STAG4D: Spatial-Temporal Anchored Generative 4D Gaussians

Yifei Zeng[1], Yanqin Jiang[2], Siyu Zhu[3], Yuanxun Lu[1], Youtian Lin[1], Hao Zhu[1], Weiming Hu[2], Xun Cao[1], and Yao Yao[1]

[1] Nanjing University
[2] CASIA
[3] Fudan University

Abstract. Recent progress in pre-trained diffusion models and 3D generation have spurred interest in 4D content creation. However, achieving high-fidelity 4D generation with spatial-temporal consistency remains a challenge. In this work, we propose STAG4D, a novel framework that combines pre-trained diffusion models with dynamic 3D Gaussian splatting for high-fidelity 4D generation. Drawing inspiration from 3D generation techniques, we utilize a multi-view diffusion model to initialize multi-view images anchoring on the input video frames, where the video can be either real-world captured or generated by a video diffusion model. To ensure the temporal consistency of the multi-view sequence initialization, we introduce a simple yet effective fusion strategy to leverage the first frame as a temporal anchor in the self-attention computation. With the almost consistent multi-view sequences , we then apply the score distillation sampling to optimize the 4D Gaussian point cloud. The 4D Gaussian spatting is specially crafted for the generation task, where an adaptive densification strategy is proposed to mitigate the unstable Gaussian gradient for robust optimization. Notably, the proposed pipeline does not require any pre-training or fine-tuning of diffusion networks, offering a more accessible and practical solution for the 4D generation task. Extensive experiments demonstrate that our method outperforms prior 4D generation works in rendering quality, spatial-temporal consistency, and generation robustness, setting a new state-of-the-art for 4D generation from diverse inputs, including text, image, and video.

Keywords: 4D Generation · 3D Gaussian Splatting · Diffusion Model

1 Introduction

Recent advancements in large-scale pre-trained diffusion models have shown remarkable progress in producing high-quality and diverse visual content, including images, videos, and 3D asserts [5, 13, 16, 22, 24, 27, 31, 38]. The progress naturally extends to the realm of dynamic 3D generation [1, 6, 12, 28, 32, 36, 39, 40]. This endeavor has gained prominence in computer vision and generative AI research, as high-quality 4D content generation is key to a broad range of applications such as autonomous driving simulation, game and film industries, digital Avatar

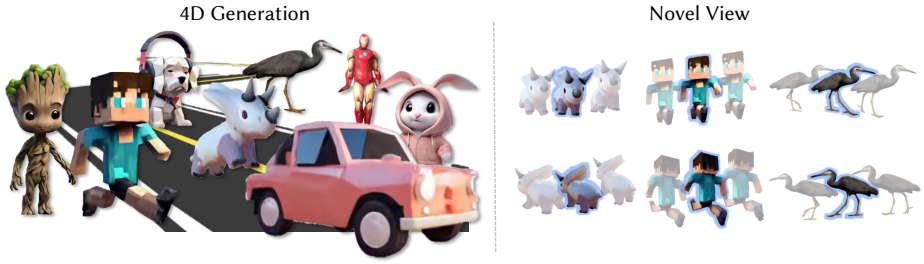


Fig. 1: Visualization of the generated 4D asserts. Our approach can generate diverse 4D content from various inputs, including text, image, and video.

creation, and spatial video production. However, previous 4D generation methods face challenges including blurry rendering, spatial-temporal inconsistency, and slow generation speed. Generating high-quality 4D content efficiently and practically remains a significant challenge.

Given the promising strides in 3D generation from pre-trained diffusion models, there has been a concerted focus on generalized dynamic 3D generation from text or uncalibrated monocular video [1, 6, 12, 28, 32, 40]. MAV3D [28] is the pioneering work of text-to-4D generation which creates 4D scenes from textual descriptions through Score Distillation Sampling (SDS) from a video diffusion model. Alternatively, Consistent4D [6] takes a video as input, and focuses on the task of video-to-4D generation. The pipeline is more flexible in that it can take advantage of high-quality input videos, dividing the difficult problem of realistic 4D content creation into a practical pipeline of text-to-video and video-to-4D generation. However, current video-to-4D generation methods still struggle to generate high-fidelity 4D asserts with spatial and temporal rendering consistency. Meanwhile, the commonly used Neural Radiance Field(NeRF) representation would suffer from an over-saturated appearance and a long optimization time.

In this paper, we address two important factors in 4D content creation: a suitable 4D representation and spatial-temporal consistency imposed during pseudo-projection generation (i.e., multi-view video generation). Our approach employs a 4D Gaussian splatting approach specially tailored for the generation task. In the generation process, we first generate the multi-view images at all timestamps of the 4D scene. To enforce the 4D consistency of the multi-view video initialization, we design a direct fusion approach for attention computation in both spatial and temporal domains. This approach implicitly addresses texture degradation and geometric misalignment, thereby avoiding the need for adding an explicit multi-view or temporal consistency loss during the Score Distillation Sampling(SDS) [22] optimization. Then, the reference photometric loss and the SDS loss conditioned on the initial multi-view video are applied for 4D Gaussian point cloud optimization. To tackle the instability in optimizing the 4D Gaussian points, we introduce an adaptive densification strategy that is informed by the

Gaussian gradient distribution. This approach yields high-quality and robust 4D scene generation from a monocular video input.

We have conducted comprehensive experiments to demonstrate the effectiveness of each component of the proposed method. Our approach achieves a 2x faster generation speed compared with the previous video-to-4D approaches (e.g., [6]) and a significantly better generation quality than prior state-of-the-art methods. It is also noteworthy that the 4D content generated by our method can be rendered in real-time, opening up a wide range of possibilities for practical applications. In summary, our primary contributions are as follows:

- We introduce a holistic 4D generation pipeline that streamlines the generation process into sequential stages: video generation, multi-view video initialization, and 4D optimization using multi-view video conditioned SDS.
- We harness the dynamic 3D Gaussian representation for 4D generation, complemented by an adaptive densification strategy. This combination enables highly precise and efficient 4D generation from monocular video inputs.
- We present a novel training-free attention fusion module to effectively integrate temporal anchor frames into the multi-view diffusion process, significantly enhancing the 4D consistency of the generated multi-view videos.
- Our method outperforms previous approaches in optimization efficiency, rendering quality, and 4D consistency, establishing a new benchmark for 4D generation across various input types, such as text, images, and videos.

2 Related work

2.1 3D Generation

Dreamfusion [22] first proposes the SDS loss [22, 30] for NeRF optimization, which stands for the most prevalent technique for nowadays text-to-3D approaches. To mitigate the multi-view Janus problem, Zero123 [13] and Syncdreammer [14] fine-tune 2D diffusion models to grant the image generator with the ability of viewpoint control, while later works [15, 17, 26, 27] explicitly generate fixed multi-view images in one diffusion pass. Building upon the success of multi-view diffusion methods, we integrate a multi-view generation module into the 4D generation task, with meticulous handling of 4D spatial-temporal consistency for enhanced 4D generation quality.

Except for equipping diffusion models with multi-view or depth awareness, another direction of progress focuses on 3D representations. Following the idea of differentiable optimization, Magic3D [10] applies a two-stage generation pipeline and changes the 3D representation to instant-NGP [20] and DMTet [25], achieving faster runtime and better generation quality. Direct2.5 [17] applies an explicit mesh representation and uses differentiable rasterization for fast mesh optimization. Recently, with the newly developed 3DGS, DreamGaussian [29] demonstrates the ability to generate 3D objects within several minutes by substituting NeRF with the 3D Gaussian representation, showing the potential of using an

explicit point-based representation in 3D generation tasks. In this work, we introduce a novel 4D Gaussian representation and a specially tailored optimization scheme for the 4D generation task.

2.2 4D Reconstruction

Dynamic 3D reconstruction is yet another heated research topic in the field of computer vision and graphics. By extending the static NeRF [19] framework to dynamic scenes, dynamic NeRFs [3, 4, 9, 34] have demonstrated remarkable progress on dynamic 3D reconstruction. However, limited by the implicit neural representation and the complex nature of dynamic 3D information, dynamic NeRFs still suffer from slow optimization speed and low reconstruction quality. With the development of 3DGS [7], researchers have quickly extended the 3D Gaussian to dynamic scene representation, which achieves faster training and rendering speeds compared to D-NeRFs based on implicit neural representations. Dynamic 3D Gaussian [18] applies the per-frame 3DGS optimization for 4D scene reconstruction. Some studies [33, 35] attempt to amalgamate the explicit point-based 3DGS with an implicit neural field for dynamic information modeling, while Gaussian-Flow [11] introduces an explicit per-point motion model to represent a 4D scene without using implicit neural networks. It is noteworthy that it is non-trivial to extend the Gaussian reconstruction to 4D reconstruction as each reconstruction pipeline requires heuristic tuning of hyperparameters at each stage. In this paper, we target combining the 4DGS for the challenging task of 4D generation.

2.3 4D Generation

Recent developments in 3D content generation and video diffusion technologies have triggered researchers' interest in exploring 4D content generation from various input conditionings. MAV3D [28] presents the early attempt at text-to-4D generation. By utilizing score distillation sampling derived from video diffusion models, MAV3D optimizes a dynamic NeRF based on textual prompts. In a parallel vein, Consistent4D [6] has introduced the task of video-to-4D. This method capitalizes on the pre-trained knowledge from image diffusion models to optimize dynamic NeRFs via SDS optimization, showcasing the potential of high-quality 4D content generation from pre-trained 2D diffusion models. Concurrently with our work, 4DGen [36] has introduced a framework for generating dynamic 3D models, which employs spatial-temporal pseudo labels on keyframes within a multi-view diffusion model. However, the overall quality of its generation can be further improved, as its rendering fidelity is limited by the implicit neural representation and inadequate spatial-temporal information exchange exists in the diffusion generation process. In contrast, our method applies an efficient 4D Gaussian representation with spatial-temporal attention from reference- and the first-frame anchors, achieving substantially better generation results than previous works.

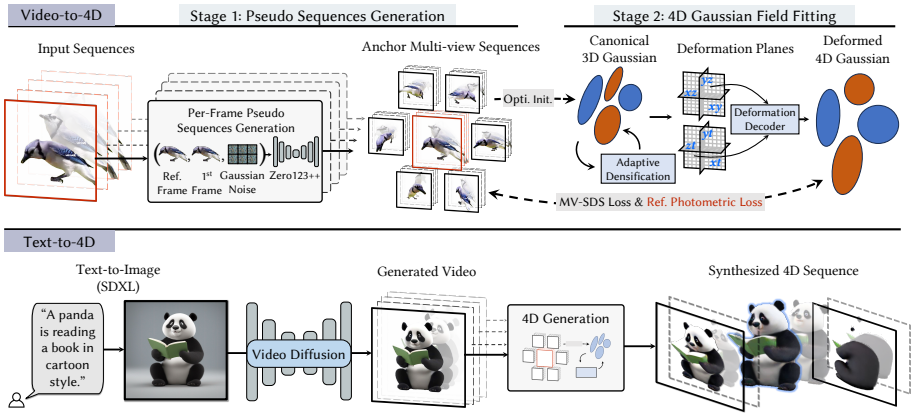


Fig. 2: Overall pipeline. Given a video input, we apply a multi-view diffusion model to produce coherent multi-view sequences, which serve as spatial and temporal anchors. Next, we train a deformable 3D Gaussian using multi-view SDS loss and reference loss. For a text-to-4D generation, our pipeline can be naturally extended to accept the text input by integrating with an off-the-shelf text-to-video module.

3 Method

The framework of the proposed approach, as depicted in Figure 2, outlines the generation and optimization process of a dynamic scene from a given video input. Also, our approach can be easily extended to a text-to-4D generation pipeline by utilizing an off-the-shelf video generation module. Section 3.1 presents the 4D representation used in the proposed method, along with an adaptive densification strategy that adjusts the densification threshold based on the relative motion gradient of the scene points. Subsequently, Section 3.2 provides a brief overview of the multi-view diffusion model used in the pipeline and introduces a novel spatial-temporal attention fusion mechanism designed to produce the almost consistent multi-view image sequences. Finally, as detailed in Section 3.3, the generated multi-view sequences are leveraged with spatial and temporal anchor frames for consistent 4D generation with the SDS optimization.

3.1 4D Representation

4D Gaussian Splatting The concept of 3D Gaussian Splatting was initially introduced in the work of [7], where an explicit point-based representation is used to model the 3D scene. Specifically, the 3D Gaussian point cloud, denoted as $\mathcal{S}$ for a static scene, is characterized by the tuple:

$$\mathcal{S} = [\mathcal{X}, s, r, \sigma, \zeta], \tag{1}$$

where $\mathcal{X} = (x, y, z)$ denotes the positions of the 3D Gaussian points, and s, r, σ and ζ represent the scale, rotation, opacity, and spherical harmonics (SH) coeffi-

cients of the radiance, respectively. In 3D reconstruction, the 3D Gaussian point cloud will be optimized through a point-based differentiable volume rendering.

This representation can be naturally extended to 4D scenes by presenting the continuous scene dynamics as a 3D motion field. Inspired by [33], we present a 4D scene as a 3D Gaussian point cloud with a hex-plane-based deformation field, denoted as $\mathcal{F}(\mathcal{S}, t)$. The 3D Gaussian point cloud at time t can be expressed as:

$$\mathcal{F}(\mathcal{S}, t) = [\mathcal{X}_t, s_t, r_t, \sigma, \zeta], \tag{2}$$

where $\mathcal{X}_t = (x_t, y_t, z_t)$, s_t and r_t represent the updated information of Gaussian position, scale, and rotation at time t.

Adaptive Densification The standard 3D Gaussian Splatting technique typically employs a point cloud densification control strategy to dynamically adjust the number of Gaussians and their density within a unit volume. This adaptive approach allows for the transition from an initial sparse Gaussian set to a denser configuration to better represent the 3D scene. The 4D Gaussian Splatting method introduced in the work [33] utilizes a densification strategy similar to that presented in vanilla 3DGS [7], which involves the use of a fixed densification threshold based on the view-space position gradient. However, while the fixed gradient threshold strategy demonstrates efficacy in the context of reconstruction, particularly when multiple views afford robust and redundant coverage of the target scene, it performs sub-optimally in the generative setting. This limitation arises from the constraints imposed by the input single image or monocular videos, leading to significant uncertainty in the spatial and scale dimensions for each training object. This will lead to different optimal thresholds for different cases, as illustrated in the ablation study.

To address the issue, we propose an adaptive threshold approach, wherein only candidates with relatively large gradients are selected for densification. Our approach is based on the statistical analysis of the accumulated gradient of each point, which follows a log distribution of similar shapes throughout the training process. We apply an adaptive threshold that filters out Gaussians with a relatively small gradient and only densifies those with a large gradient. The adaptive threshold is set to select a fixed percentage (top $\lambda\%$) of points with the highest gradient in each densification operation. This ensures that the threshold adapts to the distribution of the gradient and maintains a stable relative position. Experimental results demonstrate that the proposed simple strategy can significantly enhance the quality and robustness of the 4D generation.

3.2 Temporal and Multi-view Consistent Diffusion

Multi-view Consistent Diffusion Our pipeline adopts Score Distillation Sampling(SDS) [22] from image diffusion models for the optimization of dynamic Gaussian. Specifically. Image-to-image diffusion model Zero123 [13] is utilized

and the SDS loss gradient could be formulated as follows:

$$\begin{aligned} \nabla_\theta \mathcal{L}_{SDS}(\phi, \mathbf{x}) &= \mathbb{E}_{t,\epsilon}\left[\omega(t)(\hat{\epsilon}_\theta(\mathbf{z}_t; \mathbf{I}_{in}, \mathbf{R}, \mathbf{T}, t) - \epsilon)\frac{\partial \mathbf{x}}{\partial \theta}\right], \\ \hat{z} &= z_t - \sigma_t \hat{\epsilon}_t(z; \mathbf{I}_{in}, \mathbf{R}, \mathbf{T}). \end{aligned} \tag{3}$$

where θ represents the parameters of the 3D representation, $\mathbf{x}$ the rendered image at the current view, t the timestamp in the diffusion process, ϵ the ground truth noise, $\hat{\epsilon}$ the predicted noise from the noisy image $\mathbf{z_t}$ conditioned on an initial input $\mathbf{I}_{in}$, and the relative camera pose between input view and target view $(\mathbf{R}, \mathbf{T})$. Zero123 can generate the target view image at any relative camera location but only one target view at a time, thus good at optimizing the 3D object from all views whilst bad at generating spatially consistent images from multiple target views. In contrast, Zero123++ [26] leverages reference attention [37] to model the relationship of the images from multiple target views as well as the input view, resulting in multi-view consistency output yet at the cost of fixed target camera locations. We combine the advantages of both models by taking the multi-view consistent output of Zero123++ as the input images of Zero123 when calculating SDS loss.

Temporally Consistent Diffusion Following the view generation philosophy of multi-view diffusion based 3D generation, a practical solution to 4D content creation is to generate the multi-view videos of the 4D scene, and then apply the multi-view video conditioned SDS loss to optimize the scene. However, it is rather difficult to generate temporally consistent multi-view videos by the separate per-frame generation. Inspired by zero-shot video generation method [8], we propose a training-free temporal attention module to enable Zero123++ [26] with temporal-awareness.

In our design, we obtain the attention information during denoising the multiview images of the first frame. Then we apply the recorded attention to the later denoising process. Specifically, during the denoising process of the latent code, the self-attention layer computes three essential components—queries (Q), keys (K), and values (V), which is formulated as:

$$Self\text{-}Attn(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{c}})V. \tag{4}$$

We first compute the key K_0 and value V_0 for the initial frame, T_0. Subsequently, for each subsequent frame T_t, where $t \in \{1, ..., N\}$, we perform a mixing operation on the key K_t and value V_t with the initial key K_0 and value V_0 derived from the first frame T_0. This process can be formally represented as follows:

$$\begin{cases} K_t = \gamma \mathbf{K}_0 + (1-\gamma)\mathbf{K}_t \\ V_t = \gamma \mathbf{V}_0 + (1-\gamma)\mathbf{V}_t, \end{cases} \tag{5}$$

where the parameter γ serves as a weighting factor that governs the influence of the initial frame, as illustrated in the ablation part. Additionally, we employ

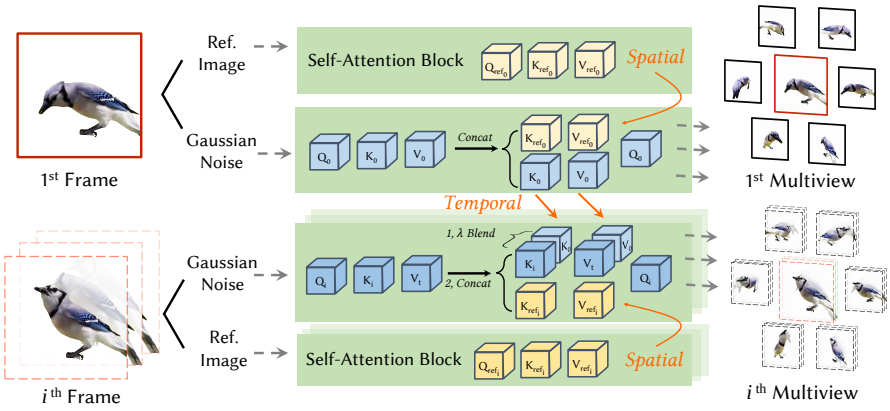


Fig. 3: Illustration of the proposed spatial and temporal attention fusion during multi-view sequence inference.

reference attention following the approach outlined in [26, 37], to extract local conditioning from the input image. In this reference attention mechanism, both the key and the value are obtained through the concatenation of features from the reference image and the noisy input image. Our design of temporal attention module is simple yet effective with superior temporal consistency, image quality and 3D consistency to vanilla cross-frame attention proposed in [8], as demonstrated in experiment section.

3.3 Training Objectives

Following the generation of multi-view sequences from a monocular reference video, we can acquire 6 anchor views $\{I_t^i\}_{i\in\{1\ldots6\}}$ and a reference view I_t^{ref} at each timestep t. In our optimization process, we employ multi-view score distillation sampling (SDS) utilizing the generated images $\{I_t^i\}_{i=1\ldots6}$ in conjunction with the reference images I_t^{ref}. The multi-view score distillation loss function $\mathcal{L}_{MVSDS}$ can be defined as:

$$\begin{aligned}\mathcal{L}_{MVSDS} &= \lambda_1\mathcal{L}_{SDS}^{i} + \lambda_2\mathcal{L}_{SDS}^{ref} \\ &= \lambda_1\mathcal{L}_{SDS}(\phi, I_t^i) + \lambda_2\mathcal{L}_{SDS}(\phi, I_t^{ref}),\end{aligned} \tag{6}$$

where λ_1 and λ_2 are two weighting factors, the index i is determined based on the proximity of the rendering viewpoint to the viewpoint of the generated images. This selection process, which we refer to as multi-view score distillation sampling, involves choosing the nearest reference image to the rendered camera view for calculating the SDS loss.

Following the approach [29], we utilize the reference image to compute both the reconstruction loss $\mathcal{L}_{rec}$ and the foreground mask loss $\mathcal{L}_{mask}$. Thus, the final

optimization objective is:

$$\mathcal{L} = \mathcal{L}_{MVSDS} + \lambda_3 \mathcal{L}_{rec} + \lambda_4 \mathcal{L}_{mask}, \tag{7}$$

where λ_3 and λ_4 are the weighting parameters . During training, we first use $\mathcal{L}$ to supervise a fixed frame to get static canonical 3D Gaussian. Then we use all anchors and reference images to train the dynamic 4D Gaussian point cloud.

3.4 Text-to-4D Extension

Our method is designed to be readily adaptable to text and image inputs, offering novel capabilities such as directly generating video sequences from textual descriptions or static images. We initiate this process using a 2D diffusion model SDXL [21], which generates the image from textual input. This 2D output is then transformed into a video sequence via a video diffusion model, such as SVD [2], adding temporal coherence and motion to the static image. Finally, the aforementioned pipeline further lifts the video sequence to a 4D scene, introducing an extra spatial dimension that creates immersive experiences beyond conventional 2D or 3D media. The generation pipeline with the given text, image, and video inputs is visualized in Fig. 2.

4 Experiment

4.1 Experiment Setup

Dataset For the video-to-4D task, we utilize the dataset provided by Consistent4D [6] for quantitative evaluation, which comprises multi-view videos depicting 7 dynamic objects. Additionally, for qualitative evaluation, we curate a set of challenging videos from online sources to assess the robustness and generalization capabilities of each method. For text/image-to-4D tasks, we follow the data settings in 4Dfy [1] and DreamGaussian4D [23] to create the corresponding results. Notably, we use the same input image and input video with DreamGaussian4D for a fair comparison.

Evaluation metrics We employ CLIP, LPIPS, and FVD for Video-to-4D evaluation as in the previous approach [6]. Specifically, CLIP and LPIPS serve as image-level metrics, assessing the semantic similarity between rendered images and ground truths; FVD is a video-level metric commonly used in video generation tasks, considering not only single-frame quality but also temporal coherence; we also consider the FID-VID metric to measure the temporal consistency. As 4D-Gen [36] does not support a 30-frame setting we report its FVD under a 16-frame reconstruction, namely FVD-16 as an additional metrics. For text-to-4D generation, we follow previous works and provide a user study for quantitative comparisons between different methods. We also provide a user study for video-to-4D and both studies are detailed in the supplementary material.

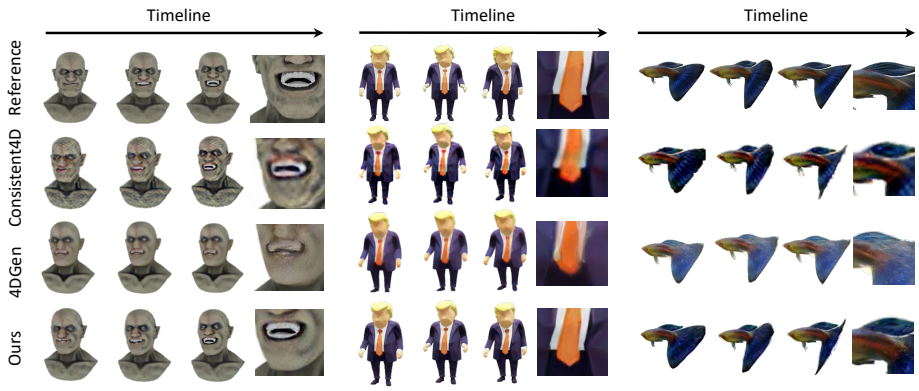


Fig. 4: Qualitative comparison on video-to-4D generation. Our method achieves better faithfulness to the input video with higher quality on the overall reconstruction.

Baselines For video-to-4D generation, we compare our method with two baselines: Consistent4D [6] and the concurrent work 4DGen [36]. We utilize the official code released by the authors to generate comparison results. For text&image-to-4D generation, we make comparisons with two state-of-art models that are open-sourced: 4Dfy [1] and DreamGaussian4D [23]. Their results are generated using codes from their official GitHub repository.

4.2 Implementation Details

In the training of the video-to-4D generative model, we adopt a two-stage approach. Initially, the model is initialized in canonical space and trained for 1000 steps. Subsequently, the deformation fields are learned over 7000 additional steps to accurately capture the dynamic scene. For the deformation process, we employ multi-layer perceptrons (MLPs) with 64 hidden layers and 32 hidden features per layer. The initial learning rate for the deformation MLPs is set to 1.6×10^{-4} and is decayed to 1.6×10^{-6} by the end of the training. In the context of adaptive densification, we opt to densify the top 2.5% of points with the most accumulated gradient. In the generative phase, we ultilize the spatial-temporal consistent videos for the multiview score distillation sampling optimization. Empirically, we set λ to 0.5 for our temporally consistent diffusion. Regarding the loss functions, we maintain the SDS loss weight at 1 and adjust the weights of other losses accordingly. Specifically, the reconstruction loss is assigned to 4×10^4, the mask loss is weighted at 1×10^4. The training process requires approximately 1 hour on an RTX 3090 GPU, and the rendering process can be performed at 150 FPS in real time. For Text&Image-to-4D, we use SDXL [21] for image generation, and apply SVD [2] to the generated image to create a corresponding video. Then we use the same setting as Video-to-4D to transfer the video into 4D content.

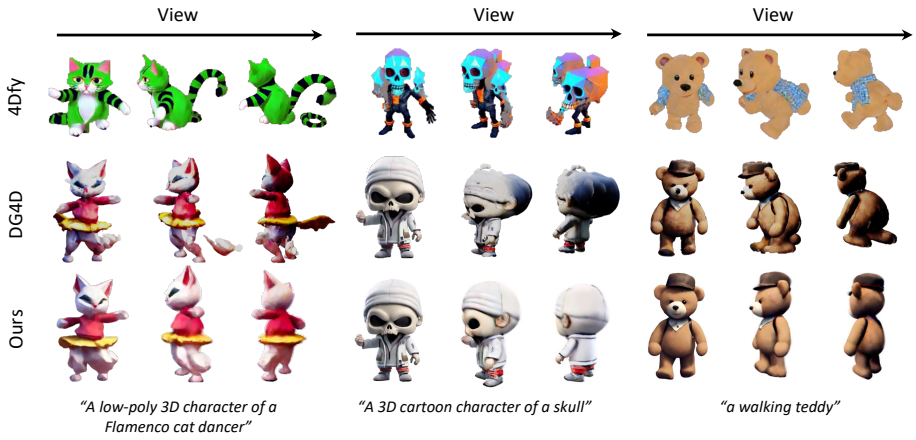


Fig. 5: Qualitative comparison on text-to-4D generation. Our method achieves the best results on the text alignment and visual quality.

4.3 Comparison with State-of-the-art Approaches

In this section, we conduct a comprehensive comparison with the aforementioned baselines using synthetic and in-the-wild data. The superior performance demonstrated in both quantitative and qualitative evaluations serves to substantiate the effectiveness and robustness of our proposed method.

Quantitative Results on Video-to-4D The experimental results presented in Table 1 offer a comparative analysis of our method against state-of-the-art techniques, namely Consistent4D and 4DGen, across several metrics. Our method outperforms Consistent4D and 4DGen in CLIP, showing better semantic consistency with the target content. It also achieves the lowest LPIPS score, meaning it generates more realistic images than the others. For the video quality and smoothness, our method beats 4DGen in FID-VID and FVD-16, and Consistent4D in FVD, implying that our videos are closer to real videos and have less temporal artifacts. Overall, the numbers confirm that our method is superior in generating semantically aligned and perceptually convincing videos with high fidelity and coherence.

Qualitative Comparison on Video-to-4D In our investigation of the video-to-4D task, we have employed various methodologies to conduct qualitative comparisons. Figure 4 shows the rendered outcomes at distinct temporal intervals and perspectives. A critical observation of our findings reveals that Gaussian splatting is inadequate in producing credible and temporally consistent images during per-frame reconstruction. Furthermore, we have observed that Consistent4D often produces over-saturated and unrealistic patterns, which may stem from the over-restrictive geometry representation in their cascade Dynerf. Additionally, our analysis indicates that 4DGen struggles to generate reasonable motion for the Gaussians, leading to its points lacking deformation over time. Moreover, 4DGen fails to generate detailed textures for the surface, which appear

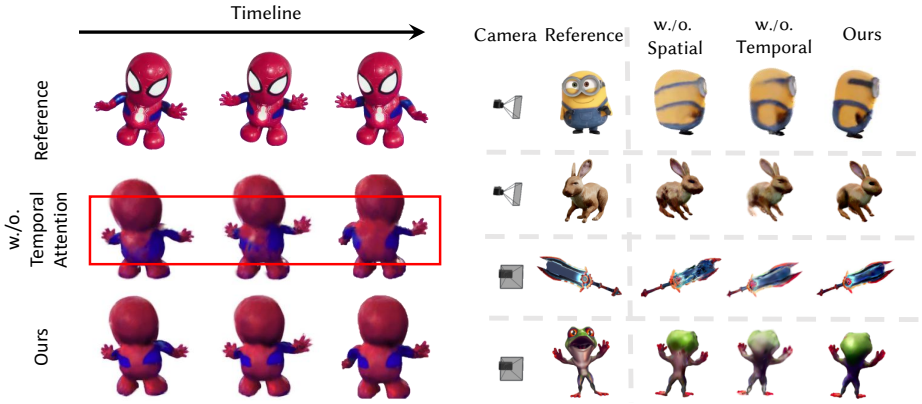


Fig. 6: Ablation on spatial and temporal attention. We evaluate the effect of spatial and temporal attention on the reconstruction results.

to be over-smoothed in some cases. Conversely, our methodology demonstrates substantial enhancements in comparison to existing strategies, particularly in aspects of reconstruction fidelity and stability. Notably, our technique facilitates seamless color transitions, as exemplified in the Fish case, and renders a convincing low-poly effect, as demonstrated in the Trump case, underscoring our method's adaptability to diverse dynamic environments.

Qualitative Comparison on Text-to-4D We also evaluate our method on the Text&Image-to-4D task, which aims to generate 4D content from text and image inputs. Figure 5 shows the 4D generation results of our method and two baselines, 4Dfy [1] and DreamGaussian4D [23], under different views. As can be seen, 4Dfy [1] produces coarse results and fails to capture the complex semantics of the text input, e.g. in the Cat case. DreamGaussian4D [23] generates meshes with fine-grained details in the front view, but suffers from severe artifacts and distortions in the side view or back view. In contrast, our method synthesizes Gaussians with smooth and consistent geometry and realistic texture across different views, thanks to our attention mechanism that leverages spatial and temporal information. Our method demonstrates its superiority and versatility on the Text&Image-to-4D task, which can be seen as a direct application of our Video-to-4D framework.

Table 1: Comparison with state-of-the-art methods.

	CLIP ↑	LPIPS ↓	FID-VID ↓	FVD ↓	FVD-16 ↓
Consistent4D	0.877	0.134	/	1133.93	/
4DGen	0.894	0.130	71.99	/	1005.72
Ours	**0.909**	**0.126**	**52.58**	**992.21**	**952.41**

4.4 Ablation Study

The experimental analysis presented in Table 2 provides a comprehensive evaluation of the impact of various components on the performance of a diffusion model with a 4D representation. The components under consideration include baseline diffusion, spatial anchor, spatial-temporal anchor, and adaptive densification. In the first configuration of Table 2, we regard the model utilizing the diffusion process [13] with adaptive densification as the baseline.

Table 2: Ablation study on effectiveness of each novel component.

Diffusion Model			4D Gaussian	Evaluation Metrics			
Baseline	Spat.	Temp.	Adaptive Dens.	CLIP ↑	LPIPS ↓	FID-VID ↓	FVD ↓
✓			✓	0.895	0.135	77.88	1369.65
✓	✓		✓	0.899	0.130	67.04	1198.23
✓	✓	✓		0.890	0.136	90.30	1453.53
✓	✓	✓	✓	**0.909**	**0.126**	**52.58**	**992.21**

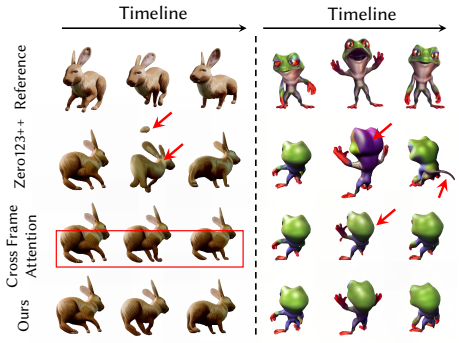
Table 3: Ablation study on different attention mechanisms.

	Evaluation Metrics			
	CLIP ↑	LPIPS ↓	FID-VID ↓	FVD ↓
Spat. only	0.899	0.130	67.04	1198.23
Cross-Frame	0.901	0.129	63.32	1053.25
Spat.-Temp.	**0.909**	**0.126**	**52.58**	**992.21**

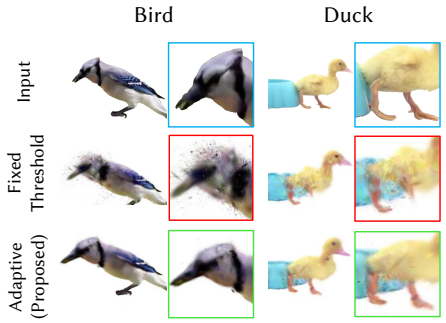
Spatial and Temporal Anchors Table 2 shows that adding the spatial anchor to the baseline model leads to a marginal improvement in all metrics, especially FID-VID and FVD. The final model, which uses both spatial-temporal anchor and adaptive densification, improves all metrics significantly. It has the best CLIP, LPIPS, FID-VID, and FVD scores, meaning it generates the most semantically consistent, realistic, and smooth videos.

Figure 6 shows the importance of our spatial-temporal anchor for improving the 4D content quality. The baseline 4D Gaussian method, labeled as "w./o. spatial," fails to deform the point clouds properly, leading to unreasonable shapes without spatial-temporal constraints. A model that uses only the spatial anchor, "w./o. temporal," struggles to deform solid shapes, as seen in the rabbit's leg and the frog's back. Our method, which combines both spatial and temporal anchors, overcomes these issues and produces visually consistent and accurate 4D content.

Adaptive Densification We present the statistical analysis of the adaptive densification module in the third and final configurations of Table 2. The results show that the adaptive densification module improves the overall quality of the generative model, especially in terms of the FID-VID and FVD metrics. We also compare the proposed adaptive densification with a fixed threshold approach in Figure 7b. We observe that a fixed threshold can lead to either under-densification or over-densification, depending on the case. For example, in the Bird case, the fixed threshold is too low for densification, which results in excessive Gaussian points. On the other hand, in the Duck case, the fixed threshold is too high for densification, which causes the foot to be blurry due to insufficient Gaussian points. Our method overcomes this limitation by adapting the gradient threshold to different cases, thus producing robust and stable 4D Gaussian results.



(a) Ablation on Multiview Diffusion Models We evaluate the effect of different attention mechanism on Zero123++. Our attention mechanism can achieve both temporal coherence and geometric fidelity given a monocular input video.



(b) Ablation on Adaptive Densification We evaluate the effect of using different densify mechanism during reconstruction. Our adaptive threshold provides reasonable densification guidance for each case.

Different Attention Mechanisms In our study, we evaluated three distinct attention mechanisms within the context of images generated by our diffusion model: spatial attention, cross-frame attention, and spatial-temporal attention. The visual outcomes of these varied approaches are illustrated in Figure 7a. We employed a parameter λ to modulate the extent of spatial and temporal attention exerted, as delineated in Equation 5. Our findings indicate that exclusive reliance on spatial attention does not guarantee consistency across sequential frames. While cross-frame attention enhances frame-to-frame coherence, it fails to achieve a satisfactory alignment with reference images. Conversely, the integration of spatial-temporal attention ensures consistency across both dimensions. We quantitatively assessed the quality of images generated under different attention regimes using four established metrics: CLIP, LPIPS, FID-VID, and FVD. The comparative results are tabulated in Table 3. It was observed that images generated with spatial-temporal attention outperformed others across all evaluation metrics, which underscores the efficacy of our proposed mechanism.

5 Conclusion

The paper presents a novel approach for dynamic 3D content generation from monocular videos, addressing the challenges of 4D representation and spatial-temporal consistency. By leveraging specially tailored 4D Gaussian splatting and a novel information fusion module, the proposed method achieves high-quality and robust 4D scene generation. Comprehensive experiments demonstrate the method's effectiveness, showcasing a faster generation speed and significant improvements in rendering quality and temporal consistency compared to prior state-of-the-art methods. Overall, the proposed method sets a new benchmark for training speed, rendering quality, and 4D consistency in dynamic 3D content generation from monocular videos, opening up possibilities for real-world applications.

Acknowledgements

This work was supported by the National Key R&D Program of China (2022YFF0902200) and NSFC grant 62441204.

References

1. Bahmani, S., Skorokhodov, I., Rong, V., Wetzstein, G., Guibas, L., Wonka, P., Tulyakov, S., Park, J.J., Tagliasacchi, A., Lindell, D.B.: 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7996–8006 (2024)
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets (2023)
3. Du, Y., Zhang, Y., Yu, H.X., Tenenbaum, J.B., Wu, J.: Neural radiance flow for 4d view synthesis and video processing. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14304–14314. IEEE Computer Society (2021)
4. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5712–5721 (2021)
5. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
6. Jiang, Y., Zhang, L., Gao, J., Hu, W., Yao, Y.: Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. arXiv preprint arXiv:2311.02848 (2023)
7. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (ToG) **42**(4), 1–14 (2023)
8. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439 (2023)
9. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
10. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 300–309 (2023)
11. Lin, Y., Dai, Z., Zhu, S., Yao, Y.: Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. arXiv:2312.03431 (2023)
12. Ling, H., Kim, S.W., Torralba, A., Fidler, S., Kreis, K.: Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. arXiv preprint arXiv:2312.13763 (2023)
13. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9298–9309 (2023)
14. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453 (2023)

15. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. arXiv preprint arXiv:2310.15008 (2023)
16. Lu, Y., Zhang, J., Li, S., Fang, T., McKinnon, D., Tsin, Y., Quan, L., Cao, X., Yao, Y.: Direct2. 5: Diverse text-to-3d generation via multi-view 2.5 d diffusion. arXiv preprint arXiv:2311.15980 (2023)
17. Lu, Y., Zhang, J., Li, S., Fang, T., McKinnon, D., Tsin, Y., Quan, L., Cao, X., Yao, Y.: Direct2.5: Diverse text-to-3d generation via multi-view 2.5d diffusion (2023)
18. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. arXiv preprint arXiv:2308.09713 (2023)
19. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
20. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**(4), 102:1–102:15 (Jul 2022). https://doi.org/10.1145/3528223.3530127, https://doi.org/10.1145/3528223.3530127
21. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis (2023)
22. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
23. Ren, J., Pan, L., Tang, J., Zhang, C., Cao, A., Zeng, G., Liu, Z.: Dreamgaussian4d: Generative 4d gaussian splatting. arXiv preprint arXiv:2312.17142 (2023)
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2022)
25. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
26. Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., Su, H.: Zero123++: a single image to consistent multi-view diffusion base model. arXiv preprint arXiv:2310.15110 (2023)
27. Shi, Y., Wang, P., Ye, J., Mai, L., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv:2308.16512 (2023)
28. Singer, U., Sheynin, S., Polyak, A., Ashual, O., Makarov, I., Kokkinos, F., Goyal, N., Vedaldi, A., Parikh, D., Johnson, J., et al.: Text-to-4d dynamic scene generation. arXiv preprint arXiv:2301.11280 (2023)
29. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023)
30. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12619–12629 (2023)
31. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
32. Wang, X., Wang, Y., Ye, J., Wang, Z., Sun, F., Liu, P., Wang, L., Sun, K., Wang, X., He, B.: Animatabledreamer: Text-guided non-rigid 3d model generation and reconstruction with canonical score distillation. arXiv preprint arXiv:2312.03795 (2023)

33. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. arXiv preprint arXiv:2310.08528 (2023)
34. Xian, W., Huang, J.B., Kopf, J., Kim, C.: Space-time neural irradiance fields for free-viewpoint video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9421–9431 (2021)
35. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. arXiv preprint arXiv:2309.13101 (2023)
36. Yin, Y., Xu, D., Wang, Z., Zhao, Y., Wei, Y.: 4dgen: Grounded 4d content generation with spatial-temporal consistency. arXiv preprint arXiv:2312.17225 (2023)
37. Zhang, L.: Reference-only control (2023), https://github.com/Mikubill/sd-webui-controlnet/discussions/1236
38. Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qing, Z., Wang, X., Zhao, D., Zhou, J.: I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models (2023)
39. Zhao, Y., Yan, Z., Xie, E., Hong, L., Li, Z., Lee, G.H.: Animate124: Animating one image to 4d dynamic scene. arXiv preprint arXiv:2311.14603 (2023)
40. Zheng, Y., Li, X., Nagano, K., Liu, S., Hilliges, O., De Mello, S.: A unified approach for text-and image-guided 4d scene generation. arXiv preprint arXiv:2311.16854 (2023)