This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Gaussian Variation Field Diffusion for High-fidelity Video-to-4D Synthesis

Bowen Zhang[1*] Sicheng Xu[2] Chuxin Wang[1] Jiaolong Yang[2]
Feng Zhao[1†] Dong Chen[2†] Baining Guo[2]
[1]University of Science and Technology of China [2]Microsoft Research Asia

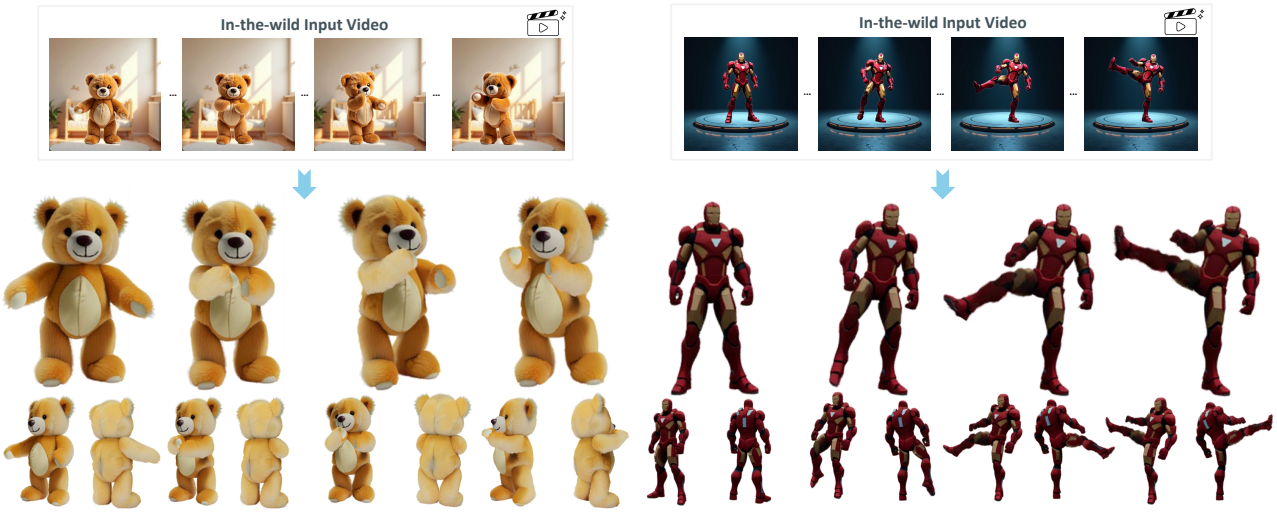


Figure 1. Our model is capable of creating high-fidelity 4D objects from in-the-wild video inputs. ***Best viewed with zoom-in.***

Abstract

In this paper, we present a novel framework for video-to-4D generation that creates high-quality dynamic 3D content from single video inputs. Direct 4D diffusion modeling is extremely challenging due to costly data construction and the high-dimensional nature of jointly representing 3D shape, appearance, and motion. We address these challenges by introducing a Direct 4DMesh-to-GS Variation Field VAE *that directly encodes canonical Gaussian Splats (GS) and their temporal variations from 3D animation data without per-instance fitting, and compresses high-dimensional animations into a compact latent space. Building upon this efficient representation, we train a* Gaussian Variation Field diffusion model *with temporal-aware Diffusion Transformer conditioned on input videos and canonical GS. Trained on carefully-curated animatable 3D objects from the Objaverse dataset, our model demonstrates superior generation quality compared to existing methods. It also exhibits remarkable generalization to in-the-wild video inputs despite being trained exclusively on synthetic data, paving the way for generating high-quality animated 3D content. Project page: GVFDiffusion.github.io.*

1. Introduction

Recent advances in generative models have demonstrated remarkable capabilities across various modalities, including image [21, 23, 30, 31, 44, 54, 88], video [2, 3, 19], and 3D content [24, 65, 67, 80, 91, 93]. While these achievements mark important milestones, they naturally lead to the next frontier: 4D generation, which aims to create dynamic 3D content. The challenge of generating such content—a fundamental aspect of representing our inherently four-dimensional world—remains largely unexplored. This gap is particularly significant given that real-world phenomena inherently combine spatial and temporal dynamics, from the subtle object movements to complex character articulations.

Despite diffusion models [15, 44, 58] having demonstrated strong modeling capabilities in both 2D and 3D domains, training a robust 4D diffusion model for dynamic 3D content generation presents two main technical challenges. First, obtaining a large-scale 4D dataset is time-consuming. A straightforward approach involves fitting individual dynamic Gaussian Splatting (4DGS) representations [75] for each 3D animation sequence, but this solution typically requires tens of minutes per instance, making it computationally expensive and less scalable as the number of instances increases. Second, the higher-dimensional nature of the

*Intern at Microsoft Research Asia. †Corresponding authors.

problem necessitates a large number of parameters (usually exceeding 100K tokens) to represent 3D shape, appearance, and motion simultaneously, making direct modeling with diffusion approaches extremely challenging. These limitations have significantly hindered the development of efficient and high-quality 4D generative models.

Motivated by the effectiveness of diffusion models applied to compact latent spaces in recent 2D and 3D generation works [3, 57, 58, 80, 93], we present a novel framework for 4D generative modeling that comprises a *Direct 4DMesh-to-GS Variation Field VAE* and a *Gaussian Variation Field diffusion model*. Our VAE framework encodes the canonical 3D Gaussian Splatting (3DGS) of objects and compresses each Gaussian's attribute variations (*i.e.*, *Gaussian Variation Fields*) into a compact latent space from 4D mesh data, thereby bypassing costly per-instance reconstructions. Inspired by previous works [5, 89], we employ a perceiver-style transformer network [26, 27, 71] with displacements of mesh points to effectively encode motion information. To bridge the gap between Gaussian Splatting representation and mesh-based ground truth motion, we introduce a *mesh-guided loss* that aligns the motion of Gaussian points with the corresponding mesh vertices. Our VAE is trained end-to-end with this mesh-guided loss and an image-level loss, enabling faithful compression of complex Gaussian Variation Fields. This approach reduces high-dimensional motion sequences to a compact 512-dimensional latent space, thus facilitating efficient diffusion modeling for 4D content generation.

Following the construction of our VAE, the 4D generative modeling naturally decomposes into canonical 3DGS generation and Gaussian Variation Field modeling. We leverage state-of-the-art 3D generative models [80] for the canonical component while focusing on modeling the Gaussian Variation Fields. To achieve this, we train a diffusion model to learn the latent space distribution of variation fields conditioned on the input video and canonical 3DGS, enabling controlled 4D content generation. Leveraging the compact nature of our latent space, we employ the Diffusion Transformer (DiT) architecture [51], augmented with temporal self-attention layers to capture smooth temporal dynamics across animations. The video frame features [47] and the canonical 3DGS are taken as conditions for the diffusion model via cross-attention layers. Additionally, we incorporate positional priors into the diffusion model, enhancing its awareness of correspondences between canonical GS and their variation fields during the denoising process, thereby improving generation quality.

We train our model on a carefully curated diverse collection of animatable 3D objects from the Objaverse [13] and Objaverse-XL [12]. Extensive evaluations demonstrate the superior video-to-4D generation quality of our method compared to existing approaches. Despite being trained on synthetic data, our model exhibits remarkable generalization capabilities when applied to in-the-wild video inputs, effectively creating impressive animations from in-the-wild animation sequences. We believe that our approach represents a notable step toward narrowing the gap between static 3D generation and 4D content creation, paving the way for generating high-quality 4D content.

2. Related Work

3D generation. Early GAN-based 3D generation approaches [7, 14, 17, 63, 76, 79, 96, 98] laid the foundation for 3D content synthesis, while diffusion-based methods [6, 9, 22, 29, 43, 61, 68, 73, 85, 90, 91] advanced generation quality. Recent approaches have focused on latent space generation, either separating geometric modeling and appearance synthesis [37, 57, 70, 77, 89, 93, 95, 97] or jointly modeling both [10, 20, 29, 35, 45, 80, 82, 83]. Alternative methods [8, 40, 52, 64, 66, 74] leverage pretrained 2D models [58] through optimization techniques. Recent works [80, 93] have achieved high-quality 3D asset generation with detailed geometry and appearance, establishing a foundation for 4D content creation.

4D reconstruction. Early 4D reconstruction methods [4, 16, 49, 50, 53] extended neural volumetric techniques for dynamic scenes, while recent Gaussian Splatting-based approaches [11, 25, 32, 36, 41, 75] offer improved efficiency. Typical 4D reconstruction methods often require significant optimization time per instance (*e.g.*, 6 minutes for 4DGaussians [75] and over 30 minutes for K-planes [16]), making them impractical to use as a preliminary step in fitting 4D representations for generation. In this paper, we explore an efficient approach to directly encode 4D mesh data for generative modeling in a single pass.

Video-to-4D generation. Early attempts [1, 28, 55, 62] at video-to-4D generation predominantly relied on optimization-based approaches, utilizing pre-trained generative priors [40, 59, 60, 72] as guidance. These methods typically employ score distillation [52] techniques to optimize either neural volumetric representations [42] or 3D Gaussian Splatting [32]. These methods suffer from lengthy optimization times and SDS-related issues [39, 74] such as spatial-temporal inconsistency or poor input alignments. While some works [86, 87] use pseudo-labels for better consistency, recent approaches [38, 48, 56, 81, 84, 92] directly reconstruct 4D content from multiview images or videos. Notably, the introduction of large-scale 4D reconstruction models [56] has significantly reduced the generation time from hours to seconds. However, most of these approaches often struggle to maintain consistent quality across temporal sequences due to inherent multiview inconsistency in 2D generation results.

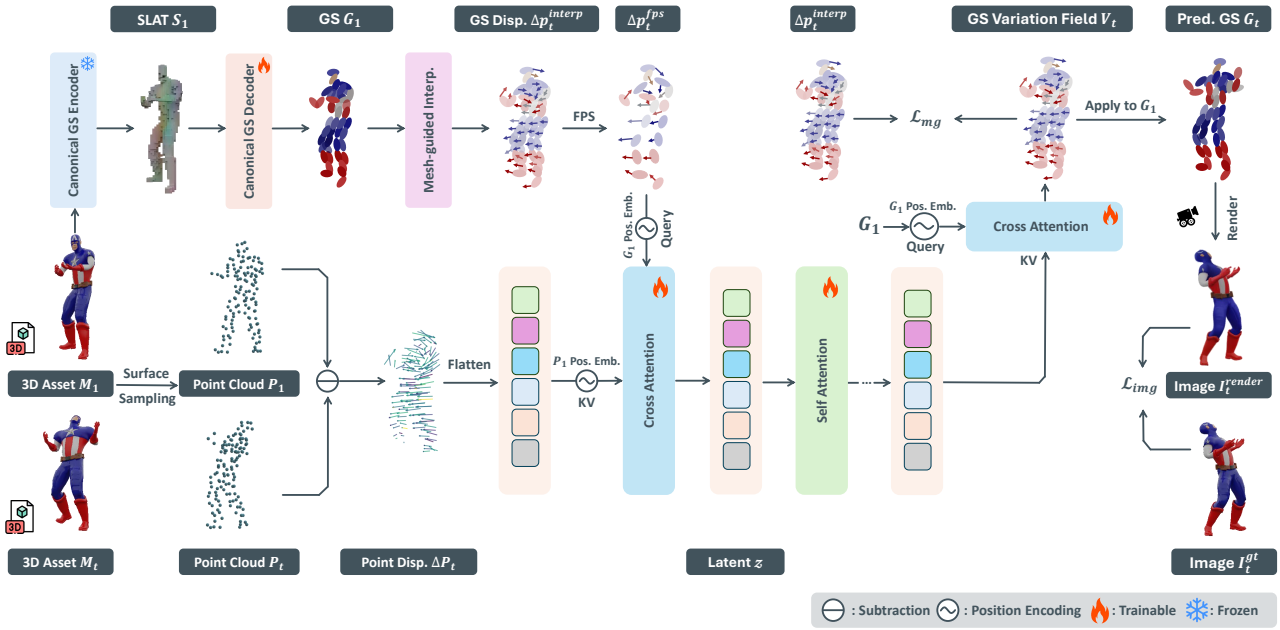


Figure 2. **Framework of 4DMesh-to-GS Variation Field VAE.** Our VAE directly encodes 3D animation data into Gaussian Variation Fields within a compact latent space, optimized through image-level reconstruction loss and the proposed ***mesh-guided loss***.

3. Method

Given an input video sequence $\mathcal{I} = \{I_t\}_{t=1}^T$ of an object, our goal is to generate a sequence of 3DGS models $\mathcal{G} = \{G_t\}_{t=1}^T$ that captures both the shape, appearance, and motion of the object. We decompose this task into canonical GS G_1 creation (using the first frame as canonical) and Gaussian Variation Fields $\mathcal{V} = \{\Delta G_t\}_{t=1}^T$ generation, where $\mathcal{V}$ describes each Gaussian's attribute variations relative to G_1 over time. Our framework comprises two main components: (1) a *direct 4DMesh-to-GS Variation Field VAE* that efficiently encodes 3D animation sequences into a compact latent space, and (2) a *Gaussian variation field diffusion model* that learns the latent distribution of variation fields conditioned on the input video and canonical GS. The following sections detail each component.

3.1. Direct 4DMesh-to-GS Variation Field VAE

Extending 3DGS to generative modeling of dynamic content presents significant challenges. Fitting individual dynamic 3DGS representations for each animation instance is computationally expensive and scales poorly. Additionally, directly modeling temporal deformation of GS sequences with diffusion models is challenging due to the high dimensionality of both Gaussian quantities (*e.g.*, typically over 100K in [32]) and the time dimension. Therefore, we propose an efficient autoencoding framework that directly encodes 3D animation data into Gaussian Variation Fields with a compact latent space, facilitating subsequent diffusion modeling.

Gaussian Variation Field encoding. Given a sequence of mesh animations $\mathcal{M} = \{M_t\}_{t=1}^T$, we first convert them to point clouds $\mathcal{P} = \{P_t | P_t \in \mathbb{R}^{N\times 3}\}_{t=1}^T$ through uniform surface sampling, where each point cloud contains N points. The displacement fields $\{\Delta P_t | \Delta P_t \in \mathbb{R}^{N\times 3}\}_{t=1}^T$ are computed as temporal differences of corresponding points between frames:

$$\Delta P_t = P_t - P_1, \tag{1}$$

where P_1 is the canonical frame's point cloud. We then leverage a pretrained mesh-to-GS autoencoder $\mathcal{E}_{GS}$ and $\mathcal{D}_{GS}$ in [80] to obtain the canonical GS representation from canonical mesh M_1:

$$\begin{aligned} S_1 &= \mathcal{E}_{GS}(M_1), \\ G_1 &= \mathcal{D}_{GS}(S_1), \end{aligned} \tag{2}$$

where $G_1 \in \mathbb{R}^{N_G \times 14}$ denotes the Gaussian parameters including positions $\boldsymbol{p}_1$, scales $\boldsymbol{s}_1$, rotation $\boldsymbol{q}_1$, colors $\boldsymbol{c}_1$, and opacity α_1, with N_G being the total number of canonical Gaussians. S_1 is the structured latent (SLAT) representation for canonical GS (more details are included in the supplementary). We finetune $\mathcal{D}_{GS}$ to ensure coherent canonical GS reconstruction with their variation fields, while keeping $\mathcal{E}_{GS}$ frozen to leverage pretrained canonical GS diffusion models.

Inspired by 3DShape2VecSet [89], we employ a cross-attention layer to aggregate motion information from 3D animation sequences into a fixed-length latent representation. While directly using G_1 as query vectors is a straightforward approach, we find it leads to poor motion awareness.

To enhance the network's sensitivity to mesh deformation, we introduce a ***mesh-guided interpolation*** mechanism that generates motion-aware query vectors based on the spatial correspondence between G_1 and P_1.

Specifically, for each canonical Gaussian position $\boldsymbol{p}_1^i$, we identify its K nearest neighbors in the canonical point cloud P_1 and compute their distances $\boldsymbol{d}_{i,k}$. To handle varying point densities across the mesh-sampled point cloud, we introduce an adaptive radius r_i that adjusts the influence region based on the local point distribution. The interpolation weight $\boldsymbol{w}_{i,k}$ and adaptive radius r_i are formulated as:

$$\boldsymbol{w}_{i,k} = \exp(-\frac{\beta \boldsymbol{d}_{i,k}}{r_i^2}), \quad r_i = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{d}_{i,k}}, \tag{3}$$

where β is a hyperparameter controlling the decay rate of interpolation weights with distance, with larger values producing more localized influence regions. We set $\beta = 7.0$ in this paper.

We then interpolate the displacement fields ΔP_t for the i-th Gaussian at time t:

$$\Delta \boldsymbol{p}_{t,i}^{interp} = \sum_{k=1}^{K} \frac{\boldsymbol{w}_{i,k}}{\sum_k \boldsymbol{w}_{i,k}} \Delta P_{t,n(i,k)} \tag{4}$$

where $n(i,k)$ denotes the k-th nearest neighbor index. We perform farthest point sampling to $\Delta \boldsymbol{p}_t^{interp}$ based on their canonical positions to formulate our motion-aware query $\Delta \boldsymbol{p}_t^{fps} \in \mathbb{R}^{L\times 3}$ with reduced sequence length. The point cloud displacement fields ΔP_t serve as keys and values in the cross attention encoder. To preserve spatial relationships, we incorporate positional embedding $PE(\cdot)$ based on the canonical positions:

$$\begin{aligned} Q_e &= f_{disp}(\Delta \boldsymbol{p}_t^{fps}) + PE(G_1), \\ K_e = V_e &= f_{disp}(\Delta P_t) + PE(P_1), \\ \boldsymbol{z} &= \text{CrossAttn}(Q_e, K_e, V_e), \end{aligned} \tag{5}$$

where f_{disp} is the displacement embedding layer. This process yields a latent representation $\boldsymbol{z} \in \mathbb{R}^{T\times L\times C}$, where T is the number of temporal frames, L is the latent size, and C is the feature dimension. Notably, our encoding procedure compresses the sequence length from $N = 8192$ to $L = 512$, significantly reducing the subsequent diffusion modeling space.

Gaussian Variation Field decoding. The decoding procedure first transforms the latent representation through n layers of self-attention blocks to enable thorough motion information exchange. The decoder then maps this processed latent to a Gaussian Variation Field $\boldsymbol{\mathcal{V}}$, defined by the variations of Gaussian attributes $\Delta G_t = \{\Delta \boldsymbol{p}_t, \Delta \boldsymbol{s}_t, \Delta \boldsymbol{q}_t, \Delta \boldsymbol{c}_t, \Delta \alpha_t\}_{t=1}^T$. To ensure the decoder is aware of all canonical Gaussian attributes, we use all parameters of G_1 to query the latent output through a cross attention layer:

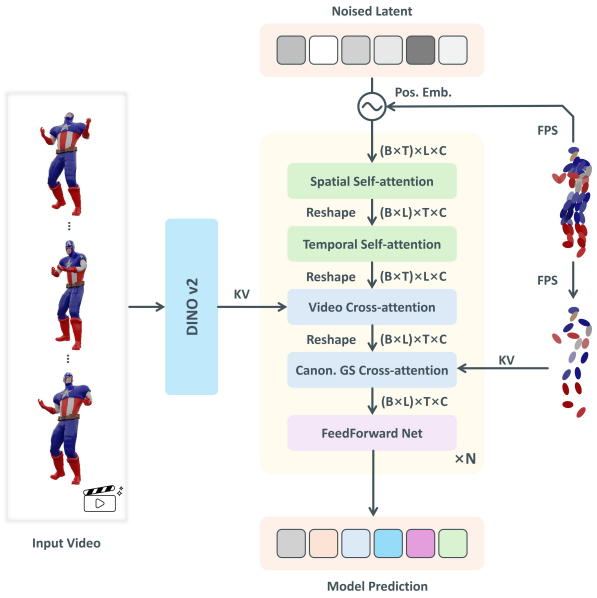


Figure 3. **Architecture of Gaussian Variation Field diffusion model.** Our model is built upon diffusion transformer, which takes noised latent as input and gradually denoises it conditioned on the video sequence and canonical GS.

$$\begin{aligned} Q_d &= f_{gs}(G_1) + PE(G_1), \quad K_d = V_d = \boldsymbol{z}_n, \\ \Delta G_t &= \text{CrossAttn}(Q_d, K_d, V_d), \end{aligned} \tag{6}$$

where f_{gs} is the embedding layer for the canonical Gaussians and $\boldsymbol{z}_n$ is the final self attention layer output. The final 3DGS sequence is obtained by:

$$\boldsymbol{\mathcal{G}} = \{G_t\}_{t=1}^T = \{G_1 + \Delta G_t\}_{t=1}^T. \tag{7}$$

Training objective. Our training objective consists of three main components. First, we employ image-level reconstruction loss between the rendered images I_t^{render} from final predicted Gaussians and ground-truth images I_t^{gt}:

$$\mathcal{L}_{img} = \mathcal{L}_1 + \lambda_{lpips}\mathcal{L}_{lpips} + \lambda_{ssim}\mathcal{L}_{ssim}, \tag{8}$$

where $\lambda_{lpips}, \lambda_{ssim}$ are loss weights for perception loss [94] and SSIM loss, respectively. To ensure faithful motion reconsturction, we introduce a ***mesh-guided loss*** that aligns the predicted Gaussian displacements with pseudo ground-truth Δp_t^{interp} obtained through ***mesh-guided interpolation***:

$$\mathcal{L}_{mg} = \sum_{t=1}^{T} \|\Delta \mathbf{p}_t - \Delta \boldsymbol{p}_t^{interp}\|_2^2, \tag{9}$$

which we find is crucial for motion reconstruction quality. Finally, to facilitate subsequent diffusion training, we also

Table 1. **Quantitative comparison of video-to-4D generation results.** Our method demonstrates consistent performance improvements across all metrics while maintaining efficient generation speed. The generation time is measured on a single A100 GPU.

Method	PSNR↑	LPIPS↓	SSIM↑	CLIP↑	FVD↓	Time↓
Consistent4D [28]	16.20	0.146	0.880	0.910	935.19	∼1.5 hr
SC4D [78]	15.93	0.164	0.872	0.870	833.15	∼20 min
STAG4D [87]	16.85	0.144	0.887	0.893	1008.40	∼1 hr
DreamGaussian4D [55]	15.24	0.162	0.868	0.904	799.56	∼15 min
L4GM [56]	17.03	0.128	0.891	0.930	529.10	**3.5 s**
Ours	**18.47**	**0.114**	**0.901**	**0.935**	**476.83**	4.5s

regularize the latent distribution with a KL divergence loss $\mathcal{L}_{kl}$. The total loss is: $\mathcal{L}_{total} = \mathcal{L}_{img} + \lambda_{mg}\mathcal{L}_{mg} + \lambda_{kl}\mathcal{L}_{kl}$, where $\lambda_{mg}, \lambda_{kl}$ are respective loss weights.

3.2. Gaussian Variation Field Diffusion

The diffusion process can be formalized as the inversion of a discrete-time Markov forward process. Let $\boldsymbol{z}^0 \in \mathbb{R}^{T\times L\times C}$ denote our initial latent of Gaussian Variation Field from the distribution $p(\boldsymbol{z})$. During the forward phase, we progressively corrupt this latent sequence by adding Gaussian noise over diffusion steps $s \in [0, S]$, following $\boldsymbol{z}^s := \alpha_s \boldsymbol{z}^0 + \sigma_s \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and α_s, σ_s define the noise schedule. After sufficient diffusion steps, $\boldsymbol{z}^S$ approaches pure Gaussian noise. Generation is achieved by reversing this process, starting from random Gaussian noise $\boldsymbol{z}^S \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and progressively denoising it to recover $\boldsymbol{z}^0$.

The compact latent space enables us to build our diffusion model upon the powerful Diffusion Transformer (DiT) architecture [51]. As illustrated in Figure 3, the model takes noise-corrupted latent as input, and processes them through a series of transformer blocks for denoising. Each transformer block incorporates diffusion timestep information through adaptive layer normalization (adaLN) and a gating mechanism. Beyond the standard spatial self attention layers, we introduce dedicated temporal self-attention layers to ensure coherent motion generation across the sequence.

To condition the generation process, we inject two types of features through cross-attention layers: (1) visual features $\boldsymbol{\mathcal{C}}^v = \{\boldsymbol{C}_t^v\}_{t=1}^T$ extracted from input video frames using DINOv2 [47], and (2) geometric features $\boldsymbol{\mathcal{C}}^{GS} = G_1^{fps}$ fartherest sampled from the static GS. We further incorporate positional embeddings based on canonical GS positions $\boldsymbol{p}_1^{fps}$ in our diffusion transformer, which strengthens the model's awareness of correspondences between canonical GS and their variation fields during the denoising process, thereby effectively improving the generation quality.

We parameterize our diffusion model $\hat{\boldsymbol{v}}_\theta$ to predict the velocity $v^s := \alpha_s \boldsymbol{\epsilon} - \sigma_s \boldsymbol{z}^0$ at each diffusion step s. The diffusion model is trained using:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{s,\boldsymbol{z}^0,\boldsymbol{\epsilon}}\left[\left\|\hat{\boldsymbol{v}}_\theta\left(\alpha_s \boldsymbol{z}^0 + \sigma_s \boldsymbol{\epsilon}, s, \boldsymbol{\mathcal{C}}\right) - \boldsymbol{v}^s\right\|_2^2\right], \quad (10)$$

where $\boldsymbol{\mathcal{C}} = \{\boldsymbol{\mathcal{C}}^v, \boldsymbol{\mathcal{C}}^{GS}\}$ represents the conditional features of both $\boldsymbol{\mathcal{C}}^v$ and $\boldsymbol{\mathcal{C}}^{GS}$.

3.3. Inference Pipeline

During inference, our framework operates in a sequential pipeline. First, we obtain the canonical GS G_1 for the first frame using a pretrained 3D diffusion model [80]. Given an input video sequence $\{I_t\}_{t=1}^T$, we extract visual features and combine them with the farthest sampled canonical Gaussians as conditioning signals for our diffusion model. The diffusion model generates latent codes $\boldsymbol{z}$, which are subsequently decoded to obtain the Gaussian Variation Field $\boldsymbol{\mathcal{V}}$. The final animated Gaussian representation G_t for each frame is obtained by applying these variations to the canonical Gaussians, effectively creating high-fidelity temporally coherent 4D animations.

4. Experiments

4.1. Dataset and Metrics

We conduct our experiments on Objaverse-V1 and Objaverse-XL [13], following previous work in 4D content generation. After filtering for objects with high-quality animations, we utilize 34K objects for training. To evaluate the video-to-4D generation quality, we construct a comprehensive test set of 100 objects by combining 7 instances from the widely-used Consistent4D [28] testset with 93 additional test instances from Objaverse-XL, ensuring a thorough evaluation with previous works. We render 4 novel views of each timestep for each instance. We assess the generation quality using multiple metrics: PSNR, LPIPS [94], and SSIM for frame-wise quality, and FVD [69] for temporal consistency of the generated sequences. All evaluations are performed on renderings at 512×512 resolution.

4.2. Implementation Details

For our VAE implementation, the canonical Mesh-to-GS autoencoder is builds upon TRELLIS [80], with training conducted in two stages: finetuning the sparse GS decoder $\mathcal{D}_{GS}$ on canonical 3D only for 150K iterations, followed by joint training with other modules for 200K iterations on 4D animation data. The VAE architecture employs point cloud

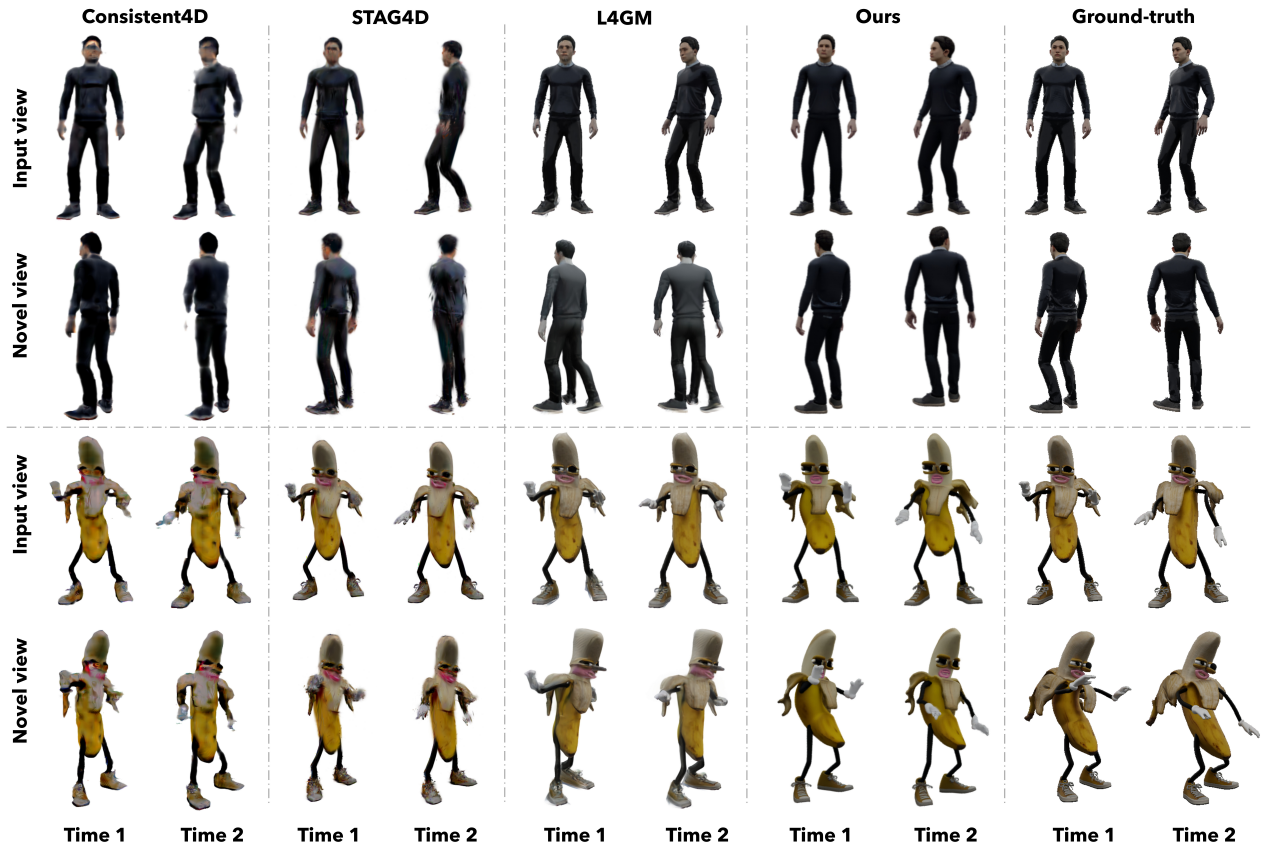


Figure 4. **Qualitative comparison with previous state-of-the-art methods.** Our model directly learns the distribution of Gaussian Variation Fields, enabling high-fidelity 4D generation with coherent temporal dynamics.

size $N = 8192$, latent size $L = 512$, and feature dimension $C = 16$. We optimize the VAE using AdamW with a learning rate $5e-6$ and $5e-5$ for $\mathcal{D}_{GS}$ and other modules respectively, using batch size 32. The diffusion model is trained on 24-frame sequences using AdamW optimizer with identical learning rate and batch size over 1300K iterations. We apply the cosine noise schedule [44] with 1000 timesteps for training the diffusion model. We set $T = 24$ for training, and $T = 32$ during inference to compare with prior works. For more implementation details, please refer to the supplementary materials.

4.3. Main Results

Quantitative comparisons. We compare the video-to-4D generation results of our model with previous state-of-the-art methods including both optimization-based approaches [28, 55, 78, 87] and feedforward approach [56]. As shown in Table 1, our method consistently outperforms existing approaches across all quality metrics, demonstrating both superior reconstruction fidelity and better temporal coherence. Unlike some prior works [28, 55, 78, 87] require minutes to hours of optimization, our approach is also more efficient, taking 4.5 seconds to generate a 4D animation sequence (3.0 seconds for canonical GS creation and 1.5 seconds for Gaussian Variation Field diffusion), only slightly slower than feedforward reconstruction method L4GM [56]. These quantitative results collectively validate both the effectiveness and efficiency of our proposed method.

Qualitative comparisons. We also provide qualitative comparisons with previous state-of-the-art methods in Figure 4. The SDS-based approaches [28, 87] turn to generate results with blurry textures and poor geometry. The feedforward method L4GM leverages multiview images generated from 2D generative prior [60] to reconstruct the 4DGS sequences. However, the results of L4GM suffer from 3D inconsistency of the generated multiview images. In contrast, our model directly generates the canonical GS and the Gaussian Variation Fields, capable of creating high-fidelity 3D consistent animations with coherent temporal dynamics.

More visualization of generated results. Figure 5 presents additional generation results from our method, including examples conditioned on both in-the-wild videos (left two cases) and test set videos (right two cases). Our model demonstrates high-quality generation capability with faithful



Figure 5. More generation results of our model including in-the-wild videos (left) and videos from test set (right).

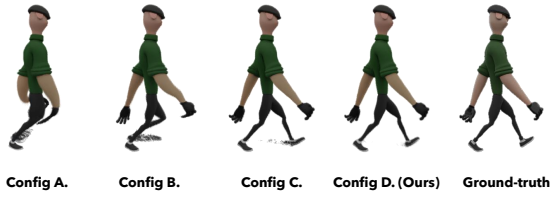


Figure 6. Visual ablation of VAE.

motion reproduction. Despite being trained on synthetic data, the model exhibits strong generalization capability by effectively capturing motion patterns from in-the-wild video inputs. Furthermore, the model successfully handles challenging multi-object scenarios, highlighting the robustness of our approach.

4.4. Ablation Study

Ablation of our VAE. In Table 2 and Figure 6, we analyze key components of our VAE. Our baseline (Config. A) starts with using positions of farthest sampled canonical GS $\boldsymbol{p}_t^{fps}$ as query for the encoder's cross-attention layer, with variation attributes limited to positions $\Delta\boldsymbol{p}_t$, scaling $\Delta\boldsymbol{s}_t$, and rotation $\Delta\boldsymbol{q}_t$ following previous 4DGS works [75]. Since we do not have ground-truth GS motion for explicit supervision, the VAE initially struggles with motion learning. After equipped with our ***mesh-guided loss***, the motion reconstruction capability is effectively improved through pseudo displacement supervision (Config. B). We then replace the encoding query with motion-aware $\Delta\boldsymbol{p}_t^{fps}$ using ***mesh-guided interpolation***, which successfully handles most of the motion sequences (Config. C). Finally, to give the model more flexibility to handle complex motion sequence, we incorporate color $\Delta\boldsymbol{c}_t$ and opacity $\Delta\alpha$ of Gaussian attributes to the variation fields, which further enhance the

Table 2. Ablation study of key factors in our VAE.

Config.	Encoder Query Type	Mesh-guided Loss	Variation Attrs.	PSNR↑	LPIPS↓	SSIM↑
A.	$\boldsymbol{p}_t^{fps}$	✗	$\Delta\boldsymbol{p}_t, \Delta\boldsymbol{s}_t, \Delta\boldsymbol{q}_t$	23.25	0.0678	0.936
B.	$\boldsymbol{p}_t^{fps}$	✓	$\Delta\boldsymbol{p}_t, \Delta\boldsymbol{s}_t, \Delta\boldsymbol{q}_t$	26.17	0.0544	0.950
C.	$\Delta\boldsymbol{p}_t^{fps}$	✓	$\Delta\boldsymbol{p}_t, \Delta\boldsymbol{s}_t, \Delta\boldsymbol{q}_t$	28.58	0.0478	0.958
D. (Ours)	$\Delta\boldsymbol{p}_t^{fps}$	✓	$\Delta\boldsymbol{p}_t, \Delta\boldsymbol{s}_t, \Delta\boldsymbol{q}_t, \Delta\boldsymbol{c}_t, \Delta\alpha_t$	**29.28**	**0.0439**	**0.964**

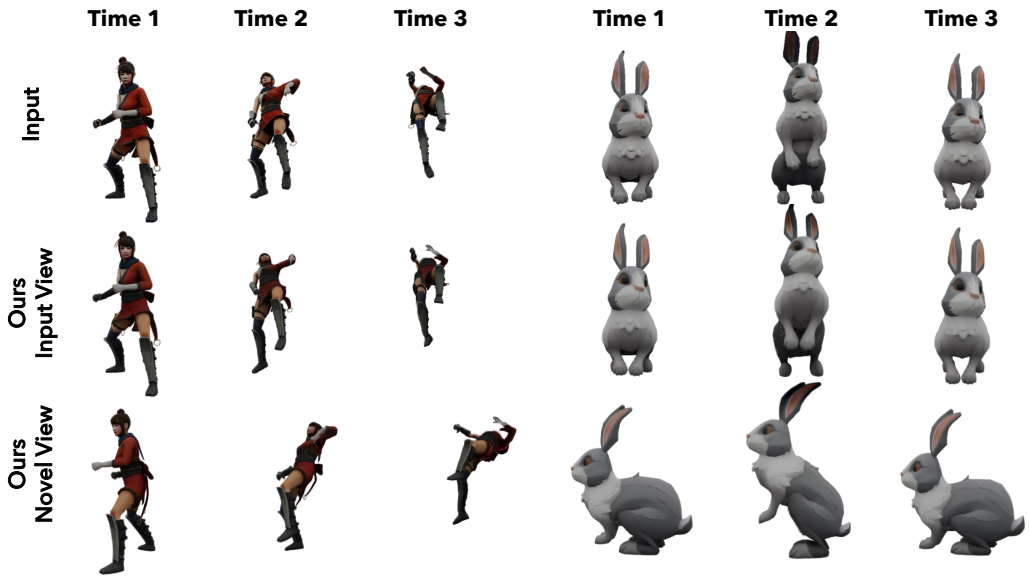


Figure 7. Our model is also capable of creating animations for existing 3D assets with conditional videos.

Table 3. Ablation study of our diffusion model.

Method	PSNR↑	LPIPS↓	SSIM↑	CLIP↑	FVD↓
Ours w/o Pos. Emb.	17.86	0.121	0.897	0.931	547.20
Ours	**18.47**	**0.114**	**0.901**	**0.935**	**476.83**

reconstruction capability of VAE.

Ablation of our diffusion model. We examine the importance of positional embeddings in our diffusion model training in Table 3. By incorporating positional prior based on canonical GS positions $\boldsymbol{p}_1^{fps}$, the diffusion transformer better captures the correspondence between spatial positions and their variations. Removing these positional embeddings leads to a significant performance drop, demonstrating their crucial role in achieving high-quality results.

4.5. Application

Despite being trained on single video inputs, our model is effective at animating existing 3D models according to the motions depicted in conditioned videos. More details are included in the supplementary materials. As demonstrated in Figure 7, this approach produces high-quality animations that faithfully reproduce the target motions. Therefore, for real-world applications, the users can first generate 2D animations from the rendered images of their 3D models using off-the-shelf video diffusion models [18, 33, 34, 46], then employ our model to create corresponding 4D animations.

5. Conclusion

In this paper, we introduce a novel framework to address the challenging task of 4D generative modeling. To efficiently construct the large-scale training dataset and reduce the modeling difficulty for diffusion, we first introduce a *Direct 4DMesh-to-GS Variation Field VAE*, which is able to efficiently compress complex motion information into a compact latent space without requiring costly per-instance fitting. Then, a *Gaussian Variation Field diffusion* model that generates high-quality dynamic variation fields conditioned on input videos and canonical 3DGS. By decomposing 4D generation into canonical 3DGS generation and Gaussian Variation Field modeling, our method significantly reduces computational complexity while maintaining high fidelity. Quantitative and qualitative evaluations demonstrate that our approach consistently outperforms existing methods. Furthermore, our model exhibits remarkable generalization capability with in-the-wild video inputs, advancing the state of high-quality animated 3D content generation.

Acknowledgments. We extend our gratitude to all the reviewers for their constructive feedback. We also appreciate Jiaqi Lou for the assistance with chart refinement and the production of the supplementary video.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7996–8006, 2024. 2

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 1, 2

[4] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 2

[5] Wei Cao, Chang Luo, Biao Zhang, Matthias Nießner, and Jiapeng Tang. 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*, pages 20496–20506, 2024. 2

[6] Ziang Cao, Fangzhou Hong, Tong Wu, Liang Pan, and Ziwei Liu. Large-vocabulary 3d diffusion model with transformer. *arXiv preprint arXiv:2309.07920*, 2023. 2

[7] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 2

[8] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *ArXiv preprint*, abs/2303.13873, 2023. 2

[9] Zhaoxi Chen, Fangzhou Hong, Haiyi Mei, Guangcong Wang, Lei Yang, and Ziwei Liu. Primdiffusion: Volumetric primitives diffusion for 3d human generation. *Advances in Neural Information Processing Systems*, 36:13664–13677, 2023. 2

[10] Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, et al. 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. *arXiv preprint arXiv:2409.12957*, 2024. 2

[11] R James Cotton and Colleen Peyton. Dynamic gaussian splatting from markerless motion capture reconstruct infants movements. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 60–68, 2024. 2

[12] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023. 2

[13] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 5

[14] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *IEEE/CVF International Conference on Computer Vision*, 2022. 2

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1

[16] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 2

[17] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *arXiv preprint arXiv:2209.11163*, 2022. 2

[18] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Factorizing text-to-video generation by explicit image conditioning. In *European Conference on Computer Vision*, pages 205–224. Springer, 2024. 8

[19] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1

[20] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023. 2

[21] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. *arXiv preprint arXiv:2303.09556*, 2023. 1

[22] Xianglong He, Junyi Chen, Sida Peng, Di Huang, Yangguang Li, Xiaoshui Huang, Chun Yuan, Wanli Ouyang, and Tong He. Gvgen: Text-to-3d generation with volumetric representation. *arXiv preprint arXiv:2403.12957*, 2024. 2

[23] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[24] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 1

[25] Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4220–4230, 2024. 2

[26] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021. 2

[27] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 2

[28] Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848*, 2023. 2, 5, 6

[29] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 2

[30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1

[31] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[32] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2, 3

[33] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. 8

[34] Kuaishou. Kling. *https://klingai.kuaishou.com*, 2024. 8

[35] Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *ECCV*, 2024. 2

[36] Mengtian Li, Shengxiang Yao, Zhifeng Xie, Keyu Chen, and Yu-Gang Jiang. Gaussianbody: Clothed human reconstruction via 3d gaussian splatting. *arXiv preprint arXiv:2401.09720*, 2024. 2

[37] Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024. 2

[38] Hanwen Liang, Yuyang Yin, Dejia Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models. *arXiv preprint arXiv:2405.16645*, 2024. 2

[39] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6517–6526, 2024. 2

[40] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2

[41] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 2

[42] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[43] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4328–4338, 2023. 2

[44] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1, 6

[45] Evangelos Ntavelis, Aliaksandr Siarohin, Kyle Olszewski, Chaoyang Wang, Luc Van Gool, and Sergey Tulyakov. Autodecoding latent 3d diffusion models. *arXiv preprint arXiv:2307.05445*, 2023. 2

[46] OpenAI. Video generation models as world simulators. *https://openai.com/index/video-generation-models-as-world-simulators*, 2024. 8

[47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 5

[48] Zijie Pan, Zeyu Yang, Xiatian Zhu, and Li Zhang. Efficient4d: Fast dynamic 3d object generation from a single-view video. *arXiv preprint arXiv:2401.08742*, 2024. 2

[49] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo

Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[50] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 2

[51] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2, 5

[52] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[53] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2

[54] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3, 2022. 1

[55] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 2, 5, 6

[56] Jiawei Ren, Cheng Xie, Ashkan Mirzaei, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, Huan Ling, et al. L4gm: Large 4d gaussian reconstruction model. *Advances in Neural Information Processing Systems*, 37:56828–56858, 2025. 2, 5, 6

[57] Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4209–4219, 2024. 2

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2

[59] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 2

[60] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2, 6

[61] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20875–20886, 2023. 2

[62] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023. 2

[63] Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 3d generation on imagenet. *arXiv preprint arXiv:2303.01416*, 2023. 2

[64] Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. *arXiv preprint arXiv:2310.16818*, 2023. 2

[65] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 1

[66] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22819–22829, 2023. 2

[67] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 1

[68] Zhicong Tang, Shuyang Gu, Chunyu Wang, Ting Zhang, Jianmin Bao, Dong Chen, and Baining Guo. Volumediffusion: Flexible text-to-3d generation with efficient volumetric encoder. *arXiv preprint arXiv:2312.11459*, 2023. 2

[69] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 5

[70] Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35:10021–10039, 2022. 2

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[72] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 2

[73] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4563–4573, 2023. 2

[74] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *ArXiv preprint*, abs/2305.16213, 2023. 2

[75] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20310–20320, 2024. 1, 2, 7

[76] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in Neural Information Processing Systems*, 29, 2016. 2

[77] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *arXiv preprint arXiv:2405.14832*, 2024. 2

[78] Zijie Wu, Chaohui Yu, Yanqin Jiang, Chenjie Cao, Fan Wang, and Xiang Bai. Sc4d: Sparse-controlled video-to-4d generation and motion transfer. In *European Conference on Computer Vision*, pages 361–379. Springer, 2024. 5, 6

[79] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. *arXiv preprint arXiv:2206.07255*, 2022. 2

[80] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 1, 2, 3, 5

[81] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*, 2024. 2

[82] Bojun Xiong, Si-Tong Wei, Xin-Yang Zheng, Yan-Pei Cao, Zhouhui Lian, and Peng-Shuai Wang. Octfusion: Octree-based diffusion models for 3d shape generation. *arXiv preprint arXiv:2408.14732*, 2024. 2

[83] Haitao Yang, Yuan Dong, Hanwen Jiang, Dejia Xu, Georgios Pavlakos, and Qixing Huang. Atlas gaussians diffusion for 3d generation with infinite number of points. *arXiv preprint arXiv:2408.13055*, 2024. 2

[84] Zeyu Yang, Zijie Pan, Chun Gu, and Li Zhang. Diffusion ^{2}: Dynamic 3d content generation via score composition of video and multi-view diffusion models. *arXiv preprint arXiv:2404.02148*, 2024. 2

[85] Lior Yariv, Omri Puny, Oran Gafni, and Yaron Lipman. Mosaic-sdf for 3d generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4630–4639, 2024. 2

[86] Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*, 2023. 2

[87] Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. Stag4d: Spatial-temporal anchored generative 4d gaussians. In *European Conference on Computer Vision*, pages 163–179. Springer, 2024. 2, 5, 6

[88] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11304–11314, 2022. 1

[89] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–16, 2023. 2, 3

[90] Bowen Zhang, Yiji Cheng, Chunyu Wang, Ting Zhang, Jiaolong Yang, Yansong Tang, Feng Zhao, Dong Chen, and Baining Guo. Rodinhd: High-fidelity 3d avatar generation with diffusion models. *arXiv preprint arXiv:2407.06938*, 2024. 2

[91] Bowen Zhang, Yiji Cheng, Jiaolong Yang, Chunyu Wang, Feng Zhao, Yansong Tang, Dong Chen, and Baining Guo. Gaussiancube: A structured and explicit radiance representation for 3d generative modeling. *arXiv preprint arXiv:2403.19655*, 2024. 1, 2

[92] Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. 4diffusion: Multi-view video diffusion model for 4d generation. *Advances in Neural Information Processing Systems*, 37:15272–15295, 2025. 2

[93] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43 (4):1–20, 2024. 1, 2

[94] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 4, 5

[95] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[96] Xinyang Zheng, Yang Liu, Pengshuai Wang, and Xin Tong. Sdf-stylegan: Implicit sdf-based stylegan for 3d shape generation. In *Computer Graphics Forum*, pages 52–63, 2022. 2

[97] Xin-Yang Zheng, Hao Pan, Peng-Shuai Wang, Xin Tong, Yang Liu, and Heung-Yeung Shum. Locally attentional sdf diffusion for controllable 3d shape generation. *ACM Transactions on Graphics (ToG)*, 42(4):1–13, 2023. 2

[98] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. *Advances in neural information processing systems*, 31, 2018. 2