SC4D: Sparse-Controlled Video-to-4D Generation and Motion Transfer

Zijie Wu[1,2], Chaohui Yu[2], Yanqin Jiang[2], Chenjie Cao[2], Fan Wang[2], and Xiang Bai[1†]

[1] Huazhong University of Science and Technology
[2] DAMO Academy, Alibaba Group
{zjw1031,xbai}@hust.edu.cn,
{huakun.ych,jiangyanqin.jyq,caochenjie.ccj,fan.w}@alibaba-inc.com
https://sc4d.github.io/

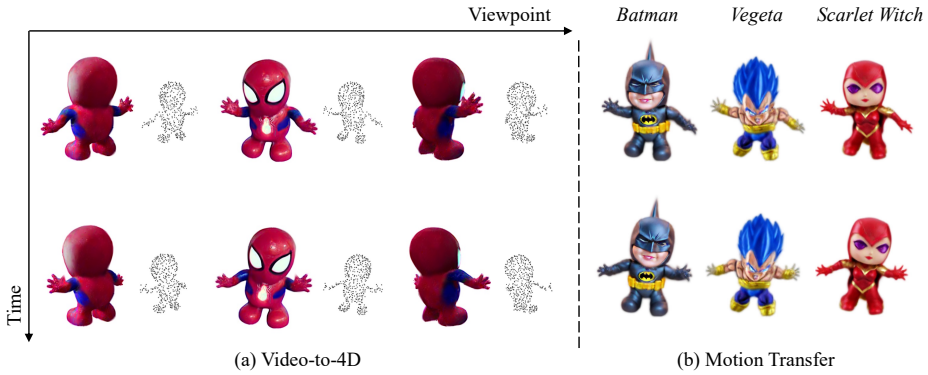

Fig. 1: We illustrate: (a) video-to-4D results of SC4D and corresponding control points visualizations, and (b) examples of our motion transfer applications in the figure.

Abstract. Recent advances in 2D/3D generative models enable the generation of dynamic 3D objects from a single-view video. Existing approaches utilize score distillation sampling to form the dynamic scene as dynamic NeRF or dense 3D Gaussians. However, these methods struggle to strike a balance among reference view alignment, spatio-temporal consistency, and motion fidelity under single-view conditions due to the implicit nature of NeRF or the intricate dense Gaussian motion prediction. To address these issues, this paper proposes an efficient, sparse-controlled video-to-4D framework named SC4D, that decouples motion and appearance to achieve superior video-to-4D generation. Moreover, we introduce Adaptive Gaussian (AG) initialization and Gaussian Alignment (GA) loss to mitigate shape degeneration issue, ensuring the fidelity of the learned motion and shape. Comprehensive experimental results demonstrate that our method surpasses existing methods in both quality and efficiency. In addition, facilitated by the disentangled modeling of motion

† Corresponding author

and appearance of SC4D, we devise a novel application that seamlessly transfers the learned motion onto a diverse array of 4D entities according to textual descriptions.

Keywords: Video-to-4D generation · Dynamic Gaussian splatting · Motion transfer

1 Introduction

In recent years, with the advancement of generative AI, we have witnessed significant progress in 3D generation techniques, which include the generation of static objects' shape, texture, and even an entire scene from a text prompt or a single image. Compared to static 3D assets, dynamic 3D (4D) content offers greater spatio-temporal flexibility and thus harbors more substantial potential for applications in AR/VR, filming, animation, simulation, and other domains. However, generating 4D objects from text descriptions or video references is rather formidable due to the difficulties of maintaining spatio-temporal consistency and ensuring motion fidelity. Nevertheless, as humans, we are adept at resolving the above challenging tasks, a capability attributable to our possession of extensive prior knowledge of the real world.

Very recently, building upon the foundations laid by extant text-to-3D [25,38, 57,60,70] and image-to-3D [26,28,41,52,54] pipelines, several methods [16,36,69] distill prior knowledge from novel view synthesis models [29,30] to generate the target 4D object as dynamic NeRF [16] or dynamic 3D Gaussians [36, 69] and impose constraints to ensure the temporal consistency among frames. Despite the commendable progress achieved by these methods, they still struggle to strike a balance among reference view alignment, spatio-temporal consistency, and motion fidelity. We argue that the representation is critical for video-to-4D generation. On the one hand, dynamic NeRF [2,10,33,39] cannot sustain temporal coherence under novel views without additional restrictions due to its implicit nature and stochasticity inherent in Score Distillation Sampling (SDS) [38]. On the other hand, it is intricate to predict accurate trajectories and rotations for dense Gaussians [17,31,62,66] with only single-view conditions.

To tackle the aforementioned problems, inspired by a recent dynamic scene reconstruction method [14], we decouple the appearance and motion of the dynamic 3D object into dense Gaussians (∼50k) and sparse control points (∼512) and design an efficient two-stage video-to-4D generation framework, named **SC4D**. Specifically, in the coarse stage, we initialize a set of sparse control points as spherical Gaussians and a Multilayer Perceptron (MLP) conditioned on time and location to predict the motion of these sparse Gaussians. Then, we optimize the parameters of these Gaussians and the MLP under the guidance of reference view reconstruction and novel view score distillation. In the fine stage, we utilize the sparse control Gaussians as implicit control points and perform Linear Binding Skinning (LBS) [51] to drive the dense Gaussians. In this stage, we jointly optimize the parameters of control points, dense Gaussians, and deformation

MLP to obtain the final results. Notably, since only a single-view ground truth video is provided, we empirically find that there is a proclivity for shape degeneration issues in the fine stage, which results in thickening, position displacements, and texture blur of the dynamic 3D object. To address these challenges, we devise Adaptive Gaussian (AG) initialization that inherits the shape and motion of the control Gaussians in the coarse stage with a random amount of Gaussians. Moreover, we present Gaussian Alignment (GA) loss to ensure shape and motion fidelity in the fine stage.

Benefiting from the disentangled modeling of appearance and motion within our method, SC4D effectively mitigates the ambiguity of these two attributes during optimization. Additionally, since the motion of the dynamic object is governed by a set of sparse control points rather than dense Gaussians, our approach simplifies the learning of motion and naturally exhibits enhanced local rigidity. Comprehensive evaluations reveal that our method surpasses existing video-to-4D generation methods [16,69] in both quality and efficiency. Moreover, after obtaining the motions of implicit control points, we introduce a novel application that transfers the learned motion onto other entities under the guidance of text-to-image models [44, 71] and dense Gaussians' depth. In conclusion, our contributions can be summarized as follows:

1. We propose SC4D, a sophisticated video-to-4D generation framework based on sparse points control, which generates dynamic 3D objects with superior quality and efficiency compared to existing methods.
2. We devise an Adaptive Gaussian (AG) initialization approach and Gaussian Alignment (GA) loss that effectively mitigate the shape degeneration issue, ensuring accurate motion and shape learning.
3. We propose a novel application based on the control points' motion and design a motion transfer pipeline that maps the learned motion onto distinct entities, as directed by textual descriptions.

2 Related Works

Optimization-based 3D Generation. Optimization-based 3D generation methods typically optimize a NeRF [33] or 3D Gaussians [17] utilizing prior knowledge from image-text matching model [42] or diffusion-based generative models [13, 29, 44, 45]. CLIP-based text-to-3D methods [15, 19, 34, 65] generally employ CLIP [42] to align each viewpoint of the target 3D scene with the given text description for optimization. DreamFusion [38] substitutes the guidance model from CLIP to a 2D diffusion model [45], and introduces Score Distillation Sampling (SDS) to distill prior knowledge from text-to-image models. As a concurrent work, SJC [57] shares a similar idea, which distills scores in a Perturb-and-Average manner. Following the paradigm of SDS, a series of methods aim at bringing finer texture details [4, 25, 60] or alleviating the Janus problem [46, 70] utilizing DMTet [47] representation, Variational Score Distillation [60], Point-E [35] condition, etc. Recently, a succession of methods further enhanced the

view quality [21, 75] and multi-view consistency [48]. There are also several 3D Gaussian-based methods [5,53,68] that achieve comparable results. As for image-to-3D, RealFusion [32] performs textual inversion [11] before 3D generation to match the intended concept. Make-it-3D [54] improves the view quality and consistency in an inpainting manner. Zero123 [29] utilizes large-scale multi-view data from Objaverse dataset [7,8] to turn Stable Diffusion (SD) [44] into a novel-view generator. Based on Zero123, a bunch of methods [28,41,52] achieves high-fidelity image-to-3D generation. There are also several approaches [20, 26, 30, 58] acquire notable progress in enhancing multi-view consistency.

4D Representation. Current 4D representations predominantly bifurcate into two principal categories: dynamic NeRF [33] and dynamic 3D Gaussians [17]. Dynamic NeRF-based methods can be further divided into deformable NeRF [37, 39, 56, 63] and time-varying NeRF [2, 9, 10, 12, 23]. Recently, a variety of methods predicated on dynamic 3D Gaussians [3, 22, 24, 31, 62, 66, 67] have emerged, leveraging Gaussians' explicit nature and real-time rendering capabilities. There are also some methods that ameliorate the dense motion prediction by learning a set of sparse trajectories [18], control points [14], or basis vectors [6].

4D Generation. Compared to 3D generation, high-quality 4D generation is even more challenging since the temporal dimension is involved. Existing text-to-4D methods [1, 27, 43, 50, 59, 73, 74] distill geometry and temporal information from diffusion-based text-to-image models [44] and text-to-video models [49] with SDS [38]. In recent developments, several video-to-4D frameworks [16,36,69] have been introduced. These pipelines endeavor to recover the dynamic 3D objects from single-view video inputs, facilitated by Zero123 [29]. However, these video-to-4D methods struggle to strike a balance among reference view alignment, spatio-temporal consistency, and motion fidelity.

3 Method

Given a single-view reference video, video-to-4D methods [16, 36, 69] aim to recover a plausible dynamic 3D object that aligns with the video source. Inspired by [14], we propose an efficient video-to-4D framework based on sparse control points (shown in Fig. 2), named **SC4D**, which utilizes separated modeling of appearance and motion to yield superior outcomes. To ensure the fidelity of learned shape and motion, we introduce Adaptive Gaussian (AG) initialization based on control points, and Gaussian Alignment (GA) loss as an additional constraint. Furthermore, we devise a novel application that enables motion transfer with text descriptions after acquiring the control point motions.

3.1 Preliminaries

Score Distilltion Sampling. Score Distillation Sampling (SDS) [38] is widely adopted to distill prior knowledge from image generation models [29, 44]. In this work, we utilize Zero123 [29] as the source of novel view information. Given a

reference image $\mathrm{I_r}$, a relative camera extrinsic (R, T) between the queried and input views, the 3D model as θ, Zero123 as ϕ, then the SDS loss is as follows:

$$\bigtriangledown_\theta L_{SDS}(\phi, x) = \mathbb{E}_{t,\epsilon}[\omega(t)(\hat{\epsilon}_\phi(z_t; \mathrm{I_r}, \mathrm{R}, \mathrm{T}, t) - \epsilon)\frac{\partial x}{\partial \theta}], \tag{1}$$

where t is the randomly sampled timestep in the diffusion process, x is the rendered image, and $\omega(t)$ is a weighting function depending on the timestep t.

3D Gaussian Splatting. 3D Gaussian Splatting (3DGS) [17] represents a scene as a set of explicit 3D Gaussians. Each Gaussian G has a center position μ, a rotation quaternion q, a scaling parameter s, opacity σ and sphere harmonic (SH) coefficients sh. It can be defined as $G(x) = e^{-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)}$, where $\sum$ is the 3D covariance matrix, calculated by $\sum = RSS^TR^T$ (R, S is equivalent to q, s). The color of a pixel u is rendered using α-blending:

$$Color(u) = \sum_i SH(sh_i, v_i)\alpha_i \prod_{j=1}^{i-1}(1 - \alpha_j), \tag{2}$$

where $\alpha_i = \sigma_i G(u)$, v_i is the view direction, and SH is the spherical harmonic function. To enhance the accuracy of fitting across diverse scenes, densification and pruning are adopted based on gradient accumulation during optimization.

Sparse-Controlled Gaussian Splatting (SC-GS). SC-GS [14] is an effective dynamic scene reconstruction method, which decouples the appearance and motion of a 4D scene as 3D Gaussians and control points. SC-GS utilizes a time-condition MLP $\mathbf{\Psi}$ to predict the translation T_i^t and rotation R_i^t for each control point. Then, 3D Gaussians are driven by control points following LBS [51]. For each Gaussian G_j, the warped location μ_j^t and rotation q_j^t can be computed as a weighted sum of its KNN control points M_j:

$$\mu_j^t = \sum_{k \in M_j} \mathrm{w}_{jk}(R_k^t(\mu_j - p_k) + p_k + T_k^t), \tag{3}$$

$$q_j^t = (\sum_{k \in M_j} \mathrm{w}_{jk} r_k^t) \otimes q_j, \tag{4}$$

where w_{jk} is a weighting ratio depending on the distance d_{jk} between Gaussian G_j center and its neighboring control point p_k, and a learned control radius o_k. p_k, r_k^t denote the position and rotation quaternion for the control point.

3.2 Coarse Stage: Sparse Control Points Initialization

As illustrated in SC-GS [14], sparse control points initialization is critical for decoupling motion and appearance of the dynamic object/scene. Joint optimization of sparse control points and dense Gaussians directly can result in uneven distribution of control points and may even lead to training collapse. In this stage, we aim to obtain a good initialization for sparse control points' locations and motion that align with the reference video.

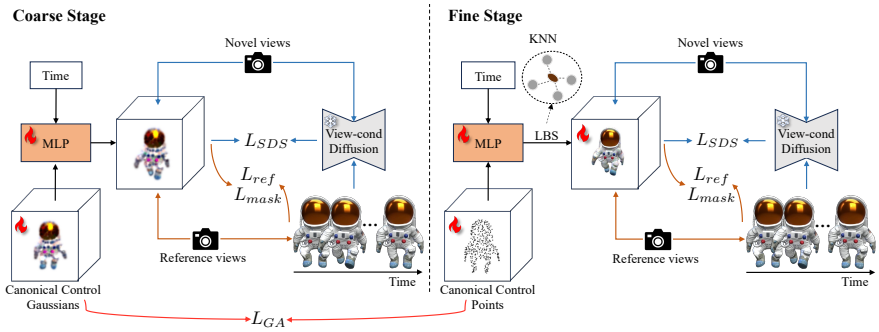

Fig. 2: Overall pipeline of the proposed SC4D. In the coarse stage, SC4D learns a proper shape and motion initialization with a set of sparse control Gaussians. Then, in the fine stage, we propose Adaptive Gaussian (AG) initialization, and Gaussian Alignment (GA) loss to prevent shape and motion degeneration, and jointly optimize control points, dense Gaussians, and deformation MLP for the final results.

As shown in Fig. 2, since no ground truth 3D data is available, we initialize M control points (denoted as control Gaussians in this stage) as 3D Gaussians with randomly sampled positions inside a sphere. In order to obtain control Gaussians that are more evenly distributed, we constrain them as spheres with the same scaling parameter (s). We denote the control Gaussians as $C_i : (p_i, r_i, s, \sigma_i, sh_i)$, and the reference image sequence as $\{I_r^f, f = 1, 2, \cdots, F\}$, where F stands for the number of frames of the reference video. Then, for a randomly sampled timestep $t = \frac{f-1}{F}$, we predict the control Gaussians' movement using MLP $\mathbf{\Psi}$. To be noted, since the control Gaussians are spherical, we only need to compute their new position as follows:

$$p_i^t = p_i + T_i^t, \tag{5}$$

where p_i is the position of C_i in the canonical space. After obtaining the deformed object at timestep t, we project it from the reference view to get $\hat{I}^t$ following Equ. (2), and compute the reconstruction loss as:

$$L_{ref} = \left\| \hat{I}^t - I_r^f \right\|_2^2. \tag{6}$$

To better leverage the information from the reference images, we additionally introduce a mask loss term:

$$L_{mask} = \left\| \alpha^t - \mathbf{M}_r^f \right\|_2^2, \tag{7}$$

where $\mathbf{M}_r$ is the foreground mask of the reference image, and α^t is the accumulated opacity obtained during rendering. As for novel view optimization, we sample B viewpoints randomly within the pitch angle range of $-ver$ to ver degrees and the yaw angle range of -180 to 180 degrees, and compute the average

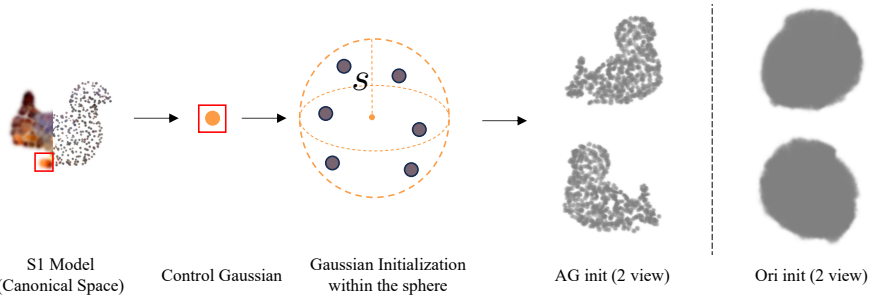

Fig. 3: Illustration of Adaptive Gaussian (AG) initialization. s is the scaling parameter of control Gaussians in the coarse stage. *Ori init* represents randomly initializing all the dense Gaussians within a sphere in the canonical space.

SDS loss following Equ. (1). The overall objective in this stage is a weighted combination of the above three losses:

$$L_{total} = \lambda_{ref} L_{ref} + \lambda_{mask} L_{mask} + \lambda_{SDS} L_{SDS}. \tag{8}$$

In this stage, we perform densification and pruning following 3DGS [17] in the first n iterations. Then we sample M (same amount as initialization) control Gaussians utilizing Farthest Point Sampling (FPS) [40] and continue training without densification in the remaining procedure.

3.3 Fine Stage: Dense Gaussians Optimization

After the coarse stage, we get a reasonable estimation for the dynamic 3D object's motion and shape. Then in the fine stage, we aim to optimize the texture details to match the source video and to further refine the motion and shape for better fidelity. To be noted, the explicit control Gaussians in the coarse stage have transitioned to implicit control points in this stage, denoted as $C_i : (p_i, o_i)$, where o_i is the control radius of each control point, initialized with the scaling parameter s of control Gaussians in the coarse stage.

Adaptive Gaussian Initialization. In this stage, dense Gaussians are driven by neighboring sparse control points, as illustrated in Equ. (3), (4). We empirically find that the dense Gaussian initialization significantly impacts the quality of the final result. One straightforward initialization approach is to initialize the dense Gaussians uniformly within a sphere in the canonical space as in the coarse stage. However, we observe that this initialization fails to generate promising results. The target object is prone to increased thickness, and issues such as diminished texture and positional drift may arise (as shown in Fig. 7).

Instead, we propose Adaptive Gaussian (AG) initialization based on the learned control Gaussians. As shown in Fig. 3, we have learned M control Gaussians in the coarse stage with the same scaling parameter s. Then, for each control Gaussian, we consider it as a sphere with a radius of s, and randomly

initialize K Gaussians within it following DreamGaussian [53]. In total, we get $N = M \times K$ Gaussians as initialization. As shown in Fig. 3, our designed initialization approach successfully inherits the shape and motion modeled in the coarse stage. The dense Gaussians are distributed near the surface of the object, which facilitates the subsequent optimization. In comparison, if directly optimizing all the dense Gaussians uniformly within a sphere in the canonical space, the deformed shape misaligns with that in the coarse stage.

Gaussian Alignment Loss. Even with a good dense Gaussian initialization, the shape and motion of the dynamic 3D object are still prone to degeneration in the later phases of training. The main reason is that: when employing Score Distillation Sampling (SDS) [38] to distill prior knowledge of novel views from Zero123 [29], a larger noise timestep biases SDS towards ensuring the plausibility of the overall shape. Conversely, with a smaller timestep, SDS tends to focus more on optimizing textures, while its capability to preserve shape diminishes.

To allow refining texture without degrading motion and shape, we propose the Gaussian Alignment (GA) loss as an additional constraint. At the beginning of this stage, we preserve the control Gaussians' parameters and the deformation MLP to query the initial positions (denoted as $\overline{p}^t$) of those control points at random timestep t. Then, we compute the Gaussian Alignment loss as:

$$L_{GA} = \left\| p^t - \overline{p}^t \right\|_2^2, \tag{9}$$

where p^t denotes the position of the current control point at timestep t. Although the proposed GA loss is quite simple, it can effectively mitigate the shape degeneration issue encountered during the latter training procedure. The Chamfer loss is another commonly used metric for constraining the distances between point clouds. However, compared to GA loss, we observe that the Chamfer loss can sometimes result in the current control points aggregating towards certain target points, thereby compromising the uniform distribution of the control points (See Sec. 4.4 of *Supp.* for the comparison of GA loss and the Chamfer loss).

In this stage, the overall training objective is formulated as follows:

$$L_{total} = \lambda_{ref} L_{ref} + \lambda_{mask} L_{mask} + \lambda_{SDS} L_{SDS} + \lambda_{GA} L_{GA}. \tag{10}$$

3.4 Motion Transfer Application

Utilizing our video-to-4D framework, we can successfully extract the resultant dynamic 3D object as well as its motion represented by a set of moving control points. Leveraging this capability, we devise an application tailored for motion transfer predicated on the trajectories of these sparse control points. This application aims to synthesize dynamic objects of distinct entities that exhibit identical motion patterns, all instantiated through text descriptions.

As shown in Fig. 4, we employ the same initialization method as outlined in Sec. 3.3. During training, we fix the parameters of the learned control points

Supp.: supplementary file.

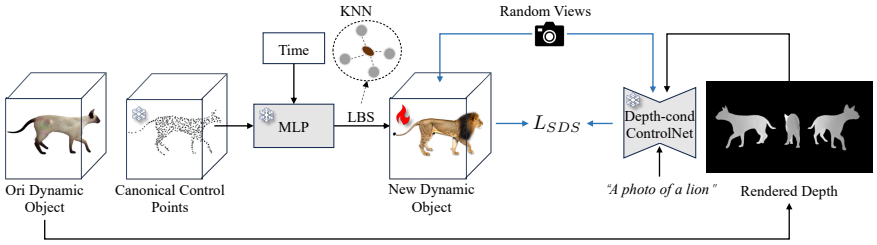

Fig. 4: Illustration of the pipeline for our motion transfer application.

as well as those of the deformation MLP, enabling us to preserve the motion of the target dynamic object and acquire its appearance from any viewpoint at any given time. To prevent the degeneration of motion and shape due to the displacement of dense Gaussians away from the control points during optimization, we elect to use a depth-condition ControlNet [71] as supervision, and query the depth from the dynamic object resultant from the video-to-4D pipeline. Since no ground truth reference is involved, we only utilize SDS [38] loss for optimization. We find that a small guidance scale can lead to plausible shape changes without degenerating the learned motion. Therefore, we first utilize a reduced guidance scale during the initial phase of the training regimen to capture the shape of the new entity that is congruent with the textual input, and its blurring appearance. Then, we save the intermediate dense Gaussians and employ the updated depth information as new conditions. In the remaining training period, we increase the guidance scale and decrease the noise scale of SDS to refine the texture, culminating in the final outcome. Please refer to Sec. 4.1 for more details.

4 Experiments

4.1 Experimental Settings

Implementation Details. We employ an MLP with a similar structure utilized in SC-GS [14] to predict the motion of control points. The MLP receives the positional embeddings [33] of time and control points' positions as input, with frequencies of 6 and 10. In the coarse stage, we initialize $M = 512$ control Gaussians uniformly within a sphere with the same scaling parameter. We perform densification and pruning following 3DGS [17] for the first 1,000 iterations with an interval of 100. Then we perform Farthest Point Sampling (FPS) [40] to sample M control Gaussians and train for another 500 iterations without densification. In the fine stage, we optimize all the parameters of control points, MLP, and dense Gaussians together. During training, we sample 4 random camera poses and the fixed pose of time t at a fixed radius of 2, with the azimuth in [-180, 180] degrees and elevation in [-ver, ver] degrees, where $ver = 30$. We set the loss weight of $\lambda_{ref}, \lambda_{mask}, \lambda_{SDS}, \lambda_{GA}$ to $5000.0, 500.0, 1.0, 10000.0$ by default. As for our motion transfer application, during the initial 1,000 iterations,

it is optimized with a guidance scale of 7.5 using the dynamic object's depth obtained from the video-to-4D pipeline as a condition. Subsequently, the depth from the intermediate result is utilized as the new condition, and we continue to optimize the remaining 1,000 iterations with a guidance scale of 30.0. We will release the source code later. Please refer to Sec. 1 of *Supp.* for more details.

Dataset. For fair comparisons, we utilize the dataset from Consistent4D [16] for evaluations. The dataset consists of 12 synthetic and 12 in-the-wild videos, all captured with a stationary camera oriented perpendicularly to the dynamic objects. Each video has 32 frames and spans around 2 seconds.

Evaluation Metrics. We have summarized the criteria for evaluating the quality of video-to-4D generation into three main aspects: alignment with the reference video, spatio-temporal consistency, and motion fidelity. Following NeRF [33], we utilize PSNR and SSIM [61] as measurements of the reference view alignment. We also add LPIPS [72] as a view quality metric, which is akin to Consistent4D [16]. As for multi-view (spatial) consistency, we adopt the commonly used CLIP [42] score to measure the visual similarity of two different renderings. To evaluate the temporal consistency and motion fidelity of the generated dynamic object, we follow [64] to utilize RAFT [55] to compute the optical flows of the reference image sequence, and warp the reference-view projections to calculate the temporal error (Temp), which can effectively reflect the temporal consistency under the reference view and motion accuracy (check Sec.1 of *Supp.* for details of the temporal error metric). We also include human evaluation, which is more representative in generation tasks. To do so, we randomly choose 10 videos to train the compared methods and render the dynamic objects from the reference view and a random novel view. Then, we invite 20 participants to select their preferred reference and novel view videos based on the reference view alignment, spatio-temporal consistency, and motion fidelity. Overall, we get $10 \times 20 = 200$ votes for reference view and novel view, respectively. Then, we calculate the percentage of votes as a measurement for user preference, as shown in Fig. 6.

4.2 Comparisons

To evaluate the effectiveness of the proposed video-to-4D framework, we compare our method with the only two open-source methods: Consistent4D [16] and 4DGen [69]. We train the dataset provided by Consistent4D with these methods and conduct comprehensive evaluations of the outcomes. All the results of Consistent4D and 4DGen are obtained using their official code and settings. Qualitative and quantitative comparisons are shown in Fig. 5 and Tab. 5.

Qualitative Comparison. As demonstrated in Fig. 5, we show four instances (*skull*, *triceratops*, *elephant*, *egret*) and the results generated by the compared methods. Consistent4D [16] generates results with artifacts and color distortions in some cases (*triceratops* and *egret*), and there is also a noticeable temporal discontinuity between frames of the same viewpoint, which is attributable to the limited capacity of low-resolution NeRF [33] and the diversity of time-varying NeRF under single view conditions. As for 4DGen [69], the generated dynamic objects often exhibit minor motion and present discernible discrepancies from the

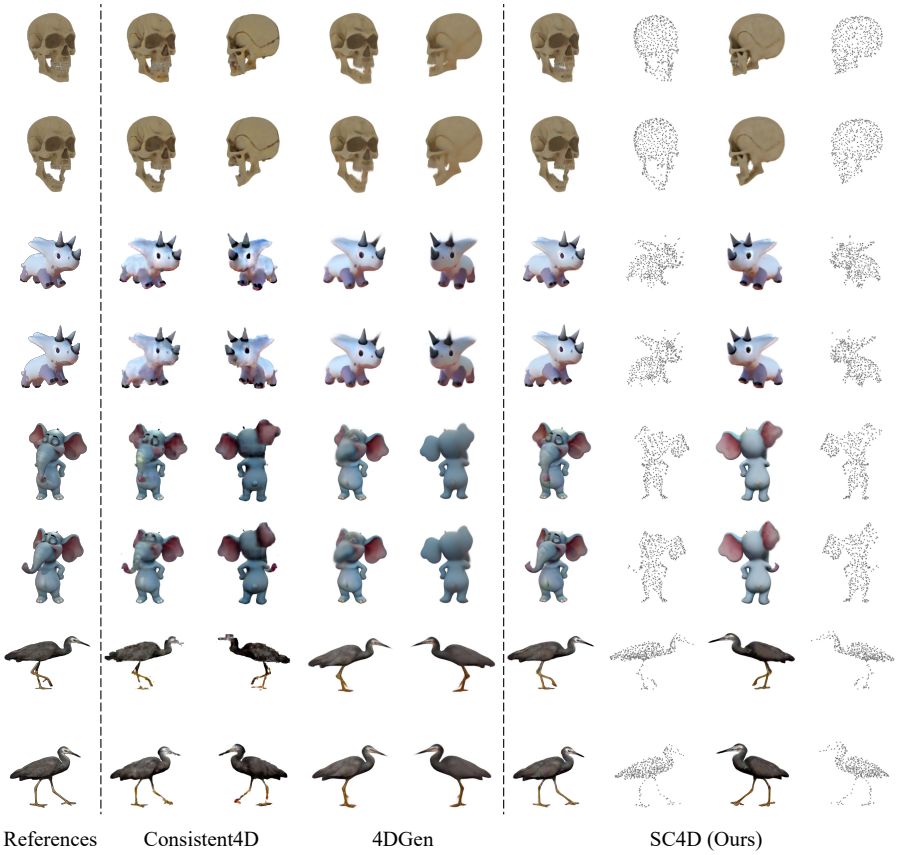

Fig. 5: Qualitative Comparisons. We compare our method with Consistent4D [16] and 4DGen [69]. For each instance, we render two viewpoints at two timesteps. We also visualize the sparse control points to show their correspondence with dense Gaussians.

reference video (*skull* and *triceratops*). Regions with large motion are subject to blurring or may even experience loss of detail (nose of *elephant*, claw of *egret*). In comparison, the results generated by our method surpass the compared methods in terms of alignment with the reference video, spatio-temporal consistency, and motion fidelity. This is also evidenced by the visualized sparse control points, from which it can be discerned that our method has learned a set of evenly distributed control points that accurately capture the dynamics and shape of the subject. Please check Sec. 5 of *Supp.* for more qualitative comparisons.

Quantitative Comparison. To quantitatively evaluate the performance of the compared methods, we randomly choose 10 reference videos from the dataset provided by Consistent4D [16], and optimize the dynamic objects using the compared methods, respectively. After optimization, we render one input view and

Table 1: Quantitative comparison of different methods. Temp represents the temporal error introduced in Sec. 4.1.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP ↑	Temp ↓	Training time ↓
Consistent4D [16]	23.97	0.91	0.09	0.89	0.0089	1.9h
4DGen [69]	21.80	0.90	0.10	0.87	0.0089	3.0h
SC4D (Ours)	**29.50**	**0.95**	**0.08**	**0.90**	**0.0081**	**1.0h**

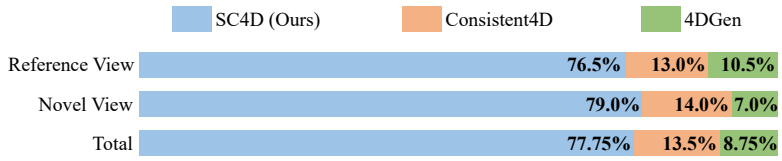

Fig. 6: User preference for video-to-4D generation methods.

four testing views (azimuth degrees: [72, 144, 216, 288], elevation degree: [0]) for every timestep of the reference video to calculate the metrics mentioned in Sec. 4.1. As reported in Tab. 1, our method outperforms Consistent4D [16] and 4DGen [69] in reference view alignment (PSNR, SSIM, LPIPS), multi-view consistency (CLIP), temporal consistency and motion fidelity (Temp), revealing the effectiveness of SC4D. As depicted in Fig. 6, user study also reveals that our proposed SC4D is more favored in human evaluations, in which the novel view preference metric also reveals the remarkable spatio-temporal consistency and motion fidelity of SC4D. Besides, our method exhibits a significant advantage in terms of training duration. All experiments are conducted on a single Tesla V100 GPU with 32 GB of graphics memory. We also utilize the test set and metrics in Consistent4D [16] to evaluate the compared methods. The results align with the conclusion drawn above. We attach them in Sec. 3 of *Supp.*.

4.3 Ablation Studies

As mentioned in Sec. 1, we propose the Adaptive Gaussian (AG) initialization and Gaussian Alignment (GA) loss to alleviate the shape degeneration issue in the fine stage. As shown in Fig. 7, without GA loss and AG initialization, the generated object is prone to suffer from over-thickness and texture blurring problems. Besides, the control points scatter, which indicates that the learned shape and motion in the coarse stage also degenerate significantly. When training with GA loss, the general shape of control points can be preserved, yet there can be observable instances of increased thickness. Additionally, when a viewpoint is excessively close to the camera, a corresponding attenuation in texture detail is also noticeable. Upon utilizing AG initialization, our method has experienced a pronounced enhancement in its final output. The dynamic objects generated

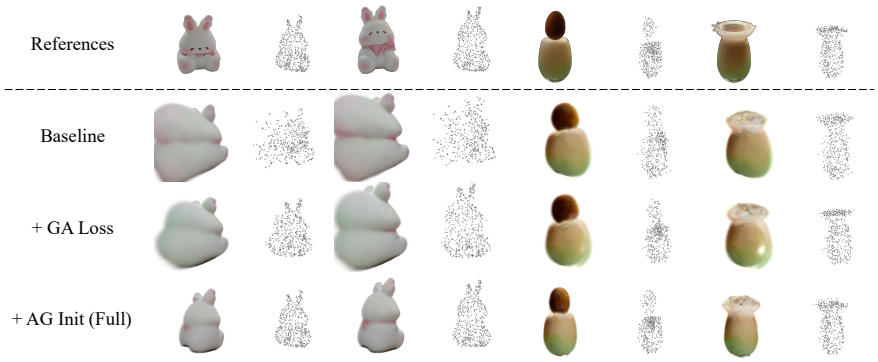

Fig. 7: Ablation studies of the proposed GA loss and AG initialization. In the first row, we show the source frames and control points from the coarse stage as references.

exhibit tangible advancements in both motion fidelity and shape plausibility. Moreover, the textural details have also undergone considerable refinement.

Table 2: Quantitative evaluations of the proposed GA loss and AG initialization.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP ↑	Temp ↓
Baseline	29.81	0.95	0.10	0.82	**0.0016**
+ GA Loss	30.19	0.96	0.09	0.83	**0.0016**
+ AG Init (Full)	**31.35**	**0.96**	**0.08**	**0.89**	**0.0016**

Tab. 2 also illustrates the effectiveness of our proposed techniques. The most noticeable improvement in metrics is observed in the CLIP score, indicating a substantial enhancement in the quality of the generated novel views. This further reveals the effectiveness of the proposed GA loss and AG initialization in mitigating the challenges associated with shape degeneration. Please refer to Sec. 4 of *Supp.* for more ablation studies.

4.4 Application Results

As mentioned in Sec. 3.4, we design an innovative application that is capable of flexibly generating diverse 4D entities with the same motion, based on control points learned during our video-to-4D pipeline, in conjunction with a depth-condition ControlNet [71] that operates upon textual descriptions.

In Fig. 8, we sequentially present: the reference frames, video-to-4D results generated using SC4D and corresponding control points visualizations, and the new entities with identical motion synthesized based on text descriptions. From the figure, it is observable that the newly generated 4D objects possess vivid

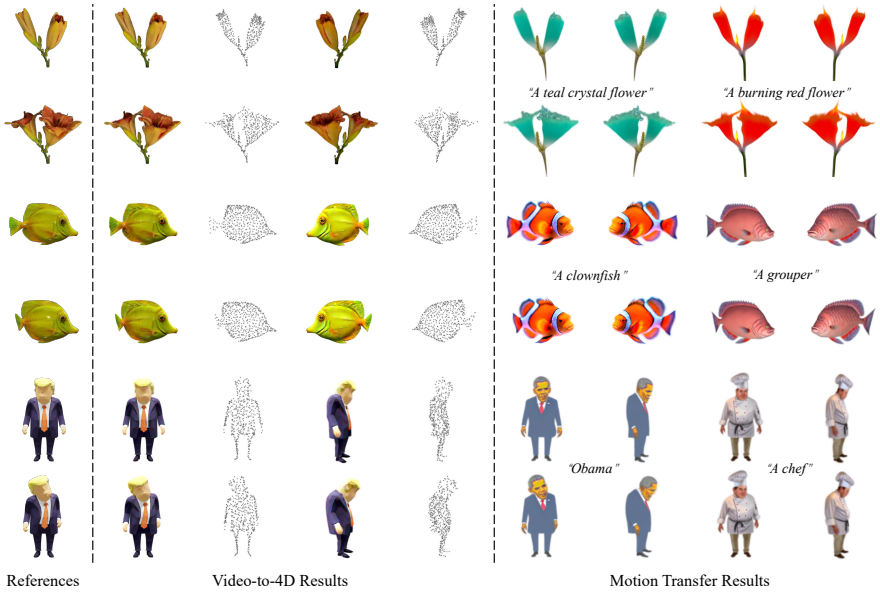

Fig. 8: Motion transfer application of SC4D. The prompt is attached between lines.

textures and align well with the motion patterns of the reference video. Moreover, the synthesized dynamic objects exhibit appreciable variations in shape, yet the motion remains coherent, which further illustrates the robustness and flexibility of our method. Please refer to Sec. 6 of *Supp.* for more application examples.

5 Limitations

The main limitations of our method are twofold: First, our approach relies on models such as Zero123 [29] to provide novel view information, and the capability of these viewpoint synthesis models is still limited, often underperforming on many complex objects. Second, similar to existing video-to-4D methods [16, 36, 69], we have not taken into account the 4D generation under moving camera scenarios, which will be one of the directions for our future research.

6 Conclusion

In this work, we propose a sophisticated video-to-4D pipeline, named SC4D, which decouples the motion and appearance of dynamic objects as sparse control points and dense 3D Gaussians. SC4D excels over contemporary approaches in generating dynamic objects with better reference view alignment, spatio-temporal consistency, and motion fidelity. Moreover, we craft an innovative application utilizing the control points learned by SC4D, which allows seamless motion transfer onto new entities, as directed by textual descriptions.

Acknowledgements

This work was supported by the National Science Fund for Distinguished Young Scholars of China (Grant No.62225603).

References

1. Bahmani, S., Skorokhodov, I., Rong, V., Wetzstein, G., Guibas, L., Wonka, P., Tulyakov, S., Park, J.J., Tagliasacchi, A., Lindell, D.B.: 4d-fy: Text-to-4d generation using hybrid score distillation sampling. arXiv preprint arXiv:2311.17984 (2023)
2. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 130–141 (2023)
3. Chen, G., Wang, W.: A survey on 3d gaussian splatting. arXiv preprint arXiv:2401.03890 (2024)
4. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. arXiv preprint arXiv:2303.13873 (2023)
5. Chen, Z., Wang, F., Liu, H.: Text-to-3d using gaussian splatting. arXiv preprint arXiv:2309.16585 (2023)
6. Das, D., Wewer, C., Yunus, R., Ilg, E., Lenssen, J.E.: Neural parametric gaussians for monocular non-rigid object reconstruction. arXiv preprint arXiv:2312.01196 (2023)
7. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. Advances in Neural Information Processing Systems **36** (2024)
8. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13142–13153 (2023)
9. Fang, J., Yi, T., Wang, X., Xie, L., Zhang, X., Liu, W., Nießner, M., Tian, Q.: Fast dynamic radiance fields with time-aware neural voxels. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)
10. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12479–12488 (2023)
11. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
12. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5712–5721 (2021)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Huang, Y.H., Sun, Y.T., Yang, Z., Lyu, X., Cao, Y.P., Qi, X.: Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. arXiv preprint arXiv:2312.14937 (2023)

15. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 867–876 (2022)
16. Jiang, Y., Zhang, L., Gao, J., Hu, W., Yao, Y.: Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. arXiv preprint arXiv:2311.02848 (2023)
17. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
18. Kratimenos, A., Lei, J., Daniilidis, K.: Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting. arXiv preprint arXiv:2312.00112 (2023)
19. Lee, H.H., Chang, A.X.: Understanding pure clip guidance for voxel grid nerf models. arXiv preprint arXiv:2209.15172 (2022)
20. Li, J., Tan, H., Zhang, K., Xu, Z., Luan, F., Xu, Y., Hong, Y., Sunkavalli, K., Shakhnarovich, G., Bi, S.: Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. arXiv preprint arXiv:2311.06214 (2023)
21. Li, W., Chen, R., Chen, X., Tan, P.: Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. arXiv preprint arXiv:2310.02596 (2023)
22. Li, Z., Chen, Z., Li, Z., Xu, Y.: Spacetime gaussian feature splatting for real-time dynamic view synthesis. arXiv preprint arXiv:2312.16812 (2023)
23. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6498–6508 (2021)
24. Liang, Y., Khan, N., Li, Z., Nguyen-Phuoc, T., Lanman, D., Tompkin, J., Xiao, L.: Gaufre: Gaussian deformation fields for real-time dynamic novel view synthesis. arXiv preprint arXiv:2312.11458 (2023)
25. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 300–309 (2023)
26. Lin, Y., Han, H., Gong, C., Xu, Z., Zhang, Y., Li, X.: Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. arXiv preprint arXiv:2309.17261 (2023)
27. Ling, H., Kim, S.W., Torralba, A., Fidler, S., Kreis, K.: Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. arXiv preprint arXiv:2312.13763 (2023)
28. Liu, M., Xu, C., Jin, H., Chen, L., Varma T, M., Xu, Z., Su, H.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. Advances in Neural Information Processing Systems **36** (2024)
29. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9298–9309 (2023)
30. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453 (2023)
31. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. arXiv preprint arXiv:2308.09713 (2023)
32. Melas-Kyriazi, L., Laina, I., Rupprecht, C., Vedaldi, A.: Realfusion: 360deg reconstruction of any object from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8446–8455 (2023)

33. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
34. Mohammad Khalid, N., Xie, T., Belilovsky, E., Popa, T.: Clip-mesh: Generating textured meshes from text using pretrained image-text models. In: SIGGRAPH Asia 2022 conference papers. pp. 1–8 (2022)
35. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022)
36. Pan, Z., Yang, Z., Zhu, X., Zhang, L.: Fast dynamic 3d object generation from a single-view video. arXiv preprint arXiv:2401.08742 (2024)
37. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5865–5874 (2021)
38. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
39. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10318–10327 (2021)
40. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
41. Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.Y., Skorokhodov, I., Wonka, P., Tulyakov, S., et al.: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. arXiv preprint arXiv:2306.17843 (2023)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
43. Ren, J., Pan, L., Tang, J., Zhang, C., Cao, A., Zeng, G., Liu, Z.: Dreamgaussian4d: Generative 4d gaussian splatting. arXiv preprint arXiv:2312.17142 (2023)
44. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
45. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
46. Seo, J., Jang, W., Kwak, M.S., Ko, J., Kim, H., Kim, J., Kim, J.H., Lee, J., Kim, S.: Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. arXiv preprint arXiv:2303.07937 (2023)
47. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. Advances in Neural Information Processing Systems **34**, 6087–6101 (2021)
48. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023)
49. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)

50. Singer, U., Sheynin, S., Polyak, A., Ashual, O., Makarov, I., Kokkinos, F., Goyal, N., Vedaldi, A., Parikh, D., Johnson, J., et al.: Text-to-4d dynamic scene generation. arXiv preprint arXiv:2301.11280 (2023)
51. Sumner, R.W., Schmid, J., Pauly, M.: Embedded deformation for shape manipulation. In: ACM siggraph 2007 papers, pp. 80–es (2007)
52. Sun, J., Zhang, B., Shao, R., Wang, L., Liu, W., Xie, Z., Liu, Y.: Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. arXiv preprint arXiv:2310.16818 (2023)
53. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023)
54. Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., Chen, D.: Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. arXiv preprint arXiv:2303.14184 (2023)
55. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 402–419. Springer (2020)
56. Tretschk, E., Tewari, A., Golyanik, V., Zollhöfer, M., Lassner, C., Theobalt, C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12959–12970 (2021)
57. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12619–12629 (2023)
58. Wang, P., Shi, Y.: Imagedream: Image-prompt multi-view diffusion for 3d generation. arXiv preprint arXiv:2312.02201 (2023)
59. Wang, X., Wang, Y., Ye, J., Wang, Z., Sun, F., Liu, P., Wang, L., Sun, K., Wang, X., He, B.: Animatabledreamer: Text-guided non-rigid 3d model generation and reconstruction with canonical score distillation. arXiv preprint arXiv:2312.03795 (2023)
60. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in Neural Information Processing Systems **36** (2024)
61. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
62. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. arXiv preprint arXiv:2310.08528 (2023)
63. Wu, T., Zhong, F., Tagliasacchi, A., Cole, F., Oztireli, C.: D^ 2nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. Advances in Neural Information Processing Systems **35**, 32653–32666 (2022)
64. Wu, Z., Zhu, Z., Du, J., Bai, X.: Ccpl: contrastive coherence preserving loss for versatile style transfer. In: European Conference on Computer Vision. pp. 189–206. Springer (2022)
65. Xu, J., Wang, X., Cheng, W., Cao, Y.P., Shan, Y., Qie, X., Gao, S.: 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. pp. 20908–20918 (2023)

66. Yang, Z., Yang, H., Pan, Z., Zhu, X., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. arXiv preprint arXiv:2310.10642 (2023)
67. Yang, Z., Yang, H., Pan, Z., Zhu, X., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. arXiv preprint arXiv:2310.10642 (2023)
68. Yi, T., Fang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., Wang, X.: Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. arXiv preprint arXiv:2310.08529 (2023)
69. Yin, Y., Xu, D., Wang, Z., Zhao, Y., Wei, Y.: 4dgen: Grounded 4d content generation with spatial-temporal consistency. arXiv preprint arXiv:2312.17225 (2023)
70. Yu, C., Zhou, Q., Li, J., Zhang, Z., Wang, Z., Wang, F.: Points-to-3d: Bridging the gap between sparse points and shape-controllable text-to-3d generation. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 6841–6850 (2023)
71. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
72. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
73. Zhao, Y., Yan, Z., Xie, E., Hong, L., Li, Z., Lee, G.H.: Animate124: Animating one image to 4d dynamic scene. arXiv preprint arXiv:2311.14603 (2023)
74. Zheng, Y., Li, X., Nagano, K., Liu, S., Hilliges, O., De Mello, S.: A unified approach for text-and image-guided 4d scene generation. arXiv preprint arXiv:2311.16854 (2023)
75. Zhu, J., Zhuang, P.: Hifa: High-fidelity text-to-3d with advanced diffusion guidance. arXiv preprint arXiv:2305.18766 (2023)