# DiTex4D: Direct Text-Driven 4D Generation with Structured Latent Diffusion

Xiaozhe Chen[1,2], Mengqi Rong[2], Jian Liu[3], and Shuhan Shen[1,2]✉

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences
[2] Institute of Automation, Chinese Academy of Sciences
[3] School of Space and Earth Sciences, Beihang University
chenxiaozhe2025@ia.ac.cn; liujian@lynmaxtech.com;
{mengqi.rong, shshen}@nlpr.ia.ac.cn

https://github.com/3dv-casia/DiTex4D


**Fig. 1:** Our framework **DiTex4D** generates high-quality 4D content from text prompts, supporting both 4D generation from scratch and 3D animation from static mesh. More visual results are available at our project page.

**Abstract.** Despite recent progress, direct text-driven 4D object generation remains challenging yet highly desirable. In this paper, we introduce DiTex4D, a native text-to-4D generation framework that enables both text-driven 4D generation from scratch and 3D animation from static mesh. Built upon large-scale pre-trained 3D generation models, our framework DiTex4D avoids intermediate text-to-video pipelines and costly per-object optimization. Specifically, (i) we achieve 4D spatiotemporal consistency via inflating 3D attention with mixed-4D RoPE and tailored correlated noise injection strategy. (ii) To enable 3D animation, we introduce a mask-based diffusion model conditioned on multi-view global context to maintain strict consistency with the initial frame.

✉ Corresponding authors

We further fine-tune the framework for 4D interpolation to synthesize high-frame-rate sequences with smoother motion. Extensive experiments demonstrate that DiTex4D can achieve higher-quality, semantically aligned, and spatiotemporally coherent 4D object generation, surpassing most existing state-of-the-art text-to-4D generation methods.

**Keywords:** 4D Generation · 3D Animation · Generation Model

## 1 Introduction

High-quality 3D content generation has received widespread attention in virtual/augmented reality, gaming, and embodied applications. Recent advances in text/image-to-3D, such as Score Distillation Sampling (SDS) [42, 51, 56], Large Reconstruction Models [21,50], and Diffusion Models adapted for 3D content [60,68,72], have made it possible to create detailed static 3D objects. Building on these strong 3D priors, a few works extend 3D generation to 4D by injecting temporal modules to model motion across frames, demonstrating promising results in video-to-4D generation [29, 65]. In contrast, directly transferring video-to-4D to text-to-4D is full of challenges. The task is inherently challenging because (a) text lacks explicit geometric and motion control priors, making it difficult to synthesize plausible dynamics while maintaining appearance and geometric consistency, and (b) it must bridge the substantial gap between the simplicity of text prompts and the complexity of dynamic 3D outputs.

Previous text-driven 4D generation methods mainly fall into two categories. The first line of work [30,44,57] generates each individual 4D content by utilizing pre-trained text-to-video generation models and optimizing the 4D representation via complex SDS, as shown in Fig. 2(a). This approach suffers from long optimization times and over-saturation artifacts. The other line of work focuses on two stage video-based methods, which primarily decompose text-to-4D generation into a text-to-video stage [2, 20] followed by either 4D reconstruction or native video-to-4D generation [47], as shown in Fig. 2(b). Methods relying on 4D reconstruction [33, 45, 61, 64] typically fine-tune video models into multi-view video generators, and then directly recover 4D content via traditional or feed-forward reconstruction techniques. While this paradigm is faster than SDS-based methods, it suffers from significant cumulative errors across the two stages, leading to suboptimal content quality. Alternatively, native video-to-4D generation approaches [29, 65, 69] based on 3D generation can produce high-quality 4D content from monocular videos, such as GVFDiffusion [69], SS4D [29]. However, directly integrating them into a two-stage text-to-4D pipeline often yields unsatisfactory results and leads to failure cases. This is primarily because such methods impose strong selective constraints on the video generated conditioned on text (e.g., requiring no motion occlusion, low motion speed, etc.).

Recently, to bypass the constraints of video generation, another line of work directly generates trajectories conditioned on text to drive input meshes, as shown in Fig. 2(c). This approach produces 4D content with stronger semantic alignment, such as AnimateAnyMesh [58]. Furthermore, it not only reduces the

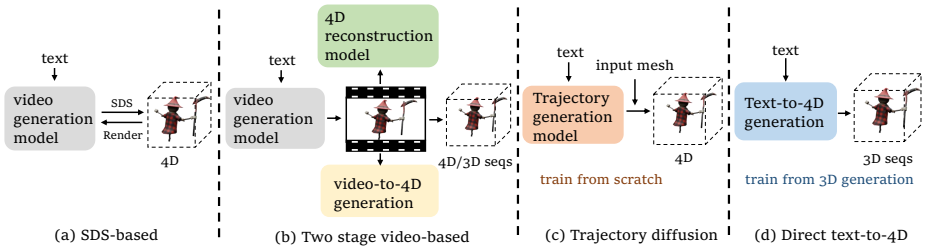


**Fig. 2:** Our method is different from previous works: (a) is the SDS-based; (b) is the video-based two stage method; (c) is the trajectory generation model not based on 3D generation; (d) is our direct text-driven 4D generation build on 3D prior.

two-stage cumulative error but also maintains geometric and texture consistency during motion. However, this strategy suffers from a critical limitation: failing to build upon existing 3D foundation models makes it lose text-to-3D priors. Given the extreme scarcity of 4D data, it struggles to generate diverse, large-deformation motions for out-of-distribution input meshes.

In this paper, we propose **DiTex4D**, a direct text-driven 4D generation framework that generates 3D Gaussian Splatting (3DGS) sequences **via two pathways**: (a) 4D generation from scratch, and (b) 3D animation from static mesh. Our model is built upon TRELLIS [60], utilizing the powerful text-to-3D knowledge of the pre-trained model and the characteristics of structured latents to achieve text-to-4D generation. Unlike previous works in Fig. 2, we allow semantic information to interact directly with temporal 3D objects, which avoids two-stage cumulative errors and improves the geometric consistency of the motion. Moreover, since our model is built on strong 3D priors, it enables more effective text–4D interaction and performs well even when large-scale 4D datasets [31] are scarce.

To implement this framework, we realize these two paths of 4D generation through several key designs: (a) inflating 3D attention to capture the spatio-temporal relationships in 4D space; (b) introducing a mask mechanism in the diffusion model and strengthening the conditioning to enforce consistency with the initial frame, thereby enabling 3D animation. And then we design a 4D interpolation model to improve motion smoothness. To the best of our knowledge, DiTex4D is the first direct text-to-4D framework that directly injects text into structured 4D latent diffusion. (visualized in Fig. 1).

Overall, our main contributions are:

- We propose **DiTex4D**, a direct text-driven 4D generation framework with two complementary pathways: (i) 4D generation from scratch and (ii) text-conditioned 3D animation from static mesh.
- We inflate 3D attention with mixed-4D RoPE and design dedicated correlated noise injection strategy to preserve spatio-temporal consistency.

– We introduce a masked diffusion mechanism and leverage multi-view global information to enforce consistency between the generated content and the input mesh, thereby enabling 3D animation.

## 2 Related Work

**3D Generation.** Early 3D generation approaches primarily relies on Generative Adversarial Networks (GANs) [16] and Score Distillation Sampling (SDS) [42]. While GAN-based methods [6, 13, 14] provide the groundwork, they struggle to scale across diverse, open-world distributions. Conversely, SDS-based methods [32,56,73] leverage geometric priors from pre-trained 2D generative models, offering better generalization. However, they frequently suffer from multi-view inconsistencies due to a lack of explicit 3D awareness. Furthermore, the time-consuming nature of per-scene optimization limit their widespread application. To mitigate these inconsistencies, subsequent works fine-tune 2D diffusion models [46] to generate multi-view images [35,36,54,62], yet these still require cumbersome post-processing reconstruction steps.

In contrast, recent feed-forward methods [18,28,59,72] directly generate 3D representations from text or image guidance. These end-to-end architectures can produce 3D assets in seconds, drastically improving efficiency. These methods broadly fall into two streams. The first encodes point clouds into unordered latent vectors (VecSet) [68], that generate generalizable, high-quality geometry. The second stream leverages voxel grids and structured latents [60], which support decoding into meshes, 3DGS [27], or Neural Radiance Fields (NeRF) [40], facilitating scaling to large models. Notably, voxel-based structured latents possess inherent spatial determinism, offering a distinct advantage for spatiotemporal modeling. Unlike ShapeGen4D [65], which require explicit latents alignment to eliminate temporal jitter, voxel-based representations inherently provide structural stability, making them highly suitable for our work.

**Video-to-4D Generation.** Similar to 3D generation, early video-to-4D generation work rely heavily on optimization-based methods [25, 33, 66, 67]. These approaches predominantly utilize SDS and its 4D variants to optimize 4D representations like HexPlane [4] or 4DGS [57] using priors from video generation model. Like SDS-based 3D generation, they suffer from multi-view inconsistencies and efficiency bottlenecks. To accelerate generation, L4GM [45] extends feed-forward 3D reconstruction models into the 4D domain. By introducing temporal layers to maintain spatiotemporal coherence, it directly reconstructs 4D content from multi-view videos [22,38,70]. Nonetheless, this approach still struggles with view inconsistencies and produced weak geometric structures.

Following the emergence of large-scale foundational 3D models, V2M4 [8] employs state-of-the-art 3D models to extract meshes frame-by-frame, relying on global mesh registration and geometry optimization. Meanwhile, GVFDiffusion [69] compresses per-frame variations in Gaussian attributes into latent

space, using a diffusion model to generate deformation fields. Yet, these methods underutilize the strong 3D priors of foundational models. More recently, methods such as motion2vecsets [5], SS4D [29] and ShapeGen4D [65] perform spatiotemporal consistent fine-tuning on 3D diffusion models, achieving high-quality video-to-4D generation using limited 4D data. Although these models successfully generate dynamic 3D content from videos, directly adapting them for text-to-4D generation remains significantly challenging.

**Text-to-4D Generation.** Generating dynamic 4D content from textual descriptions can primarily be approached via two pathways: **generating 4D content from scratch** or **animating static 3D content**. The former requires recovering complex 4D motion from brief text prompts while strictly maintaining geometric and appearance consistency. Currently, this relies mostly on SDS-based methods [10, 24, 44] that use video generation model to optimize 4D representations per object, plagued by significant stability issues. Conversely, The latter pathway benefits from strong 3D information but is constrained by the quality of the static 3D input. For instance, Animate3D [24] generates multi-view videos and a static mesh from text, then optimizes 4D content using a pre-trained MV-Video generation model while enforcing correspondence with the initial mesh. Inspired by feed-forward models, AnimateAnyMesh [58] employs a diffusion model to generate text-conditioned vertex motion trajectories. By incorporating a shape-guided mechanism, this method inherently ensures spatial consistency to achieve high-quality mesh animation. Nevertheless, constrained by the scarcity of 4D datasets, it suffers from limited generalization capabilities. In contrast, our framework is built upon a powerful 3D foundational model. Its robust 3D priors reduce the difficulty of 4D generation, while our key design unifies both pathways for text-to-4D generation.

## 3 Method

To address the challenge of text-to-4D object generation, we propose DiTex4D, a flow-based structured latent diffusion framework designed to generate 3DGS sequences that align with textual prompts. As illustrated in Fig. 3, our framework supports 4D content generation via two pathways: text-driven 4D generation from scratch, and 3D animation from static mesh. The former achieves spatiotemporal consistency by effectively inflating 3D attention mechanisms to bridge text and 4D space. Building upon this, the latter incorporates a masking mechanism that leverages known geometry and appearance, enabling animation while preserving the initial frame. Additionally, to improve motion smoothness, we implement a 4D interpolation model based on this framework. This section first introduces TRELLIS [60], a 3D generation model via structured latent that serves as our foundation model (Sec. 3.1), followed by a detailed description of the two pathways in our framework (Sec. 3.2 and Sec. 3.3).

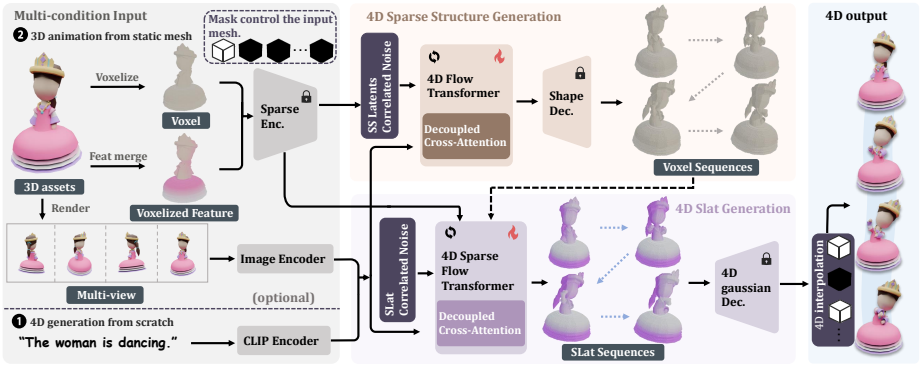


**Fig. 3: DiTex4D Overview.** Our framework supports two pathways for 4D generation, conditioned on text and static mesh inputs, while utilizing masks for mesh control. The 4D structured latent diffusion model implements a coarse-to-fine two-stage generation of multi-frame SLat, with correlated noise injection for consistency and Decoupled Cross-Attention for multi-condition guidance. The final SLat is decoded into high-quality 3DGS sequences, which are interpolated as the 4D output.

### 3.1 Preliminaries

TRELLIS [60] is a state-of-the-art 3D generation model that provides us with powerful 3D prior knowledge. It is built upon Structured Latent (SLat) that can effectively capture both the geometry and appearance of 3D objects and enables decoding into multiple 3D representations. SLat maps features from multi-view images onto activated voxels, expressed as:

$$z = \{(z_i, p_i)\}_{i=1}^{L}, z_i \in \mathbb{R}^C, p_i \in \{0, 1, ..., N-1\}^3, \tag{1}$$

where $p_i$ denotes the positional index of an activated voxel in 3D grid. The feature $z_i$ attached with each voxel is aggregated DINOv2 [41] features extracted from multi-view. $N$ denotes the spatial length of the 3D grid, and $L$ is the total number of active voxels.

The generation process employs a coarse-to-fine two-stage generation pipeline: generating the sparse structure (SS), represented as sparse voxels $\{p_i\}_i^L$, via SS Flow and then predicting Structured Latents (SLat), represented as $z = \{(z_i, p_i)\}_{i=1}^{L}$, via SLAT Flow. Modeling the backward process as a time-dependent vector field $v(x,t) = \nabla_t(x)$, the transformers $v_\theta$ in both stages are trained by minimizing the conditional flow matching (CFM) [34] objective:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,x_0,\epsilon}\|v_\theta(x,t) - (\epsilon - x_0)\|_2^2. \tag{2}$$

Compared to VecSet [68], SLat offers a more stable encoding process by avoiding the random sampling typical of point cloud processing—a crucial advantage for spatiotemporal modeling. As observed in ShapeGen4D [65], independently encoding per-frame meshes into VecSet features often leads to temporal jitter

in the latent space, hindering the diffusion model's ability to learn smooth dynamics. Conversely, the SLat encoding process involves no stochastic processes, facilitating the modeling of temporal dynamics. Furthermore, the two-stage generation pipeline of TRELLIS enables the decoupled temporal modeling of geometry and texture, providing superior controllability. These attributes motivate our selection of TRELLIS as the pre-trained foundation model.

### 3.2 Generate 4D Content from Scratch

The objective of our model is to generate dynamic 3DGS [27] sequences directly from text prompts. Unlike two stage video-based methods, we eliminate the video intermediate [2, 20, 63] to prevent error accumulation and poor text alignment caused by video limitations. By directly conditioning 4D space on text, DiTex4D bypasses the bottlenecks of video generation models, and produces 4D objects that adhere more closely to text prompt. Specifically, inspired by text-to-video generation [2, 15], we integrate temporal alignment into the 3D diffusion model and employ correlated noise injection to ensure consistency and smoothness.

**3D Attention Inflation.** Although a pre-trained 3D generator [60,72] produces strong *per-frame* content, it lacks explicit temporal modeling and thus cannot reliably capture coherent motion across frames, often yielding near-static outputs when directly applied to text-driven 4D synthesis. To address this issue, we draw inspiration from text-to-video models [3,17] and inflate the spatial self-attention layers of the pre-trained model into spatiotemporal self-attention layers. Similar to SS4D [29], this enables full attention interaction between time and space. Specifically, our inflation process involves reshaping the latent tensor, applying full self-attention, and restoring the original dimensions for subsequent layers, as follows:

$$z = reshape^{-1}(FullAttn(reshape(z))) \tag{3}$$

where $z \in \mathbb{R}^{B\times T\times L\times C}$ represents structured spatiotemporal latents, $B$ is the batch size, $T$ is the number of temporal frames, $L$ and $C$ are the sequences length and the feature dimension. The reshaping process primarily involves the merging and splitting of $T$ and $L$.

Since the text condition lacks explicit motion constraints, we modify the critical positional embedding component within the full attention mechanism. SS4D [29] retains spatial absolute position embeddings (APE) [53] while incorporating a 1D temporal Rotary Positional Embedding (RoPE) [49]. While this design works well for video-to-4D by recovering and aligning 3D geometry from per-frame observations, we find it insufficient for text-to-4D where such visual priors are absent. Inspired by video generation models, such as Wan [55, 76], we extend the original RoPE to 4D RoPE by independently applying 1D RoPE to each of the spatial and temporal coordinates $(x, y, z, t)$ and concatenating them along the channel dimensions. Furthermore, supported by findings on RoPE for Vision Transformers [19], we retain the pre-trained 3D spatial APE to reuse the learned spatial priors and combine it with our 4D RoPE, forming a **mixed-4D**

**RoPE**. Spatially, this combination of absolute and relative positional information effectively maintains geometric consistency in the absence of video constraints, even under large motions. Temporally, the relative encoding captures robust dynamic relationships and has certain extrapolation capabilities.

**Correlated Noise Injection.** Moreover, we observe that, during diffusion sampling, the additive Gaussian noise is typically sampled independently. As a result, relying exclusively on text prompts to denoise each frame leads to unstable motion. We hypothesize that the structured latents of adjacent frames should remain relatively close in distribution. When the variance of the randomly sampled noise for each frame is too large, text conditions struggle to unify the denoising process, resulting in unavoidable flickering in the generated content.

Inspired by early text-to-video generation [15, 26], we employ temporally-correlated noise injection to reduce inter-frame flickering. Specifically, we generate two noise vectors, $\epsilon_{shared}$, a common noise vector shared among all frames, and $\epsilon_{ind}$, the individual noise for each frame. The final combination of these two vectors is used as the correlated noise $\epsilon$, formulated as:

$$\epsilon_{\text{shared}} \sim \mathcal{N}\left(\mathbf{0}, \frac{\alpha^2}{1+\alpha^2}\mathbf{I}\right), \epsilon^i_{\text{ind}} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{1+\alpha^2}\mathbf{I}\right) \tag{4}$$

$$\epsilon^i = \epsilon_{\text{shared}} + \epsilon^i_{\text{ind}}$$

where $\alpha$ denotes noise sharing strength. Specifically, $\alpha = 0$ corresponds to independent noise, $\alpha \to +\infty$ represents copying noise. As proven by PYoCo [15], $\alpha = 1$ is the optimal setting for mixed noising, and used in our model.

Since the SLat-based generation follows a coarse-to-fine two-stage process, we need to perform this operation on both the SS latents and SLat. **(i) SS latents correlated noise,** as SS latents are continuous vectors in feature space without the geometric sparse structure, we can directly apply the linear superposition of noise as described above. However, for **(ii) SLat correlated noise,** since SLat is built upon sparse active voxels where both the voxel count and 3D spatial coordinates vary across frames, direct cross-frame noise addition is infeasible. Consequently, we design the tailored correlated noise technique:

$$\epsilon^i_{SLat} = \epsilon^{k_i}_{\text{shared}} + \epsilon^i_{\text{ind}} \tag{5}$$

$$\forall i, j, \quad \text{if } k_i = k_j \implies \epsilon^{k_i}_{\text{shared}} = \epsilon^{k_j}_{\text{shared}}$$

where the spatial key is defined as $k_i = (x_i, y_i, z_i), i \in \{0, 1, \ldots, T\}$, which implies that voxels at the same spatial coordinates $(x, y, z)$ share an identical key $k_i$ irrespective of time $t$. Consequently, $\epsilon^{k_i}_{\text{shared}}$ ensures that voxels with the same spatial location share the same noise, while $\epsilon^i_{\text{ind}}$prevents static content from appearing at fixed coordinates. By employing this strategy consistently during training and inference, we enhance the geometric consistency and temporal smoothness of the generated content.

### 3.3 3D Animation From Static Mesh

The task of 4D generation via 3D mesh animation focuses on synthesizing 4D motion from a static input mesh, guided by text prompts. This approach significantly enhances controllability by anchoring the generated 4D output to a specific initial frame, offering greater practical value. We design a text-driven 3D Animation model based on the foundational framework of the generation from scratch, preserving all optimization mechanisms designed for the 4D spatiotemporal domain. Simultaneously, it facilitates 3D animation by incorporating a masked diffusion mechanism and multi-view CLIP conditioning.

**Masked Diffusion Model.** The 3D animation approach is highly promising for practical applications. Benefiting from the success of image-to-video generation models [2,55,63,71], mask generation mechanisms can effectively synthesize high-quality content from static conditions. Consequently, we formulate 3D animation as a 3D-to-4D problem by introducing an additional conditional 3D mesh as the controlling initial frame. Specifically, we replicate the single-frame conditional 3D mesh $T$ times along the temporal axis to form a guidance sequence of length $T$. These guidance frames are compressed by the Structure VAE of TRELLIS into conditional latent vectors $z_c \in \mathbb{R}^{B\times T\times L\times C}$. Furthermore, we introduce a binary mask $M \in \{0,1\}^{B\times T\times L\times 1}$, where 1 indicates frames to be preserved and 0 indicates frames to be generated. The spatial and temporal dimensions of $M$ align with those of the conditional latents $z_c$. The noisy latents $z_t$, conditional latents $z_c$ and the mask $M$ are concatenated along the channel axis and then passed through the spatiotemporal full attention diffusion model. Given that the input channel dimension for the 3D animation model exceeds that of the base 4D generation model (i.e., $2 \times C + 1$ v.s. $C$), we employ an additional zero-initialized projection layer. This masking mechanism allows us to effectively preserve the explicit geometric information of the static 3D mesh while utilizing text conditions to guide the synthesis of dynamic 4D content.

**Multi-view Image CLIP Conditioning.** Although the pretrained base 4D generation model provides a framework for static mesh-based 3D animation, its lack of necessary context and semantic depth makes it difficult to fully capture the temporal consistency of dynamic 3D sequences, particularly when restricted to a single-frame input mesh. This discrepancy presents significant challenges in maintaining consistency between the initial frame and subsequent frames during the 3D animation. Therefore, in addition to text conditioning, we incorporate multi-view CLIP conditioning, inspired by ReconviaGen [7], SceneGen [39] and SAM-3D [9]. Specifically, we first render the static 3D mesh into multi-view images at azimuth angles of $\{0°, 90°, 180°, 270°\}$. We then utilize the CLIP [43] image encoder to extract feature representations from these multi-view images. These extracted features are projected by a three-layer multi-layer perceptron (MLP) to serve as the global context. Finally, this global context is injected into the diffusion model via decoupled cross-attention.

In practice, utilizing the masking mechanism and multi-view CLIP conditioning enables us to animate static 3D meshes conditioned on text prompts. The proposed mask delineates the input mesh and specifies exactly which frames are to be generated. Consequently, the aforementioned approach unifies the two pathways of 4D generation from scratch and 3D animation from static mesh, realizing a comprehensive text-driven 4D object generation framework.

### 3.4 4D Interpolation Model

To increase the temporal resolution of the generated 4D content, we fine-tune a 4D interpolation model built upon our base text-driven 4D generation framework. Motivated by the success of the masking mechanism in our 4D generation framework, we extend this strategy to frame interpolation. Specifically, we apply masking to the input sequence frames $I$ and $I + 2$ to predict the intermediate frame $I + 1$, achieving a $T \rightarrow 2T - 1$ interpolation, like Fig. 6. Consequently, this enhances the temporal smoothness of the 4D outputs and yields rendered videos with higher frame rates. Notably, our 4D interpolation model can achieve excellent performance with a relatively small number of iterations. Please refer to the supplementary material for specific details.

## 4 Experiments

### 4.1 Setting

**Datasets.** We curate a dataset containing 20,000 animated 3D objects from Objaverse [12] and ObjaverseXL [11], filtering out samples with insignificant motion amplitudes. For each motion sequence, we process every frame using TRELLIS's official script. Then we use the Qwen-3-VL model [1] to generate high-quality captions. To enable the model to generate content with larger motion amplitudes, we apply a simple data augmentation strategy by splitting long 4D data into multiple short segments to expand the dataset. Detailed information about our dataset can be found in the supplementary material.

**Implementation Details.** To optimize data efficiency and reduce training costs, we employ a progressive training strategy, scaling from 8 frames to 16 frames. As outlined in Sec. 3, our framework mainly consists of two models: the base 4D generation model and the 3D animation model. Both are trained using the AdamW optimizer [37] with a learning rate of $1 \times 10^{-4}$ under FP16 mixed precision, utilizing a batch size of 32 distributed across 8 GPUs. Further details are available in the supplementary material.

**Evaluation Metrics.** To comprehensively evaluate text-driven 4D objects generation without references, we adhere to the evaluation protocol established by Animate3D [24]. Specifically, we render dynamic 3D objects from 4 views {0°,

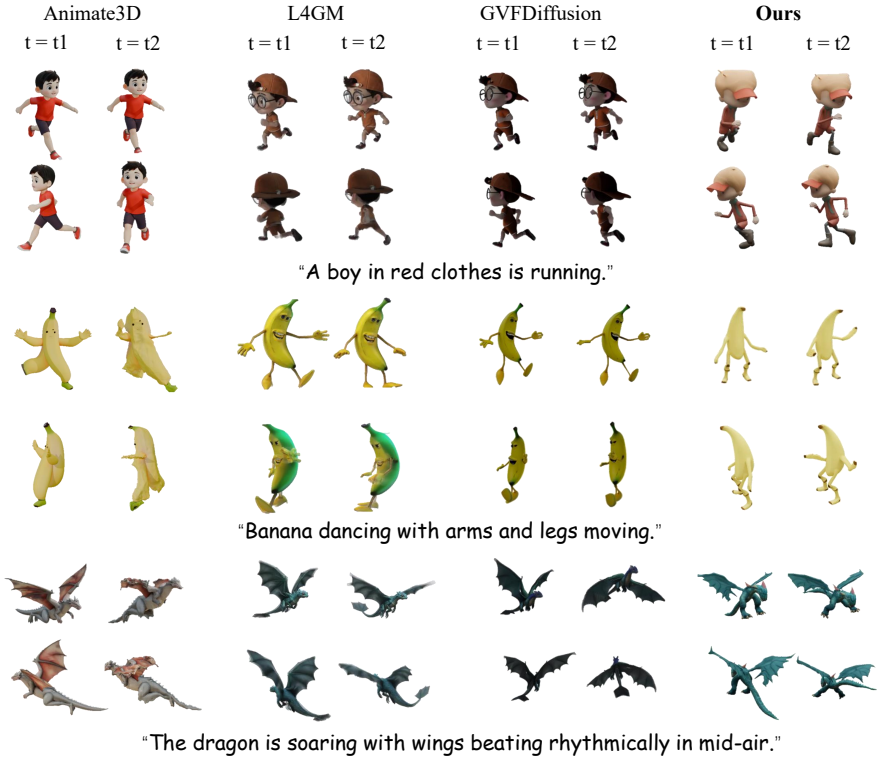


**Fig. 4:** Qualitative comparison with SOTA methods about 4D generation from scratch.

90°, 180°, 270°}. For each view, we calculate various perceptual metrics from VBench [23, 74], including I2V, Motion Smoothness, Aesthetic Quality and Dynamic Degree (denoted as I2V, M.sm, Aest.Q and Dy.Deg respectively), and report the average across all views. Notably, since base 4D generation lacks a reference image, we utilize the CLIP Score (CLIP) to replace the I2V metric. User study and further details are provided in the supplementary material.

### 4.2 Comparison

**Baseline.** We conduct comprehensive evaluations against representative state-of-the-art methods spanning different text-to-4D generation paradigms. Specifically, we compare with the SDS-based model Dream-in-4D [24] and Animate3D [24], the feed-forward reconstruction model L4GM [45], the video-to-4D model GVFDiffusion [69] and ActionMesh [47], and the text-to-3D animation method AnimateAnyMesh [58]. All baselines are evaluated using their official implementations and recommended settings. Additionally, we use Kling [52] as the video generator for the video-to-4D methods, and TripoSG [28] as the 3D generator.

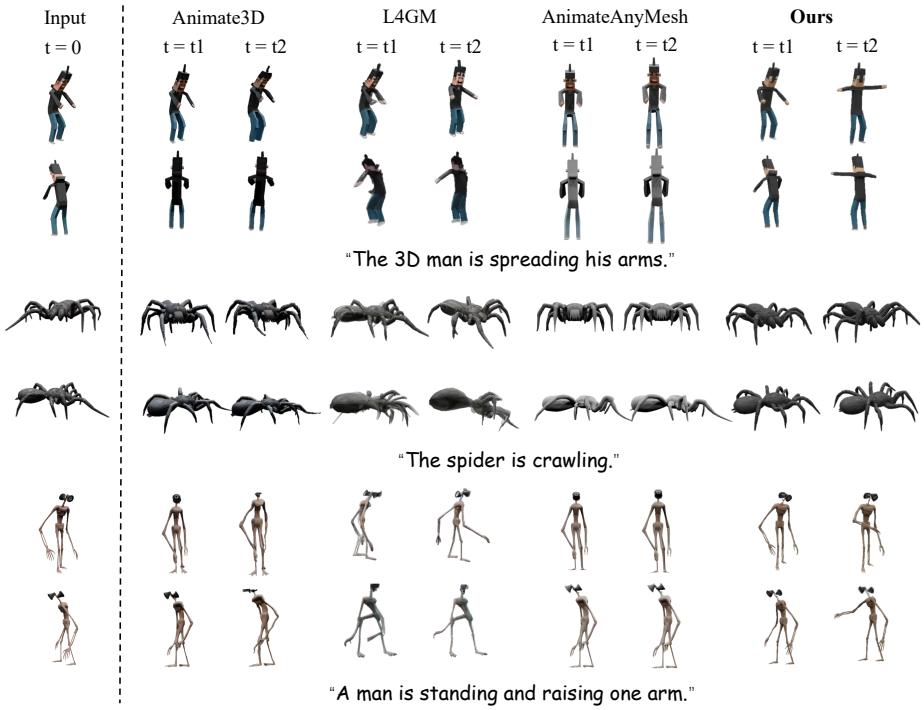


**Fig. 5:** Qualitative comparison with SOTA methods about 3D animation from static mesh.

**Qualitative Comparison.** Our proposed DiTex4D framework demonstrates superior performance in semantic alignment, motion naturalness, and geometric consistency. **4D generation from scratch,** as shown in Fig. 4, existing methods exhibit various defects in the task of text-to-4D generation. Animate3D [24] converts the initial mesh into a Gaussian splatting representation and optimizes motion trajectories via multi-view SDS and ARAP [48] regularization. However, this optimization pipeline suffers from error accumulation, often necessitating manual tuning to achieve high-fidelity results. While L4GM attempts to reconstruct 4D content directly from single-view videos but is inherently limited by the video generator's performance, particularly when processing isolated object sequences without natural backgrounds. Furthermore, the lack of explicit multi-view constraints makes the generated content prone to multi-view inconsistencies. And GVFDiffusion [69] relies heavily on input video type; consequently, its performance degrades significantly in scenarios involving mutual occlusion or large motion amplitudes.

For **3D animation from static mesh,** as shown in Fig. 5, both Animate3D [24] and L4GM [45] exhibit similar limitations in this setting. Although AnimateAnyMesh [58] achieves feed-forward high-quality 3D animation tasks,

**Table 1:** Quantitative comparisons with state-of-the-art methods of 4D Generation from scratch and 3D Animation from static mesh.

| Methods | 4D Generation from scratch | | | | 3D Animation from static mesh | | | |
|---|---|---|---|---|---|---|---|---|
| | CLIP↑ | M.Sm↑ | Aest.Q↑ | Dy.Deg↑ | I2V↑ | M.Sm↑ | Aest.Q↑ | Dy.Deg↑ |
| Dream-in-4D [75] | 24.17 | 0.973 | 0.416 | 0.187 | – | – | – | – |
| Animate3D [24] | 23.96 | 0.986 | 0.509 | 0.428 | 0.877 | 0.987 | 0.514 | 0.398 |
| L4GM [45] | 24.97 | 0.991 | 0.482 | **0.578** | 0.848 | 0.992 | 0.498 | **0.618** |
| GVFDiffusion [69] | 25.03 | 0.989 | 0.539 | 0.161 | 0.902 | 0.989 | 0.535 | 0.182 |
| ActionMesh [47] | 25.38 | 0.992 | 0.528 | 0.364 | 0.934 | 0.994 | 0.557 | 0.364 |
| AnimateAnyMesh [58] | 25.42 | 0.994 | 0.527 | 0.152 | **0.948** | **0.997** | 0.546 | 0.100 |
| **Ours** | **25.51** | **0.995** | **0.559** | 0.281 | 0.946 | **0.997** | **0.562** | 0.250 |

since it is not built upon high-quality 3D priors, leading to a high dependency on 4D datasets, which results in relatively small and unnoticeable generated motions under certain prompts. In contrast, our framework is directly built upon robust 3D models, inheriting strong text-to-3D priors while reducing data requirements. By enabling direct interaction between text conditions and 4D content, our framework achieves superior results with two key advantages: (1) the generated content is better aligned with text prompts; (2) the generated content exhibits relatively strong generalization capabilities even with small datasets.

**Quantitative Comparison.** For quantitative comparison, we curate an extra test set for each task comprising 32 randomly selected objects spanning diverse categories (human, animals, objects, etc.). Using identical text prompts as conditions, we computed the metrics across all baseline methods.

As shown in Tab. 1, both text-driven 4D generation pathways implemented by our framework demonstrate superior performance across all CLIP and VBench metrics (I2V, M.Sm, Aest.Q), highlighting their effectiveness in semantic alignment, and aesthetic quality. The user study results further validate the advantages of our method, showing significant improvements in the accuracy of text-4D alignment and motion amplitudes. These quantitative evaluations align closely with the visual results demonstrated in our qualitative observations.

### 4.3 Ablations.

We conduct controlled ablation experiments to isolate the contribution of each design component, removing one at a time from the full method while training on the same dataset. To reduce training costs, we use 8 frames instead of 16, and we train each model variant for same steps under identical settings in any tasks. Given that the 3D animation model is fine-tuned from the base 4D generation model, we evaluate the Mixed-4D RoPE and correlated noise injection on the base model, while assessing the multi-view image conditioning on the animation model. The results are presented in Tab. 2 and Fig. 7.

**Mixed-4D RoPE.** We found that mixed-4D RoPE is crucial for maintaining spatial coherence in long sequences. By simultaneously encoding both absolute

**Table 2:** Ablation study of different components.

| Method | CLIP↑ | M.Sm↑ | Aest.Q↑ | Dy.Deg↑ |
|---|---|---|---|---|
| w/o mixed-4D RoPE (1D RoPE + 3D APE) | 25.48 | 0.994 | 0.553 | 0.312 |
| w/o mixed-4D RoPE (only 4D RoPE) | 23.64 | 0.964 | 0.469 | **0.531** |
| w/o Correlated Noise Injection | 25.17 | 0.977 | 0.544 | 0.406 |
| w/o 4D interpolation | 25.52 | 0.986 | 0.556 | 0.281 |
| **Ours (base 4D generation model)** | **25.54** | **0.996** | **0.559** | 0.281 |
| **Method** | **I2V↑** | **M.Sm↑** | **Aest.Q↑** | **Dy.Deg↑** |
| w/o Multi-View Condition | 0.918 | 0.991 | 0.557 | 0.377 |
| **Ours (3D animation model)** | **0.948** | **0.997** | **0.563** | 0.250 |


**Fig. 6:** The example of 4D interpolation.

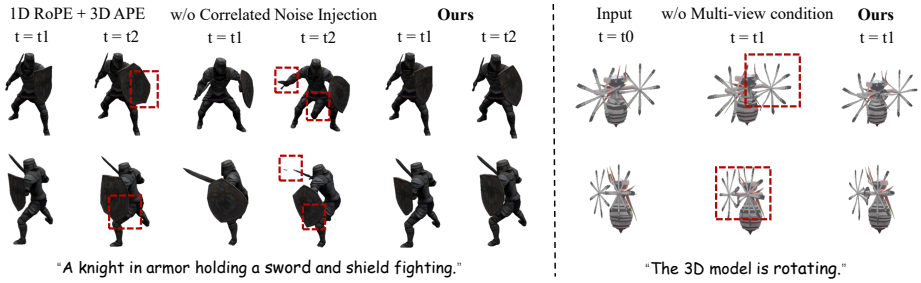


**Fig. 7:** Qualitative comparison of ablations.

and relative spatial positions, it effectively mitigates positional drift and structural inconsistencies during temporal generation, in Fig. 7. Conversely, the only 4D RoPE variant loses the pre-trained 3D absolute positional information and fails to achieve 4D generation within the same number of training steps.

**Correlated Noise Injection.** We found that applying temporally-correlated noise injection significantly reduces inter-frame flickering compared to independent noise sampling, resulting in more stable motion for 4D content in Fig. 7.

**Multi-view Image Condition.** Compared to text-only conditioning, multi-view conditions provide richer global context, enabling enhanced control over the temporal consistency between the generated content and the input mesh. As shown in Fig. 7. By utilizing multi-view images to define appearance ("what it looks like") and text to guide motion ("how it moves"), the model achieves superior generation quality compared to relying on text for all attributes.

**4D Interpolation.** We found that 4D interpolation can effectively improve Motion Smoothness in Fig. 6. This is mainly because our 4D generation model targets large motion amplitudes, interpolating intermediate states effectively alleviates the resulting jitter, producing smooth results.

## 5 Conclusion

We present **DiTex4D**, a direct feed-forward text-driven 4D generation framework addressing both 4D generation from scratch and 3D animation from static mesh. It solves the challenge of directly applying video-to-4D techniques to text-to-4D. By building upon 3D generation foundation models, we utilize mixed-4D RoPE and temporally-correlated noise injection to expand 3D attention into spatiotemporal full attention. Furthermore, we extend this architecture into a masked diffusion model, incorporating multi-view image conditioning to fully enable the text-guided animation of static 3D meshes. Extensive experiments demonstrate that the DiTex4D framework consistently outperforms existing baselines in text-to-4D generation tasks.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. U22B2055, No. 62273345, No. 62402494, No. 62572470), the Beijing Natural Science Foundation (No. L223003), and the Shandong Provincial Natural Science Foundation (No. ZR2025LGY009).

## References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
3. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023)
4. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 130–141 (2023)
5. Cao, W., Luo, C., Zhang, B., Nießner, M., Tang, J.: Motion2vecsets: 4d latent vector set diffusion for non-rigid shape reconstruction and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20496–20506 (2024)
6. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16123–16133 (2022)
7. Chang, J., Ye, C., Wu, Y., Chen, Y., Zhang, Y., Luo, Z., Li, C., Zhi, Y., Han, X.: Reconviagen: Towards accurate multi-view 3d object reconstruction via generation. arXiv preprint arXiv:2510.23306 (2025)

8. Chen, J., Zhang, B., Tang, X., Wonka, P.: V2m4: 4d mesh animation reconstruction from a single monocular video. arXiv preprint arXiv:2503.09631 (2025)
9. Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., et al.: Sam 3d: 3dfy anything in images. arXiv preprint arXiv:2511.16624 (2025)
10. Dai, S., Su, X., Wan, B., Hu, R., Xu, K.: Textmesh4d: High-quality text-to-4d mesh generation. arXiv preprint arXiv:2506.24121 (2025)
11. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. Advances in Neural Information Processing Systems **36**, 35799–35813 (2023)
12. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13142–13153 (2023)
13. Deng, Y., Yang, J., Xiang, J., Tong, X.: Gram: Generative radiance manifolds for 3d-aware image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10673–10683 (2022)
14. Gao, J., Shen, T., Wang, Z., Chen, W., Yin, K., Li, D., Litany, O., Gojcic, Z., Fidler, S.: Get3d: A generative model of high quality 3d textured shapes learned from images. Advances in neural information processing systems **35**, 31841–31854 (2022)
15. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22930–22941 (2023)
16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
17. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
18. He, X., Zou, Z.X., Chen, C.H., Guo, Y.C., Liang, D., Yuan, C., Ouyang, W., Cao, Y.P., Li, Y.: Sparseflex: High-resolution and arbitrary-topology 3d shape modeling. arXiv preprint arXiv:2503.21732 (2025)
19. Heo, B., Park, S., Han, D., Yun, S.: Rotary position embedding for vision transformer. In: European Conference on Computer Vision. pp. 289–305. Springer (2024). https://doi.org/10.1007/978-3-031-72684-2_17
20. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
21. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400 (2023)
22. Huang, Z., Feng, H., Sun, Y.T., Guo, Y.C., Cao, Y.P., Sheng, L.: Animax: Animating the inanimate in 3d with joint video-pose diffusion models. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–13 (2025)
23. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)

24. Jiang, Y., Yu, C., Cao, C., Wang, F., Hu, W., Gao, J.: Animate3d: Animating any 3d model with multi-view video diffusion. Advances in Neural Information Processing Systems **37**, 125879–125906 (2024)
25. 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)
26. Kara, O., Kurtkaya, B., Yesiltepe, H., Rehg, J.M., Yanardag, P.: Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6507–6516 (2024)
27. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
28. Li, Y., Zou, Z.X., Liu, Z., Wang, D., Liang, Y., Yu, Z., Liu, X., Guo, Y.C., Liang, D., Ouyang, W., et al.: Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. arXiv preprint arXiv:2502.06608 (2025)
29. Li, Z., Zhang, M., Wu, T., Tan, J., Wang, J., Lin, D.: Ss4d: Native 4d generative model via structured spacetime latents. ACM Transactions on Graphics (TOG) **44**(6), 1–12 (2025)
30. Li, Z., Chen, Y., Liu, P.: Dreammesh4d: Video-to-4d generation with sparse-controlled gaussian-mesh hybrid representation. Advances in Neural Information Processing Systems **37**, 21377–21400 (2024)
31. Liang, H., Yin, Y., Xu, D., Liang, H., Wang, Z., Plataniotis, K.N., Zhao, Y., Wei, Y.: Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models. arXiv preprint arXiv:2405.16645 (2024)
32. 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)
33. 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. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8576–8588 (2024)
34. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
35. Liu, M., Xu, C., Jin, H., Chen, L., Varma T, M., Xu, Z., Su, H.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. Advances in Neural Information Processing Systems **36**, 22226–22246 (2023)
36. 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)
37. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
38. Ma, Z., Chen, X., Yu, S., Bi, S., Zhang, K., Ziwen, C., Xu, S., Yang, J., Xu, Z., Sunkavalli, K., et al.: 4d-lrm: Large space-time reconstruction model from and to any view at any time. arXiv preprint arXiv:2506.18890 (2025)
39. Meng, Y., Wu, H., Zhang, Y., Xie, W.: Scenegen: Single-image 3d scene generation in one feedforward pass. arXiv preprint arXiv:2508.15769 (2025)
40. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)

41. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
42. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
44. 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)
45. Ren, J., Xie, C., Mirzaei, A., Kreis, K., Liu, Z., Torralba, A., Fidler, S., Kim, S.W., Ling, H., et al.: L4gm: Large 4d gaussian reconstruction model. Advances in Neural Information Processing Systems **37**, 56828–56858 (2024)
46. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
47. Sabathier, R., Novotny, D., Mitra, N.J., Monnier, T.: Actionmesh: Animated 3d mesh generation with temporal 3d diffusion. arXiv preprint arXiv:2601.16148 (2026)
48. Sorkine, O., Alexa, M.: As-rigid-as-possible surface modeling. In: Symposium on Geometry processing. vol. 4, pp. 109–116 (2007)
49. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
50. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In: European Conference on Computer Vision. pp. 1–18. Springer (2024). https://doi.org/10.1007/978-3-031-73235-5_1
51. 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)
52. Team, K., Chen, J., Ci, Y., Du, X., Feng, Z., Gai, K., Guo, S., Han, F., He, J., He, K., et al.: Kling-omni technical report. arXiv preprint arXiv:2512.16776 (2025)
53. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
54. Voleti, V., Yao, C.H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In: European Conference on Computer Vision. pp. 439–457. Springer (2024)
55. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
56. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in neural information processing systems **36**, 8406–8441 (2023)
57. 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. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20310–20320 (2024)

58. Wu, Z., Yu, C., Wang, F., Bai, X.: Animateanymesh: A feed-forward 4d foundation model for text-driven universal mesh animation. arXiv preprint arXiv:2506.09982 (2025)
59. Xiang, J., Chen, X., Xu, S., Wang, R., Lv, Z., Deng, Y., Zhu, H., Dong, Y., Zhao, H., Yuan, N.J., et al.: Native and compact structured latents for 3d generation. arXiv preprint arXiv:2512.14692 (2025)
60. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21469–21480 (2025)
61. Xie, Y., Yao, C.H., Voleti, V., Jiang, H., Jampani, V.: Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. arXiv preprint arXiv:2407.17470 (2024)
62. Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., Shan, Y.: Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191 (2024)
63. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
64. Yao, C.H., Xie, Y., Voleti, V., Jiang, H., Jampani, V.: Sv4d 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4d generation. arXiv preprint arXiv:2503.16396 (2025)
65. Yenphraphai, J., Mirzaei, A., Chen, J., Zou, J., Tulyakov, S., Yeh, R.A., Wonka, P., Wang, C.: Shapegen4d: Towards high quality 4d shape generation from videos. arXiv preprint arXiv:2510.06208 (2025)
66. Yin, M., Cao, Y., Peng, S., Han, K.: Splat4d: Diffusion-enhanced 4d gaussian splatting for temporally and spatially consistent content creation. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–10 (2025)
67. Zeng, Y., Jiang, Y., Zhu, S., Lu, Y., Lin, Y., Zhu, H., Hu, W., Cao, X., Yao, Y.: Stag4d: Spatial-temporal anchored generative 4d gaussians. In: European Conference on Computer Vision. pp. 163–179. Springer (2024)
68. Zhang, B., Tang, J., Niessner, M., Wonka, P.: 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. ACM Transactions On Graphics (TOG) **42**(4), 1–16 (2023)
69. Zhang, B., Xu, S., Wang, C., Yang, J., Zhao, F., Chen, D., Guo, B.: Gaussian variation field diffusion for high-fidelity video-to-4d synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12502–12513 (2025)
70. Zhang, H., Chen, X., Wang, Y., Liu, X., Wang, Y., Qiao, Y.: 4diffusion: Multi-view video diffusion model for 4d generation. Advances in Neural Information Processing Systems **37**, 15272–15295 (2024)
71. Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., Zhou, J.: I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv:2311.04145 (2023)
72. Zhao, Z., Lai, Z., Lin, Q., Zhao, Y., Liu, H., Yang, S., Feng, Y., Yang, M., Zhang, S., Yang, X., et al.: Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. arXiv preprint arXiv:2501.12202 (2025)
73. Zheng, C., Lin, Y., Liu, B., Xu, X., Nie, Y., He, S.: Recdreamer: Consistent text-to-3d generation via uniform score distillation. In: The Thirteenth International Conference on Learning Representations (2025)

74. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Gu, L., Zhang, Y., He, J., Zheng, W.S., et al.: Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755 (2025)
75. Zheng, Y., Li, X., Nagano, K., Liu, S., Hilliges, O., De Mello, S.: A unified approach for text-and image-guided 4d scene generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7300–7309 (2024)
76. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)