

## T2Bs: Text-to-Character Blendshapes via Video Generation

Jiahao Luo<sup>1</sup> Chaoyang Wang<sup>2</sup> Michael Vasilkovsky<sup>2</sup> Vladislav Shakhrai<sup>2</sup> Di Liu<sup>3</sup>  
 Peiye Zhuang<sup>2</sup> Sergey Tulyakov<sup>2</sup> Peter Wonka<sup>4</sup> Hsin-Ying Lee<sup>2</sup> James Davis<sup>1</sup> Jian Wang<sup>2</sup>  
<sup>1</sup>University of California, Santa Cruz <sup>2</sup>Snap Inc. <sup>3</sup>Rutgers University <sup>4</sup>KAUST  
 {jluo53, davisje}@ucsc.edu, jwang4@snapchat.com



Figure 1. Text-to-character blendshapes (T2Bs) is capable of creating animatable blendshapes to synthesize diverse expressions of a virtual character generated solely from text prompts.

### Abstract

We present T2Bs, a framework for generating high-quality, animatable character head morphable models from text by combining static text-to-3D generation with video diffusion. Text-to-3D models produce detailed static geometry but lack motion synthesis, while video diffusion models generate motion with temporal and multi-view geometric inconsistencies. T2Bs bridges this gap by leveraging deformable 3D Gaussian splatting to align static 3D assets with video outputs. By constraining motion with static geometry and employing a view-dependent deformation MLP, T2Bs (i) outperforms existing 4D generation methods in accuracy and expressiveness while reducing video artifacts and view inconsistencies, and (ii) reconstructs smooth, coherent, fully registered 3D geometries designed to scale for building morphable models with diverse, realistic facial motions. This enables synthesizing expressive, animatable character heads that surpass current 4D generation techniques. Project Page: <https://snap-research.github.io/T2Bs/>.

### 1. Introduction

The creation of animatable 3D virtual character head avatars [19, 50, 52, 66, 81, 111] has become increasingly important due to their wide-ranging applications in social media, gaming, and entertainment. These avatars enable expressive, engaging, and highly personalized digital representations, enhancing both creative storytelling and interactive experiences. 3D human face models [4, 14] have been extensively studied [6, 26, 36, 48] and widely applied in realistic [49, 60, 67, 82, 93] and virtual character [19, 52, 81, 111] animation. However, the development of virtual character head models that deviate significantly from the human head distribution, such as animals, remains relatively underexplored. The unique challenges in this domain arise from the complexity of anatomy, the vast diversity of species, and the stylization often required for cartoon-like representations. Creating such virtual character models remains a labor-intensive process, typically requiring skilled artists and significant manual effort [33].

Recent advancements in text-to-3D generation, particu-

larly those using diffusion models [9, 58, 74, 86, 102, 112], have shown remarkable progress in creating high-quality static virtual character, including meshes and Gaussian representations [30, 41, 43, 68, 69]. These methods are highly effective in generating detailed and realistic 3D geometry. However, their primary limitation lies in their inability to generate **diverse motion dynamics** that encompass the full range of movements a virtual character could exhibit.

Video diffusion models, including image-to-video frameworks [51, 57, 78, 101], have shown promise in generating motion information. Several recent methods propose solutions on how to employ video models to generate dynamic 3D objects [37, 62, 63, 92, 95] combining techniques from video generation and 3D reconstruction. There are two main drawbacks we would like to improve upon. First, the methods have inherent limitations for generalization (model trained with limited 4D data) or efficiency (e.g., using score distillation sampling). Second, the methods only tackle the problem of generating a single animation. This is not sufficient to obtain a 3D animatable model.

To address the first limitation, we propose a new framework that builds on recent 4D video generation techniques [80, 90]. 4D video methods directly generate a multi-view video in the form of a grid of images. This enables a cleaner separation of motion generation and 3D reconstruction techniques and, ultimately, much higher visual quality. However, 4D video still faces certain challenges, such as color inconsistencies, geometric distortions, and limited control over viewpoints and fine-grained details. To address the 3D inconsistency issues inherent in existing 4D generation methods, we introduce View-Conditioned Deformable Gaussian Splatting (VCDGS), which maintains structural coherence across views.

To address the second limitation, we use our framework multiple times with different text prompts to generate a larger variety of motions for a 3D character. Each initial text prompt results in one 4D video with a corresponding 3D mesh for each frame. Using 3D meshes in a variety of poses, derived from many video sequences, we construct a blendshape model for the generated character that can be controlled to fit a variety of facial expressions. We evaluate our model and demonstrate its applicability in expression retargeting.

Our key contributions are as follows:

- We introduce a scalable pipeline via multi-view video diffusion that overcomes data limitations, allowing the creation of blendshape models for a wide range of virtual characters from text prompts.
- We propose View-Conditioned Deformable Gaussian Splatting (VCDGS) to address the 3D inconsistency issue in existing 4D video methods, ensuring structurally coherent and high-quality deformations.
- We evaluate the expressiveness of our generated blend-

shape models and the potential for downstream animation and retargeting applications.

## 2. Related Works

**3D Generation:** Text/image-to-3D generation has advanced significantly with diffusion models [22, 64]. DreamFusion [58] pioneered high-quality text-to-3D generation by introducing Score Distillation Sampling (SDS) to leverage pretrained text-image diffusion models. Subsequent works [9, 11, 38, 39, 59, 73, 74, 76, 86, 102, 112] enhanced DreamFusion by refining SDS variants and incorporating diverse 3D representations [30, 41–43, 55, 68, 69] beyond NeRF [53]. However, these methods remain computationally expensive due to iterative optimization during inference. To accelerate 3D generation, recent approaches adopt a two-stage pipeline: (1) training image/video diffusion models to generate multi-view images [16, 35, 44, 70, 71, 78, 97], followed by (2) feedforward reconstruction methods [24, 96, 113] that rapidly synthesize 3D assets. Further, single-stage methods [7, 10, 12, 25, 29, 46, 54, 56, 84, 106, 110] train diffusion models to directly generate explicit 3D representations. In this work, we build on existing 3D generation methods to create high-quality static 3D characters, while addressing an orthogonal problem: generating blendshapes for character morphing.

**4D Generation:** 4D generation is an emerging field with diverse definitions. Some approaches focus on generating videos with camera controls [3, 20, 87, 88, 100], while others generate a space-time video grid [32, 80, 95, 104, 107]. Our work aligns most closely with research that directly generates 4D representations, such as deformable Gaussian splats. This line of work [2, 27, 40, 62, 72, 103, 109] leverages priors from text-to-image [64, 65], text-to-multiview [44, 71], and text-to-video [8, 18, 23] models. However, most prior works [2, 27, 37, 40, 62, 63, 72, 92, 109] generate independent 4D objects or scenes without constructing an animatable model that is capable of interactively generating new expressions or postures. The most related works to ours [75, 105] use generative priors to create animatable human avatars, with morphable models of the human body [47] or face [36] pre-trained on real human motion capture data. In contrast, our work is the first to learn a morphable model solely from generated data.

**Human Head Avatars:** Recent methods for creating head avatars utilize monocular or multi-view video inputs to synthesize new expressions. Among these, GaussianAvatars [60] rigs 3D Gaussians to the FLAME [36] face tracking framework by anchoring them to the triangular facets of the mesh. Similarly, SplattingAvatar [67] integrates 3D Gaussians into mesh models and predicts displacements along the normal direction. FlashAvatar [93] defines Gaussians in a uniform FLAME UV space and directly predicts per-Gaussian deformation from monocular

video inputs. GaussianHead [82] employs tri-plane representations and motion fields to simulate continuous geometric changes in heads, rendering rich textures, including skin and hair. While these methods achieve impressive results, they rely heavily on face tracking and pre-existing human head models, limiting their applicability to scenarios involving characters that deviate significantly from human-like geometry. This constraint poses challenges for applications involving virtual characters or animal-based models, which fall outside the distribution of standard human datasets.

### 3. Methods

To build blendshape models from input text prompts, our pipeline consists of several stages. First, given a text prompt describing a virtual character, we use an off-the-shelf text-to-3D generator to create a textured 3D mesh. Next, a video diffusion model animates the 3D assets, generating a set of videos based on various text prompts describing different expressions. These text prompts are automatically generated using a prompt template, which substitutes the character description and the corresponding motion. In the subsequent capture stage, we employ a combination of advanced multi-view video generation models and robust reconstruction algorithms to capture the deformations of the 3D mesh from the generated videos. Finally, blending bases are computed from the captured deformed shapes. The pipeline is illustrated in Fig. 2. The following sections provide a detailed explanation of each stage of the pipeline.

#### 3.1. Video Generation of Character Animations

##### 3.1.1. Obtaining Diverse Expressions

Starting with a textured 3D mesh obtained from a 3D generator, we collect a dataset of videos showcasing different expressions of the same character. To achieve this, we first create a pool of text prompts describing various expressions, such as ‘blinking eyes,’ ‘frowning,’ and ‘opening mouth.’ Next, we render a frontal view of the 3D character, which serves as the conditional image input to a video diffusion model, which generates multiple videos corresponding to each prompt. To ensure the generated videos maintain a static camera pose and that the subject remains within the frame, we append “static camera, the camera is holding still” to the input prompts. These videos are image based only, often contain more than one animated component, and are not necessarily geometrically consistent, but they do contain a diverse range of expressions.

##### 3.1.2. Synchronized Multi-view Video Generation

The *monocular* videos generated in the previous step provide supervision on how different expressions should appear in the frontal view. We refer to this as the fixed-view video. However, learning a 3D morphable model from

frontal-view videos alone is ambiguous due to the under-constrained nature of the reconstruction problem. To address this, prior works have relied on either hand-crafted geometric or motion priors [31, 79, 83, 85], which fail to produce high-quality results for complex motions such as tongue movements, or on score distillation sampling (SDS) loss [62, 104], which is computationally expensive and typically requires several GPU hours to reconstruct even a short sequence. In this work, we explore the use of 4D video generation models capable of producing synchronized multi-view videos, offering more direct supervision for rendering expressions from multiple view points.

Specifically, the 4D video generation model produces a *space-time frame grid*,  $\mathcal{I}_{v,t}$ , where  $v$  and  $t$  are indices representing the viewpoint and time, respectively. The model takes two input videos: the first is the fixed-view video showing the expression changes of the virtual character in the frontal view, *i.e.*,  $[\mathcal{I}_{0,0}, \mathcal{I}_{0,1}, \dots, \mathcal{I}_{0,T}]$ , assuming the frontal view index is 0. The second video is a fixed-time video showing viewpoint changes. This is rendered using the 3D mesh, with the camera moving circularly around the subject, *i.e.*,  $[\mathcal{I}_{0,0}, \mathcal{I}_{1,0}, \dots, \mathcal{I}_{V,0}]$ . Based on these inputs, the model generates the remaining frames in the grid,  $\mathcal{I}_{v,t}$ ,  $\forall v > 0, t > 0$ . This process is repeated for the fixed-view video generated from each text prompt, to create a set of multi-view videos.

We experimented with two available 4D generation models, SV4D [95] and 4Real-Video [80]. We found that SV4D tends to produce blurrier results, particularly in the mouth and eye regions of the character. In contrast, 4Real-Video consistently generates more plausible results.

**Limitations of Multi-view Video Generation.** State-of-the-art 4D video model can generate visually consistent frames, however, directly applying vanilla 3D reconstruction yields noisy reconstruction due to several imperfections. First, the generated frames exhibit geometric *multi-view inconsistency* and do not strictly adhere to epipolar constraints. Additionally, because current 4D models process only short durations of time in a single pass, they lack long-term context conditioning. As a result, these models inevitably produce *inconsistent appearance* for regions that disappear and then reappear in the frame grid. To address these limitations, we propose a robust algorithm for 3D model reconstruction.

#### 3.2. Robust Reconstruction from Generated Videos

##### 3.2.1. Deformable 3D Gaussian Splats Representation

We represent the morphable model using deformable Gaussian splats, which offer high-quality rendering and fast performance. Specifically, we use static 3D Gaussian splats to represent the canonical model of the character. Each splat encodes its 3D position  $\mathbf{x} \in \mathbb{R}^3$ , orientation  $\mathbf{q} \in SO(3)$ , scale  $\mathbf{s} \in \mathbb{R}^3$ , and RGB color  $\mathbf{c} \in \mathbb{R}^3$ . To reduce reconstruc-

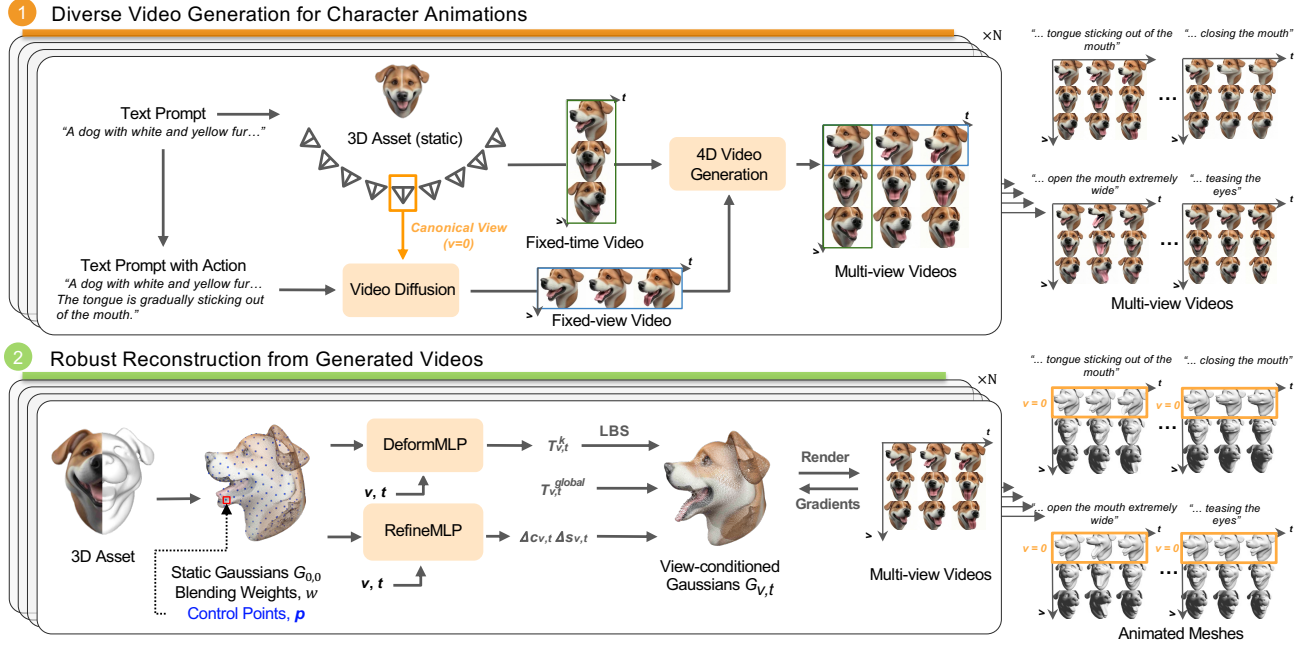


Figure 2. **Overview of T2Bs.** In the first part, we illustrate the generation of multi-view videos from text prompts. A static 3D mesh is first created using an off-the-shelf text-to-3D generator [94], followed by rendering a fixed-time video with the camera moving in a circular path. We define a canonical view ( $v=0$ ) and an augmented prompt to generate a fixed-view video. A 4D video generation method is then applied to produce multi-view videos. In the second part, starting from a static 3D asset, we define static Gaussians  $G_{0,0}$ , control points  $p$  and blending weights  $w$ . During deformation, we predict view-dependent transformations of control points to model local non-rigid deformations, along with global transformations to capture overall pose changes. We interpolate Gaussian positions and orientations with Linear Blend Skinning (LBS), with rendering optimized through image-space loss minimization. After training, we extract a mesh for each frame, defined in the canonical view ( $v=0$ ). We repeat this process with multiple prompts, and build a blendshape model (shown in Fig. 1 1st row) using hundreds of samples.

tion ambiguities, we omit higher-order spherical harmonics. Given the availability of the textured mesh, we initialize the canonical 3D Gaussian splats by cloning the vertex positions of the mesh, assigning the corresponding colors, and setting the opacity to 1. Subsequently, the scale and orientation of the splats are optimized by minimizing the rendering loss.

To deform the canonical 3D Gaussian splats, we adopted a linear blend skinning (LBS) formulation, e.g. [28, 98, 99]:

$$\mathbf{x}_t = \sum_k w_k \mathbf{T}_t^k \mathbf{x}, \quad (1)$$

where  $\mathbf{x}_t$  is the transformed position at time  $t$ ,  $w_k$  represents the blending weight associated with the  $k$ -th deformation component, and  $\mathbf{T}_t^k \in SE(3)$  denotes the corresponding rigid transformation.

We compute each control point  $\mathbf{p}_k$  and blending weight  $w_k$  in the rest pose, where  $\mathbf{p}_k$  is assigned via KNN and  $w_k$  via the Mahalanobis distance from each Gaussian to neighboring control points, using  $k = 2000$  control points as detailed in the supplementary material. While  $w_k$  remains fixed, the per-control-point transformations  $\mathbf{T}_t^k$  are

optimized by minimizing the following rendering loss:

$$\sum_{v,t} \mathcal{L}_{\text{huber}}(\mathcal{I}_{v,t} - \mathcal{I}'_{v,t}) + \mathcal{L}_{\text{LPIPS}}(\mathcal{I}_{v,t}, \mathcal{I}'_{v,t}), \quad (2)$$

which combines an image-space Huber loss and a feature-space LPIPS loss [108]. This loss compares the frames generated by the video model,  $\mathcal{I}_{v,t}$ , with the frames rendered by the deformable Gaussian splats,  $\mathcal{I}'_{v,t}$ .

### 3.2.2. View-Conditioned Deformable Gaussian Splatting

Directly optimizing the loss in Eq. (2) yields unsatisfactory reconstruction due to imperfections in the frames generated by the video model. To mitigate these imperfections and improve quality, we propose View-Conditioned Deformable Gaussian Splatting (VCDGS).

**View-Dependent Deformation for Multi-view Inconsistency.** First, we propose modeling *multi-view inconsistency* in the generated video frames as deformation. Specifically, we train a *DeformMLP* that incorporates both the time  $t$  and the view index  $v$  as inputs to predict the rigid transformation for each LBS component. This approach ensures that the transformation depends on both  $t$  and  $v$ , as described

below:

$$\mathbf{T}_{v,t}^k = \text{DeformMLP}(\mathbf{p}_k, v, t), \quad (3)$$

where  $\mathbf{p}_k \in \mathbb{R}^3$  denotes the position of the control point for the  $k$ -th deformation component. We note that the view-dependent transformations  $\mathbf{T}_{v,t}^k$  are utilized only during training, when it is necessary to match the inconsistent multi-view video data. During testing, we designate  $v = 0$  as the canonical frame and use  $\mathbf{T}_{0,t}^k$  to construct the final blendshape.

Additionally, we observe that directly training the DeformMLP can sometimes be unstable. To address this, we decompose the transformations into two parts: independent component-wise transformations predicted by DeformMLP and a shared rigid transformation representing the global  $SE(3)$  motion of the object. Specifically,

$$\tilde{\mathbf{T}}_{v,t}^k = \mathbf{T}_{v,t}^{\text{global}} \cdot \text{DeformMLP}(\mathbf{p}_k, v, t) \quad (4)$$

During the initial stages of training, we first fit the global transformation  $\mathbf{T}_{v,t}^{\text{global}}$  and then jointly optimize both the global transformation and the component-wise DeformMLP.

**Gaussian Refinement for Appearance Inconsistency.** To capture appearance changes in the same region of the character across different frames generated by the video model, we introduce RefineMLP. This module predicts offsets for the non-positional properties of the Gaussian splats, including color, scale, and orientation, to better align with the generated videos. Specifically,

$$\Delta \mathbf{c}_{v,t}, \Delta \mathbf{s}_{v,t}, \Delta \mathbf{q}_{v,t} = \text{RefineMLP}(\mathbf{x}, v, t) \quad (5)$$

Here,  $\Delta \mathbf{c}_{v,t}$ ,  $\Delta \mathbf{s}_{v,t}$ , and  $\Delta \mathbf{q}_{v,t}$  represent the refinement offsets for color, scale, and orientation, respectively. Similar to the view-dependent deformation, RefineMLP is used only during training and is not applied during inference nor in the construction of the final blendshape.

### 3.3. T2Bs Blendshape Model and Retargeting

A set of deformed Gaussian splats are generated by the reconstruction process described in Sec.3.2 for each video generated in Sec.3.1. These splats are associated with mesh vertices so a 3D mesh representation for each frame of each video sequence is directly obtainable. These meshes are useful for rendering the existing video sequences, however they are not yet a blendshape model suitable for producing new animations.

We apply principal component analysis (PCA) to the per-frame geometry of all video sequences to construct a set of orthogonal blendshapes, eliminating the need for manually selecting blendshapes from specific frames based on expression prompts. This set of blendshapes is sufficiently expressive to model all observed deformations.

To produce new animations we retarget from human facial expressions to the domain of our blendshape model. We use semantically meaningful landmarks [34] that correspond to a subset of the standard 68 human facial landmarks. However, our virtual characters, particularly animal characters, have geometric distributions that deviate significantly from human faces, making off-the-shelf human landmark detection methods unsuitable.

To address this, we first render the static asset in a frontal view and use VLM models [45, 61] to identify the left eye, right eye, and mouth regions. Next, we select eye and mouth landmarks along the edges of these segmentations at predefined angles. To map the selected 2D landmarks back to the static asset’s 3D space, we employ shadow mapping to locate the visible 3D points whose projections are closest to the identified 2D landmarks. Ultimately, we select 20 canonical landmarks that effectively cover the left eye, right eye, and mouth regions. We transfer motion from human landmarks to virtual character landmarks and then fit our orthogonal model to the deformed character landmarks.

## 4. Experiments

### 4.1. Implementation Details

When evaluating our method, we generate multi-view videos as described in Sec. 3.1, rendering the 3D asset at a resolution of  $512 \times 512$ , followed by cropping as required by video models [80, 101]. Each video prompt combines a predefined expression prompt (from a set of 20, e.g., those shown on the right side of Fig. 2 with more details in Supp. material) with the original character prompt and camera motion description. For each concatenated prompt, we generate three sets of multi-view videos and manually filter out cases with unnatural expressions, retaining those that either align with the prompt or appear naturally plausible. We train VCDGS for 60,000 iterations per set of multi-view videos. The entire process, including multi-view video generation and VCDGS optimization, takes approximately one hour on a single NVIDIA A100 GPU.

### 4.2. VCDGS Evaluation

**Comparison Baselines.** We compare the proposed View-conditioned Deformable Gaussian Splatting (VCDGS) with recent state-of-the-art 4D generation methods: *Dream-Gaussian4D* [62], *SV4D* [95], *4Real-Video* [80] and *L4GM* [63]. For a fair comparison, we initialize Dream-Gaussian4D [62] with the same static Gaussian splats as our method before optimizing the deformation field. For SV4D [95], we render a 21-frame, 360-degree video of the generated 3D asset and select 8 frames as input. For 4Real-Video [80], we follow the authors’ implementation, rendering a 15-frame input video with azimuths from -60 to 60 degrees. For L4GM [63], we adhere to the official imple-



Figure 3. Qualitative comparison of 4D generation methods. We compare the 4D generation results of our method with L4GM [63], DreamGaussian4D (DG4D) [62], SV4D [95], and 4Real-Video [80]. DG4D incorporates our accurate static Gaussian representation as input alongside a monocular video, while SV4D and 4Real rely on fixed-time renderings. Viewpoints are displayed at  $\pm 60$  degrees relative to the original perspective used for generating the monocular video. Among all methods, our approach achieves the most visually consistent and appealing results. Note that, unlike other methods, ours also produces high-quality meshes. Video results are in the Supp. material.

mentation to generate Gaussian splats from the input fixed-view video.

**Qualitative Comparison.** A qualitative comparison of 4D generation results between our method and baseline approaches is shown in Fig. 3. Each method renders views at  $\pm 60$  degrees from the original viewpoint used for monocular video generation. Reconstructing 4D assets from distant viewpoints remains highly challenging, even with access to the first frame. Although initialized with the same static Gaussians as our method, DG4D produces distorted geometry. SV4D struggles with complex structures, such as moose antlers, resulting in blurry outputs. Similarly, 4Real-Video suffers from geometric distortions, limiting its performance. Our method overcomes these challenges by refining 4Real-Video’s output through a view-dependent design and integrating geometric cues from the static scene to resolve 3D inconsistencies. As a result, it achieves the most visually consistent and high-quality reconstructions among all baselines. In addition to rendered images, our approach also generates a 3D mesh, shown on the right side of the figure.

**Quantitative Evaluation.** Evaluating generated 3D/4D assets is challenging. Due to the limited number of generated assets, commonly used metrics like FID [5] and FVD [77]

are not statistically meaningful. While CLIP scores [21] assess text-3D alignment, they lack flexibility for other criteria. Instead, we adopt the Elo rating [15] from GPTEval3D [91], which leverages GPT-4o [1] to compare 3D assets across multiple criteria and compute rankings based on pairwise comparisons. Following the official implementation, we sampled 300 pairwise comparisons and report Elo ratings for three key criteria: *3D plausibility*, *texture details*, and *text-asset alignment*, omitting criteria requiring surface normal rendering, as some baselines do not support it. As shown in Table 2, our results are predominantly preferred by GPT-4o evaluations.

**User Study.** We conducted a user study comparing DreamGaussian4D (DG4D)[62], SV4D[95], 4Real-Video [80], and our method. The study included 422 samples, each rated by 10 participants, yielding 4,220 ratings across four dimensions. *Appearance* evaluates identity and detail preservation. *Motion* measures the accuracy of motion transfer while avoiding scaling and rotation artifacts. *Geometry* assesses distortions such as squishing or shape deformations. *Overall* represents the plausibility of motion transfer while maintaining the appearance and geometry of the base asset. As shown in Table 1, VCDGS outperforms

| Criteria (%) | DG4D [62] | SV4D [95] | 4Real-Video [80] | Ours        |
|--------------|-----------|-----------|------------------|-------------|
| Appearance   | 0.0       | 0.2       | 13.0             | <b>86.8</b> |
| Motion       | 0.1       | 0.4       | 40.9             | <b>58.7</b> |
| Geometry     | 0.0       | 0.1       | 15.8             | <b>84.2</b> |
| Overall      | 0.1       | 0.1       | 15.3             | <b>84.6</b> |

Table 1. User preference distribution across different methods.

| Criteria (%)      | SV4D [95] | L4GM [63] | DG4D [62] | 4Real-Video [80] | Ours          |
|-------------------|-----------|-----------|-----------|------------------|---------------|
| Text-Asset Align. | 818.5     | 729.9     | 1136.6    | 1075.2           | <b>1301.5</b> |
| 3D Plausibility   | 799.4     | 734.1     | 1138.3    | 1126.7           | <b>1306.2</b> |
| Texture Details   | 664.1     | 624.5     | 1157.4    | 1151.4           | <b>1436.4</b> |

Table 2. GPTEval3D [91] ratings using GPT-4o, higher is better.

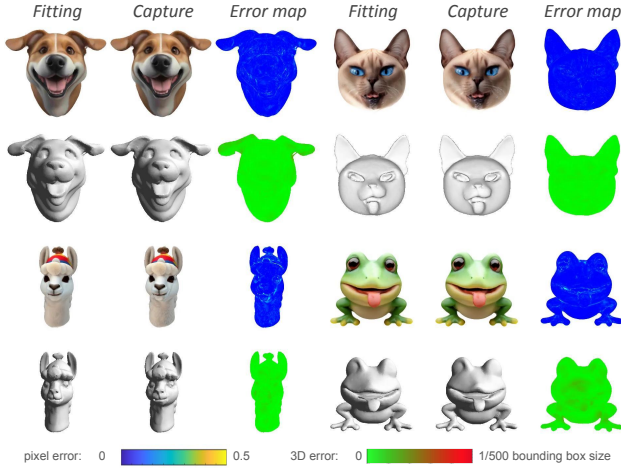


Figure 4. Fitting the learned blendshape model to held-out test set captures. The color scale from blue to yellow represents the RGB error, while the scale from green to red denotes the 3D point-to-point error.

all baselines, including 4Real-Video, which it refines. This improvement stems from our view-dependent deformations, which resolve multi-view inconsistencies.

### 4.3. Blendshape Evaluation

We evaluate the quality of blendshape models learned from VCDGS captures and demonstrate their applicability for expression transfer using a simple retargeting approach. Please refer to the project website for the video results.

**Model Expressiveness.** A robust statistical expression model should effectively generalize to new data while remaining closely aligned with the specific object it represents. We fit T2Bs model with 100 blendshapes to geometry extracted from animations outside the model’s training set by minimizing  $\mathcal{L}_{\text{geo}} + \lambda_{rgb} L_{\text{rgb}}$ , where  $L_{\text{geo}}$  is point-to-point euclidean distance and  $L_{\text{rgb}}$  is the pixel-wise color distance when rendering the capture and the reconstruction with the same camera. We set  $\lambda_{rgb} = 0.1$ .

As shown in Fig. 4, the reconstructions closely align with the captures in both geometry and appearance. For each identity, the first column shows model fitting, the second

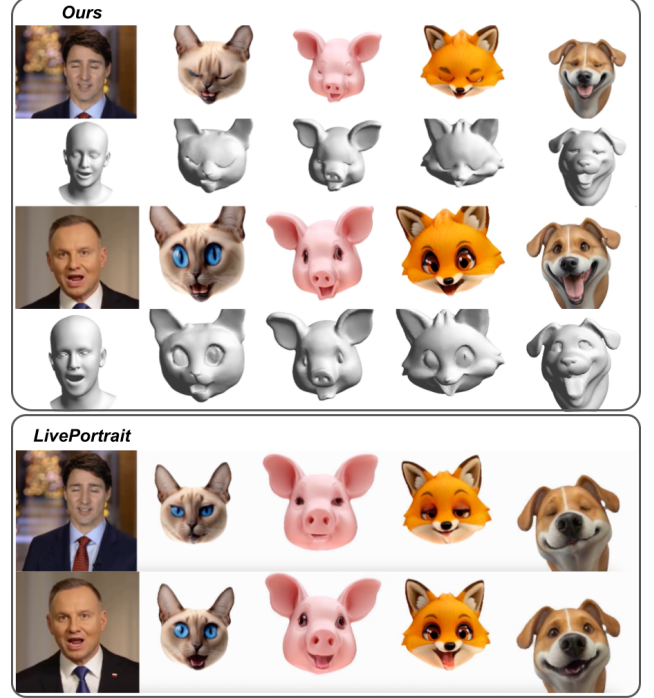


Figure 5. Real-world captures and corresponding retargeted expressions. Top: 3D-aware retargeting by T2Bs, rendered with both textured and geometry only. Bottom: 2D-only retargeting by LivePortrait [17]. T2Bs achieves better retargeting, especially for non-human features such as exaggerated eyes and mouth shapes.

displays a held-out video capture, and the third presents RGB and 3D error maps, comparing the capture and fitting results in terms of rendering and per-vertex differences. Our blendshape model generalizes effectively to new data, demonstrating its ability to faithfully reconstruct meshes of held-out expressions.

**Retargeting.** Given human face tracking, we update the virtual character’s landmarks based on the relative positions of six landmarks on each eye and eight on the mouth from the human facial landmarks. We then fit the 100 blendshape model by optimizing its weights to minimize a combination of landmark distance  $\mathcal{L}_{\text{lm}}$  and as-rigid-as-possible (ARAP) regularization  $\mathcal{L}_{\text{ARAP}}$ , with weightings of 1.0 and 0.1. For the eye regions, we introduce additional regularization.

Fig. 5 shows examples of the alignment between human and virtual character expressions. We compare with LivePortrait [17], an image base retargeting method also relies on landmarks. While LivePortrait fails to close large eyes or open wide mouths, our method, achieved by smooth 3D geometry, is more expressive to handle non-human features.

### 4.4. Ablation Studies

In this section, we conduct ablation studies on proposed components. We show qualitative comparison with our virtual characters, and show Quantitative comparison on

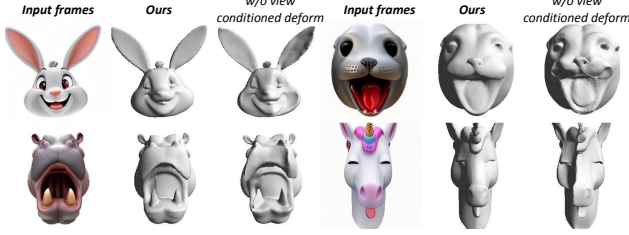


Figure 6. Ablation study on view conditioning: Without view conditioning, the model develops artifacts and fails to maintain coherent geometry due to 3D inconsistencies in the generated 4D videos.



Figure 7. Ablation study of changing 4D video generation method to SV4D. Our method remains robust to SV4D’s blurry outputs, producing improved results over its raw output.

Objaverse-XL [13] samples in Supp. material.

**View Conditioning.** We introduce view dependency into the deformation representation to address 3D inconsistencies in generated 4D videos. To validate this design, we compare reconstructions without view-conditioned deformation as shown in Fig. 6. For clarity, geometry is visualized as a mesh with the same topology as the static 3D asset. Without view conditioning, multi-view inconsistencies in the generated video frames lead to artifacts and geometric distortions in the reconstructed meshes.

**Source of Multi-view Videos.** We optimize VCDGS using videos generated by 4Real-Video. To assess whether this specific 4D video generation method is necessary, we conduct an ablation study using 4D videos from other sources. Specifically, we fit VCDGS on SV4D [95] outputs while keeping all other settings unchanged. As shown in Fig. 7, incorporating VCDGS improves SV4D’s results. SV4D alone tends to produce blurry novel views, while our model remains robust to blurry inputs and generates high-quality renderings with well-detailed geometry.

**Color Offset Prediction.** We conducted an ablation study to evaluate the impact of predicting color offsets with RefineMLP. Fig. 8 (top) provides two examples of common color and lighting inconsistency issues that arise during video generation. When the Gaussian color is fixed, the model struggles to converge in affected areas, leading to overfitting in shape deformation, as shown in Fig. 8 (bottom). Color offset prediction models the lighting changes, allowing for better geometry reconstruction.

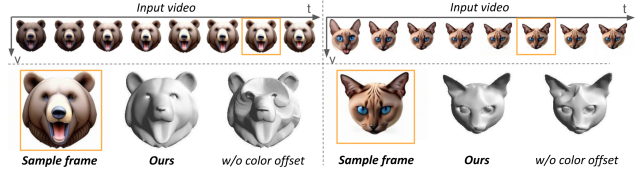


Figure 8. Ablation study on color offset prediction. (Top) Two video examples with noticeable variations in color and lighting. (Bottom) Reconstructions with and without the predicted color offset. Without color refinement, shape deformation tends to overfit, leading to suboptimal reconstruction quality.

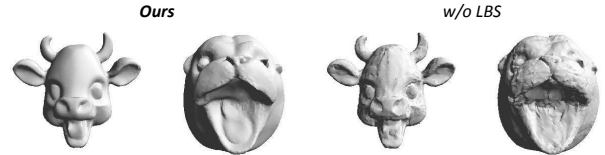


Figure 9. Ablation study on integrating Linear Blend Skinning (LBS) in deformation. Removing LBS leads to degraded reconstruction quality.

**Linear Blend Skinning.** LBS employs deformation primitives to ensure well-regularized deformations. Fig. 9 illustrates the impact of the LBS on geometry reconstruction. We compare our method with a commonly used approach [89] that directly predicts per-Gaussian deformation. Our method produces smooth, high-quality reconstructions, whereas the comparison results exhibit significant noise.

## 5. Conclusion

We introduced T2Bs a novel framework for generating high-quality, animatable 3D character blendshape models from textual descriptions. By leveraging advancements in text-to-3D generation and video diffusion models, our pipeline bridges the gap between static geometry and dynamic motion. Extensive evaluations confirm its effectiveness in producing expressive and visually coherent 3D character models.

**Limitations.** Our method is designed for head animations and does not extend to full-body motion, as blendshape models are unsuitable for non-linear articulated motions. While it effectively models challenging geometries like protruding tongues, it relies on the initial 3D asset having sufficient coverage of the mouth interior, as 4D reconstruction is sensitive to rapid topological changes. Additionally, the video diffusion model enables diverse expression generation but occasionally produces unnatural deformations or temporal inconsistencies. Future advancements in video generation may help address these limitations.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6
- [2] 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 *CVPR*, 2024. 2
- [3] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024. 2
- [4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. 2023. 1
- [5] Naresh Babu Bynagari. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Asian J. Appl. Sci. Eng*, 8(1):25–34, 2019. 6
- [6] Chen Cao, Yanlin Weng, Shun Zhou, Yiyong Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 1
- [7] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction, 2023. 2
- [8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2
- [9] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV*, 2023. 2
- [10] Zhaoxi Chen, Fangzhou Hong, Haiyi Mei, Guangcong Wang, Lei Yang, and Ziwei Liu. Primdiffusion: Volumetric primitives diffusion for 3d human generation, 2023. 2
- [11] Zilong Chen, Feng Wang, Yikai Wang, and Huaping Liu. Text-to-3d using gaussian splatting, 2024. 2
- [12] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *CVPR*, 2023. 2
- [13] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023. 8
- [14] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG)*, 39(5):1–38, 2020. 1
- [15] Arpad E Elo. The proposed uscf rating system, its development, theory, and applications. *Chess life*, 22(8):242–247, 1967. 6
- [16] Ruiqi Gao\*, Aleksander Holynski\*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srivasan, Jonathan T. Barron, and Ben Poole\*. Cat3d: Create anything in 3d with multi-view diffusion models. *Advances in Neural Information Processing Systems*, 2024. 2
- [17] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and re-targeting control. *arXiv preprint arXiv:2407.03168*, 2024. 7
- [18] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2
- [19] Xiao Han, Yukang Cao, Kai Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, Tao Xiang, and Kwan-Yee K Wong. Headsculpt: Crafting 3d head avatars with text. *Advances in Neural Information Processing Systems*, 36:4915–4936, 2023. 1
- [20] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 2
- [21] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *CoRR*, abs/2104.08718, 2021. 6
- [22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 2
- [23] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2
- [24] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d, 2024. 2
- [25] Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu. Neural wavelet-domain diffusion for 3d shape generation, 2022. 2
- [26] *A 3D Face Model for Pose and Illumination Invariant Face Recognition*, Genova, Italy, 2009. IEEE. 1
- [27] Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360 {\\deg} dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848*, 2023. 2

- [28] Pushkar Joshi, Wen C Tien, Mathieu Desbrun, and Frédéric Pighin. Learning controls for blend shape based realistic facial animation. In *ACM Siggraph 2006 Courses*, pages 17–es. 2006. 4
- [29] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions, 2023. 2
- [30] 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
- [31] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. *arXiv preprint arXiv:2405.17421*, 2024. 3
- [32] Bing Li, Cheng Zheng, Wenxuan Zhu, Jinjie Mai, Biao Zhang, Peter Wonka, and Bernard Ghanem. Vivid-zoo: Multi-view video generation with diffusion model, 2024. 2
- [33] Hao Li, Thibaut Weise, and Mark Pauly. Example-based facial rigging. *Acm transactions on graphics (tog)*, 29(4): 1–6, 2010. 1
- [34] Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph.*, 32(4):42–1, 2013. 5
- [35] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model, 2023. 2
- [36] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 1, 2
- [37] Zhiqi Li, Yiming Chen, and Peidong Liu. Dreammesh4d: Video-to-4d generation with sparse-controlled gaussian-mesh hybrid representation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2
- [38] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching, 2023. 2
- [39] 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. In *CVPR*, 2023. 2
- [40] 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 *CVPR*, 2024. 2
- [41] Di Liu, Xiang Yu, Meng Ye, Qilong Zhangli, Zhuowei Li, Zhixing Zhang, and Dimitris N Metaxas. Deformer: Integrating transformers with deformable models for 3d shape abstraction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14236–14246, 2023. 2
- [42] Di Liu, Qilong Zhangli, Yunhe Gao, and Dimitris Metaxas. Leopard: Learning explicit part discovery for 3d articulated shape reconstruction. *Advances in Neural Information Processing Systems*, 36:54187–54198, 2023.
- [43] Di Liu, Teng Deng, Giljoo Nam, Yu Rong, Stanislav Pidhorskyi, Junxuan Li, Jason Saragih, Dimitris N. Metaxas, and Chen Cao. Lucas: Layered universal codec avatars, 2025. 2
- [44] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023. 2
- [45] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 5
- [46] Zhen Liu, Yao Feng, Michael J. Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. Meshdiffusion: Score-based generative 3d mesh modeling, 2023. 2
- [47] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2
- [48] Jiahao Luo, Fahim Hasan Khan, Issei Mori, Akila De Silva, Eric Sandoval Ruezga, Minghao Liu, Alex Pang, and James Davis. How much does input data type impact final face model accuracy? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18985–18994, 2022. 1
- [49] Jiahao Luo, Jing Liu, and James Davis. Splatface: Gaussian splat face reconstruction leveraging an optimizable surface. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 774–783. IEEE, 2025. 1
- [50] Yifang Men, Hanxi Liu, Yuan Yao, Miaomiao Cui, Xuansong Xie, and Zhouhui Lian. 3dtoonify: Creating your high-fidelity 3d stylized avatar easily from 2d portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10127–10137, 2024. 1
- [51] Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, et al. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *CVPR*, 2024. 2
- [52] Mohit Mendiratta, Xingang Pan, Mohamed Elgharib, Karthik Teotia, Ayush Tewari, Vladislav Golyanik, Adam Kortylewski, and Christian Theobalt. Avatarstudio: Text-driven editing of 3d dynamic human head avatars. *ACM Transactions on Graphics (ToG)*, 42(6):1–18, 2023. 1
- [53] 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, 2020. 2
- [54] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Peter Kotschieder, and Matthias Nießner. Diffrr: Rendering-guided 3d radiance field diffusion, 2023. 2

- [55] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):1–15, 2022. 2
- [56] Alex Nichol, Heewoo Jun, Pratul Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts, 2022. 2
- [57] OpenAI. Video generation models as world simulators, 2024. Accessed: 2024-11-08. 2
- [58] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 2
- [59] 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. In *ICLR*, 2024. 2
- [60] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 1, 2
- [61] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5
- [62] Jiawei Ren, Liang Pan, Jiayang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 2, 3, 5, 6, 7
- [63] Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Hanxue Liang, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, and Huan Ling. L4gm: Large 4d gaussian reconstruction model. In *Advances in Neural Information Processing Systems*, 2024. 2, 5, 6, 7
- [64] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2
- [65] 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. *NeurIPS*, 2022. 2
- [66] Shen Sang, Tiancheng Zhi, Guoxian Song, Minghao Liu, Chunpong Lai, Jing Liu, Xiang Wen, James Davis, and Linjie Luo. Agileavatar: Stylized 3d avatar creation via cascaded domain bridging. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 1
- [67] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2
- [68] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis, 2021. 2
- [69] Tianchang Shen, Jacob Munkberg, Jon Hasselgren, Kangxue Yin, Zian Wang, Wenzheng Chen, Zan Gojcic, Sanja Fidler, Nicholas Sharp, and Jun Gao. Flexible isosurface extraction for gradient-based mesh optimization. *ACM Transactions on Graphics*, 42(4):1–16, 2023. 2
- [70] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model, 2023. 2
- [71] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *ICLR*, 2024. 2
- [72] 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
- [73] 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, 2023. 2
- [74] Jiayang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation, 2024. 2
- [75] Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B. Lindell. Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models, 2024. 2
- [76] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts, 2023. 2
- [77] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. *arXiv preprint arXiv:1812.01717*, 2018. 6
- [78] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion, 2024. 2
- [79] Chaoyang Wang, Lachlan Ewen MacDonald, Laszlo A Jeni, and Simon Lucey. Flow supervision for deformable nerf. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21128–21137, 2023. 3
- [80] Chaoyang Wang, Peiye Zhuang, Tuan Duc Ngo, Willi Menapace, Aliaksandr Siarohin, Michael Vasilkovsky, Ivan Skorokhodov, Sergey Tulyakov, Peter Wonka, and Hsin-Ying Lee. 4real-video: Learning generalizable photo-realistic 4d video diffusion. *arXiv preprint arXiv:2412.04462*, 2024. 2, 3, 5, 6, 7
- [81] Duotun Wang, Hengyu Meng, Zeyu Cai, Zhijing Shao, Qianxi Liu, Lin Wang, Mingming Fan, Xiaohang Zhan, and Zeyu Wang. Headevolver: Text to head avatars via expressive and attribute-preserving mesh deformation. *arXiv preprint arXiv:2403.09326*, 2024. 1

- [82] Jie Wang, Jiu-Cheng Xie, Xianyan Li, Feng Xu, Chi-Man Pun, and Hao Gao. Gaussianhead: Impressive head avatars with learnable gaussian diffusion. *arXiv preprint arXiv:2312.01632*, 2023. 1, 3
- [83] Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. 2024. 3
- [84] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. Rodin: A generative model for sculpting 3d digital avatars using diffusion, 2022. 2
- [85] Yikai Wang, Xinzhou Wang, Zilong Chen, Zhengyi Wang, Fuchun Sun, and Jun Zhu. Vidu4d: Single generated video to high-fidelity 4d reconstruction with dynamic gaussian surfels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3
- [86] 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. In *NeurIPS*, 2023. 2
- [87] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH*, 2024. 2
- [88] Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David J Fleet. Controlling space and time with diffusion models. *arXiv preprint arXiv:2407.07860*, 2024. 2
- [89] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, 2024. 8
- [90] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T. Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv:2411.18613*, 2024. 2
- [91] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation. In *CVPR*, 2024. 6, 7
- [92] Zijie Wu, Chaohui Yu, Yanqin Jiang, Chenjie Cao, Fan Wang, and Xiang. Bai. Sc4d: Sparse-controlled video-to-4d generation and motion transfer. *arxiv:2404.03736*, 2024. 2
- [93] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1802–1812, 2024. 1, 2
- [94] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 4
- [95] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*, 2024. 2, 3, 5, 6, 7, 8
- [96] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 2
- [97] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation, 2024. 2
- [98] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *CVPR*, 2021. 4
- [99] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 4
- [100] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH*, 2024. 2
- [101] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 5
- [102] Taoran Yi, Jiemin Fang, Junjie Wang, Guanqun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models, 2024. 2
- [103] Yuyang Yin, Dejie Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*, 2023. 2
- [104] Heng Yu, Chaoyang Wang, Peiye Zhuang, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Laszlo A Jeni, Sergey Tulyakov, and Hsin-Ying Lee. 4real: Towards photorealistic 4d scene generation via video diffusion models. In *NeurIPS*, 2024. 2, 3
- [105] Ye Yuan, Xueting Li, Yangyi Huang, Shalini De Mello, Koki Nagano, Jan Kautz, and Umar Iqbal. Gavatar: Animatable 3d gaussian avatars with implicit mesh learning. *arXiv preprint arXiv:2312.11461*, 2023. 2
- [106] Bowen Zhang, Yiji Cheng, Chunyu Wang, Ting Zhang, Jiaolong Yang, Yansong Tang, Feng Zhao, Dong Chen, and Baining Guo. Rodinhd: High-fidelity 3d avatar generation with diffusion models, 2024. 2
- [107] Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. 4diffusion: Multi-view video diffusion model for 4d generation. *arXiv preprint arXiv:2405.20674*, 2024. 2

- [108] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. [4](#)
- [109] Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023. [2](#)
- [110] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation, 2023. [2](#)
- [111] Zhenglin Zhou, Fan Ma, Hehe Fan, Zongxin Yang, and Yi Yang. Headstudio: Text to animatable head avatars with 3d gaussian splatting. In *European Conference on Computer Vision*, pages 145–163. Springer, 2024. [1](#)
- [112] Joseph Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance. In *ICLR*, 2023. [2](#)
- [113] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers, 2023. [2](#)