4DSTR: Advancing Generative 4D Gaussians with Spatial-Temporal Rectification for High-Quality and Consistent 4D Generation

Mengmeng Liu[1*], Jiuming Liu[2*], Yunpeng Zhang[3], Jiangtao Li[3], Michael Ying Yang[4], Francesco Nex[1], Hao Cheng[1†]

[1] University of Twente
[2] University of Cambridge
[3] PhiGent Robotics
[4] University of Bath
m.liu-1@utwente.nl, jl2538@cam.ac.uk, h.cheng-2@utwente.nl

Abstract

Remarkable advances in recent 2D image and 3D shape generation have induced a significant focus on dynamic 4D content generation. However, previous 4D generation methods commonly struggle to maintain spatial-temporal consistency and adapt poorly to rapid temporal variations, due to the lack of effective spatial-temporal modeling. To address these problems, we propose a novel 4D generation network called 4DSTR, which modulates generative 4D Gaussian Splatting with spatial-temporal rectification. Specifically, temporal correlation across generated 4D sequences is designed to rectify deformable scales and rotations and guarantee temporal consistency. Furthermore, an adaptive spatial densification and pruning strategy is proposed to address significant temporal variations by dynamically adding or deleting Gaussian points with the awareness of their pre-frame movements. Extensive experiments demonstrate that our 4DSTR achieves state-of-the-art performance in video-to-4D generation, excelling in reconstruction quality, spatial-temporal consistency, and adaptation to rapid temporal movements.

Introduction

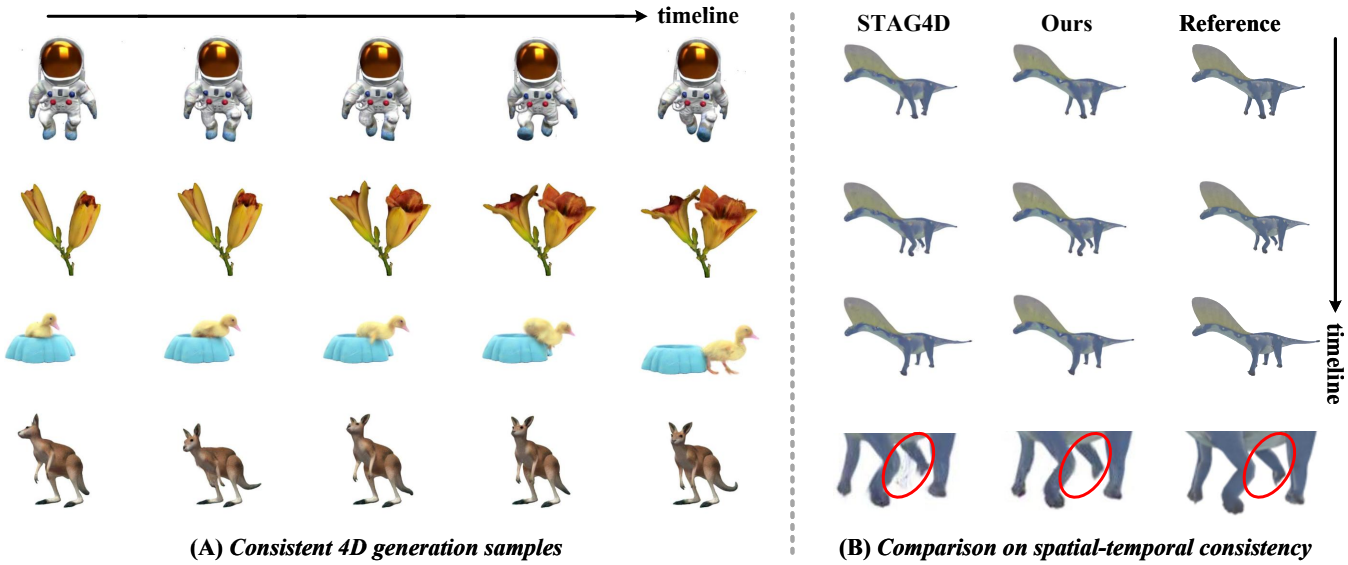
Recently, there have been advancements in generating high-quality and diverse visual contents with pre-trained diffusion models (Ho, Jain, and Abbeel 2020), including 2D images (Ho, Jain, and Abbeel 2020; Zhou et al. 2025a), 3D shapes (Liu et al. 2025b), etc. These successful experiences naturally boost the exploration using generative diffusion models for dynamic 4D content generation (Singer et al. 2023; Bahmani et al. 2024; Wu et al. 2024b), which has various applications in autonomous driving simulation (Liu et al. 2023a, 2024c; Cheng et al. 2023; Cheng, Liu, and Chen 2023; Ni et al. 2025), virtual reality (Huang et al. 2025c,b), and digital avatar animation (Wang et al. 2025a), etc.

A common research line takes text (Singer et al. 2023; Ling et al. 2024; Wang et al. 2025b) as input conditions, leveraging pre-trained text-to-image or text-to-video diffusion models as the preprocess. MAV3D (Singer et al. 2023) is a pioneering framework for text-to-4D generation that utilizes the pre-trained text-to-image and text-to-video diffusion models for the static and dynamic sequence generation, supervised by the Score Distillation Sampling (SDS) loss. Subsequent approaches design trajectory conditions (Bahmani et al. 2024) or deformable 4D Gaussian Splatting (Ling et al. 2024) to enhance the motion fidelity and appearance quality of generated 4D samples. However, recent text-to-4D generation networks fail to capture the multi-view spatial-temporal consistency and struggle with effective knowledge distillation from diffusion models as in Fig. 1.

Another research line focuses on the video-to-4D generation task. Consistent4D (Jiang et al. 2024c) firstly proposes an Interpolation-driven Consistency Loss to enforce the similarity between reconstruction samples from DyNeRF (Fridovich-Keil et al. 2023) and interpolated frames. However, the implicit nature of neural representation leads to long optimization time and poor motion controllability. With the great progress of explicit 4D Gaussian Splatting (Wu et al. 2024a), more researchers (Ren et al. 2023; Wu et al. 2024b; Li, Chen, and Liu 2024; Wu et al. 2025) formulate the deformable Gaussian Splatting as the intermediate 4D representation. However, these video-to-4D generation pipelines still encounter challenges like spatial-temporal inconsistency and suboptimal motion quality. STAG4D (Zeng et al. 2024) designs temporal anchor points for enhanced 4D consistency, but lacks temporal correlation, and their designed Gaussian densification technique fails to consider rapid texture differences in the same region across frames, as shown in Fig. 2, and overlooks the time-varying requirements in the number of Gaussian points for adaptive appearance modeling.

To address these challenges, we propose a novel 4D generation framework called **4DSTR**, which conducts the temporal correlation and spatial Gaussian rectification methods to guarantee spatial-temporal consistency and motion realism in generated 4D contents. Specifically, to ensure the temporal consistency of multi-frame generative 4D Gaussian points, the scales and rotations from all the video frames are correlated through the Mamba-based (Gu and Dao 2023; Liu et al. 2024b) temporal encoding layer. Then the per-

*These authors contributed equally.
†Corresponding author.



(A) ***Consistent 4D generation samples*** **(B)** ***Comparison on spatial-temporal consistency***

Figure 1: **Consistent 4D generation with spatial-temporal rectification.** Our method proposes a novel framework for high-quality 4D generation as in (A). Compared to the state-of-the-art method STAG4D (Zeng et al. 2024), our method has higher generation consistency and quality in the dynamic region (red circle) of generated 4D sequences (B), which demonstrates that our rectification methods significantly boost spatial-temporal consistency in generative 4D Gaussian representations.

frame scale residuals and rotation residuals are generated to rectify the original inaccurate Gaussian points, which can significantly boost the inherent motion coherence of generated 4D sequences. Additionally, a spatial rectification method with the adaptive densification and pruning strategy is proposed to add Gaussian points to regions with rich textures requiring more 4D representation tokens and delete Gaussian points in regions with less texture or unstable Gaussian attributes for each training iteration. The adaptive densification and pruning strategy can enable the generative 4D Gaussian points to adapt to rapid temporal variations and possess more photorealistic reconstruction quality. Integrating these two designs, our method can significantly boost spatial-temporal consistency and generation quality for the video-to-4D generation task in Fig. 1. When combining with a pre-trained text-to-image or text-to-video generation diffusion model, our method can also support the text-to-4D generation task.

Overall, the contributions of this paper are as follows:

- We propose 4DSTR, a novel 4D generation pipeline with spatial-temporal rectification to strengthen the spatial-temporal consistency of generated 4D videos and enhance the adaptation ability to rapid temporal variations.
- To ensure temporal consistency, a Mamba-based temporal encoding layer is designed to correlate video sequences and regress scale and rotation residuals of generative 4D Gaussian points in each frame.
- To adapt to rapid spatial variations across frames, we dynamically rectify the number of 4D Gaussian points using a per-frame adaptive Gaussian densification and pruning strategy with all-frame temporal awareness.
- Extensive experiments on the video-to-4D show the superiority of our proposed approach. Our 4DSTR outperforms the state-of-the-art 4D generation approach with a 15.1% FID-VID reduction and a 19.9% FVD reduction, which reveals our method's great potential in the spatial-temporal consistency of generated 4D sequences.

Related Work

Recently, there has been a significantly increasing research focus on 4D generation with various guidance inputs including image, video (Jiang et al. 2024c; Ren et al. 2023; Wu et al. 2024b; Zeng et al. 2024), and text (Singer et al. 2023; Ling et al. 2024). In this section, we mainly delve into descriptions for video-to-4D generation, text-to-4D generation, and also spatial-temporal modeling, respectively.

Video-to-4D Generation

Video-to-4D generation produces spatiotemporal content from uncalibrated monocular videos. Consistent4D (Jiang et al. 2024c) applies an interpolation-driven loss on DyNeRF (Fridovich-Keil et al. 2023) reconstructions but suffers from long optimization and limited motion control. DreamGaussian4D (Ren et al. 2023) and SC4D (Wu et al. 2024b) employ deformable 4D Gaussian splatting, while DreamMesh4D (Li, Chen, and Liu 2024) combines mesh representations with geometric skinning for improved surface detail; however, they all lack strong spatial–temporal coherence. STAG4D (Zeng et al. 2024) introduces temporal anchors and adaptive densification to enhance frame-to-frame correlation. CAT4D (Wu et al. 2025) designs a sampling strategy to generate an unbounded collection of consistent multi-view videos for 4D generation. In contrast, we demonstrate that spatial-temporal modulation across frames

is essential for enhanced consistency and high-quality motion.

Text-to-4D Generation

MAV3D (Singer et al. 2023) pioneers by using a text-to-image diffusion model for static objects and a text-to-video model with SDS loss for dynamic sequences. AYG (Ling et al. 2024) integrates compositional text-to-image, text-to-video, and 3D-aware multi-view diffusion to optimize 4D Gaussians, while TC4D (Bahmani et al. 2024) improves motion realism via trajectory-conditioned rigid transforms and local deformations. However, reliance on pre-trained diffusion models limits performance and leads to spatial–temporal inconsistency and domain gaps (Zhang et al. 2024). 4Real-Video (Wang et al. 2025b) proposes a two-stream grid-based 4D video generation method with both viewpoint and temporal updates.

4D Spatial–Temporal Modeling

4D spatial–temporal modeling extends static point cloud analysis to video by capturing temporal correlations (Liu et al. 2025c; Liu, Cheng, and Yang 2023; Nie et al. 2025; Liu et al. 2024d; Zhou et al. 2025b,c). PSTNet (Fan et al. 2022) employs disentangled spatial and temporal convolutions, while P4Transformer (Fan, Yang, and Kankanhalli 2021) uses Transformer for long-sequence embedding. Some researchers study scene flow methods (Liu et al. 2024a; Jiang et al. 2024b,a; Liu et al. 2023b) for 4D motion learning. Mamba4D (Liu et al. 2025a) uses the Mamba architecture (Gu and Dao 2023) to further enhance scalability. In this work, we transfer the successful experiences in 4D modeling into the spatial-temporal correlation of 4D Gaussian points for more consistent and high-quality 4D generation.

Method

In this section, we will dive into the details of our proposed 4D generation pipeline as illustrated in Fig. 3. First, we analyze the limitations of current 4D generative Gaussian representations. To mitigate these restrictions, we propose a temporal correlation module to establish temporal consistency and regress the scale and rotation residuals for each frame. Furthermore, an adaptive densification and pruning strategy is proposed to dynamically add or remove 4D Gaussian points at the frame level to strengthen the ability to adapt to rapid spatial variations over time.

Limitations of Prior 4D Representations

Dynamic NeRF-based methods (Fridovich-Keil et al. 2023) often suffer from poor motion consistency and flexibility due to their implicit nature and fixed bounding boxes (Bahmani et al. 2024). Recent works (Zeng et al. 2024; Ren et al. 2023; Li, Chen, and Liu 2024) therefore adopt deformable 4D Gaussian Splatting (Wu et al. 2024a), which extends 3D Gaussians (Kerbl et al. 2023) with a deformation field:

$$\mathcal{F}(\mathcal{S}, t) = [\mathcal{X}_t, s_t, r_t, \sigma, \zeta], \tag{1}$$

where $\mathcal{X}_t$, s_t, r_t update position, scale, and rotation, and σ, ζ correspond to the opacity and spherical harmonics (SH) coefficients of the radiance (Kerbl et al. 2023), respectively.

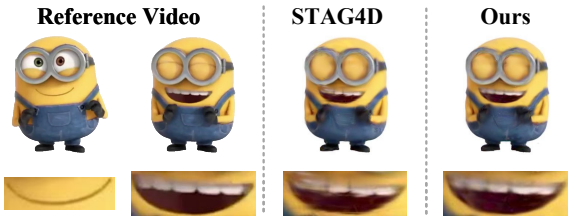


Figure 2: **Rapid temporal variations among frames.** The mouth of Minions witnesses rapid appearance variations for two different frames. Compared to STAG4D (Zeng et al. 2024), our method designs an adaptive Gaussian densification and pruning strategy, which largely enhances the adaptation capability of our 4D generative Gaussian.

However, these methods process each timestamp independently, lacking explicit temporal correlation or rectification (Zeng et al. 2024). They also keep a constant number of Gaussians across frames, which hinders adaptation to rapid texture changes. As shown in Fig. 2, suddenly changing regions with more texture details should be supplemented with more Gaussian points. We address these issues by introducing a spatial-temporal correlation and rectification module alongside adaptive densification and pruning for enhanced spatial–temporal consistency and motion fidelity.

Temporal Correlation and Rectification

To effectively guarantee spatial-temporal consistency of generated 4D videos, we design a temporal buffer to store and facilitate interactions among multi-frame Gaussian attributes through the Mamba architecture (Gu and Dao 2023) as in Fig. 3. Unlike previous 4D Gaussian Splatting methods that only query Gaussian features from the current frame (Wu et al. 2024a), our approach utilizes temporal Gaussian attributes $\mathcal{F}(\mathcal{S}, t) = [\mathcal{X}_t, s_t, r_t, \sigma, \zeta]_{t=1}^{T}$ from multiple timestamps as input for effective temporal correlation. Specifically, Gaussian attributes predicted from the current frame will correlate with the stored history Gaussian attributes in the temporal buffer to generate updated temporal features and regress the rectified Gaussian attribute residuals. The temporal buffer $M_0^G \in \mathbb{R}^{T \times d_G}$, which holds Gaussian attributes over time, is initialized empty and retains a fixed length T. Here, $d_G = 11$ corresponds to the Gaussian attribute vectors: position, scale, rotation, and opacity.

Temporal Correlation with Mamba. During the temporal correlation phase, the current Gaussian attribute feature $\mathcal{F}_t$ just generated from the deformable network will interact with the temporal buffer through a sliding window mechanism. When current Gaussian attributes are written into the temporal buffer M_{t-1}^{G}, the temporally farthest Gaussian attribute $\hat{\mathcal{F}}_{t-T-1}$ is discarded. The $T-1$ most recent Gaussian attributes $\{\hat{\mathcal{F}}_{t-T}, ..., \hat{\mathcal{F}}_{t-1}\}$ are then concatenated with the current Gaussian attribute F_t as follows:

$$\{\hat{\mathcal{F}}_{t-T-1}, \hat{\mathcal{F}}_{t-T}, ..., \hat{\mathcal{F}}_{t-1}\} \Rightarrow \{\hat{\mathcal{F}}_{t-T}, ..., \hat{\mathcal{F}}_{t-1}, \mathcal{F}_t\}, \tag{2}$$

where the braces denote the concatenation of Gaussian attributes along the temporal axis. Concatenated features

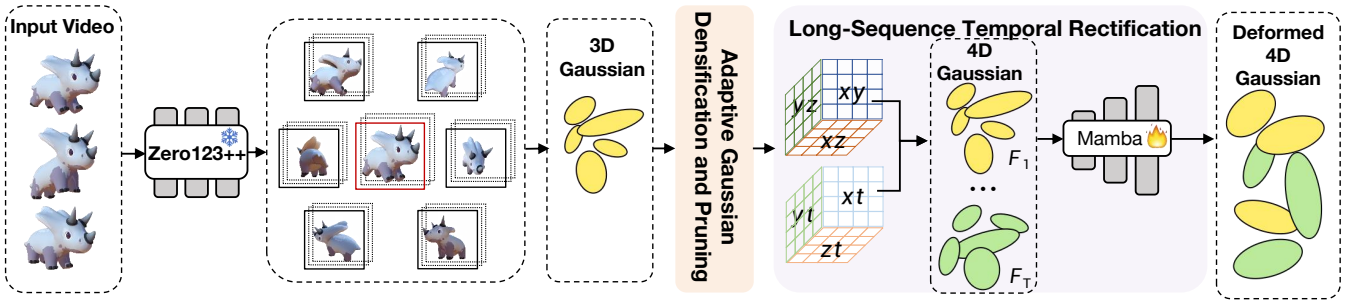


Figure 3: **The overall pipeline of our 4DSTR.** Given an input video, we use Zero123++ (Shi et al. 2023) to generate multi-view frames and initialize the first-frame 3D Gaussians. A lightweight multi-head decoder then maps voxel features to per-frame 4D Gaussian parameters. To ensure 4D coherence, our temporal correlation module regresses scale and rotation residuals, while per-frame adaptive densification and pruning dynamically adjust Gaussian counts to capture rapid spatial changes.

$M^{\mathrm{G}}_{t-1} = \{\hat{\mathcal{F}}_{t-T}, ..., \hat{\mathcal{F}}_{t-1}, \mathcal{F}_t\}$ are then passed as input tokens into the subsequent Mamba-based temporal correlation module:

$$\hat{\mathcal{F}}_t = \mathrm{Mamba}(M^{\mathrm{G}}_{t-1}), \tag{3}$$

where Mamba refers to the standard selective state-space model (Gu and Dao 2023). The temporally-correlated Gaussian attributes are then used to regress the scale and rotation residuals for the rectification.

Gaussian Attribute Rectification. After the temporal correlation, we rectify the scales and rotations of Gaussian points for each frame. Here, we only extend details of the scale rectification, and the same procedure is applied to rotation modulation. We rectify the Gaussian scales in the current frame through the feature fusion of temporally-correlated Gaussian attribute features $\hat{\mathcal{F}}_t$, current scales s_t, and history scales $\hat{s}_{t-1}$ as:

$$\Delta s_t = W(\hat{\mathcal{F}}_t \oplus s_t \oplus \hat{s}_{t-1}), \tag{4}$$

where W means the dynamic weighting method in (Aydemir, Akan, and Güney 2023). $\hat{s}_{t-1}$ indicates the nearest history scales, and Δs_t represents the regressed residuals of Gaussian scales. Then the rectified scales in the timestamp t can be represented by: $\hat{s}_t = s_t + \Delta s_t$. Finally, we update the temporal buffers by concatenating the current temporally-correlated features $\hat{\mathcal{F}}_t$ with the previous $T-1$ frames:

$$M^{\mathrm{G}}_{\mathrm{t}} = \{\hat{\mathcal{F}}_{\mathrm{t\text{-}T+1}}, ..., \hat{\mathcal{F}}_{\mathrm{t\text{-}1}}, \hat{\mathcal{F}}_{\mathrm{t}}\}. \tag{5}$$

$M^{\mathrm{G}}_{\mathrm{t}}$ will be used for the rectification process in the next timestamp $t+1$, regressing scale and rotation residuals Δs_{t+1} and Δr_{t+1}.

Adaptive Gaussian Densification and Pruning

3D Gaussian Splatting employs point densification to adjust Gaussian density for accurate 3D reconstruction, while 4D Gaussian Splatting (Wu et al. 2024a) uses a fixed view-space gradient threshold. STAG4D (Zeng et al. 2024) adds an adaptive threshold but averages gradients over all frames. As a result, all frames retain the same number of Gaussian points after these operations. We observe that the same number for all frames is suboptimal for dynamic scenes.

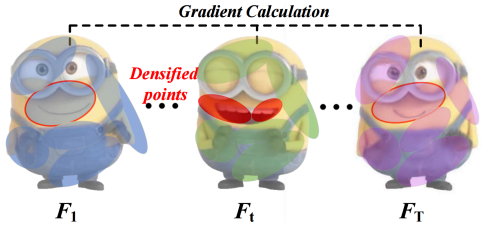


Figure 4: **Illustration of Per-frame Adaptive Gaussian Densification strategy.** We accumulate and average each Gaussian point's gradient over training steps. Then, at each timestep t, we independently apply densification or pruning based on its averaged gradient. For example, when a Minion's mouth opens at F_t, we densify that region; when it closes at F_T, we prune it.

For example, as in Fig. 4, capturing the details of Minions' mouth requires larger numbers of Gaussian points when the mouth is suddenly open. To address this, we propose per-frame adaptive densification and pruning, dynamically tuning Gaussian counts to better model rapid spatial changes.

Per-Frame Adaptive Gaussian Densification. As shown in Fig. 4, our method analyzes the accumulated gradient $\mathcal{G}(p)$ for each Gaussian point p over time, following a *log-normal* distribution throughout training. To ensure adaptive selection, we define the per-frame densification threshold τ_t as:

$$\tau_t = \mathrm{Quantile}_{(1-\lambda)}\left(\{\mathcal{G}(p) \mid p \in \hat{\mathcal{F}}_t\}\right), \tag{6}$$

where $\mathrm{Quantile}_{(1-\lambda)}(\cdot)$ represents the $(1-\lambda)$-quantile of the accumulated gradients over all Gaussian points and $\hat{\mathcal{F}}_t$ represents the set of all Gaussian points in the frame t. A Gaussian point p is densified only if $\mathcal{G}(p) \geq \tau_t$.

Per-Frame Gaussian Pruning Strategy. To maintain a balanced distribution of Gaussian points, we prune points based on opacity, screen-space size, and world-space scaling constraints (Wu et al. 2024a; Zeng et al. 2024). Specifically, a Gaussian point p is removed if its opacity σ falls below a predefined threshold τ_o. Furthermore, points are pruned if their world-space scaling exceeds a maximum threshold $s_{\max}$ or

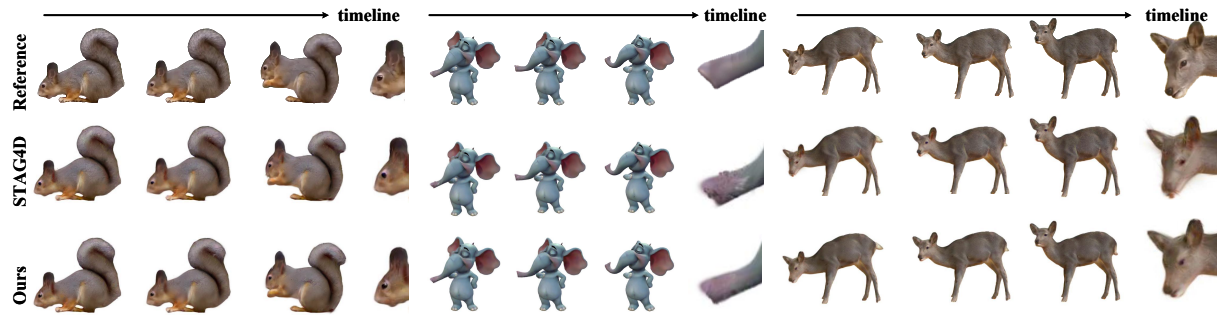


Figure 5: **Qualitative comparisons on video-to-4D generation.** Compared with the recent SOTA method STAG4D (Zeng et al. 2024), our method delivers higher-quality results in dynamic regions such as squirrel and deer heads or an elephant's trunk.

Methods	Reference	FID-VID ↓	FVD ↓	CLIP ↑	LPIPS ↓
DG4D (Ren et al. 2023)	arXiv'23	73	856	0.88	0.14
Consistent4D (Jiang et al. 2024c)	ICLR'24	/	1134	0.88	0.13
4DGen (Yin et al. 2023)	arXiv'24	72	/	0.89	0.13
SC4D (Wu et al. 2024b)	ECCV'24	/	880	0.90	0.14
STAG4D (Zeng et al. 2024)	ECCV'24	53	992	0.91	0.13
4Diffusion (Zhang et al. 2024)	NeurIPS'24	/	/	0.88	0.17
MVTokenFlow (Huang et al. 2025a)	ICLR'25	/	846	0.91	**0.12**
4DSTR (Ours)	—	**45**	**795**	**0.92**	**0.12**

Table 1: Quantitative comparisons with SOTA methods on the video-to-4D generation task. The best results are in **bold**.

falls below a minimum threshold $s_{\min}$, ensuring that only points contributing meaningfully to the reconstruction are retained as:

$$\hat{\mathcal{F}}_t' = \{p \in \hat{\mathcal{F}}_t \mid (\alpha(p) \geq \tau_o) \wedge (s_{\min} \leq \hat{s}_t \leq s_{\max})\}, \quad (7)$$

where $\hat{\mathcal{F}}_t'$ indicates the Gaussian attributes that are spatially rectified after the per-frame adaptive Gaussian pruning. By dynamically adjusting the per-frame densification threshold and selecting the top λ of Gaussian points with the highest gradients, our method continuously refines the point distribution in response to scene complexity. As shown in Fig. 2, our per-frame adaptive densification and pruning strategy effectively enhances the quality and robustness of 4D Gaussian representations, significantly outperforming the same number for all frames approaches in rapidly changing dynamic scenes (see Table 2).

Gaussian Correspondence Alignment. Per-Frame Gaussian Densification and Pruning may disrupt inter-frame correspondence of Gaussian points, but the Temporal Rectification relies on the temporal correspondence of Gaussian points across frames. To tackle the problem, we design a per-frame index to explicitly indicate and associate each densified or pruned Gaussian point with its corresponding frame. This process ensures temporal alignment between Gaussian points in the memory buffer after the Per-Frame Gaussian Densification and Pruning.

Loss Function

Given a monocular reference video, we use Zero123++ (Shi et al. 2023) to render six anchor views $\{I_t^i\}_{i=1}^6$ plus a reference I_t^{ref}, enhanced with temporally consistent diffusion from STAG4D (Zeng et al. 2024). We then optimize 4D Gaussians via multi-view SDS:

$$\mathcal{L}_{\text{MVSDS}} = \lambda_1 \mathcal{L}_{\text{SDS}}(\phi, I_t^i) + \lambda_2 \mathcal{L}_{\text{SDS}}(\phi, I_t^{\text{ref}}), \quad (8)$$

where the index i is chosen based on the closest viewpoint match between the rendered images and the reference, and λ_1, λ_2 are weighting parameters.

Following the setup in (Zeng et al. 2024), we incorporate the reference image to calculate a reconstruction term $\mathcal{L}_{\text{rec}}$ and a foreground mask term $\mathcal{L}_{\text{mask}}$. Hence, our total objective is:

$$\mathcal{L} = \mathcal{L}_{\text{MVSDS}} + \lambda_3 \mathcal{L}_{\text{rec}} + \lambda_4 \mathcal{L}_{\text{mask}}, \quad (9)$$

where λ_3 and λ_4 are additional coefficients. During training, we first apply $\mathcal{L}$ to a static frame to obtain a canonical 3D Gaussian, then use anchor and reference views to learn dynamic 4D Gaussians.

Since 4DSTR learns temporal variations directly, per-frame losses alone are insufficient. Inspired by MOTR (Zeng et al. 2022), we define a collective average loss (CAL), which aggregates losses over a sub-clip of T_s frames as $\mathcal{L}_{\text{CAL}} = \frac{1}{T_s} \sum_{t=1}^{T_s} \mathcal{L}_t$. where $\mathcal{L}_t$ denotes the total loss for frame t, computed according to Eq. (9).

Experiment

Datasets and Metrics

Datasets. In terms of the video-to-4D task, we follow Consistent4D (Jiang et al. 2024c), leveraging multi-view videos

Rectification		Evaluation Metrics			
Temporal	Spatial	FID-VID ↓	FVD ↓	CLIP ↑	LPIPS ↓
		74.24	1049.32	0.902	0.135
	✓	55.32	970.65	0.910	0.128
✓		52.21	850.32	0.912	0.126
✓	✓	**45.31**	**795.21**	**0.918**	**0.121**

Table 2: Ablation study on effectiveness of temporal and spatial rectification methods for video-to-4D generation.

Model	FID-VID ↓	FVD ↓	CLIP ↑	LPIPS ↓
STAG4D	76.00	1035.00	0.903	0.146
4DSTR (Ours)	**43.72**	**733.24**	**0.921**	**0.125**

Table 3: Performance on extended video sequences

with 7 dynamic objects to conduct the quantitative comparisons. Moreover, we also supplement the data from the online video resources created in STAG4D (Zeng et al. 2024).

Evaluation Metrics. We adopt four metrics (Jiang et al. 2024c; Zeng et al. 2024) to evaluate the performance of our models: CLIP and LPIPS for per-frame semantic and reconstruction quality, and FID-VID and FVD for video-level temporal coherence and multi-frame consistency.

Implementation Details

The deformation fields are parameterized by MLPs with 64 hidden layers of 32 units, and the temporal model uses 32 units. During training, the learning rate decays from 1.6×10^{-4} to 1.6×10^{-6}. For per-frame adaptive densification and pruning, we densify the top $\lambda = 2.5\%$ points by accumulated gradient and prune points with opacity below $\tau_o = 0.01$ or scale outside $[s_{\min} = 0.001, s_{\max} = 0.1]$ (Zeng et al. 2024). We use temporal buffers of length $T = 10$ and $T_s = 4$, SDS weights $\lambda_1, \lambda_2 = 1$, and allocate $\lambda_3 = 2 \times 10^4$, $\lambda_4 = 5 \times 10^3$ to reconstruction and mask losses. All experiments run on a single RTX 4090.

Qualitative Results

We compare to the state-of-the-art method STAG4D (Zeng et al. 2024), which leverages spatial–temporal anchors. As shown in Fig. 1, our rectification yields better spatial-temporal consistency in dynamic regions. Fig. 5 further shows our reconstructed videos offer improved rendering quality, particularly in high-dynamic regions.

Quantitative Results and Comparison

We quantitatively compare 4DSTR with DG4D (Ren et al. 2023), Consistent4D (Jiang et al. 2024c), 4DGen (Yin et al. 2023), SC4D (Wu et al. 2024b), and SOTA methods STAG4D, 4Diffusion (Zhang et al. 2024), MVTokenFlow (Huang et al. 2025a) (Table 1). 4DSTR leads on all four metrics. It surpasses MVTokenFlow and STAG4D on image reconstruction. Due to our designed spatial-temporal rectification methods, 4DSTR has much more potential advantages in video-based metrics, with a 15.1% FID-VID reduction and a 19.9% FVD reduction compared to STAG4D.

Methods	Evaluation Metrics				Latency (fps) ↑
	FID-VID ↓	FVD ↓	CLIP ↑	LPIPS ↓	
GRU	50.32	821.13	0.913	0.127	68
Attention	54.23	812.36	0.914	0.125	72
Mamba	**45.31**	**795.21**	**0.918**	**0.121**	**80**

Table 4: Ablation study on temporal correlation methods.

Temporal lengths	Evaluation Metrics			
	FID-VID ↓	FVD ↓	CLIP ↑	LPIPS ↓
T=2	57.32	855.13	0.908	0.132
T=5	53.21	823.32	0.912	0.127
T=10	45.31	**795.21**	**0.918**	0.121
T=15	**43.54**	804.32	0.917	**0.120**

Table 5: Ablation study on different temporal length.

This demonstrates the effectiveness of our temporal correlation in enhancing spatial-temporal consistency.

Ablation Studies

Significance of Temporal and Spatial Rectification. We ablate the temporal rectification and spatial densification modules in Table 2. Removing temporal correlation increases FID-VID and FVD by 22.1%, highlighting its role in maintaining temporal consistency (Fig. 7). Spatial rectification with adaptive Gaussian densification and pruning is likewise crucial for high-quality rendering of fast-moving objects (Fig. 8).

Longer Sequence Generation Results. To validate robustness over longer horizons, we construct a 60-frame test set following Consistent4D in Table 3. While STAG4D's performance degrades sharply on 60-frame inputs, 4DSTR further reduces FID-VID and FVD, confirming that our spatial-temporal rectification mechanism scales effectively to extended sequences, preserving spatial-temporal consistency.

Various Temporal Encoding Methods. We adopt a Mamba-based interaction for temporal encoding of Gaussian scales and rotations, leveraging its linear-complexity modeling of long-range dependencies. Table 4 shows that Mamba outperforms GRU (Cho et al. 2014) and attention (Vaswani et al. 2017) backbones, achieving up to a 5.01 reduction in FID-VID and running at 80 FPS, which led us to choose it for our temporal correlation module.

Varying Temporal Window Sizes. Although rapid variations are local, adjacent-frame models often fail under occlusions, appearance shifts, and large displacements. Extending the temporal window provides additional context to resolve abrupt motions and preserve spatial–temporal coherence. As shown in Table 5, increasing the window size T from 2 to 5 and then 10 consistently improves FID-VID and FVD, while further enlarging T brings negligible gains, indicating that a 10-frame window suffices to capture rapid variations. .

Results of Text-to-4D Generation

Following prior works (Zeng et al. 2024), our pipeline also supports text-to-4D generation by using SDXL (Podell

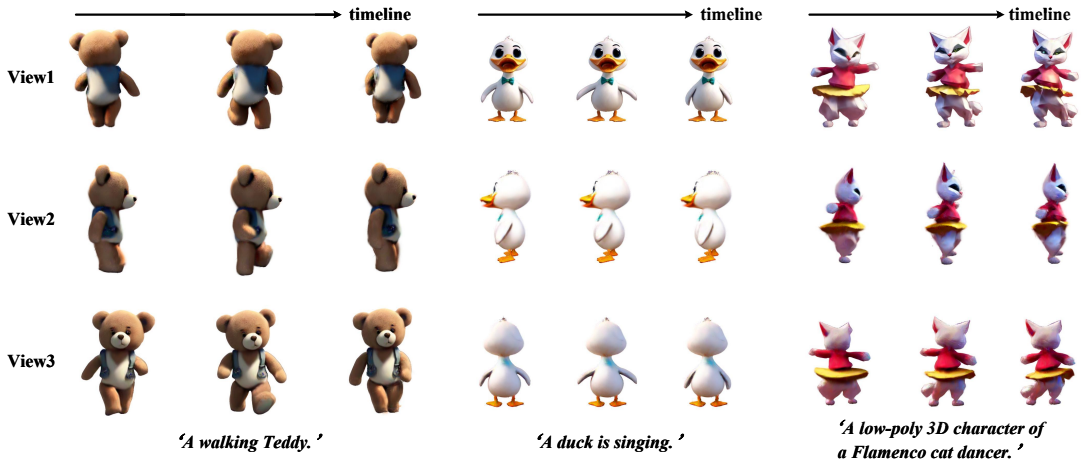


Figure 6: **Qualitative comparisons on text-to-4D generation.** Our method can also support text information as the guidance, generating consistent and high-quality 4D sequences which can be observed from diverse views.

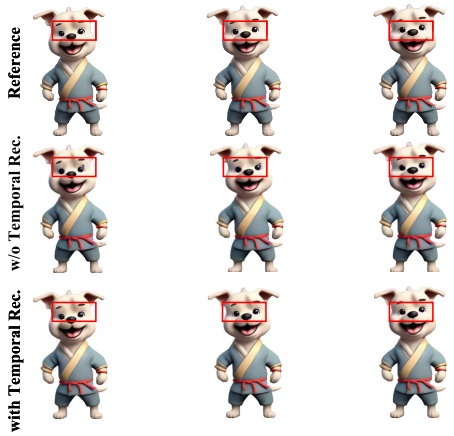


Figure 7: **Ablation on temporal rectification.** Without our temporal correlation module, the temporal consistency of generated 4D sequences is largely undermined.

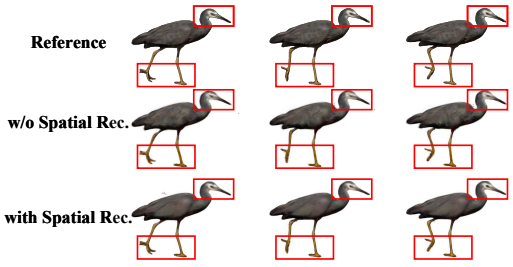


Figure 8: **Ablation on spatial rectification.** Without our designed spatial Gaussian densification and pruning module, the reconstruction quality on high-dynamic regions is poor. We highlight differences with red rectangles.

et al. 2023) for image synthesis and SVD (Blattmann et al. 2023) to create short videos, then applying the video-to-4D pipeline above. We use the same data settings as STAG4D for fair comparison and conduct a user study (Huang et al. 2025a) evaluating the performance of text-to-4D generation task. Fig. 6 shows diverse 4D samples with plausible dynamics and spatio-temporal consistency across modalities. Following (Zeng et al. 2024), the user study covers 14 test scenarios with 30 evaluators rating visual quality (Vis.), temporal consistency (Cons.), and alignment with the input text (Align.); in Table 6, our method achieves the highest scores on all metrics, demonstrating the superiority of 4DSTR on the text-to-4D generation task.

Methods	Vis.	Cons.	Align.
Consistent4D (Jiang et al. 2024c)	13.3%	20.0%	16.7%
STAG4D (Zeng et al. 2024)	33.3%	30.0%	36.7%
Ours	**53.3%**	**50.0%**	**46.7%**

Table 6: User study for the text-to-4D generation task.

Conclusion

In this paper, we introduce 4DSTR, a 4D generation network with spatial–temporal rectification. Our approach correlates deformable 4D Gaussian points across multiple frames, ensuring a consistent Gaussian representation in each frame. To handle rapid temporal changes, we introduce an adaptive spatial densification and pruning strategy, dynamically adding or removing Gaussian points based on long-range dependencies. Extensive experiments show 4DSTR sets a new SOTA in video-to-4D generation on reconstruction quality, temporal coherence, and dynamic adaptability. Furthermore, our designed pipeline can also support high-quality text-to-4D generation task when combined with existing text-to-image generators.

Acknowledgments

This work was supported by the EU HORIZON-CL4-2023-HUMAN-01-CNECT XTREME (grant no.101136006), and the Sectorplan Beta-II of the Netherlands.

References

Aydemir, G.; Akan, A. K.; and Güney, F. 2023. Adapt: Efficient multi-agent trajectory prediction with adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8295–8305.

Bahmani, S.; Liu, X.; Yifan, W.; Skorokhodov, I.; Rong, V.; Liu, Z.; Liu, X.; Park, J. J.; Tulyakov, S.; Wetzstein, G.; et al. 2024. Tc4d: Trajectory-conditioned text-to-4d generation. In *European Conference on Computer Vision*, 53–72. Springer.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Cheng, H.; Liu, M.; and Chen, L. 2023. An end-to-end framework of road user detection, tracking, and prediction from monocular images. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 2178–2185. IEEE.

Cheng, H.; Liu, M.; Chen, L.; Broszio, H.; Sester, M.; and Yang, M. Y. 2023. Gatraj: A graph-and attention-based multi-agent trajectory prediction model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 205: 163–175.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Fan, H.; Yang, Y.; and Kankanhalli, M. 2021. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14204–14213.

Fan, H.; Yu, X.; Ding, Y.; Yang, Y.; and Kankanhalli, M. 2022. Pstnet: Point spatio-temporal convolution on point cloud sequences. *arXiv preprint arXiv:2205.13713*.

Fridovich-Keil, S.; Meanti, G.; Warburg, F. R.; Recht, B.; and Kanazawa, A. 2023. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12479–12488.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, H.; Liu, Y.; Zheng, G.; Wang, J.; Dou, Z.; and Yang, S. 2025a. MVTokenFlow: High-quality 4D Content Generation using Multiview Token Flow. In *The Thirteenth International Conference on Learning Representations*.

Huang, S.; Kang, Y.; Shen, G.; and Song, Y. 2025b. AI-Augmented Context-Aware Generative Pipelines for 3D Content. *Preprints*.

Huang, S.; Shen, G.; Kang, Y.; and Song, Y. 2025c. Immersive Augmented Reality Music Interaction through Spatial Scene Understanding and Hand Gesture Recognition. *Preprints*.

Jiang, C.; Du, D.; Liu, J.; Zhu, S.; Liu, Z.; Ma, Z.; Liang, Z.; and Zhou, J. 2024a. NeuroGauss4D-PCI: 4d neural fields and gaussian deformation fields for point cloud interpolation. *arXiv preprint arXiv:2405.14241*.

Jiang, C.; Wang, G.; Liu, J.; Wang, H.; Ma, Z.; Liu, Z.; Liang, Z.; Shan, Y.; and Du, D. 2024b. 3dsflabelling: Boosting 3d scene flow estimation by pseudo auto-labelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15173–15183.

Jiang, Y.; Zhang, L.; Gao, J.; Hu, W.; and Yao, Y. 2024c. Consistent4D: Consistent 360° Dynamic Object Generation from Monocular Video. In *The Twelfth International Conference on Learning Representations*.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4): 1–14.

Li, Z.; Chen, Y.; and Liu, P. 2024. Dreammesh4d: Video-to-4d generation with sparse-controlled gaussian-mesh hybrid representation. *Advances in Neural Information Processing Systems*, 37: 21377–21400.

Ling, H.; Kim, S. W.; Torralba, A.; Fidler, S.; and Kreis, K. 2024. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8576–8588.

Liu, J.; Han, J.; Liu, L.; Aviles-Rivero, A. I.; Jiang, C.; Liu, Z.; and Wang, H. 2025a. Mamba4D: Efficient 4D Point Cloud Video Understanding with Disentangled Spatial-Temporal State Space Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17626–17636.

Liu, J.; Huang, Z.; Liu, M.; Deng, T.; Nex, F.; Cheng, H.; and Wang, H. 2025b. TopoLiDM: Topology-Aware LiDAR Diffusion Models for Interpretable and Realistic LiDAR Point Cloud Generation. *arXiv preprint arXiv:2507.22454*.

Liu, J.; Wang, G.; Jiang, C.; Liu, Z.; and Wang, H. 2023a. Translo: A window-based masked point transformer framework for large-scale lidar odometry. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1683–1691.

Liu, J.; Wang, G.; Liu, Z.; Jiang, C.; Pollefeys, M.; and Wang, H. 2023b. Regformer: An efficient projection-aware transformer network for large-scale point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8451–8460.

Liu, J.; Wang, G.; Ye, W.; Jiang, C.; Han, J.; Liu, Z.; Zhang, G.; Du, D.; and Wang, H. 2024a. Difflow3d: Toward robust uncertainty-aware scene flow estimation with iterative diffusion-based refinement. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15109–15119.

Liu, J.; Yu, R.; Wang, Y.; Zheng, Y.; Deng, T.; Ye, W.; and Wang, H. 2024b. Point mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy. *arXiv preprint arXiv:2403.06467*.

Liu, J.; Zhuo, D.; Feng, Z.; Zhu, S.; Peng, C.; Liu, Z.; and Wang, H. 2024c. Dvlo: Deep visual-lidar odometry with local-to-global feature fusion and bi-directional structure alignment. In *European Conference on Computer Vision*, 475–493. Springer.

Liu, M.; Cheng, H.; Chen, L.; Broszio, H.; Li, J.; Zhao, R.; Sester, M.; and Yang, M. Y. 2024d. Laformer: Trajectory prediction for autonomous driving with lane-aware scene constraints. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2039–2049.

Liu, M.; Cheng, H.; and Yang, M. Y. 2023. Tracing the influence of predecessors on trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3253–3263.

Liu, M.; Yang, M. Y.; Liu, J.; Zhang, Y.; Li, J.; Oude Elberink, S.; Vosselman, G.; and Cheng, H. 2025c. DVLO4D: Deep Visual-Lidar Odometry with Sparse Spatial-temporal Fusion. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 9740–9747. IEEE.

Ni, C.; Zhao, G.; Wang, X.; Zhu, Z.; Qin, W.; Huang, G.; Liu, C.; Chen, Y.; Wang, Y.; Zhang, X.; et al. 2025. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1559–1569.

Nie, J.; Xie, F.; Zhou, S.; Zhou, X.; Chae, D.-K.; and He, Z. 2025. P2P: Part-to-Part Motion Cues Guide a Strong Tracking Framework for LiDAR Point Clouds. *International Journal of Computer Vision*, 1–17.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv preprint arXiv:2307.01952*.

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*.

Shi, R.; Chen, H.; Zhang, Z.; Liu, M.; Xu, C.; Wei, X.; Chen, L.; Zeng, C.; and Su, H. 2023. Zero123++: a Single Image to Consistent Multi-view Diffusion Base Model. *arXiv preprint arXiv:2310.15110*.

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, B.; Wang, X.; Ni, C.; Zhao, G.; Yang, Z.; Zhu, Z.; Zhang, M.; Zhou, Y.; Chen, X.; Huang, G.; et al. 2025a. HumanDreamer: Generating Controllable Human-Motion Videos via Decoupled Generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12391–12401.

Wang, C.; Zhuang, P.; Ngo, T. D.; Menapace, W.; Siarohin, A.; Vasilkovsky, M.; Skorokhodov, I.; Tulyakov, S.; Wonka, P.; and Lee, H.-Y. 2025b. 4Real-Video: Learning generalizable photo-realistic 4D video diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17723–17732.

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 IEEE/CVF conference on computer vision and pattern recognition*, 20310–20320.

Wu, R.; Gao, R.; Poole, B.; Trevithick, A.; Zheng, C.; Barron, J. T.; and Holynski, A. 2025. Cat4d: Create anything in 4d with multi-view video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26057–26068.

Wu, Z.; Yu, C.; Jiang, Y.; Cao, C.; Wang, F.; and Bai, X. 2024b. Sc4d: Sparse-controlled video-to-4d generation and motion transfer. In *European Conference on Computer Vision*, 361–379. Springer.

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*.

Zeng, F.; Dong, B.; Zhang, Y.; Wang, T.; Zhang, X.; and Wei, Y. 2022. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, 659–675. Springer.

Zeng, Y.; Jiang, Y.; Zhu, S.; Lu, Y.; Lin, Y.; Zhu, H.; Hu, W.; Cao, X.; and Yao, Y. 2024. Stag4d: Spatial-temporal anchored generative 4d gaussians. In *European Conference on Computer Vision*, 163–179. Springer.

Zhang, H.; Chen, X.; Wang, Y.; Liu, X.; Wang, Y.; and Qiao, Y. 2024. 4diffusion: Multi-view video diffusion model for 4d generation. *Advances in Neural Information Processing Systems*, 37: 15272–15295.

Zhou, P.; Peng, X.; Song, J.; Li, C.; Xu, Z.; Yang, Y.; Guo, Z.; Zhang, H.; Lin, Y.; He, Y.; et al. 2025a. OpenING: A Comprehensive Benchmark for Judging Open-ended Interleaved Image-Text Generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 56–66.

Zhou, S.; Nie, J.; Zhao, Z.; Cao, Y.; and Lu, X. 2025b. Focustrack: One-stage focus-and-suppress framework for 3d point cloud object tracking. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 7366–7375.

Zhou, S.; Yuan, Z.; Yang, D.; Hu, X.; Qian, J.; and Zhao, Z. 2025c. Pillarhist: A quantization-aware pillar feature encoder based on height-aware histogram. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 27336–27345.