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

Wenqiang Sun* [1,3], Shuo Chen* [2], Fangfu Liu* [2], Zilong Chen[2,3],
Yueqi Duan[2], Jun Zhu† [2,3], Jun Zhang† [1], Yikai Wang† [2]
[1]Hong Kong University of Science and Technology [2]Tsinghua University [3]ShengShu
wsunap@connect.ust.hk, eejzhang@ust.hk, yikaiw@outlook.com

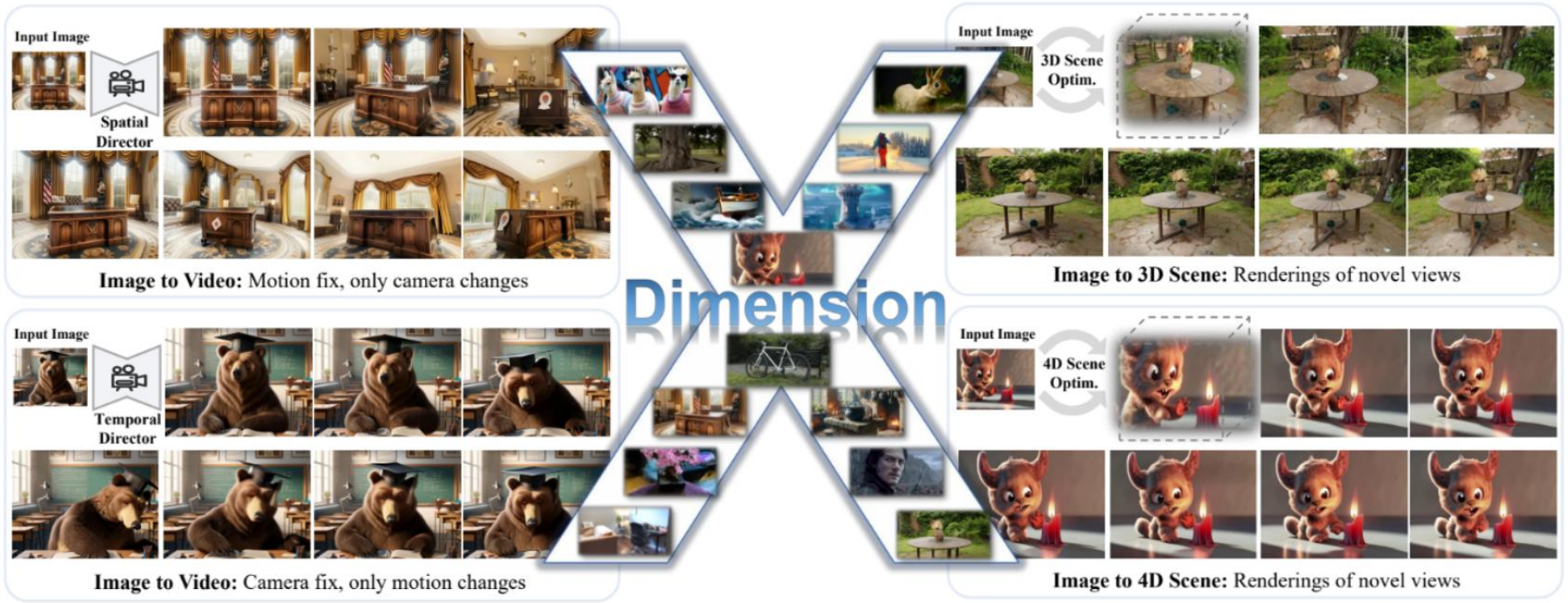


Figure 1. With just a single image as input, our proposed **DimensionX** can generate highly realistic videos and 3D/4D environments that are aware of spatial and temporal dimensions.

https://chenshuo20.github.io/DimensionX/

Contributions

In summary, our main contributions are:

- We present DimensionX, a novel framework for generating photorealistic 3D and 4D scenes from only a single image using decoupled video diffusion.
- We propose ST-Director, which decouples the spatial and temporal priors in video diffusion models by learning (spatial and temporal) dimension-aware modules with our curated datasets. We further enhance the hybrid-dimension control with a training-free composition approach according to the essence of the video diffusion denoising process.
- To bridge the gap between video diffusion and real-world scenes, we design a trajectory-aware mechanism for 3D generation and an identity-preserving denoising approach for 4D generation, enabling more realistic and controllable scene synthesis.
- Extensive experiments manifest that our DimensionX delivers superior performance in video, 3D, and 4D generation compared with baseline methods.

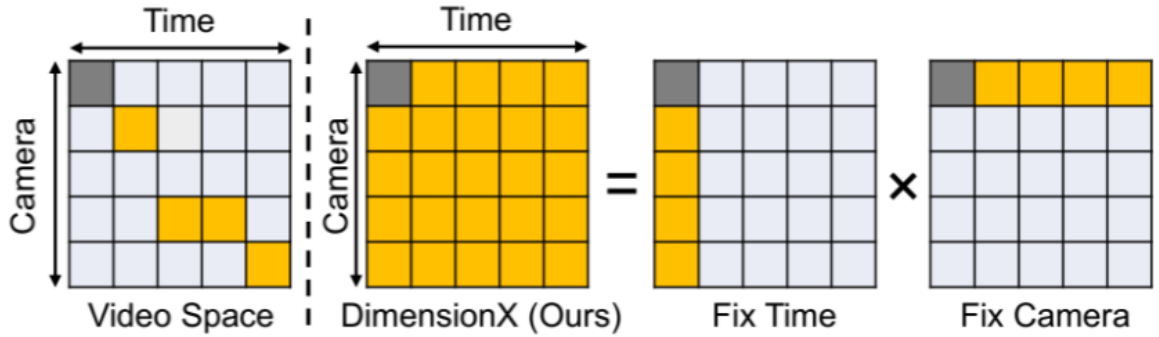


Figure 2. **Illustration of DimensionX.** Our key insight is to decouple the spatial and temporal factors in video diffusion. The figure is reproduced from CAT4D [60].

process. As shown in Fig. 2, our key insight is to decouple the temporal and spatial factors in video diffusion, thus achieving precise control over each individually and in combination. To achieve the dimension-aware control, we establish a comprehensive framework to collect datasets that vary in spatial and temporal dimensions. With these

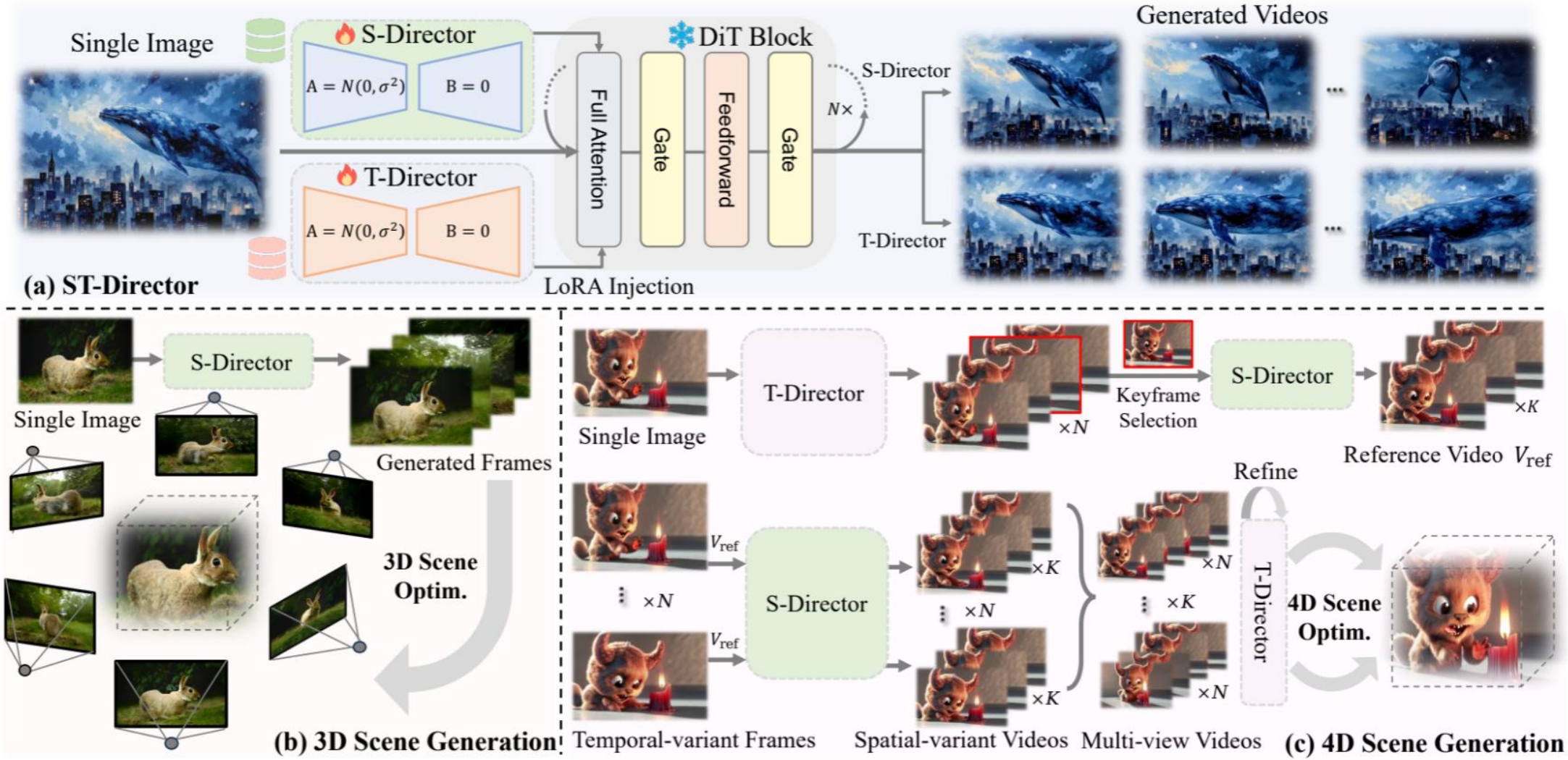


Figure 3. **Pipeline of DimensionX.** Our framework is mainly divided into three parts. **(a) Controllable Video Generation with ST-Director.** We introduce ST-Director to decompose the spatial and temporal parameters in video diffusion models by learning dimension-aware LoRAs on our collected dimension-variant datasets. **(b) 3D Scene Generation with S-Director.** Given one view, a high-quality 3D scene is recovered from the video frames generated by S-Director. **(c) 4D Scene Generation with ST-Director.** Given a single image, a temporal-variant video is produced by T-Director, from which a key frame is selected to generate a spatial-variant reference video. Guided by the reference video, per-frame spatial-variant videos are generated by S-Director, which are then combined into multi-view videos. Through the multi-loop refinement of T-Director, consistent multi-view videos are then passed to optimize the 4D scene.

Methodology - Building Dimension-variant Dataset

Trajectory planning for spatial-variant data. Camera motion in 3D space has 6 degrees of freedom (DoF), enabling 12 movement patterns through positive/negative translations and rotations. Additionally, we include the orbital motion pattern that circles subjects smoothly, offering unique views beyond standard DoF movements. Please see the visualization of our designed S-Director in Appendix E.

To acquire the spatial-variant dataset, we propose reconstructing photorealistic 3D scenes and rendering videos consistent with our spatial variation tendency. To facilitate the selection and planning of rendering paths, we need to compute the coverage range of the cameras throughout the entire scene. Specifically, we utilize the Principal Component Analysis (PCA) technique to compute the bounding box of cameras in the scene. Another key component is to acquire the occupancy field of 3D scenes. With the 3D scene, we render multi-view images and depth maps, and use TSDF [10] to extract the mesh. The bounding box and occupancy field help us to plan feasible rendering regions.

To decouple spatial and temporal parameters in video diffusion, we introduce a framework to collect spatial- and temporal-variant videos from open-source datasets. Notably, we employ a trajectory planning strategy for spatial-variant data and flow guidance for temporal-variant data.

Flow guidance for temporal-variant data. To achieve the temporal control, we aim to filter the temporal-variant data to fine-tune the video diffusion model. Specifically, we use optical flow to identify static-camera videos. Fixed cameras produce flow maps with large white areas (static background), while moving cameras show flow everywhere with almost no white regions. This clear visual difference makes optical flow an effective tool for classifying camera motion. Please see the figure illustration in Appendix E.

Also a limitation?

We conceptualize each video frame $I_t(u, v)$ as a projection from a 4D space $\mathbb{R}^3 \times \mathbb{R}^1$, where u and v are the image coordinates in the frame, and the 4D space consists of three spatial dimensions x, y, z and a temporal dimension t. In this framework, the 4D space consists of a static background and dynamic objects, which is represented as $\mathcal{S}(t)$ at time t.

Methodology - ST-Director for Decoupled Video Generation

Each video frame at time t is, therefore, a 2D projection of this 3D scene structure onto the image plane, governed by the camera's parameters at that moment. To formalize this, we define the projection function $\mathcal{P}_{C(t)}$, which projects the 3D scene $\mathcal{S}(t)$ onto a 2D image:

$$I_t(u, v) = \mathcal{P}_{C(t)}(\mathcal{S}(t)), \tag{1}$$

where $C(t)$ represents the camera parameters at time t.

cise control. Specifically, we introduce two orthogonal basis directors: **S-Director** (Spatial Director) and **T-Director** (Temporal Director), which allow us to separate spatial and

How to make them orthogonal?

3.2.1. Dimension-aware Decomposition

We systematically analyze the video content generation by decomposing it into two fundamental dimensions: spatial variations (camera movements through static scenes) and temporal variations (object motions viewed from fixed cameras). This structured decomposition enables us to separately examine and model two core aspects of video dynamics - the changing perspectives created by camera movement and the temporal evolution of objects within static views. By isolating these components, we can more ef-

Building on this decomposition framework, we map videos from our spatial-variant dataset to the spatial representation (fixed temporal content), while videos from the temporal-variant dataset are mapped to the temporal representation (fixed spatial viewpoint). In order to train the S-Director and T-Director to generate videos with these spatial and temporal structures, we employ LoRA [18], a fine-tuning method that is both parameter-efficient and computationally light, training each director separately on the datasets to decouple the video diffusion. Specifically, the S-Director is trained on the spatial-variant dataset, learning patterns in which time is held constant ($\mathcal{S}(t) = \mathcal{S}_0$), thereby generating videos with the pattern $I_t(u, v) = \mathcal{P}_{C(t)}(\mathcal{S}_0)$. Similarly, the T-Director is trained on the temporal-variant dataset, learning patterns where the camera remains stationary ($C(t) = C_0$), producing videos satisfying $I_t(u, v) = \mathcal{P}_{C_0}(\mathcal{S}(t))$.

Methodology - ST-Director for Decoupled Video Generation

How to merge them?

frame sequences along its designated axis. However, most videos naturally involve a blend of spatial and temporal elements, making it essential to combine both directors to capture multidimensional information in the 4D space, represented as $I_t(u, v) = \mathcal{P}_{C_t}(\mathcal{S}(t))$. To achieve the hybrid-dimension control, we aim to merge the S-Director and T-Director, allowing for the video generation along the spatial and temporal dimensions. In pursuit of this goal, we analyze the mechanics of the base model (See in Appendix E) and each director's denoising process by visualizing the attention maps produced by the base model and both directors (as shown in Fig. 4). We identify two key observations:

Observation 1: *The initial steps of the denoising process are critical for defining the generated video.*

From the attention maps, we observe that during the initial denoising stage, both the base model and two directors create initial outlines aligned with final results.

Observation 2: *Spatial information is constructed earlier than temporal information.*

As shown in Fig. 4, we observe that the object motion synthesis is initially underdeveloped during the early denoising stage. Specifically, with S-Director, the attention maps reveal that the structural outlines of the final video appear much earlier than with temporal control.

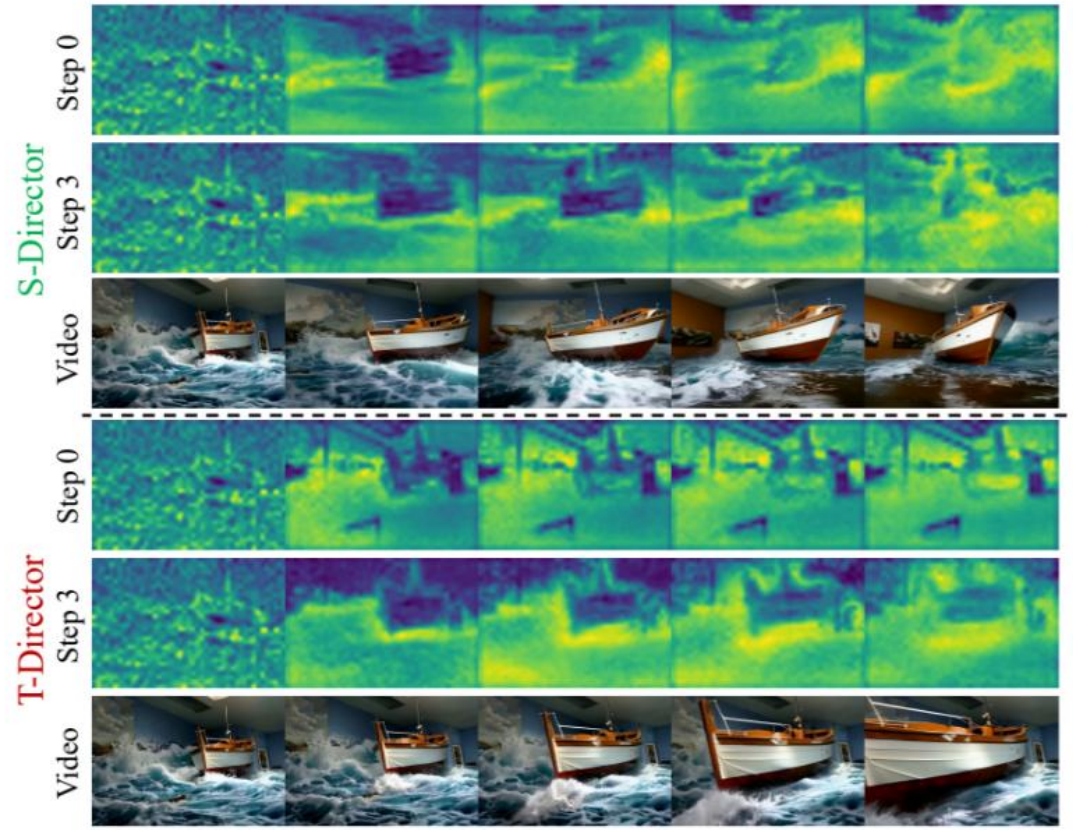


Figure 4. **Visualization of Attention Map.** Starting from step 0, the early denoising steps have determined the outline and layouts of output videos. Specifically, the spatial component is recovered earlier than the temporal information during the denoising process.

Methodology - ST-Director for Decoupled Video Generation

How to merge them?

frame sequences along its designated axis. However, most videos naturally involve a blend of spatial and temporal elements, making it essential to combine both directors to capture multidimensional information in the 4D space, represented as $I_t(u, v) = \mathcal{P}_{C_t}(\mathcal{S}(t))$. To achieve the hybrid-dimension control, we aim to merge the S-Director and T-Director, allowing for the video generation along the spatial and temporal dimensions. In pursuit of this goal, we analyze the mechanics of the base model (See in Appendix E) and each director's denoising process by visualizing the attention maps produced by the base model and both directors (as shown in Fig. 4). We identify two key observations:

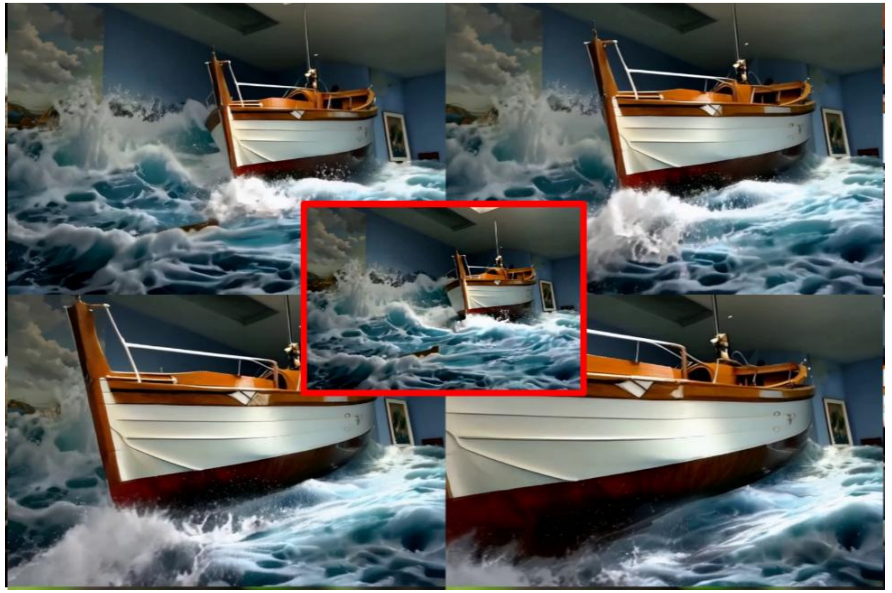
Observation 1: *The initial steps of the denoising process are critical for defining the generated video.*

From the attention maps, we observe that during the initial denoising stage, both the base model and two directors create initial outlines aligned with final results.

Observation 2: *Spatial information is constructed earlier than temporal information.*

As shown in Fig. 4, we observe that the object motion synthesis is initially underdeveloped during the early denoising stage. Specifically, with S-Director, the attention maps reveal that the structural outlines of the final video appear much earlier than with temporal control.

Based on these two observations, we propose a training-free approach, **Switch-Once**, a novel approach to compose diverse LoRAs. This approach combines the S-Director and T-Director to generate videos that seamlessly blend spatial and temporal information, achieving a balanced synthesis represented by $I_t(u, v) = \mathcal{P}_{C_t}(\mathcal{S}(t))$. Following Observation 2, we initiate the denoising process with the S-Director to establish comprehensive camera motion across the scene. Then, as indicated by Observation 1, we switch to the T-Director after the initial steps (*e.g.* 3 in our experiments) of the denoising process, thereby enhancing the dynamics of generated videos. The resulting video, in Fig. 5 (4th column), demonstrates the effectiveness of our approach.



Methodology - 3D Scene Generation with S-Director

Built upon the S-Director, our video diffusion model is able to generate controllable and consistent 3D frames from a single image, allowing for the reconstruction of photorealistic scenes. Specifically, to deal with where diverse and flexible spatial variation, we introduce a **trajectory-aware mechanism** to handle potential various camera movements, including single-view and sparse-view settings.

Single-view Scene Generation. Given a single image I, our goal is to reconstruct the 3D scene with generated video frames $\{I^i\}_{i=1}^{N}$, where N represents the frame length. Although current video diffusion models have shown potential for long video generation, the total duration still falls far short of the frame count required for real-world scene reconstruction. Specifically, the powerful open-source video diffusion model (*e.g.* CogVideoX [65]) currently generates a maximum of only 49 frames, whereas reconstructing a large scene (*e.g.* 360 degree scene) typically requires hundreds of multi-view images. To address this, we extend the video diffusion model to generate 145 frames.

Sparse-view Scene Generation. In this setting, we propose incorporating a video interpolation model and an adaptive S-Director to achieve a smooth and consistent transition between the sparse views. First, we develop a video diffusion model to generate the high-quality interpolated video, which takes two images $\{I^1, I^2\}$ as the start and end frames. The objective function for the video diffusion process is formulated as

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{z_t \sim p, \epsilon \sim \mathcal{N}(0,I), t} \left[\|\epsilon - \epsilon_\theta(z_t, t, z_1, z_2, c)\|_2^2 \right], \tag{2}$$

where z_t is the noisy latent sequence, and ϵ_θ represents the model's prediction of the noise at timestep t, conditioned on the first and last frame latent: $z_1 = \mathcal{E}(I^1)$ and $z_2 = \mathcal{E}(I^2)$. With the interpolated video diffusion model, we then train various S-Directors to provide refined camera motion guidance, ensuring smooth and consistent transition between the sparse-view images. In particular, we tailor two key strategies to fully leverage the guidance prior carried in S-Directors: early-stopping training and adaptive trajectory-planning. The early-stopping strategy prevents overfitting to target trajectories during training, while adaptive trajectory-planning is used at inference to choose the best matched S-Director based on camera pose differences, enabling the model to effectively handle input images from a wide range of angles. Following the original 3DGS pipeline [20], we adopt the loss function as follow:

$$\mathcal{L}_{\text{conf}} = \mathcal{C}\left(\lambda_1 \mathcal{L}_1 + \lambda_{\text{ssim}} \mathcal{L}_{\text{ssim}} + \lambda_{\text{lpips}} \mathcal{L}_{\text{lpips}}\right), \tag{3}$$

where $\mathcal{C}$ is confidence maps, and $\lambda_1, \lambda_{\text{ssim}}, \lambda_{\text{lpips}}$ represent coefficients. Please refer to Appendix E for more details.

Methodology - 4D Scene Generation with ST-Director

Equipped with spatial and temporal controlled video diffusion, a 4D dynamic scene can be recovered from a single image. A direct way is to stitch together the per-frame spatial-variant videos of the temporal-variant video into multi-view videos, which are then used to reconstruct the 4D scene. However, maintaining consistency in the background and object appearance across spatial-variant videos is challenging. To address this difficulty, we design an **identity-preserving denoising** strategy, including reference video latent sharing and appearance refinement, to enhance the consistency of all spatial-variant videos.

Given an input image I, our goal is to generate a photorealistic 4D scene with coherent dynamics and backgrounds. First, we employ T-Director to generate a temporal-variant video $\{I^i\}_{i=1}^{N}$ for the input image, from which a reference frame I_{ref} is chosen to produce the corresponding spatial-variant video $v_{ref} = \{I^i\}_{i=1}^{K}$, where K represents the number of cameras. Subsequently, v_{ref} is used to guide the generation of spatial-variant videos across all temporal-variant video frames $\{I^i\}_{i=1}^{N}$, which are then combined into multi-view videos $\left\{\{I_j^i\}_{i=1}^{N}\right\}_{j=1}^{K}$. $\{I_j^i\}_{i=1}^{N}$ represents the temporal-variant video from the camera j. Despite guidance from the reference video, minor shape inconsistencies still exist, causing temporal jitter and inter-view misalignment. To mitigate these issues, we introduce an appearance refinement to further enforce the consistency across multi-view videos. With consistent multi-view videos, we choose deformable 3DGS [57] to model the dynamic scene.

Methodology - 4D Scene Generation with ST-Director

Reference Video Latent Sharing. Through our empirical study, we propose choosing the reference frame based on the dynamic object's mask and the magnitude of optical flow values, allowing us to acquire a frame that best encompasses the dynamic region. With the reference frame I_{ref}, S-Director is applied to produce the corresponding spatial-variant video v_{ref}. Applying the forward diffusion process on v_{ref}, we derive the noisy latent code z_{ref} as following:

$$z_{ref} = \sqrt{\alpha_t} z_0 + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{1}), \tag{4}$$

where $z_0 = \mathcal{E}(v_{ref})$, representing the compressed latent by the encoder $\mathcal{E}$, and $\sqrt{\alpha_t}$ determines the strength of using the reference video. Starting from the same initialization latent z_{ref}, all frames are subsequently denoised to produce spatial-variant videos with strong coherence. Moreover, we propose blending the denoised z_t of each frame with the reference video latent $z_{ref,t}$ at the early denoising steps:

$$z_t = \lambda z_t + (1-\lambda) z_{ref,t}, \tag{5}$$

where λ is an adjustable parameter.

Appearance Refinement. Inspired by the image-to-image translation SDEdit [33], we apply random noise to each multiview video $v_j = \{I_j^i\}_{i=1}^N$, and perform multi-step denoising, acquiring smooth and high-quality videos with the video diffusion prior:

$$v_j^{\text{refine}} = f_\theta\left(v_j + \epsilon\left(t_0\right); t_0, c\right), \tag{6}$$

where t_0 represents the forward diffusion timestep, and v_j^{refine} is the refined video with the denoise function f_θ of T-Director. In addition, we repeat the refine process during the middle timestep to enhance the smoothness.

Having acquired consistent multi-view videos , we use the deformable 3DGS [57] to reconstruct the 4D scene.

Experiments

Implementation Details. We choose the open-source I2V model CogVideoX [65], which adopts the diffusion transformer architecture, as our video diffusion model. For the ST-Director training, we set the LoRA rank to 256, and fine-tune the LoRA layers for 3000 steps at the learning rate 1e-3 on 100 dimension-variant videos. To enlarge video frames, we modify the RoPE [44] positional embedding to extend the video length to 145 frames. For the training of video interpolation models, we first full fine-tune the base model for 2,000 steps at the learning rate 5e-5, then we train the S-Director but for only 1,000 steps.

Datasets. In our whole framework, our video diffusion model is mainly trained on three datasets: DL3DV-10K [26], OpenVid [35], and RealEstate-10K [75]. To verify the 3D generation ability of DimensionX, we compare our approach with other baselines on Tank-and-Temples [21], MipNeRF360 [5], NeRF-LLFF [34], and DL3DV-10K [26].

Experiments

Video Generation

Baselines and Evaluation Metrics. We compare our DimensionX with the original CogVideoX [65](open-source) and Dream Machine 1.6 (closed-source product). Following the previous benchmark VBench [19], we evaluate the Subject Consistency, Dynamic Degree, and Aesthetic Score.

	Consistency↑	Dynamic ↑	Aesthetic ↑
CogVideoX [65]	93.56	11.76	57.81
Dream Machine 1.6	93.69	38.24	68.96
Ours	**97.69**	**47.06**	**70.82**

Table 2. **Quantitative comparison for video generation.**

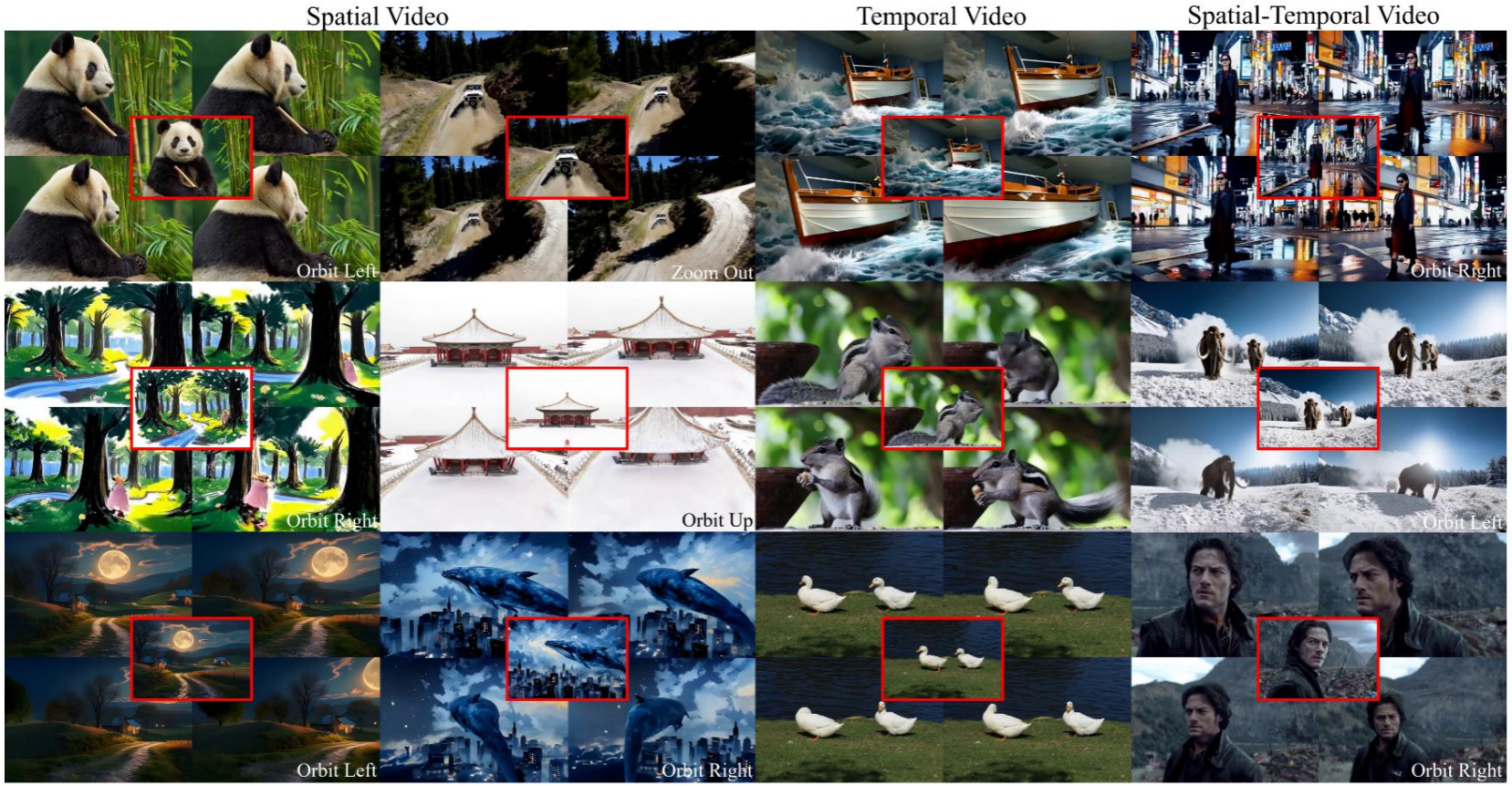


Figure 5. **Qualitative results in dimension-aware video generation.** For spatial videos, the camera moves while objects stay still; for temporal videos, objects move with a static camera. A spatial-temporal video combines both, with the camera following a trajectory as objects move.

Experiments

3D Scene Generation

Baselines and Evaluation Metrics. In the single-view setting, we compare our approach with two generative methods: ZeroNVS [40] and ViewCrafter [69]. For the sparse-view scenario, we select two sparse-view reconstruction and one sparse-view generation methods: DNGaussian [23], InstantSplat [13], and ViewCrafter [69]. We adopt PSNR, SSIM, and LPIPS as the metric for our quantitative results. Specifically, in both single-view and sparse-view settings, we begin by reconstructing the 3D scene, followed by calculating the metrics using renderings from novel views.

	Methods	Tank and Temples			MipNeRF360			LLFF			DL3DV		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Single-View	ZeroNVS [40]	12.31	0.301	0.567	15.84	0.327	0.536	15.62	0.497	0.354	12.39	0.251	0.559
	ViewCrafter [69]	15.18	0.499	0.319	15.65	0.404	0.378	17.56	0.620	0.337	14.78	0.422	0.417
	Ours	17.11	0.613	0.199	18.91	0.527	0.333	20.38	0.744	0.200	18.28	0.642	0.215
Sparse-View	DNGaussian [23]	12.13	0.292	0.511	15.21	0.127	0.632	17.51	0.586	0.409	14.99	0.286	0.432
	InstantSplat [13]	18.70	0.634	0.258	16.80	0.574	0.296	22.33	0.818	0.149	18.30	0.691	0.222
	ViewCrafter [69]	18.76	0.637	0.216	18.49	0.691	0.212	21.60	0.823	0.155	19.19	0.686	0.196
	Ours	20.42	0.668	0.185	20.21	0.713	0.184	25.11	0.913	0.067	21.69	0.780	0.124

Table 1. **Quantitative comparison of single-view and sparse-view scenarios.** Our approach outperforms other baselines in all metrics both in terms of single-view and sparse-view (two-view) settings.

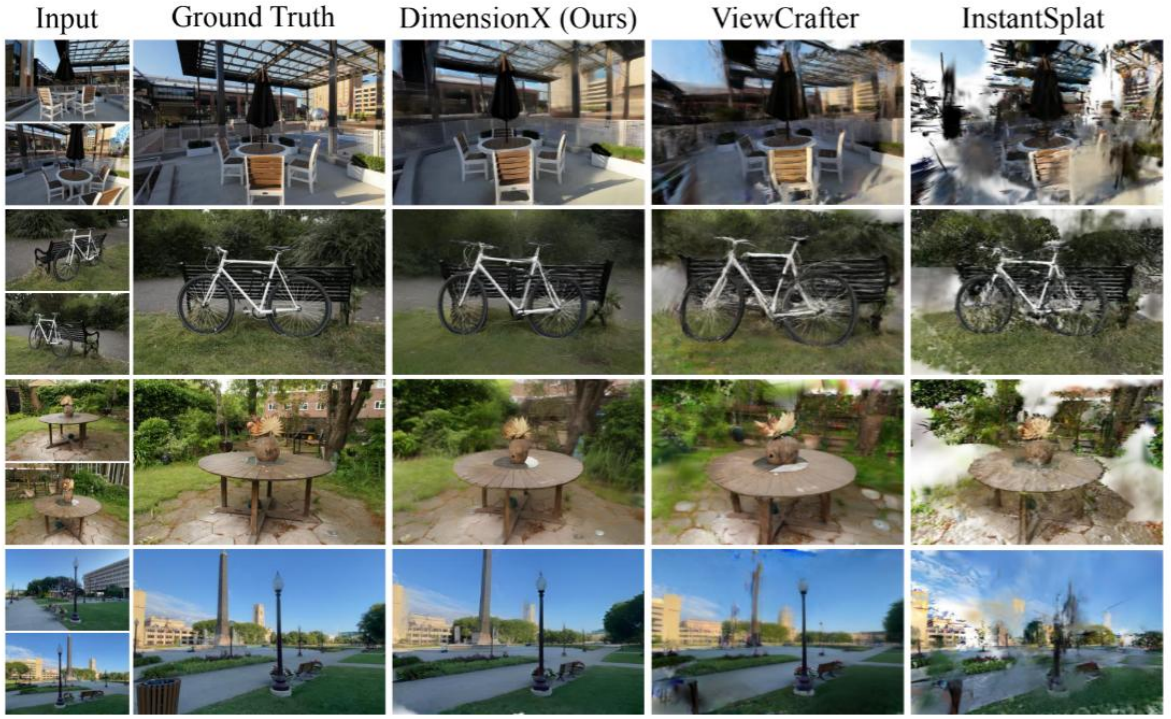


Figure 6. **Qualitative results in sparse-view 3D generation.**

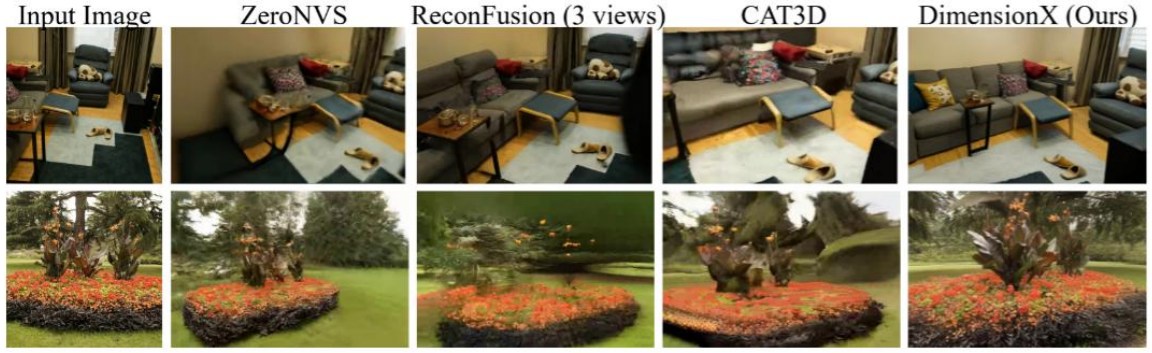


Figure 7. **Qualitative comparison with NVS methods.** The input images and results of baselines are from the project of CAT3D.

Experiments

4D Scene Generation

Baselines and Evaluation Metrics. To the best of our knowledge, there is currently no open-source work dedicated to generating multi-view videos from a single image. Following previous works [3, 48], we choose SV4D [62] and multi-view images + CogVideoX as baselines. Specifically, CLIP-T and CLIP-F are used to assess visual quality, while Mat. Pix.(K) and CLIP-V evaluate view synchronization. See Appendix E for more details on these metrics.

Method	Visual Quality		View Synchronization	
	CLIP-T ↑	CLIP-F ↑	Mat. Pix.(K) ↑	CLIP-V ↑
SV4D	30.97	98.31	294.3	87.8
M.V. Img+CogVideoX	34.05	98.67	323.0	96.1
Base	32.59	98.77	563.7	96.2
+ Ref Video	34.50	99.20	766.0	97.8
+ Both	**35.83**	**99.38**	**834.3**	**98.3**

Table 4. **4D quantitative comparison & ablation.**

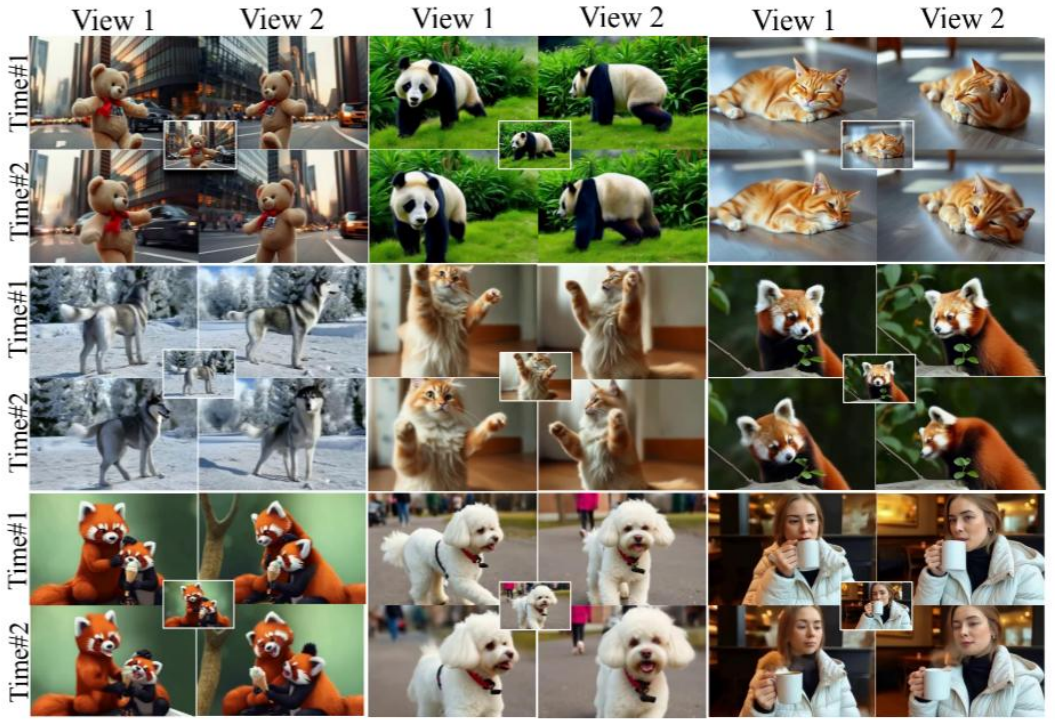


Figure 8. **Qualitative results of multi-view video generation.**

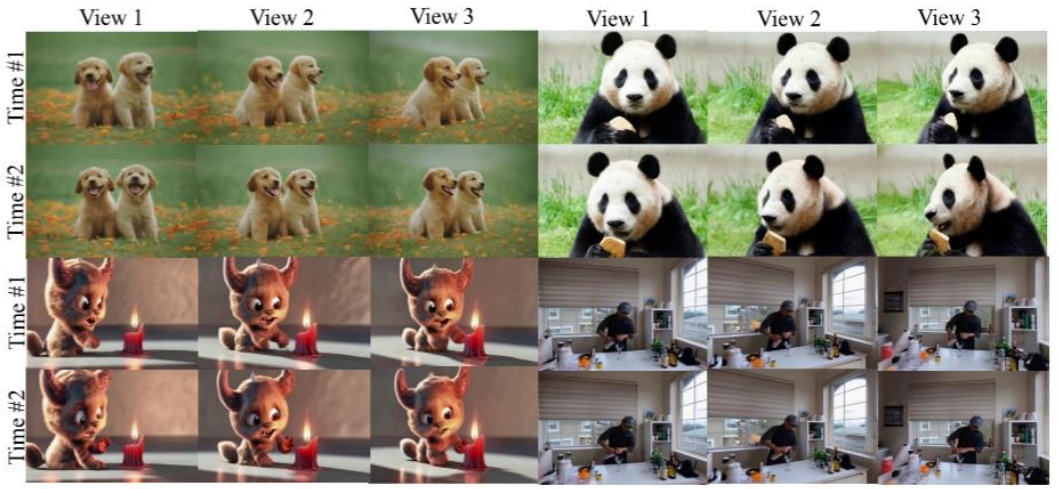


Figure 9. **4D reconstruction results from our generated multi-view videos.**

Experiments

Ablation Study

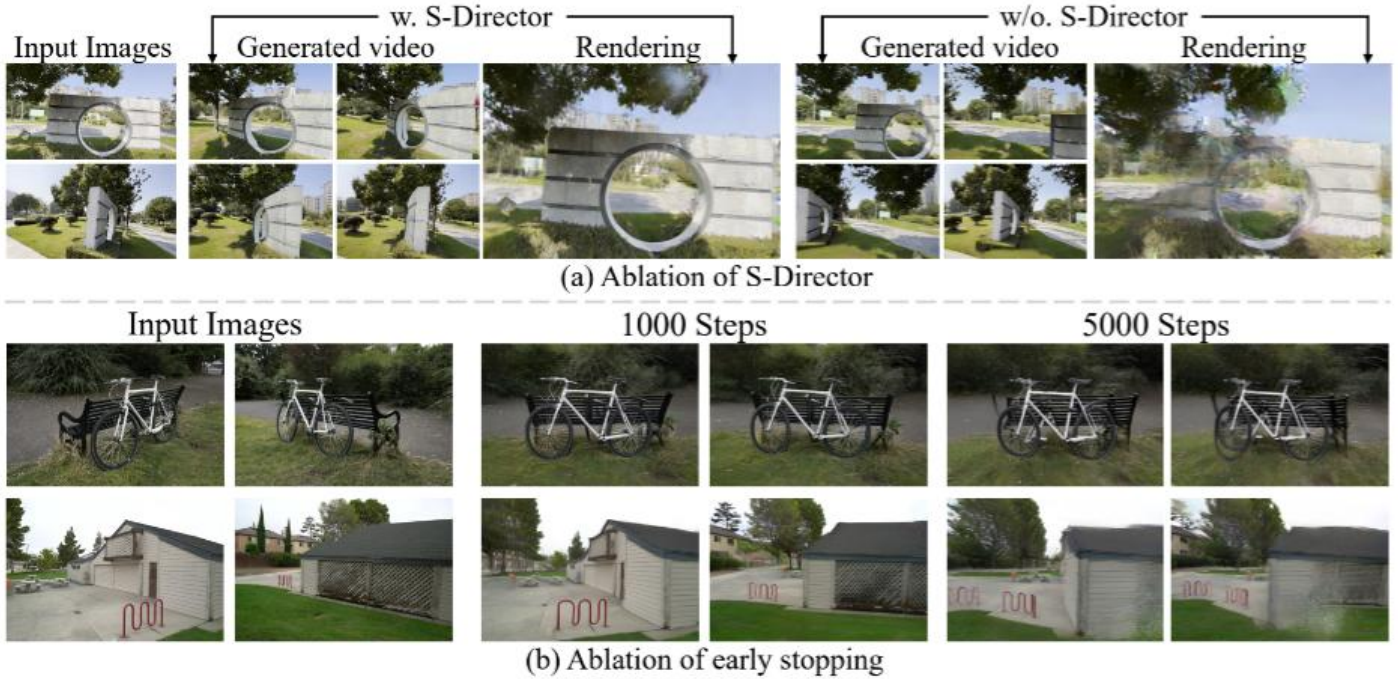


Figure 10. **Ablation study on the sparse-view 3D generation.**

Method	PSNR↑	SSIM↑	LPIPS↓
Ours	**20.42**	**0.668**	**0.185**
w/o. early stopping	17.93	0.492	0.311
w/o. S-Director	17.32	0.521	0.278

Table 3. **Ablation study for 3D generation.**

Method	Visual Quality		View Synchronization	
	CLIP-T ↑	CLIP-F ↑	Mat. Pix.(K) ↑	CLIP-V ↑
SV4D	30.97	98.31	294.3	87.8
M.V. Img+CogVideoX	34.05	98.67	323.0	96.1
Base	32.59	98.77	563.7	96.2
+ Ref Video	34.50	99.20	766.0	97.8
+ Both	**35.83**	**99.38**	**834.3**	**98.3**

Table 4. **4D quantitative comparison & ablation.**

Trajectory-aware mechanism for 3D generation. As illustrated in Fig. 10 (a), when handling the large-angle sparse view, the absence of S-Director often results in the "Janus problem", where multiple heads are generated, significantly degrading reconstruction quality (shown in Tab. 3). Moreover, the result in Fig. 10 (b) indicates that, in comparison to training for 5000 steps, only training S-Director for 1000 steps is able to handle more complex and flexible inputs, showcasing DimensionX's generalization ability.

Identity-preserving denoising for 4D generation. As presented in Tab. 4, through reference video latent sharing and appearance refinement, we achieve high consistency in global background, subject motion, and detailed appearance across frames. Please see Appendix F for more ablations.