

TextMesh4D: High-Quality Text-to-4D Mesh Generation

Sisi Dai¹ Xinxin Su¹ Boyan Wan¹ Ruizhen Hu² Kai Xu¹

¹National University of Defense Technology ²Shenzhen University



Figure 1. Given a text prompt, our method TextMesh4D can generate high-fidelity dynamic 3D mesh with the most preservation of geometry and appearance among realistic and continuous motion.

Abstract

Recent advancements in diffusion generative models significantly advanced image, video, and 3D content creation from user-provided text prompts. However, the challenging problem of dynamic 3D content generation (text-to-4D) with diffusion guidance remains largely unexplored. In this paper, we introduce TextMesh4D, a novel framework for high-quality text-to-4D generation. Our approach leverages per-face Jacobians as a differentiable mesh representation and decomposes 4D generation into two stages: static object creation and dynamic motion synthesis. We further propose a flexibility-rigidity regularization term to

stabilize Jacobian optimization under video diffusion priors, ensuring robust geometric performance. Experiments demonstrate that TextMesh4D achieves state-of-the-art results in terms of temporal consistency, structural fidelity, and visual realism. Moreover, TextMesh4D operates with a low GPU memory overhead—requiring only a single 24GB GPU—offering a cost-effective yet high-quality solution for text-driven 4D mesh generation. The code will be released to facilitate future research in text-to-4D generation.

1. Introduction

3D content generation has garnered significant attention with the popularity of various applications such as virtual reality, augmented reality, gaming and robotics simulation, etc. Recent advancements in text-to-image/video diffusion models [6, 17, 42, 47, 49, 50, 55], along with the pioneering technique of Score Distillation Sampling (SDS) [37], have driven significant progress in text-to-3D generation. However, this development has primarily focused on static representations, leaving dynamic 3D content generation, also known as 4D generation, comparatively underexplored.

Text-to-4D generation poses significant challenges due to the conflict between the simplicity of input text prompts and the complexity of dynamic 3D outputs, which requires natural motion and maintaining temporal consistency in both appearance and geometry. Some approaches [3, 48, 63] address this by adopting dynamic neural radiance fields (NeRF) [34] as the 4D representation, while others [29] employ dynamic 3D Gaussian Splatting (3DGS) [24]. However, these representations lack sufficient surface or shape constraints and often favor view-specific fidelity over precise geometry, leading to geometric distortion and appearance inconsistencies in the generated motion. In contrast, 3D triangle meshes offer explicit topological constraints that ease geometry distortion and naturally decouple shape from appearance during rendering, thereby enabling more consistent and higher-quality 4D generation. Moreover, compared to NeRF or Gaussian representations, mesh-based optimization requires significantly less GPU memory. The generated results can be directly integrated into standard CG pipelines, providing non-expert users with a straightforward method for complex 4D mesh workflows.

To this end, we introduce TextMesh4D, a novel framework for high-quality text-to-4D generation with mesh representations, using distilled priors from pre-trained diffusion models in a zero-shot manner. We tackle this complex task by splitting it into two stages: (1) generating a high-quality static 3D textured mesh, and (2) deforming that mesh to produce high-quality motion. The challenge thus lies in identifying a mesh representation that effectively supports both stages. Our key insight is that per-face Jacobians [2], which inherently exhibit low-frequency signals, mitigate shape collapse—compared to vertex displacements—and enable continuous, natural motion generation.

Building on this observation, we develop a 4D parameterization framework based on Jacobians in two parts: (1) static parameterization, which leverages Jacobians to produce a high-quality 3D textured mesh, laying the foundation for subsequent motion, and (2) dynamic parameterization, which incorporates local deformations through delta Jacobians in conjunction with global transformations. We first optimize the static parameters via Score Distillation Sampling (SDS) using a carefully evaluated combination of im-

age and 3D diffusion models, resulting in the high-quality 3D textured mesh. Next, we optimize the dynamic parameters with SDS for video diffusion priors, yielding natural and consistent motion.

Moreover, although Jacobians, as a smooth representation, provide an elegant solution and continuous performance under video diffusion priors, their high degree of freedom can make optimization challenging without direct supervision. To address this, we introduce a tailored regularization term that strikes a balance between rigidity and flexibility, thereby activating geometric performance under video score distillation sampling and ensuring robust outcomes. Experiments demonstrate that our method achieves text-driven 4D generation with state-of-the-art quality in terms of geometry, appearance, and motion.

Our contributions can be summarized as follows:

- **TextMesh4D:** we propose a novel text-to-4D generation framework that employs Jacobians as a differentiable mesh representation, introducing our Jacobian-based 4D parameterization for generating a static 3D textured object followed by vivid dynamic motion.
- **Flexibility-Rigidity Regularization:** We introduce a tailored regularization term that balances flexibility and rigidity to fully exploit the Jacobian representation under video distillation sampling.
- **Superior Performance on Commodity Hardware:** Our framework delivers state-of-the-art 4D generation results—achieving high temporal consistency, structural preservation, and visual fidelity—while running on a single 24GB GPU with low memory overhead.

2. Related Work

Text-to-Image/Video Generation. In recent years, diffusion models have achieved significant advancements in image and video generation, including text-to-image (T2I) models [40, 42, 43], as well as text-to-video (T2V) models [1, 9, 55]. These models are trained on large-scale open-domain datasets, typically including LAION-5B [44], WebVid-10M [5], HD-VG-130M [57]. Recent advancements in text-image-to-video generation (TI2V) have incorporated image-based semantic conditions into T2V models [12, 15, 58]. The latest model DynamiCrafter [60], employs a learnable image encoding network and dual cross-attention layers to effectively integrate text and image information, achieving impressive open-domain TI2V generation. Our work distills the generative power of video diffusion models for motion generation, with the belief that our method will evolve accordingly with the continued advancement of video generation technology.

Text-to-3D Generation. Early methods [10, 22, 30] for text-to-3D generation require paired data of 3D data and

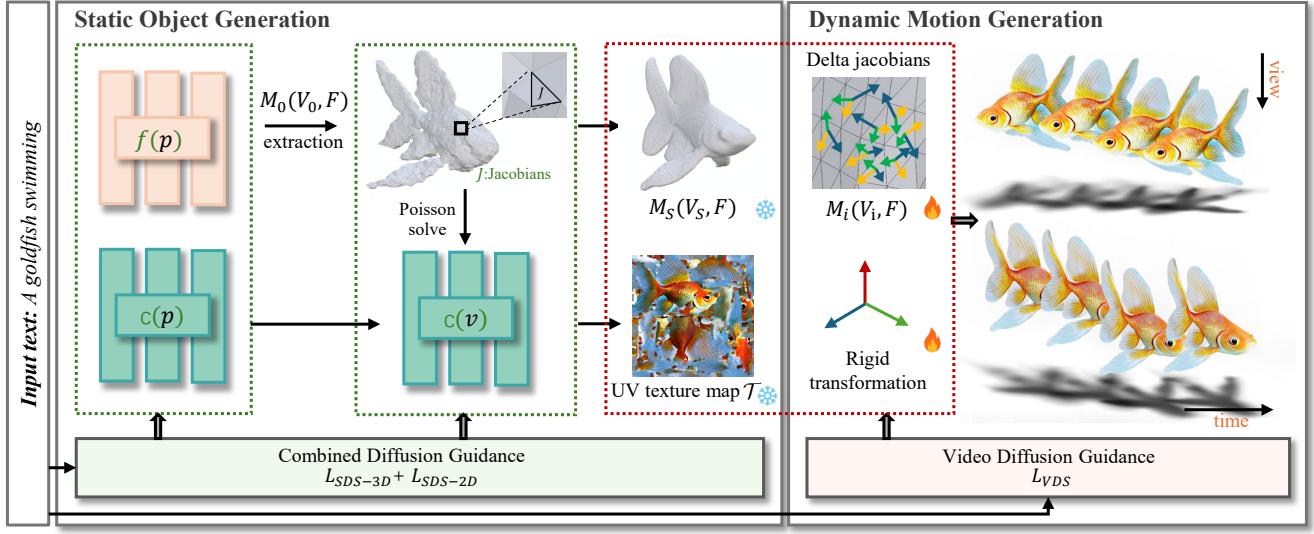


Figure 2. Overview of our TextMesh4D. Given a text prompt, TextMesh4D aims to generate high-quality 4D mesh in line with the prompt. To achieve this, TextMesh4D builds upon differentiable Jacobians for static and dynamic mesh representations. In the first static stage (Sec. 3.2), we generate a high-quality static 3D with Jacobians, initialized by NeuS with supervision provided by combined 2D and 3D diffusion priors. In the second dynamic stage (Sec. 3.3), we generate the dynamic motion with delta Jacobians for local deformation and global transformation, with tailored flexibility-rigidity geometric regularizer.

corresponding textual descriptions to learn the joint embedding space of shape and text for supervision, which limits their generality to unseen object categories. Benefiting from large pre-trained text-to-image models and differentiable rendering techniques, breakthroughs [18, 23, 26, 35] in text-to-3D content generation have been achieved. Recently, the technique SDS (Score Distillation Sampling) has been introduced in the pioneering work DreamFusion [37], enabling 3D generation by distilling guidance from pre-trained T2I diffusion models, achieving impressive results and becoming popular. There are a lot of follow-up works to improve DreamFusion. Some focus on 3D representation [11, 28]: Magic3D [28] proposes a coarse-to-fine pipeline to generate the fine-grained mesh; TextMesh [53] extends the geometry representation from NeRF to an SDF framework, enhancing detailed mesh extraction and photorealistic rendering; DreamGaussian [51] proposes to adopt 3D Gaussian Splatting to increase efficiency. Some works focus on improving SDS: SJC [54] proposes a variant of SDS while VSD are proposed in ProlificDreamer [59]; DreamTime [19] improves the generation quality by modifying the timestep sampling strategy. Others focus on inducing 3D prior into the guidance source, which effectively alleviates the Janus problem. Additional 3D prior are introduced in shape [8, 13, 20, 33, 36, 62], providing geometric initial values for optimizing NeRF. MV-Dream [46] proposes to fine-tune the diffusion model to generate multi-view images and as so explicitly embeds 3D information into a 2D diffusion model. Moreover, works

on image-based generation [32, 38, 52] and text-based editing [16, 27, 45, 64] are also boosted by utilizing these capabilities. Our first static stage performs text-to-3D generation with our Jacobian-based representation.

Text-to-4D Generation. Our research focuses on generating dynamic 4D content from textual descriptions. A pioneering effort, MAV3D [48], combines a T2V diffusion model with dynamic Neural Radiance Fields (NeRF) and HexPlane [7] to optimize both scene appearance and motion consistency. Building on this foundation, 4D-fy [3] employs a hybrid SDS approach that integrates T2I, 3D-aware T2I, and T2V diffusion models to achieve high-fidelity 4D generation. Align Your Gaussians (AYG)[29] employs dynamic 3D Gaussian splatting to reduce optimization time while enhancing temporal consistency. TC4D [4] introduces trajectory conditioning to maintain coherence between global and local motion. Although these methods have demonstrated effectiveness, they often rely heavily on NeRF and video models, leading to substantial computational costs. Comp4D [61] employs a Large Language Model (LLM) to segment input prompts into distinct entities, generating 4D objects independently and then combining them based on trajectory data provided by the LLM. Our work performs text-to-4D generation using a mesh representation, consisting of decomposed static and dynamic stages, which include both local deformation and global transformation. Additionally, incorporating geometry and texture disentanglement additionally.

3. Method

Our goal is to generate 4D mesh from text prompts, using distilled priors from pre-trained diffusion models in a zero-shot manner. The input is a given text prompt describing both the desired object and motion. The output is a sequence of textured 3D meshes, formulated as $\mathcal{M} = \{\mathcal{M}_i = (\mathcal{V}_i, \mathcal{F}, \mathcal{T})\}_{i=1}^L$, where \mathcal{V}_i denotes the vertices of the i -th mesh, which vary across sequence to capture the motion. \mathcal{F} represents the faces, \mathcal{T} indicates the UV texture map, and L is the length of the sequence.

We now first introduce our Jacobian-based parameterization (Sec. 3.1), and then explain how the parameters are optimized from static (Sec. 3.2) to dynamic (Sec. 3.3).

3.1. Jacobian-based 4D Parameterization

Jacobians. At each triangle f_j of mesh \mathcal{M} , the Jacobian $J_j \in \mathbb{R}^{3 \times 3}$ is a linear transformation from the triangle's tangent space to vertex space $\mathcal{V} \in \mathbb{R}^3$. Defining the deformation as vertex displacement $\Delta\mathcal{V}$ via Φ , a linear operator ∇_j is yield to associate each Φ with its corresponding Jacobian matrix $\nabla_j(\Phi)$. Thus, the Jacobian $\nabla_j(\Phi)$ restricts Φ within each triangle f_j , inherently providing low-frequency, smooth signals for deformation as vertex positions.

Given an target assignment of Jacobian J_j , a deformation map Φ^* can be solved following Poisson equation in a least-squares sense:

$$\Phi^* = \min_{\Phi} \sum_{f_j \in \mathcal{F}} |f_j| \|\nabla_j(\Phi) - J_j\|_2^2,$$

where $|f_j|$ is the area of triangle f_j . With deformation map Φ^* embedding the entire mesh, Φ can be optimized indirectly by optimizing the Jacobian matrices $\mathcal{J} = \{J_j\}$ for each face. We then leverage a differentiable Poisson solver layer [2] for our optimization.

4D Parameterization. The total 4D parametrization consists of decomposed parts: 1) static parameters for a textured 3D mesh, $\mathcal{M}_s = \{\mathcal{V}_s, \mathcal{F}, \mathcal{T}\}$; 2) a sequence of dynamic parameters $\Theta = \{\theta_i = \{\Delta\mathcal{V}_i, \mathcal{R}_i\}\}_{i=1}^L$, comprising the desired motion including both deformation $\Delta\mathcal{V}_i$, and rigid transformation \mathcal{R}_i with global translation and rotation, where L is the length of the sequence. Therefore, the output mesh sequence is $\mathcal{M} = \{\mathcal{M}_i = (\mathcal{R}_i(\mathcal{V}_s + \Delta\mathcal{V}_i), \mathcal{F}, \mathcal{T})\}_{i=1}^L$.

To achieve high-quality generation, rather than basing on direct vertex positions, we bulid the parametrization upon Jacobians $\mathcal{J} = \{J_j\}$ at each triangle as the mesh representation, thereby facilitating smooth, continuous, and globally consistent deformations. Thus, as illustrated in Fig. 2, our method consists of two stages: 1) First, we optimize for a high-quality static 3D model \mathcal{M}_s , the parameters are substituted by Jacobians as $\mathcal{M}_s = \{\mathcal{V}_0 + \Delta\mathcal{V}, \mathcal{F}, \mathcal{T}\} = \{\mathcal{V}_0 +$

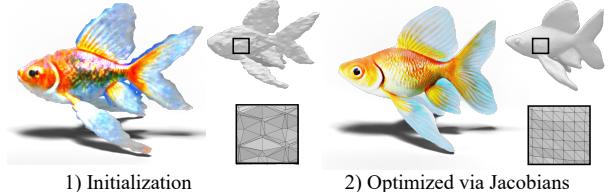


Figure 3. Comparison of geometry and texture between initialization and subsequent jacobian-based generation.

$\mathcal{J}, \mathcal{F}, \mathcal{T}\}$, where \mathcal{V}_0 and \mathcal{F} are initialized parameters by SDF network (Sec. 3.2); 2) With the static parameters fixed, dynamic parameters $\Theta = \{\theta_i = \{\Delta\mathcal{V}_i, \mathcal{R}_i\}\}_{i=1}^L$ are also represented by Jacobians as $\Theta = \{\theta_i = \{\Delta\mathcal{J}_i, \mathcal{R}_i\}\}_{i=1}^L$, then to be optimized (Sec. 3.3).

3.2. Static Object Generation

Mesh Initialization. Recall that at this stage, our objective is to generate a high-quality, textured 3D mesh solely from an input text prompt. We observe that learning large topological changes via direct mesh initialization (e.g., spot) is challenging for arbitrary inputs and often results in unsatisfactory mesh quality. To this end, we adopt NeuS [56] for mesh initialization. NeuS [56] is a volume rendering method that integrates the advantages of signed distance functions (SDF) and Neural Radiance Fields (NeRF) [34], better for extracting a 3D geometry and obtaining a mesh as the initialization for further high-quality generation. We denote the NeuS $\mathcal{N} = \{f(p), c(p)\}$, both $f(p)$ and $c(p)$ are networks implemented using MLPs, outputting the SDF and color at point p , respectively. The NeuS is then optimized with supervision provided by combined diffusion priors under input text conditioning. After optimization, we extract the surface at the zero-level set of SDF as the initial mesh \mathcal{M}_0 using the marching cubes algorithm [31].

Optimization via Jacobians. The extracted mesh \mathcal{M}_0 consisted by a set of vertices $\mathcal{V}_0 \in \mathbb{R}^{N \times 3}$, faces $\mathcal{F} \in \mathbb{R}^{M \times 3}$, which is converted to the representation of differentiable Jacobians, $\{\mathcal{V}_0 + \mathcal{J}, \mathcal{F}\}$, where the Jacobians are initialized as identity matrices for optimization. We inherit the weights of the color network from the initialization phase and continue to refine them. However, unlike the initialization phase, which utilizes random sampling points, the sampling in this phase is focused on regions near the initialized surface. Thus we denote the color network at this phase as $c(v)$. This concentrated sampling strategy allows for a more precise refinement of the color generation.

To achieve the high-quality generation, we employ combined diffusion priors from both 3D-aware and 2D-image diffusion models, following [3, 63]. the 3D-aware diffusion model, e.g. MVDream [46], is trained with multi-view embeddings along with camera parameters, providing a 3D prior and alleviating the Janus problem. As for the 2D-

image diffusion priors, we incorporate an additional loss term based on the variational score distillation (VSD) for appearance improvement. Combined SDS with them leverages the complementary strengths of 3D-aware and 2D-image diffusion models, resulting in a better generation for our static object:

$$\mathcal{L}_{\text{static}} = \lambda_{3D} \mathcal{L}_{\text{SDS-3D}} + \lambda_{2D} \mathcal{L}_{\text{SDS-2D}}, \quad (1)$$

The loss weights $\{\lambda_{3D}, \lambda_{2D}\}$ are carefully tuned for better generation. They are set to $\{0.7, 0.3\}$ during the initialization stage and adjusted to $\{0.5, 0.5\}$ subsequently. Please refer to the supplementary material for loss details.

To sum up, during the initialization stage, we optimize the networks $\{f(p), c(p)\}$. After initialization, with the initialized \mathcal{M}_0 , parameters that further need to be optimized are $\{\mathcal{J}, c(v)\}$. Finally, the UV-space texture map \mathcal{T} is extracted from $c(v)$ for the subsequent motion generation. Fig. 3 illustrates the evolution from the initialization phase to the final generation at this stage.

3.3. Dynamic Motion Generation

With the static parameters fixed, we then optimize the dynamic parameters $\Theta = \{\theta_i = \{\Delta \mathcal{J}_i, \mathcal{R}_i\}\}_{i=1}^L$ to produce the vivid 3D motion. Note that a differentiable renderer is required to project the textured mesh sequence with \mathcal{T} to the image space, thus enabling gradient steps during optimization at this stage. We implement the renderer based on NVdiffrast [25] as follows:

$$R(\cdot | C) := \mathcal{M}_i^{\theta_i, \mathcal{M}_S} \mapsto I^{\mathcal{M}_i} \in \mathbb{R}^{H \times W}, \quad (2)$$

where H and W denote the height and width of the rendered image, with C representing the camera extrinsics.

Objective Function. The overall objective function at this stage is:

$$\mathcal{L}_{\text{dynamic}} = \mathcal{L}_{\text{VDS}} + \mathcal{L}_{\text{flex}} + \mathcal{L}_{\text{rig}} + \mathcal{L}_{\text{other}} \quad (3)$$

First of all, we use video diffusion priors to provide semantic motion guidance by video score distillation sampling (VDS). This procedure queries a video diffusion model, e.g. [1], to see how a rendered video from our representation aligns with input prompt, through the noise sampling of video diffusion process. The gradients are then backpropagated to the dynamic parameters. Please refer to the supplementary material for the corresponding loss \mathcal{L}_{VDS} computation.

However, the stochastic nature of SDS and Jacobians' high degree of freedom introduce distortions and instable convergence into the optimization. To address this issue, we design the tailored regularization term, with a synergy of flexibility and rigidity, for Jacobians' robust optimization under guidance from video score distillation sampling.

$\mathcal{L}_{\text{flex}}$ is a geometric regularization term on the optimized Jacobians to prevent the divergence too far from the static object's geometry during dynamic optimization, inspired by [14]. This term ensures the global geometry is preserved while still allowing for flexible deformations that capture motion semantics. Specifically, the term penalizes the difference between the dynamic Jacobians $\{J + \Delta J_i\}$ and the static identity with a weight:

$$\mathcal{L}_{\text{flex}} = \sum_{i=0}^{\ell-1} \sum_{j=0}^{|f|-1} e^{\|\hat{J}_j - I\|} \|\hat{J}_j - I\|_2, \quad (4)$$

We then further employ As-Rigid-As-Possible (ARAP) energy [21] as the rigidity regularization term \mathcal{L}_{rig} :

$$\mathcal{L}_{\text{rig}} = \sum_{i=0}^{\ell-1} \sum_{j=0}^{n-1} \sum_{k \in \mathcal{N}_{v_j}} w_{j,k} \|(v_j^i - v_k^i) - R_j(v_j^s - v_k^s)\|^2, \quad (5)$$

where \mathcal{N}_{v_j} represents the one-ring neighborhoods of vertex v_j . $w_{j,k} = (\cot \alpha_{jk} + \cot \beta_{jk}) / 2$, measuring the impact of v_k on v_j . α_{jk} and β_{jk} are the angles on the faces adjacent to the edge (v_j, v_k) , which are opposite the edge itself. R_j represents the optimal rotation estimated by Singular Value Decomposition (SVD) [21]. This term encourages the generation to maintain locally rigid during the deformation.

The rigidity-flexibility integrated regularizer strikes a balance between preserving rigid motion and maintaining deformation fidelity, thereby enhancing the optimization results, as demonstrated in our ablation studies. We leave additional details of the other regularizers, e.g. smoothness term and Jacobian's dof regularization, in the supplementary material.

4. Experiments

In this section, we first compare our method with several state-of-the-art baselines and then conduct extensive ablation studies to verify the design choices of our method.

4.1. Results and Comparisons

Baselines. We compare with the publicly available 4Dfy [3] in text-to-4D generation. Additionally, since the source code for AYG [29] is not available, we further designed a variant of DreamGaussian4D [41], by inputting the text prompt to Kling to generate a high-quality image, which then serves as the input for DreamGaussian4D.

Evaluation metrics. It is a common difficulty to quantitatively evaluate text-to-4D generation results due to the lack of ground truth. We thus adopt the evaluation metrics following [3]: CLIP score and user study. The CLIP score assesses how well the generated output aligns with the input text prompt by computing the cosine similarity between the

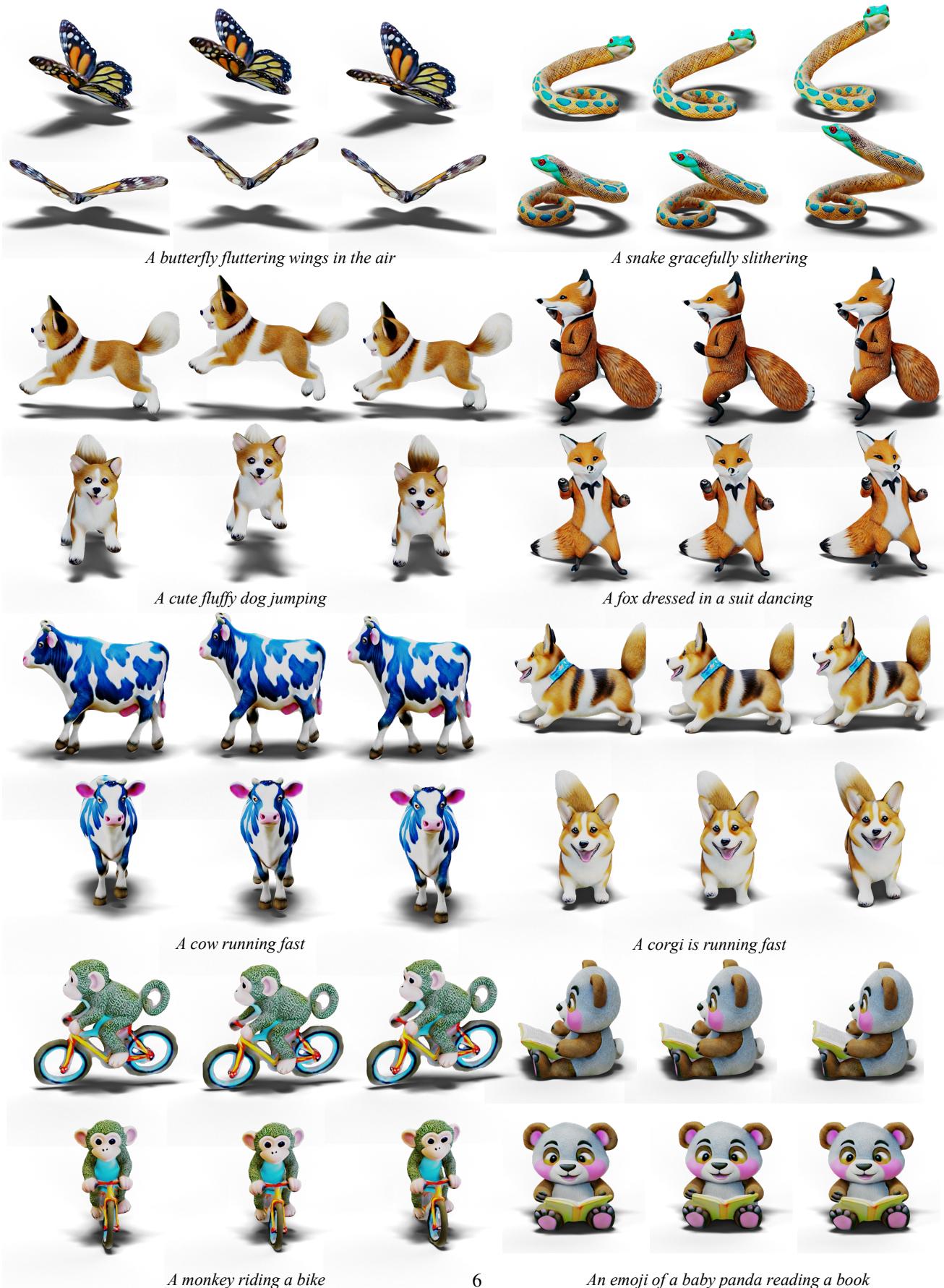


Figure 4. Text-to-4D generation results of our method, demonstrating high quality, consistency and realistic motion. Dynamic motions are presented in the supplementary video.

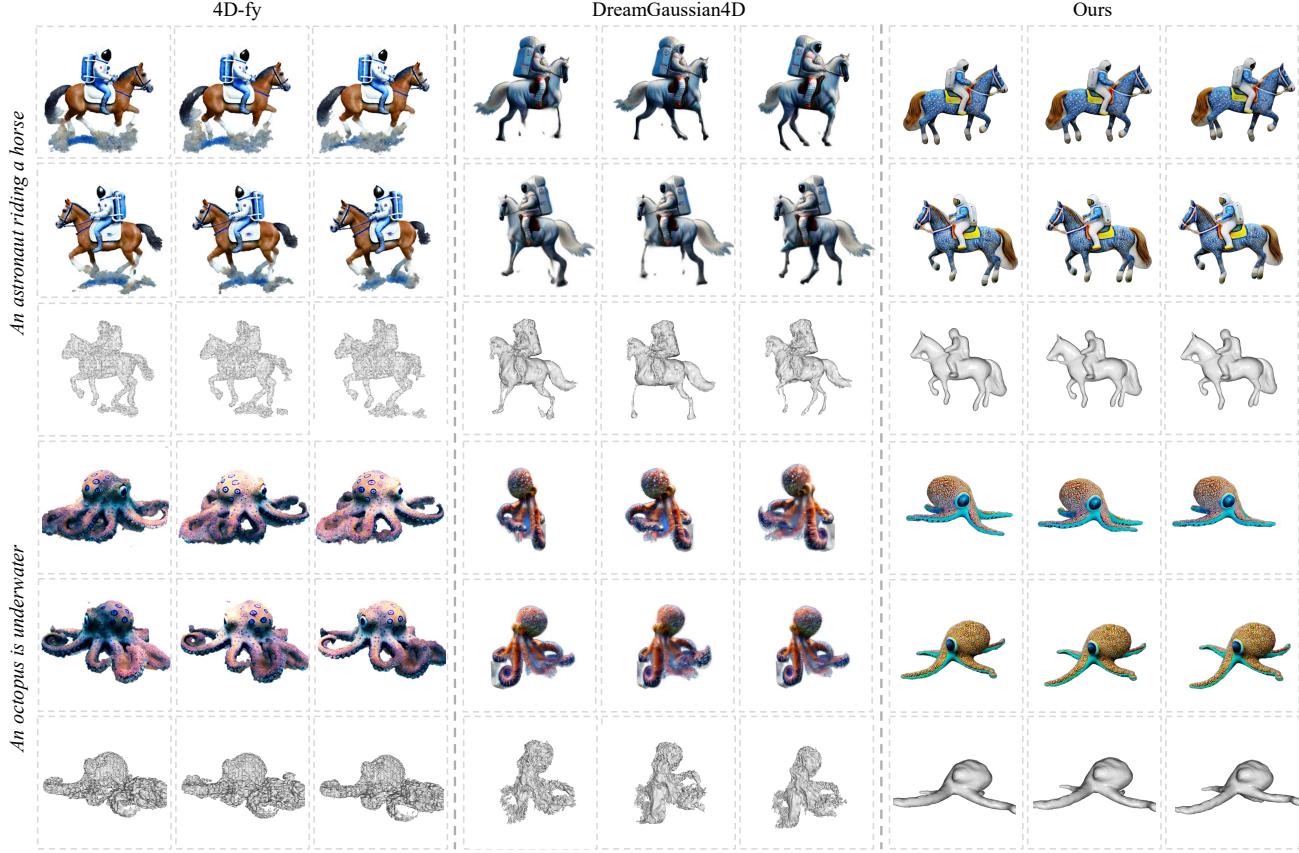


Figure 5. We compare our method with 4D-fy and the variation of DreamGaussian4D. Our method achieves significantly higher quality in both texture and geometry, while maintaining consistency during motion.

Table 1. Quantitative comparisons with 4d-fy and variations of DreamGaussian4D. The methods are evaluated in terms of CLIP Score (CLIP) and metrics of user study, where the numbers are the percentages of users who voted for the corresponding method.

Method	CLIP	User Study				
		AQ	SQ	MQ	TA	Overall
4D-fy	31.03	6.9%	4.4%	1.0%	13.4%	4.7%
DreamGaussian4D	29.17	0.3%	0.5%	0.8%	3.1%	1.8%
Ours	32.32	92.7%	95.1%	98.2%	83.5%	93.5%

textual CLIP embedding and visual CLIP embedding [39]. Multiple camera views and frames over time are sampled during the computation of the CLIP score. Moreover, we conduct a user preference study to evaluate sample quality along the dimensions of *appearance quality (AQ)*, *3D structure quality (SQ)*, *motion quality (MQ)*, *text alignment (TA)*, and *overall preference (Overall)*, as in 4D-fy [3]. There are 31 participants involved to rate the 20 prompts’ results from our method as well as from the baselines. Please refer to the supplementary document for more details.

Comparison results. As shown in Fig. 5, our method significantly outperforms the baselines in terms of motion vividness, geometric preservation, and appearance consistency. The implicit representation of baselines, such as NeRF in 4D-fy and 3DGs in DreamGaussian4D, results in most of the generated motion manifesting as floating artifacts in the geometric field, making it visually challenging to interpret. Furthermore, since these baselines do not model global transformations, the generated motion is confined to local regions of the object, further limiting their performance. In contrast, our method leverages a mesh representation that incorporates both local deformation and global transformation, enabling successful spatial movement with only an input text. Additionally, the mesh representation allows our method to run efficiently on a 24GB GPU, achieving motion convergence within 1.5 hours. In comparison, 4D-fy requires an 80GB GPU and over 15 hours to complete, while DreamGaussian4D needs a 48GB GPU. The quantitative results presented in Table 1 further demonstrate our superiority. The user study indicates that users generally favor our method when directly comparing with the results obtained by the baselines, with the number indicating the percentage of participants choosing the corre-

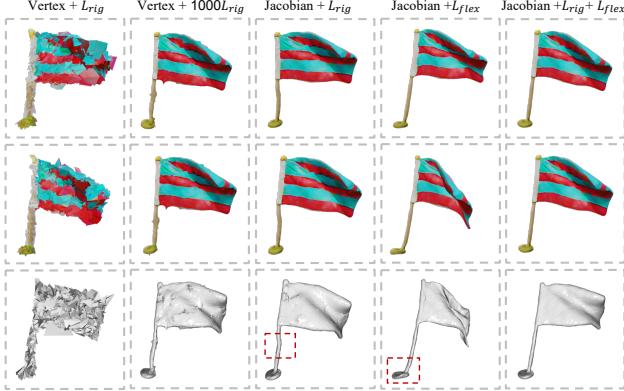


Figure 6. Qualitative ablations with the given text prompt “a flag fluttering in the air”. The top two rows showcase two frames, while the bottom row illustrates the geometry corresponding to the second row.

sponding method. More of our text-to-4D generation results are shown in Fig. 4. Please refer to the supplementary video for a better viewing experience.

4.2. Ablation Study

We conduct ablation studies to verify the necessity of designed choices. Qualitative comparisons are shown in Fig. 6 and quantitative results are reported in Tab. 2.

Vertex displacements. To verify the importance of the per-face Jacobians, we substituted them with vertex displacement optimization, which is the most straightforward approach for mesh optimization. Under the same VDS guidance, in addition to ARAP, we introduced regularization constraints, including normal consistency and Laplacian regularization.

With a light ARAP energy weight (the same as ours, shown in the first column of Fig. 6), vertex displacements align motion with the text semantics but lead to a large number of self-intersections, resulting in a loss of shape fidelity. With a heavier ARAP energy weight (shown in the second column of Fig. 6), although the shape is preserved, there is still some shape collapse, and no motion is generated.

The results under vertex displacement representation exhibit a granular effect, similar to the results obtained using NeRF [3] or Gaussian-based methods [41]. These methods focus on optimizing spatial points, and due to the randomness of VSD score distillation sampling method, discrete spatial point representations are unstable under VDS guidance, leading to difficulties in convergence and introducing instability and artifacts in the final results. In contrast, with the same ARAP weight (shown in the third column of Fig. 6), the continuous per-face Jacobian representation ensures smooth optimization by mitigating the noisy gradients produced by VSD, thus demonstrating its superiority.

Flexibility regularizer. To evaluate the effectiveness of

Table 2. Ablation Study. The choices are also evaluated in terms of metrics of user study, where the numbers are the percentages of users who voted for the corresponding method.

Settings	User Study				Overall
	AQ	SQ	MQ	TA	
Vertex+ \mathcal{L}_{rig}	1.9%	1.0%	1.5%	1.8%	1.1%
Vertex+1000 \mathcal{L}_{rig}	1.1%	7.3%	0.5%	1.3%	3.9%
Jacobian+ \mathcal{L}_{rig}	15.2%	9.0%	3.4%	5.6%	7.4%
Jacobian+ \mathcal{L}_{flex}	5.6%	6.9%	39.7%	26.3%	10.3%
Ours	76.2%	75.8%	55.0%	65.0%	77.3%

our designed geometry regularizers, we perform dynamic motion optimization by omitting the flexibility regularization term and leaving only the rigidity term. As shown in the third column of Fig. 6, the motion optimization becomes trapped in a local optimum due to the sole rigidity constraint. For example, to simulate fluttering in the air, the flagpole undergoes bending, causing the flag to appear to move in an unnatural way.

Rigidity regularizer. We also perform dynamic motion optimization by omitting the rigidity regularization terms, leaving only the flexibility term. As shown in the fourth column of Fig. 6, although the motion is most extensive, the shape appears to be stretched (similar to clay), highlighting the necessity of the local rigidity loss. With our integrated flexibility-rigidity regularizer, the realistic motion of “a flag fluttering in the air” is achieved.

5. Conclusion

We propose TextMesh4D, a novel framework for high-fidelity 4D mesh generation from a text prompt. By building our 4D differentiable mesh representation based on per-face Jacobians, our method first generates a high-quality static 3D mesh, and then learns dynamic motion, including both local deformation and global transformation, through supervision provided by video diffusion priors. Furthermore, we introduce flexibility and rigidity regularizers to stabilize Jacobian optimization under video diffusion priors, ensuring robust geometric performance. TextMesh4D achieves superior results in text-to-4D generation, with realistic motion, geometry preservation, and appearance consistency.

Limitations & Future Work. As our framework is designed to first generate static content and then dynamic motion, the dynamic generation may fail if the static generation is unsatisfactory, leading to incorrect accumulation. We believe this issue could be addressed with further advances in 2D and 3D diffusion models. Moreover, the optimization space for global transformations in our method is still constrained by the camera space of differentiable rendering. We believe that combining our method with the latest video diffusion models under camera control offers a promising direction for future work.

References

- [1] Zeroscope text-to-video model. https://huggingface.co/cerspense/zeroscope_v2_576w. Accessed: 2023-10-31. 2, 5
- [2] Noam Aigerman, Kunal Gupta, Vladimir G Kim, Siddhartha Chaudhuri, Jun Saito, and Thibault Groueix. Neural jacobian fields: Learning intrinsic mappings of arbitrary meshes. *arXiv preprint arXiv:2205.02904*, 2022. 2, 4
- [3] 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, 3, 4, 5, 7, 8
- [4] Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, et al. Tc4d: Trajectory-conditioned text-to-4d generation. In *European Conference on Computer Vision*, pages 53–72. Springer, 2025. 3
- [5] Max Bain, Arsha Nagrani, GÜl Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proc. ICCV*, 2021. 2
- [6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voeleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2
- [7] 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. 3
- [8] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 958–968, 2024. 3
- [9] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2
- [10] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III* 14, pages 100–116. Springer, 2019. 2
- [11] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22246–22256, 2023. 3
- [12] I2VGen-XL contributors. I2vgen-xl. Accessed October 15, 2023 [Online] <https://modelscope.cn/models/damo/Image-to-Video/summary>. 2
- [13] Sisi Dai, Wenhao Li, Haowen Sun, Haibin Huang, Chongyang Ma, Hui Huang, Kai Xu, and Ruizhen Hu. Interfusion: Text-driven generation of 3d human-object interaction. *arXiv preprint arXiv:2403.15612*, 2024. 3
- [14] William Gao, Noam Aigerman, Thibault Groueix, Vova Kim, and Rana Hanocka. Textdeformer: Geometry manipulation using text guidance. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 5
- [15] Xianfan Gu, Chuan Wen, Weirui Ye, Jiaming Song, and Yang Gao. Seer: Language instructed video prediction with latent diffusion models. *arXiv preprint arXiv:2303.14897*, 2023. 2
- [16] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19740–19750, 2023. 3
- [17] 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. 2
- [18] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022. 3
- [19] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023. 3
- [20] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [21] Takeo Igarashi, Tomer Moscovitch, and John F Hughes. As-rigid-as-possible shape manipulation. *ACM transactions on Graphics (TOG)*, 24(3):1134–1141, 2005. 5
- [22] Tansin Jahan, Yanran Guan, and Oliver Van Kaick. Semantics-guided latent space exploration for shape generation. In *Computer Graphics Forum*, pages 115–126. Wiley Online Library, 2021. 2
- [23] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022. 3
- [24] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2
- [25] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakkio Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (ToG)*, 39(6):1–14, 2020. 5
- [26] Han-Hung Lee and Angel X Chang. Understanding pure clip guidance for voxel grid nerf models. *arXiv preprint arXiv:2209.15172*, 2022. 3

- [27] Yuhang Li, Yishun Dou, Yue Shi, Yu Lei, Xuanhong Chen, Yi Zhang, Peng Zhou, and Bingbing Ni. Focaldreamer: Text-driven 3d editing via focal-fusion assembly. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3279–3287, 2024. 3
- [28] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022. 3
- [29] Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. 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*, pages 8576–8588, 2024. 2, 3, 5
- [30] Zhengzhe Liu, Yi Wang, Xiaojuan Qi, and Chi-Wing Fu. Towards implicit text-guided 3d shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17896–17906, 2022. 2
- [31] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 4
- [32] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8446–8455, 2023. 3
- [33] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 3
- [34] 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, 4
- [35] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 3
- [36] Ryan Po and Gordon Wetzstein. Compositional 3d scene generation using locally conditioned diffusion. In *2024 International Conference on 3D Vision (3DV)*, pages 651–663. IEEE, 2024. 3
- [37] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 3
- [38] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 3
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7
- [40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2
- [41] 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. 5, 8
- [42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 2022. 2
- [43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2
- [44] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Proc. NeurIPS*, 2022. 2
- [45] Etai Sella, Gal Fiebelman, Peter Hedman, and Hadar Averbuch-Elor. Vox-e: Text-guided voxel editing of 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 430–440, 2023. 3
- [46] 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. 3, 4
- [47] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2
- [48] 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, 3
- [49] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2
- [50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2
- [51] 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. 3
- [52] 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 *Proc.*

- ceedings of the IEEE/CVF international conference on computer vision*, pages 22819–22829, 2023. 3
- [53] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. In *2024 International Conference on 3D Vision (3DV)*, pages 1554–1563. IEEE, 2024. 3
- [54] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 3
- [55] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2
- [56] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 4
- [57] Wenjing Wang, Huan Yang, Zixi Tuo, Huigu He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023. 2
- [58] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 2
- [59] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [60] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 2
- [61] Dejia Xu, Hanwen Liang, Neel P Bhatt, Hezhen Hu, Hanxue Liang, Konstantinos N Plataniotis, and Zhangyang Wang. Comp4d: Llm-guided compositional 4d scene generation. *arXiv preprint arXiv:2403.16993*, 2024. 3
- [62] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20908–20918, 2023. 3
- [63] Yufeng Zheng, Xuetong Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text- and image-guided 4d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7300–7309, 2024. 2, 4
- [64] Jingyu Zhuang, Chen Wang, Liang Lin, Lingjie Liu, and Guanbin Li. Dreameditor: Text-driven 3d scene editing with neural fields. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 3