Orthogonal Spatial-temporal Distributional Transfer for 4D Generation

Wei Liu[1], Shengqiong Wu[2*], Bobo Li[2], Haoyu Zhao[3], Hao Fei[2], Mong-Li Lee[2], Wynne Hsu[2]

[1] School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu, China,
[2] National University of Singapore,
[3] Wuhan University
liuwei628@aufe.edu.cn, swu@u.nus.edu, libobo@nus.edu.sg

Abstract

In the AIGC era, generating high-quality 4D content has garnered increasing research attention. Unfortunately, current 4D synthesis research is severely constrained by the lack of large-scale 4D datasets, preventing models from adequately learning the critical spatial-temporal features necessary for high-quality 4D generation, thus hindering progress in this domain. To combat this, we propose a novel framework that transfers rich spatial priors from existing 3D diffusion models and temporal priors from video diffusion models to enhance 4D synthesis. We develop a spatial-temporal-disentangled 4D (STD-4D) Diffusion model, which synthesizes 4D-aware videos through disentangled spatial and temporal latents. To facilitate the best feature transfer, we design a novel Orthogonal Spatial-temporal Distributional Transfer (Orster) mechanism, where the spatiotemporal feature distributions are carefully modeled and injected into the STD-4D Diffusion. Furthermore, during the 4D construction, we devise a spatial-temporal-aware HexPlane (ST-HexPlane) to integrate the transferred spatiotemporal features, thereby improving 4D deformation and 4D Gaussian feature modeling. Experiments demonstrate that our method significantly outperforms existing approaches, achieving superior spatial-temporal consistency and higher-quality 4D synthesis.

1 Introduction

As one of the key fields of computer vision, AIGC has witnessed rapid advancements in the latest decade, evolving from synthesizing static images (Rombach et al. 2022; Wu et al. 2023, 2024b, 2025b) to generating dynamic video content (Singer et al. 2022; Zuo et al. 2024; Fei et al. 2024, 2025), and achieving comprehensive 3D scene understanding & generation (Chen et al. 2024b; Shi et al. 2023). Recently, the focus has extended to understanding and generating 4D content (Ren et al. 2023; Zeng et al. 2025; Bahmani et al. 2024; Liang et al. 2024), representing the next frontier in visual modeling. 4D synthesis (Singer et al. 2023b,a; Wu et al. 2024c; Miao et al. 2025; Zheng et al. 2024) has emerged as an important research topic due to its significant potential in practical applications, including animation production (Huang et al. 2015), gaming (Li, Chen, and Liu 2024), and the AR/VR industry (Li et al. 2024).

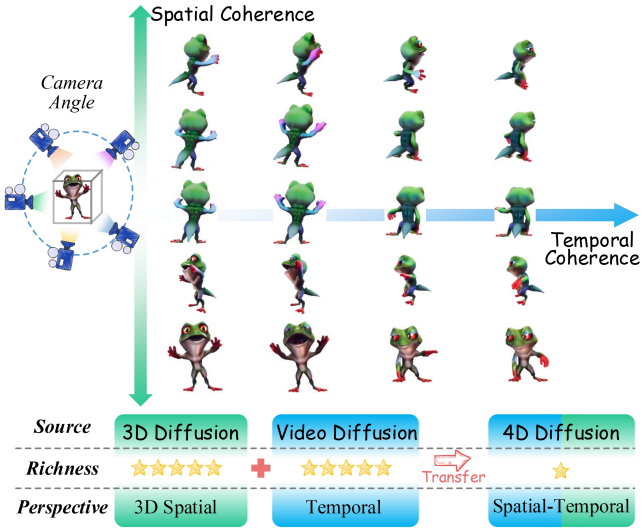

Figure 1: 4D generation requires spatial and temporal coherence simultaneously. While data resources in 4D are scarce, this paper proposes transferring the 3D spatial prior and temporal prior feature learning from existing resources-rich 3D diffusion and video diffusion, respectively.

Compared to images, videos, and 3D visual content, 4D visual data encompasses the most comprehensive characteristics, integrating both stringent spatial and temporal properties (Jiang et al. 2023). This makes high-quality 4D content generation particularly challenging, as it requires robust spatial-temporal feature modeling (Rahamim et al. 2024; Chen et al. 2024a; Yuan et al. 2024). Unfortunately, one of the greatest obstacles to progress in this area is the scarcity of labeled 4D datasets (Jiang et al. 2023), i.e., training a powerful 4D generation model inherently relies on access to large-scale 4D data. Earlier methods directly train models using the limited available 4D data (Deitke et al. 2023). However, the lack of sufficient supervision results in suboptimal spatial-temporal feature modeling, leading to limited 4D generation performance. To address this challenge, researchers (Liang et al. 2024) consider leveraging a pre-trained 3D-aware video diffusion model, where the extensive annotated resources of spatial priors from 3D diffusion

*Corresponding author.

models (Poole et al. 2023) and temporal priors from video diffusion models (Zuo et al. 2024) are integrated simultaneously. Fig. 1 illustrates this intuition. By performing final fine-tuning on a small amount of 4D data, the model can achieve an overall improvement in 4D generation.

Nevertheless, this research (Liang et al. 2024) straightforwardly injects the temporal video prior into a 3D backbone diffusion model. This method of integrating spatial and temporal priors still faces quite critical issues, which prevent current approaches from fully exploiting the spatial priors from 3D diffusion and the temporal priors from video diffusion. On the one hand, directly overlaying the temporal features onto the 3D spatial features leads to a catastrophic forgetting problem, where the later temporal representation dominates the original spatial features in the backbone 3D diffusion. Moreover, this integration method merely transfers feature representations without considering the disentanglement of temporal and spatial features. In the 4D generation, the characteristics of time and space follow completely different distributions, which are heterogeneous and orthogonal. For example, from the spatial aspect, the distribution describes different parts of a frog's geometry, while at the temporal level, it depicts the motions. Conversely, a completely different object with a distinct spatial distribution could perform the same action, thereby sharing the same temporal distribution. Thus, during the transfer process, spatial and temporal aspects should be appropriately modeled according to their respective distributions.

To combat these bottlenecks, this work proposes a novel 4D generation framework that fully leverages the abundant, rich static spatial and dynamic temporal features from 3D and video diffusions, respectively. As shown in Fig. 2, the overall 4D generation pipeline consists of a 4D video diffusion process and a 4D construction process. Technically, we first develop a spatial-temporal-disentangled 4D-aware Diffusion (**STD-4D** Diffusion) model to synthesize 4D-aware videos, wherein a disentangled representation mechanism for spatial and temporal latents is devised. The spatial-temporal disentangled 4D-UNet maintains the spatial and temporal feature representations in an interleaved manner, facilitating the separate transfer of spatial and temporal features. This is followed by explicit 4D construction using 4D Gaussian Splatting (4DGS) (Wu et al. 2024a) to generate high-quality 4D assets. From the generated 4D video, we decompose **i)** the spatial hexplane corresponding to the static 3D GS, and **ii)** the spatiotemporal hexplane corresponding to dynamic sequences, based on which we integrate the transferred spatial and temporal features to form the desired 4DGS feature, and finally decode it to obtain sequences of 4D gaussians, i.e., 4D assets.

Alongside the above system, we design a four-step spatial-temporal-enhanced 4D training process. Step-❶, the 4D diffusion model is preliminarily pre-trained on limited 4D data to establish a foundational understanding. Step-❷, we propose an Orthogonal Spatial-temporal Distributional Transfer (**Orster**) learning based on a knowledge distillation technique (Hinton 2015). We perform spatial transfer learning and temporal transfer learning simultaneously to inject spatial-temporal knowledge into our spatial-temporal-disentangled 4D-UNet from the host 3D and video diffusions, respectively. Step-❸, we further perform disentangled spatial-temporal consistency alignment on multi-view video data, ensuring learned spatial and temporal features are fully aligned, thereby guaranteeing the spatial-temporal consistency of the generated 4D content. Step-❹, we introduce a phase of conditional 4D generation training, enabling the synthesis of high-quality 4D assets based on various conditions, such as text prompts, images, or static 3D inputs.

To sum up, our main contributions are threefold: **i)** We present a novel framework to generate high-quality 4D content by transferring spatial-temporal priors from 3D and video diffusion models. **ii)** We develop a novel spatially-temporally disentangled 4D-aware diffusion model for 4D generation, incorporating an Orthogonal Spatial-temporal Distributional Transfer (Orster) learning mechanism, achieving highly effective knowledge transfer. **iii)** Extensive qualitative and quantitative experiments demonstrate that our method significantly outperforms existing approaches, generating 4D contents with superior spatial-temporal consistency and rich detail.

2 Related Work

Data Scarcity of 4D Generation. Recent advances in 3D content generation (Xiang et al. 2024; Shi et al. 2023; Sun et al. 2024; Li et al. 2024; Xu et al. 2025) have demonstrated remarkable success, driven by techniques such as NeRF (Mildenhall et al. 2022), which revolutionized 3D reconstruction by modeling high-quality static 3D scenes from sparse input views. While 3D generation focuses primarily on spatial consistency, 4D generation (Ren et al. 2024; Xu et al. 2024; Wang et al. 2024) extends these challenges by incorporating temporal dynamics, making it essential to model both spatial and temporal features simultaneously. Unfortunately, one significant challenge in advancing 4D generation is the scarcity of annotated 4D datasets. Unlike 3D data, which can often be synthesized or labeled using existing tools, collecting large-scale, high-quality 4D datasets is a labor-intensive process. Recent works (Singer et al. 2023a; Yu et al. 2024) attempt to address this by first learning static 3D shapes and then introducing motion using video diffusion models (Wu et al. 2025a; Chu et al. 2025; Yang et al. 2025). However, such approaches often result in suboptimal spatiotemporal modeling due to insufficient supervision from limited data. In contrast, we leverage well-trained 3D diffusion models and video diffusion models to learn rich spatial priors and temporal priors independently. By transferring such knowledge into 4D synthesis, we enable the generation of high-quality 4D content with superior spatial-temporal consistency, effectively overcoming data scarcity.

Spatial-temporal Modeling of 4D Generation. Diffusion models (Ho, Jain, and Abbeel 2020; Croitoru et al. 2023) have emerged as a dominant framework for generative tasks, achieving state-of-the-art (SoTA) performance across multiple domains. Notable works in video diffusion include video diffusion models (Zhang et al. 2024), which excel in generating temporally coherent video sequences. On the other hand, 3D-aware diffusion models, such as

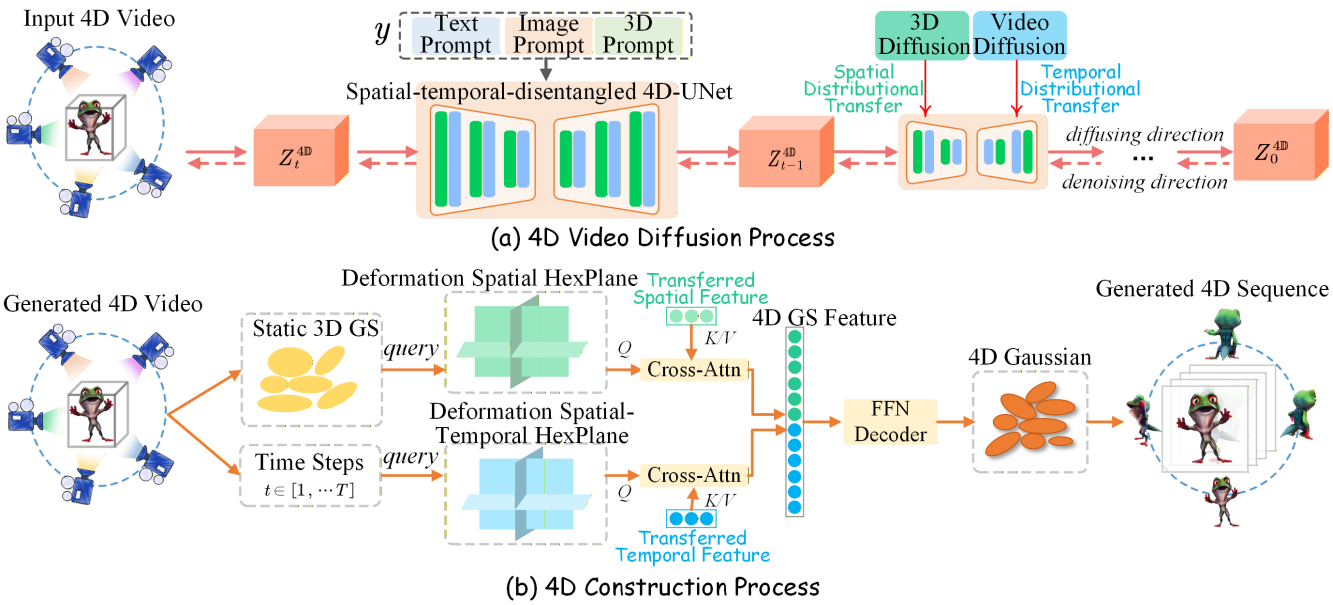

Figure 2: Overall pipeline of our 4D generation framework, including (a) 4D diffusion stage, and (b) 4D construction stage.

Latent-NeRF (Metzer et al. 2023) and 3DiM (Watson et al. 2023), ensure spatial consistency by generating multi-view-consistent images of static 3D objects. While these advancements represent significant progress, adapting these techniques to 4D generation presents unique challenges due to the need for integrating spatial and temporal features across dynamic scenes (Wang et al. 2025). Recent attempts like Diffusion4D (Liang et al. 2024) tackle this by combining 3D-aware and video diffusion models, but they often struggle with latent entanglement and temporal inconsistencies. Our work builds upon these advancements by proposing a novel 4D diffusion model that disentangles spatial and temporal latents. This allows for dedicated modeling of each aspect while maintaining its coherence in the generation process. Through feature distillation, we inject spatial priors from pre-trained 3D diffusion models and temporal priors from pre-trained video diffusion models into our 4D diffusion framework. Also, our 4D construction incorporates the SoTA dynamic 4DGS (Wu et al. 2024a) with HexPlane deformation technology (Cao and Johnson 2023), enabling the synthesis of high-quality 4D assets from generated videos.

3 Architecture of 4D Synthesis System

Task Definition and System Pipeline. Our system consists of two components: a 4D Diffusion module and a 4D Construction module. As seen in Fig. 2, given a prompt y (can be a text, image or static 3D content), the 4D Diffusion generates an orbital video $\mathcal{V} = \{I_i \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^{T}$ around a dynamic 3D asset during the denoising process. $\mathcal{V}$ consists of T multi-view images, $\mathcal{T} = \{\tau_i\}_{i=1}^{T}$ along a predefined camera trajectory, where H and W represent the image height and width. Then, the 4D Construction generates a high-quality 4D asset $\mathcal{G}^{4D}$ from the orbital video $\mathcal{V}$.

3.1 Spatial-temporal-disentangled 4D Diffusion

The core of our 4D-aware diffusion model is a spatial-temporal-disentangled 4D-UNet framework. Fig. 3(b) illustrates this mechanism. The denoising process begins by encoding the 4D input data into a latent representation using a pre-trained Variational Autoencoder (VAE) (Kingma and Welling 2013). Specifically, the VAE first encodes a 4D input X^{4D} into a 4D latents Z_t^{4D}. We then introduce a disentanglement block, another VAE, to disentangle X^{4D} into spatial representation Z_t^S and temporal representation Z_t^T:

$$Z_t^S = \text{Disentangle}_{\text{spatial}}(Z_t^{4D}), \tag{1}$$

$$Z_t^T = \text{Disentangle}_{\text{temporal}}(Z_t^{4D}). \tag{2}$$

Each disentangled latent embedding is processed separately through a 4D-UNet, respectively. The spatial latent Z_t^S is processed via spatial denoising:

$$Z_{t-1}^S = \epsilon_\theta^S(Z_t^S, t), \tag{3}$$

and the Z_t^T is processed via temporal denoising:

$$Z_{t-1}^T = \epsilon_\theta^T(Z_t^T, t), \tag{4}$$

where ϵ_θ^S and ϵ_θ^T are spatial and temporal denoising, respectively. The disentangled denoised embeddings are then used to update the 4D latent, incorporating the conditional input y:

$$Z_{t-1}^{4D} = \text{FFN}(Z_{t-1}^S, Z_{t-1}^T | y). \tag{5}$$

Finally, the denoised latent embedding Z_{t-1}^{4D} is decoded back into the 4D domain via the VAE decoder:

$$\hat{X}^{4D} = \text{VAE}_{\text{dec}}(Z_{t-1}^{4D}). \tag{6}$$

This disentangled modeling ensures that spatial and temporal dynamics are effectively learned and aligned, during which we can inject the spatial and temporal features into the above process, respectively. This will be elaborated later.

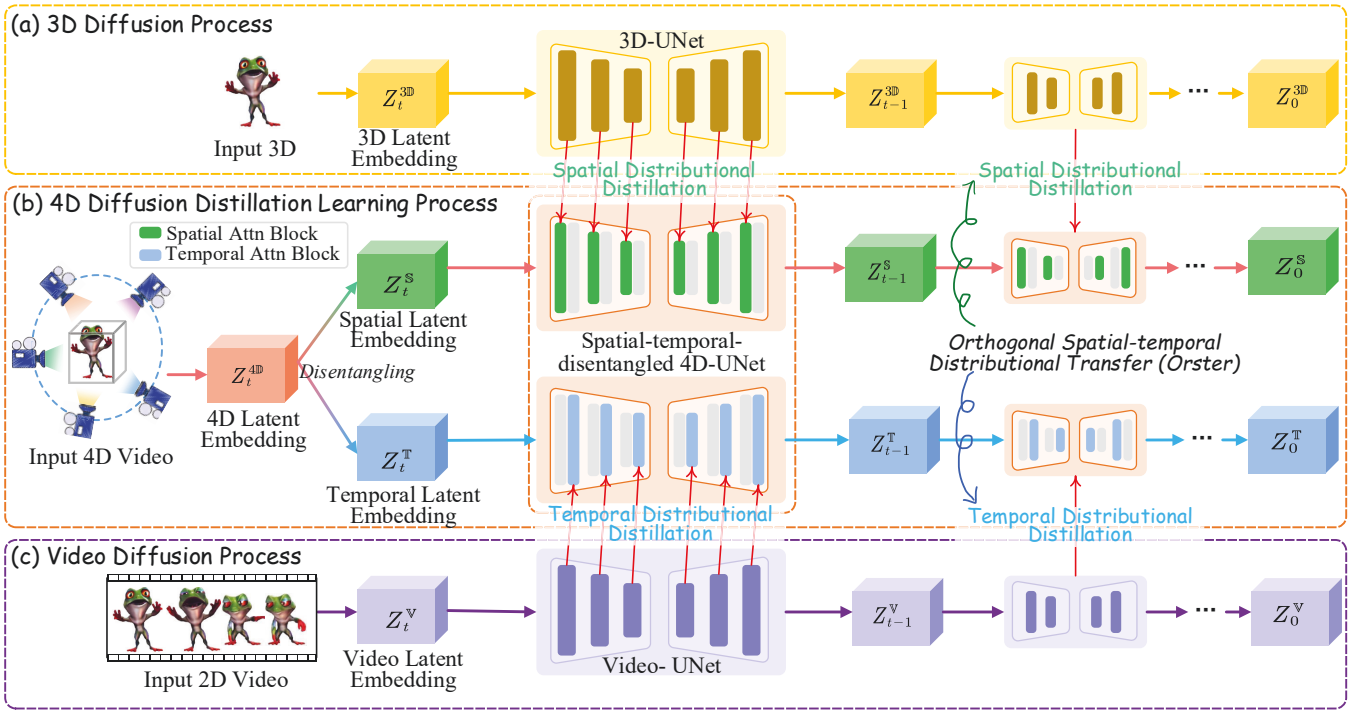

Figure 3: Overview of the knowledge transfer process between external 3D/video Diffusions and our STD-4D Diffusion, where STD-4D Diffusion disentangles the 4D latent into spatial and temporal channels, during which the spatial and temporal features from 3D diffusion and video diffusion are distilled into the spatial and temporal blocks of the 4D-UNet, respectively, via the Orthogonal Spatial-temporal Distributional Transfer (Orster) mechanism (cf. Fig. 5).

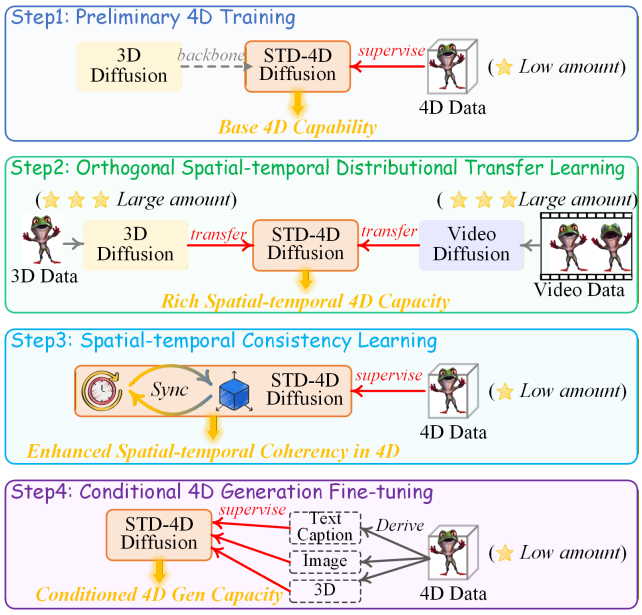

Figure 4: Four-stage STD-4D Diffusion training.

3.2 Gaussian Deformation for 4D Construction with Spatial-temporal-aware HexPlane

The construction module incorporates the SoTA dynamic 4DGS (Wu et al. 2024a) for 4D synthesis. After generating an orbital video $\mathcal{V}$, we extract spatial and temporal anchors, represented by static 3D Gaussians and their associated temporal sequence. To model the dynamic 4D representation, we employ a HexPlane (Cao and Johnson 2023) structure that encodes 4D spatial-temporal information by decomposing the 4D field into six deformation feature planes, spanning each pair of coordinate axes. Given a query point (x, y, z, t), the HexPlane predicts the transformation parameters, position displacement $\Delta\mathbf{p}$, rotation $\mathbf{R}$, and scale $\mathbf{s}$, for each Gaussian anchor $\mathcal{G}$:

$$\Delta\mathbf{p}, \mathbf{R}, \mathbf{s} \leftarrow \text{HexPlane}(D(x, y, z, t)), \tag{7}$$

where $D(\cdot)$ represents the transformations within HexPlane, and the process dynamically adjusts the Gaussians to capture object motion across time. To obtain these parameters more accurately, we consider a spatial-temporal-aware HexPlane (ST-HexPlane), which utilizes the transferred spatial prior $\mathbf{O}^s$ and temporal prior $\mathbf{O}^t$ from the existing 3D diffusion and video diffusion.

$$\mathbf{V}^{s/t} = \text{Attn}(Q, K, V) = \text{Attn}(D(x, y, z, t), \mathbf{O}^{s/t}, \mathbf{O}^{s/t}), \tag{8}$$

$$\Delta\hat{\mathbf{p}}, \hat{\mathbf{R}}, \hat{\mathbf{s}} \leftarrow \text{ST-HexPlane}([\mathbf{V}^s; \mathbf{V}^t]), \tag{9}$$

The updated Gaussians are then decoded for the 4D asset:

$$\mathcal{G}^{4D} = \text{Decoder}(\mathcal{G} + \Delta\mathbf{p}, \mathbf{R}, \mathbf{s}). \tag{10}$$

Refer to (Ren et al. 2023) for more decoding technical details. The process, as illustrated in Fig. 2(b), enables pre-

cise modeling of 4D dynamics by leveraging HexPlane's efficient spatial-temporal factorization and 4DGS's ability to represent high-fidelity dynamic structures.

4 Spatial-temporal-enhanced 4D Learning

Overview. Building upon the above 4D synthesis system, particularly with the Spatially-temporally Disentangled 4D diffusion, in this section, we propose spatial-temporal aware learning. Our key idea is to transfer the rich spatial and temporal features learned by 3D Diffusions and Video Diffusions, which benefit from extensive training signals, into our 4D diffusion framework (specifically the 4D U-Net), compensating for the lack of sufficient spatial-temporal signals required for 4D Diffusion generation. Technically, we design four training stages: 1) preliminary 4D training, 2) orthogonal spatial-temporal distributional transferring, 3) spatial-temporal consistency training, and 4) conditional 4D generation training. Fig. 4 illustrates the overall stages.

4.1 Preliminary 4D Training

As the first step, before performing transfer learning, we enable the backbone diffusion model with preliminary dynamic 4D generation capabilities. Considering that 3D and 4D are closely related modalities, we adopt a pre-trained 3D-aware video diffusion model (Zuo et al. 2024) as our backbone. Then, we train it on Objaverse (Deitke et al. 2023), consisting of multi-view dynamic orbital image sequences. This object is marked as $\mathcal{L}_{\text{ldm}} = ||\epsilon_t - \epsilon(Z_t, Y, t)||$, where Z_t is the 4D latent. Due to the very limited data size, the 4D features learned in this stage are quite restricted, serving primarily as foundational training to establish a baseline for subsequent transfer learning.

4.2 Orthogonal Spatial-temporal Distributional Transfer (Orster) Learning

Based on the distillation learning (Hinton 2015), we now perform Orster learning to gain rich spatiotemporal features. As shown in Fig. 3, building on our STD-4D Diffusion framework, we distill the spatial and temporal features from external well-trained 3D diffusion and video diffusion models, respectively. Via our Orster mechanism, the spatial geometry features from the UNet of the 3D diffusion model are injected into the spatial blocks of the 4D-UNet, while the temporal features from the UNet of the video diffusion model are fused into the temporal blocks of the 4D-UNet.

As mentioned earlier, the spatial and temporal distributions should be modeled separately during the distillation process. However, within the same 4D scene, time and space also exhibit a reasonable joint distribution. To capture the intricate interaction between spatial and temporal feature distributions, we propose Orster as shown in Fig. 5. First, for any spatial embedding f_s from host 3D Diffusion and temporal embedding f_t from host video Diffusion, we define the joint spatiotemporal distribution Gaussian Kernel as:

$$\kappa(f_s, f_t) = \exp\left(\frac{-1}{2}\left(\frac{\|f_s - g_s\|^2}{\sigma_s^2} + \frac{\|f_t - g_t\|^2}{\sigma_t^2} + \frac{2\alpha\,\langle f_s - g_s,\, f_t - g_t\rangle}{\sigma_{st}^2}\right)\right), \tag{11}$$

where (g_s, g_t) are the mean spatial and temporal features of all (f_s, f_t), scale parameters $\sigma_s, \sigma_t, \sigma_{st}$ and α are learnable. Then, we compute the resulting feature representation via Spatial/Temporal Cross-Attention, respectively:

$$\bar{f_s^{3d}} = \text{Spt-Attn}(f_s^{3d}, \kappa, f_s^{3d}), \tag{12}$$

$$\bar{f_t^{v}} = \text{Tmpr-Attn}(f_t^{v}, \kappa, f_t^{v}). \tag{13}$$

Lastly, we conduct the distillation over $\bar{f}$ and f:

$$\mathcal{L}_{\text{orster}} = \underbrace{\lambda_o||f_s^{4d} - \bar{f_s^{3d}}||}_{\mathcal{L}^s_{\text{orster}}} + \underbrace{(1-\lambda_o)||f_t^{4d} - \bar{f_t^{v}}||}_{\mathcal{L}^t_{\text{orster}}}, \tag{14}$$

where λ_o is a term weight. This approach helps the effective transfer of spatial-temporal features into the 4D diffusion model. Once the transfer learning is completed, we merge the spatial and temporal blocks from the two channels into a unified 4D-UNet in a complete and integrated view.

4.3 Spatial-temporal Consistency Learning

Despite obtaining sufficiently high-quality spatiotemporal features in the previous step via sophisticated modeling, inconsistencies may still arise because the spatial and temporal features are extracted from different sources. To address this, we perform spatial-temporal consistency training to further refine and align the learned features. We leverage the same multi-view 4D video dataset to jointly tune the spatial and temporal alignment, ensuring coherent integration across both dimensions. Our overall objective is given by:

$$\mathcal{L}_{\text{const}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}, \tag{15}$$

where the reconstruction loss $\mathcal{L}_{\text{rec}} = \|I_{\text{pred}} - I_{\text{gt}}\|_1$ ensures accurate view-wise reconstruction, and the perceptual loss $\mathcal{L}_{\text{perc}} = \sum_l \|\phi_l(I_{\text{pred}}) - \phi_l(I_{\text{gt}})\|_2^2$ (with $\phi_l(\cdot)$ denoting features from a pretrained network) captures high-level visual fidelity. Also, the temporal smoothness loss $\mathcal{L}_{\text{temp}} = \|F_t^{(t+1)} - F_t^{(t)}\|_2^2$ promotes consistency across consecutive frames, and the feature alignment loss $\mathcal{L}_{\text{align}} = 1 - \frac{\langle F_s, F_t\rangle}{\|F_s\|\cdot\|F_t\|}$ enforces coherent integration between spatial and temporal features. Iterative optimization of $\mathcal{L}_{\text{const}}$ gradually refines the spatiotemporal representations, ensuring precise alignment and overall coherence in the generated 4D content.

4.4 Conditional 4D Generation Fine-tuning

Our framework supports various prompts y as generation conditions, including texts, images, or static 3D content. For text conditions, text embeddings T are extracted using the CLIP model and injected into the 4D-UNet via a cross-attention mechanism. For image conditions, the first-view image I_0 of the orbital video V, captured at timestamp 0, is used as the reference image, which is injected into the 4D U-Net through the cross-attention mechanism. For static

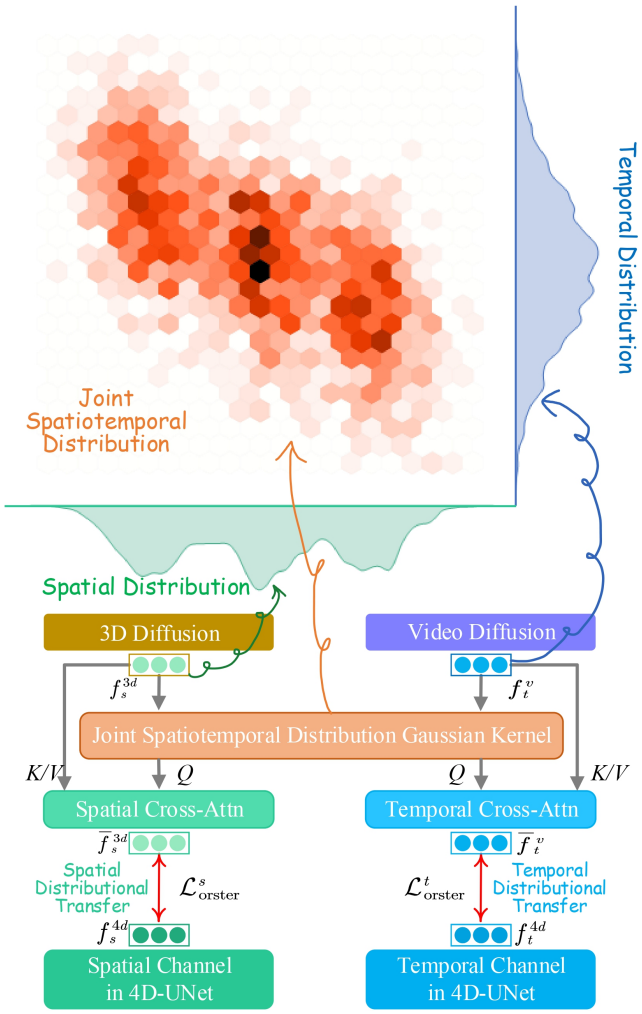

Figure 5: A closer look at the Orthogonal Spatial-temporal Distributional Transfer (Orster) learning module.

3D conditions, the video V is used as the reference, where video features are extracted by a pre-trained encoder and fed into the diffusion model. Thus, we perform diverse Conditional 4D Generation Training, enabling the model to learn to synthesize high-quality 4D assets from various conditions. Specifically, we optimize the following objective:

$$\mathcal{L}_{\text{cond}} = \mathbb{E}_{Z_0,(T|I_0|V),t,\epsilon}\left[\|\epsilon - \epsilon_\theta(Z_t,(T|I_0|V),t)\|^2\right]. \tag{16}$$

4.5 4D Construction Optimization

In addition to training our STD-4D Diffusion module, the 4D Construction process (cf. Fig. 2(b)) also requires optimization. First, for the first-view image I_0 in the video V, we use a pretrained 3D-aware video Diffusion model to generate an orbital-view video $\bar{V}'$ of a static 3D object. As the 4D-aware video Diffusion model is finetuned from that pretrained model, there is a high 3D geometric consistency between $\bar{V}'$ and V. Hence, we train with this data. For the ST-HexPlane's Deformation Field optimization, we follow (Ren et al. 2023), fix the camera to the reference view, and minimize the Mean Squared Error (MSE) between the rendered image and the driving video frame at each timestamp,

$$\mathcal{L}_{\text{hex}} = \frac{1}{\tau}\sum_{t=1}^{\tau}\|f(\phi(S,\tau),\, o_{\text{ref}}) - I^{\tau}_{\text{ref}}\|^2. \tag{17}$$

We then refine the 4DGS using only V to enhance spatiotemporal awareness; thanks to the 4D consistency in our generated videos, precise pixel-level matching across different views and timestamps can be achieved. Following (Liang et al. 2024), we adopt L_1 and L_{lpips} (Zhang et al. 2018) losses for optimization and introduce a depth smoothness loss (Niemeyer et al. 2022) as a regularizer to reinforce geometric smoothness. The total loss is formulated as:

$$\mathcal{L}_{\text{gs}} = \lambda_{l1}\,\mathcal{L}_{l1} + \lambda_{\text{lpips}}\,\mathcal{L}_{\text{lpips}} + \lambda_{\text{dep}}\,\mathcal{L}_{\text{dep}} + \lambda_{\text{hex}}\,\mathcal{L}_{\text{hex}}. \tag{18}$$

5 Experiment

5.1 Settings

Implementation Details. We adopt VideoMV (Zuo et al. 2024) as the backbone 3D diffusion model, and ModelScopeT2V (Wang et al. 2023), and I2VGen-XL (Zhang et al. 2023) as text-to-video and image-to-video diffusion backbone, respectively. Our 4D-aware video diffusion model is trained for 5,000 iterations. During the sampling phase, we condition the model on text prompts, front-view images, or orbital videos of static 3D assets starting from the front view. In the 4D construction stage, we optimize the 4DGS representation in two phases: a coarse optimization phase consisting of 6,000 iterations and a fine optimization phase consisting of 3,000 iterations. We employ the Consistent4D (Jiang et al. 2023) as a test set.

Evaluations. We consider both automatic quantitative evaluation and human qualitative evaluation. Following (Liang et al. 2024), we adopt CLIP-O, where 36 uniformly rendered orbital views of the generated 4D assets are used as targets, and CLIP-F, which evaluates scores using only front-view images as targets. Also, we utilize LPIPS, PSNR, SSIM, and FVD to assess the appearance, texture quality, and the spatial-temporal consistency of the generated visuals. The baselines include 4DFY (Bahmani et al. 2024), Animate124 (Zhao et al. 2023), Diffusion4D (Liang et al. 2024), 4DGen (Yin et al. 2023), and STAG4D (Zeng et al. 2025).

5.2 Overall Results and Ablation Study

Overall Results. We compare our approach with several SoTA baselines under text-to-4D, image-to-4D, and 3D-to-4D settings. Table 1 presents the comprehensive results, from which we can derive the following key observations. First, we observe that the overall performance varies across different settings, with the general trend being 3D condition $>$ Image $>$ Text. This is because the former provides richer priors and initial information for feature generation. Second, our proposed system consistently outperforms all strong-performing baselines across all settings and metrics, directly demonstrating its effectiveness comprehensively.

Ablation Study on System Architecture. In Table 2, we present the system ablation to explore the fine-grained con-

Figure 6: 4D generation comparisons with SoTA baseline. Best viewing via zooming in.

Model	CLIP-F↑	CLIP-O↑	SSIM↑	PSNR↑	LPIPS↓	FVD↓
• **Text-to-4D**						
4DFY	0.78	0.61	-	14.2	0.23	1042.3
Animate124	0.75	0.58	-	15.0	0.21	720.5
Diffusion4D	0.82	0.69	-	15.3	0.20	684.0
Ours	**0.85**	**0.72**	-	**16.8**	**0.19**	**523.4**
• **Image-to-4D**						
4DGen	0.84	0.71	0.69	14.4	0.31	736.6
STAG4D	0.86	0.72	0.76	15.2	0.27	675.4
Diffusion4D	0.90	0.80	0.82	16.8	0.19	490.2
Ours	**0.93**	**0.82**	**0.84**	**17.3**	**0.17**	**477.7**
• **3D-to-4D**						
Diffusion4D	0.91	0.81	0.83	17.2	0.18	482.4
Ours	**0.95**	**0.84**	**0.85**	**17.6**	**0.16**	**465.3**

Table 1: Main results under different task settings.

	CLIP-F↑	CLIP-O↑	PSNR↑	LPIPS↓	FVD↓
Ours (complete)	0.93	0.82	17.3	0.17	477.7
• **4D Diffusion**					
w/o Disentangling	0.77	0.65	15.0	0.31	596.7
w/o Spt Channel	0.87	0.77	16.3	0.25	501.1
w/o Tmp Channel	0.86	0.78	16.5	0.24	549.3
• **4D Construction**					
w/o Spt Feat. (Eq. 8)	0.86	0.76	16.6	0.24	522.4
w/o Tmp Feat. (Eq. 8)	0.89	0.81	16.9	0.21	578.5

Table 2: Ablation study on system architecture.

tributions of different modules. First, for the backbone 4D Diffusion, we analyze the disentangling mechanism, the roles of the spatial block, and the temporal block. When either of these blocks is removed, we observe varying degrees of performance degradation, highlighting their essential impacts, where the overall spatiotemporal disentangling mechanism shows the biggest influence. Also, for the 4D Construction module, we see that both the spatial feature and temporal feature exhibit varied yet non-negligible impacts.

Ablation Study on Learning Strategy. Next, we analyze the impact of training strategies, cf. Table **??**. Removing the preliminary training ($\mathcal{L}_{\text{ldm}}$), or spatiotemporal consistency learning ($\mathcal{L}_{\text{const}}$) and 4D construction training ($\mathcal{L}_{\text{gs}}$) result in varied performance drop. Notably, the proposed

	CLIP-F↑	CLIP-O↑	PSNR↑	LPIPS↓	FVD↓
Ours (complete)	0.93	0.82	17.3	0.17	477.7
w/o Pre. 4D ($\mathcal{L}_{\text{ldm}}$)	0.82	0.60	15.7	0.38	601.6
w/o Orster ($\mathcal{L}_{\text{orster}}$)	0.40	0.32	12.7	0.36	668.3
w/o Spt (λ_o=0)	0.67	0.50	15.0	0.40	503.9
w/o Tmp (λ_o=1)	0.59	0.55	14.3	0.34	557.7
w/o ST Kernal (Eq. 11)	0.84	0.67	16.3	0.44	566.4
w/o Spt Attn. (Eq. 12)	0.80	0.70	15.7	0.40	536.8
w/o Tmp Attn. (Eq. 13)	0.78	0.71	16.0	0.41	560.3
w/o ST Consis. ($\mathcal{L}_{\text{const}}$)	0.85	0.72	16.1	0.23	542.2
w/o 4D Constr. ($\mathcal{L}_{\text{gs}}$)	0.68	0.60	15.3	0.37	587.4

Table 3: Ablation study on learning strategies.

Orster learning ($\mathcal{L}_{\text{orster}}$) contributes most significantly. Stepping into Orster, we see that the joint spatiotemporal distribution Gaussian kernel and also the spatial & temporal attention mechanisms play an important role.

5.3 Qualitative Results with Visualizations

We provide visual comparisons to better aid the understanding of how advanced our system can be in 4D generation. As shown in Fig. 6, our system generates accurate and very realistic 4D assets with much more excellent spatial and temporal consistency. While the baseline Diffusion4D model produces 4D assets, their results often exhibit quite low-quality geometry of appearances or nearly imperceptible motion. In contrast, our approach delivers high-fidelity and dynamic actions, and very fine details and more refined appearances

6 Conclusion

This paper proposes a novel framework for high-quality 4D generation by transferring rich spatial priors from 3D diffusion models and temporal priors from video diffusion models. We develop a Spatial-Temporal-Disentangled 4D Diffusion model that synthesizes 4D-aware videos through disentangled spatial and temporal latent spaces. To facilitate effective cross-modal prior transfer, we introduce the Orster mechanism, which injects spatiotemporal feature distributions into the STD-4D Diffusion. Extensive experiments demonstrate that our approach achieves superior spatial–temporal consistency and overall higher 4D generation quality than existing baseline methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62202001), and the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001).

References

Bahmani, S.; Skorokhodov, I.; Rong, V.; Wetzstein, G.; Guibas, L.; Wonka, P.; Tulyakov, S.; Park, J. J.; Tagliasacchi, A.; and Lindell, D. B. 2024. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of CVPR*, 7996–8006.

Cao, A.; and Johnson, J. 2023. Hexplane: A fast representation for dynamic scenes. In *Proceedings of CVPR*, 130–141.

Chen, C.; Huang, S.; Chen, X.; Chen, G.; Han, X.; Zhang, K.; and Gong, M. 2024a. Ct4d: Consistent text-to-4d generation with animatable meshes. *arXiv preprint arXiv:2408.08342*.

Chen, S.; Chen, X.; Zhang, C.; Li, M.; Yu, G.; Fei, H.; Zhu, H.; Fan, J.; and Chen, T. 2024b. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the CVPR*, 26428–26438.

Chu, W.; Ke, L.; Liu, J.; Huo, M.; Tokmakov, P.; and Fragkiadaki, K. 2025. Robust Multi-Object 4D Generation for In-the-wild Videos. In *Proceedings of the CVPR*, 22067–22077.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10850–10869.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the CVPR*, 13142–13153.

Fei, H.; Wu, S.; Ji, W.; Zhang, H.; and Chua, T.-S. 2024. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the CVPR*, 7641–7653.

Fei, H.; Zhou, Y.; Li, J.; Li, X.; Xu, Q.; Li, B.; Wu, S.; Wang, Y.; Zhou, J.; Meng, J.; et al. 2025. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the ICML*.

Hinton, G. 2015. Distilling the Knowledge in a Neural Network. *Proceedings of the NeurIPS Deep Learning Workshop*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Proceedings of the NeuraIPS*, 33: 6840–6851.

Huang, P.; Tejera, M.; Collomosse, J.; and Hilton, A. 2015. Hybrid Skeletal-Surface Motion Graphs for Character Animation from 4D Performance Capture. *ACM Trans. Graph.*, 34(2).

Jiang, Y.; Zhang, L.; Gao, J.; Hu, W.; and Yao, Y. 2023. Consistent4D: Consistent 360° Dynamic Object Generation from Monocular Video. In *Proceedings of the ICLR*.

Kingma, D. P.; and Welling, M. 2013. Auto-Encoding Variational Bayes. *arXiv e-prints*, arXiv–1312.

Li, R.; Pan, P.; Yang, B.; Xu, D.; Zhou, S.; Zhang, X.; Li, Z.; Kadambi, A.; Wang, Z.; Tu, Z.; et al. 2024. 4k4dgen: Panoramic 4d generation at 4k resolution. *arXiv preprint arXiv:2406.13527*.

Li, Z.; Chen, Y.; and Liu, P. 2024. DreamMesh4D: Video-to-4D Generation with Sparse-Controlled Gaussian-Mesh Hybrid Representation. In *Proceedings of the NeurIPS*.

Liang, H.; Yin, Y.; Xu, D.; Liang, H.; Wang, Z.; Plataniotis, K. N.; Zhao, Y.; and Wei, Y. 2024. Diffusion4D: fast spatial-temporal consistent 4D generation via video diffusion models. In *Proceedings of the NeuraIPS*, 110854–110875.

Metzer, G.; Richardson, E.; Patashnik, O.; Giryes, R.; and Cohen-Or, D. 2023. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures. In *Proceedings of the CVPR*, 12663–12673.

Miao, Q.; Li, K.; Quan, J.; Min, Z.; Ma, S.; Xu, Y.; Yang, Y.; and Luo, Y. 2025. Advances in 4D Generation: A Survey. *arXiv preprint arXiv:2503.14501*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2022. NeRF: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1): 99–106.

Niemeyer, M.; Barron, J. T.; Mildenhall, B.; Sajjadi, M. S.; Geiger, A.; and Radwan, N. 2022. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the CVPR*, 5480–5490.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *Proceedings of the ICLR*.

Rahamim, O.; Malca, O.; Samuel, D.; and Chechik, G. 2024. Bringing Objects to Life: 4D generation from 3D objects. *arXiv preprint arXiv:2412.20422*.

Ren, J.; Pan, L.; Tang, J.; Zhang, C.; Cao, A.; Zeng, G.; and Liu, Z. 2023. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*.

Ren, J.; Xie, C.; Mirzaei, A.; Kreis, K.; Liu, Z.; Torralba, A.; Fidler, S.; Kim, S. W.; Ling, H.; et al. 2024. L4gm: Large 4d gaussian reconstruction model. *Proceedings of the NeuraIPS*, 37: 56828–56858.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the CVPR*, 10684–10695.

Shi, Y.; Wang, P.; Ye, J.; Long, M.; Li, K.; and Yang, X. 2023. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Singer, U.; Sheynin, S.; Polyak, A.; Ashual, O.; Makarov, I.; Kokkinos, F.; Goyal, N.; Vedaldi, A.; Parikh, D.; Johnson, J.; and Taigman, Y. 2023a. Text-To-4D Dynamic Scene Generation. In *Proceedings of the ICML*, 31915–31929.

Singer, U.; Sheynin, S.; Polyak, A.; Ashual, O.; Makarov, I.; Kokkinos, F.; Goyal, N.; Vedaldi, A.; Parikh, D.; Johnson, J.; et al. 2023b. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*.

Sun, W.; Chen, S.; Liu, F.; Chen, Z.; Duan, Y.; Zhang, J.; and Wang, Y. 2024. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*.

Wang, C.; Zhuang, P.; Ngo, T. D.; Menapace, W.; Siarohin, A.; Vasilkovsky, M.; Skorokhodov, I.; Tulyakov, S.; Wonka, P.; and Lee, H. 2025. 4Real-Video: Learning Generalizable Photo-Realistic 4D Video Diffusion. In *Proceedings of the CVPR*, 17723–17732.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023. ModelScope Text-to-Video Technical Report. arXiv:2308.06571.

Wang, X.; Ma, W.; Wang, A.; Chen, S.; Kortylewski, A.; and Yuille, A. 2024. Compositional 4D Dynamic Scenes Understanding with Physics Priors for Video Question Answering. *arXiv preprint arXiv:2406.00622*.

Watson, D.; Chan, W.; Brualla, R. M.; Ho, J.; Tagliasacchi, A.; and Norouzi, M. 2023. Novel View Synthesis with Diffusion Models. In *Proceedings of the ICLR*.

Wu, G.; Yi, T.; Fang, J.; Xie, L.; Zhang, X.; Wei, W.; Liu, W.; Tian, Q.; and Wang, X. 2024a. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. In *Proceedings of the CVPR*, 20310–20320.

Wu, R.; Gao, R.; Poole, B.; Trevithick, A.; Zheng, C.; Barron, J. T.; and Holynski, A. 2025a. CAT4D: Create Anything in 4D with Multi-View Video Diffusion Models. In *Proceedings of the CVPR*, 26057–26068.

Wu, S.; Fei, H.; Qu, L.; Ji, W.; and Chua, T.-S. 2024b. Next-gpt: Any-to-any multimodal llm. In *Proceedings of the ICML*.

Wu, S.; Fei, H.; Yang, J.; Li, X.; Li, J.; Zhang, H.; and Chua, T.-s. 2025b. Learning 4d panoptic scene graph generation from rich 2d visual scene. In *Proceedings of the CVPR*, 24539–24549.

Wu, S.; Fei, H.; Zhang, H.; and Chua, T.-S. 2023. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. *Proceedings of the NeuraIPS*, 36: 79240–79259.

Wu, Z.; Yu, C.; Jiang, Y.; Cao, C.; Wang, F.; and Bai, X. 2024c. Sc4d: Sparse-controlled video-to-4d generation and motion transfer. In *Proceedings of the ECCV*, 361–379. Springer.

Xiang, J.; Lv, Z.; Xu, S.; Deng, Y.; Wang, R.; Zhang, B.; Chen, D.; Tong, X.; and Yang, J. 2024. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*.

Xu, D.; Liang, H.; Bhatt, N. P.; Hu, H.; Liang, H.; Plataniotis, K. N.; and Wang, Z. 2024. Comp4d: Llm-guided compositional 4d scene generation. *arXiv preprint arXiv:2403.16993*.

Xu, H.; Xu, G.; Zheng, Z.; Zhu, X.; Ji, W.; Li, X.; Guo, R.; Zhang, M.; Fei, H.; et al. 2025. VimoRAG: Video-based Retrieval-augmented 3D Motion Generation for Motion Language Models. In *Proceedings of the NeuraIPS*.

Yang, L.; Liu, C.; Zhu, Z.; Liu, A.; Ma, H.; Nong, J.; and Liang, Y. 2025. Not All Frame Features Are Equal: Video-to-4D Generation via Decoupling Dynamic-Static Features. *arXiv preprint arXiv:2502.08377*.

Yin, Y.; Xu, D.; Wang, Z.; Zhao, Y.; and Wei, Y. 2023. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*.

Yu, H.; Wang, C.; Zhuang, P.; Menapace, W.; Siarohin, A.; Cao, J.; Jeni, L.; Tulyakov, S.; and Lee, H.-Y. 2024. 4real: Towards photorealistic 4d scene generation via video diffusion models. *Proceedings of the NeuraIPS*, 37: 45256–45280.

Yuan, Y.-J.; Kobbelt, L.; Liu, J.; Zhang, Y.; Wan, P.; Lai, Y.-K.; and Gao, L. 2024. 4dynamic: Text-to-4d generation with hybrid priors. *arXiv preprint arXiv:2407.12684*.

Zeng, Y.; Jiang, Y.; Zhu, S.; Lu, Y.; Lin, Y.; Zhu, H.; Hu, W.; Cao, X.; and Yao, Y. 2025. Stag4d: Spatial-temporal anchored generative 4d gaussians. In *Proceedings of the ECCV*, 163–179.

Zhang, H.; Chen, X.; Wang, Y.; Liu, X.; Wang, Y.; and Qiao, Y. 2024. 4diffusion: Multi-view video diffusion model for 4d generation. *Proceedings of the NeuraIPS*, 37: 15272–15295.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the CVPR*, 586–595.

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

Zhao, Y.; Yan, Z.; Xie, E.; Hong, L.; Li, Z.; and Lee, G. H. 2023. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*.

Zheng, Y.; Li, X.; Nagano, K.; Liu, S.; Hilliges, O.; and De Mello, S. 2024. A unified approach for text-and image-guided 4d scene generation. In *Proceedings of the CVPR*, 7300–7309.

Zuo, Q.; Gu, X.; Qiu, L.; Dong, Y.; Zhao, Z.; Yuan, W.; Peng, R.; Zhu, S.; Dong, Z.; Bo, L.; et al. 2024. Videomv: Consistent multi-view generation based on large video generative model. *arXiv preprint arXiv:2403.12010*.