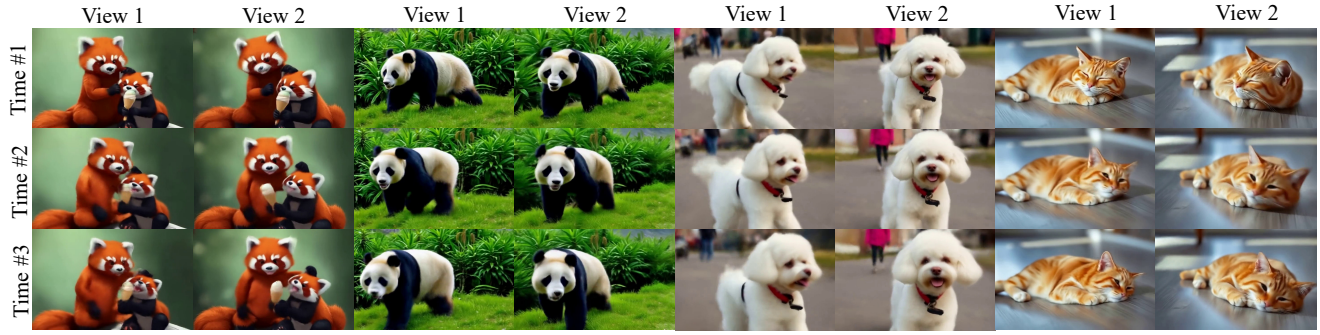

Figure 18. **More results of 4D scene generation.** DimensionX generates consistent and high-quality 4D scenes.