MotionDreamer: Exploring Semantic Video Diffusion Features for Zero-Shot 3D Mesh Animation

Lukas Uzolas Elmar Eisemann Petr Kellnhofer
Delft University of Technology
The Netherlands
{l.uzolas, e.eisemann, p.kellnhofer}@tudelft.nl

Abstract

Animation techniques bring digital 3D worlds and characters to life. However, manual animation is tedious and automated techniques are often specialized to narrow shape classes. In our work, we propose a technique for automatic re-animation of various 3D shapes based on a motion prior extracted from a video diffusion model. Unlike existing 4D generation methods, we focus solely on the motion, and we leverage an explicit mesh-based representation compatible with existing computer-graphics pipelines. Furthermore, our utilization of diffusion features enhances accuracy of our motion fitting. We analyze efficacy of these features for animation fitting and we experimentally validate our approach for two different diffusion models and four animation models. Finally, we demonstrate that our time-efficient zero-shot method achieves a superior performance re-animating a diverse set of 3D shapes when compared to existing techniques in a user study.

1. Introduction

Animation is an important component of video games, simulators, and movies. It makes otherwise rigid environments come to life and is often a result of a tedious motion-data capture coupled to skilled manual editing [17]. However, this does not scale well for applications involving large virtual worlds with thousands of individual entities or for individual objects that are difficult to motion capture due to their physical size or real-world inaccessibility. For this reason, we propose an end-to-end generative method that re-animates static 3D objects using a pre-trained Video Diffusion Model [5, 21, 24, 91, 93] (VDM) without any additional training (Fig. 1).

We build on the success of Diffusion models [23, 79]. Beyond producing nearly photo-realistic 2D images [52, 66, 71, 74], diffusion was also adapted for 3D [35, 64] and 4D shape synthesis [2, 28, 42, 68, 78, 90, 96, 100, 101]. However, the associated methods suffer from either a high optimization cost and low diversity [40] of the mode-seeking Stochastic Distillation Sampling [64] (SDS), or, as we show, they are susceptible to the visual artifacts in RGB outputs of existing VDMs. Furthermore, our method generates a unique animation as a sequence of object poses in a matter of minutes rather than hours common for end-to-end 4D generative methods. This is a feature crucial for processing of larger sample sets with subsequent filtering based on subjective preferences. Therefore, we position our approach into a category distinct from end-to-end 4D generation.

Instead of iterative SDS, we leverage the surprising versatility of semantic features extracted from diffusion models for down-stream tasks such as one-shot segmentation [32] or semantic feature matching [12, 80], which we adapt for motion fitting. We rely on a classical surface mesh representation in combination with diverse animation models [1, 38, 45, 103] to obtain animated 3D shapes that are fast to render, compatible with existing rendering frameworks and versatile across object classes.

In summary, we present the following contributions: 1. We introduce a novel zero-shot generative method for 3D mesh animation based on rendering in the semantic feature space of pre-trained VDMs. 2. We analyze effectiveness of VDM features for pose estimation to validate our method and design choices. 3. We evaluate two VDMs and four animation models and demonstrate a preference of our 3D animations to existing generative approaches in a user study.

2. Related Work

Our method exploits VDMs to create novel animations of 3D objects. Here we discuss relevant work on video generation and existing approaches for 3D shape representation and animation.

2.1. Video generation

Generative visual models have advanced rapidly from Variational Auto-Encoders [34], Normalizing Flows [10, 69]

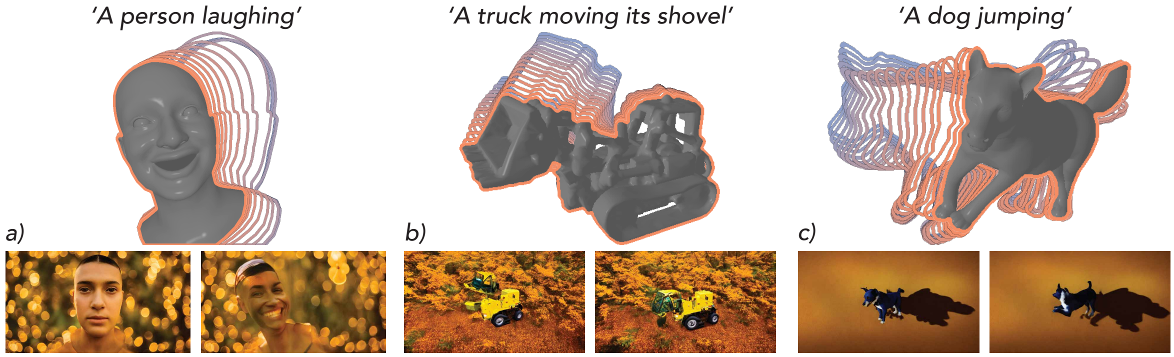

Figure 1. Our Zero-shot 3D mesh animations. From top to bottom: The desired motion description, the resulting animated mesh with motion contours, the driving video from a pre-trained video diffusion model. Notice robustness of our method to the temporal identity shift (a) and to the geometric distortions (b). Diverse shapes are supported through a range of animation models including a) FLAME [38], b) Neural Jacobian Fields [1] and c) SMAL [103]. Examples are shown on our project page: https://graphics.tudelft.nl/MotionDreamer.

and Generative Adversarial Models [18] to Diffusion Models [23, 79] and Continuous Normalizing Flows [43] achieving a nearly photorealistic image synthesis [52, 66, 71, 74] as well as state-of-the-art video synthesis [5, 21, 24, 91, 93]. Surprisingly, the features learned by the U-Net [73] of many diffusion models exhibit semantic properties useful for downstream tasks such as segmentation [32] and feature matching [46, 80, 97]. Consequently, we analyze utility of two such models [91, 93] for our motion fitting, while we leave opportunities presented by recent large VDMs [7, 9] utilizing Visual Transformers [11] as an avenue for future research.

2.2. Shape and pose representations

There exist many ways for representing 3D shapes from classical explicit representations including point-clouds, voxels or surface meshes favored in real-time applications, to implicit neural shape representations [51, 55, 84] enabling photorealistic 3D scene reconstruction. In the middle, 3D Gaussians [31] have been shown to combine advantages of both at an increased storage cost. In this paper we focus on surface meshes for their fast rendering, efficient storage and wide application support.

While animation of object poses can be encoded as a sequence of static representations [48], specialized representations ease editing for both arbitrary and class-specific shapes. In the first category, deformation fields offer maximal flexibility for dense volumetric optimization [65], Neural Jacobian Fields (NJFs) [1] offer space-time continuity and smoothness for surface optimization and external cages reduce the control space for easier editing [94]. In the second category, low-dimensional templates support manual animation and motion capture by combining Linear Blend Skinning [36] and Blend Shapes [56, 57] for specific classes of shapes such as faces [4, 38], bodies [45, 59], hands [72], or even animals [103]. Our method is agnostic to the choice of an animation model, which we test on high-dimensional NJFs [1] and on low-dimensional templates [38, 45, 103].

2.3. 3D motion and animation

Capture Motion, most often for humans, can be directly captured [54] using sparse inertial sensors [81] or dense visual observations [25] either with tracking markers [77] or without them [76]. For a monocular video, we can estimate 2D poses [8, 60, 85] and uplift them to 3D [6, 47, 50, 75, 99] thanks to data priors [59, 63] based on large motion datasets [26, 29, 49]. However, the specific training for each class limits generalization. In contrast, recent advances in neural rendering [31, 51, 84] enabled class-agnostic 4D reposable reconstructions [53, 86, 95]. Our method is similarly based on class-agnostic differentiable pose optimization but differently from a direct image supervision, we exploit diffusion features of a monocular video rather than multi-view observations.

Generation Learned priors can also be used for text-conditioned motion synthesis [102]. However, this is in practice limited to human domain [20, 27, 82, 83] where annotated 3D motion datasets exist [19, 61] or to other skeletal shapes [30] if at least 2D annotations are available. Alternatively, image and video generative models enabled class-agnostic joint shape and motion 4D generation [2, 28, 42, 68, 78, 90, 96, 100, 101] is usually based on Stochastic Distillation Sampling (SDS) [64] which, however, narrows the sampled distribution [40] due it its mode-seeking behavior. Closest to us, Ren et al. [68] extract motion from a full video input. Our method shares the idea of extracting motion from a video model but thanks to utilizing the feature space it produces more natural motion with fewer visual artifacts. Furthermore, we do not use 3D uplifting methods requiring background masks such as Zero-1-to-3 [44]. Additionally, by focusing on motion alone we achieve faster sampling. Finally, both captured or generated motion can be transferred from one shape to another [16], either based on morphological similarity [41, 88] or data-driven domain matching [39, 70]. We experimentally show that our method is preferable when neither of the two conditions can be satisfied.

3. Preliminaries

Our method exploits internal representation of VDMs. Here, we provide a brief summary of these models and semantic information encoded in their internal features.

3.1. Video Diffusion Models

VDMs are a type of a generative model producing video sequences by gradual denoising [23, 79] of a Gaussian-noise image sequence $\mathbf{z} \in \mathbb{R}^{L \times H \times W \times D_{\text{lat}}}$, where L is the frame count, H, W the spatial dimensions, and D_{lat} is 3 for RGB models or the latent feature dimension for Latent Diffusion [71]. The forward diffusion process $q(\mathbf{z}_t | \mathbf{z}_0, t)$ gradually transports $\mathbf{z_0} \equiv \mathbf{z}$ to the Gaussian-noise prior over T steps such that $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This is used to learn a denoising function $f_\theta(\mathbf{z}_t, t, \mathbf{c})$ as a θ-parameterized network approximating the reverse process $p_\theta(\mathbf{z}_{t-1} | \mathbf{z}_t, t, \mathbf{c})$. A commonly used ϵ-prediction training procedure minimizes an objective $\sum_{t,\mathbf{c},\mathbf{z},\epsilon} \|\epsilon - f_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2$ across data and noise samples $\mathbf{z} \sim p_{\text{data}}$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Finally, sampling the noise prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and reversing the diffusion yields video generation. The conditioning vector $\mathbf{c} \in \mathbb{R}^N$ is often a text embedding, image embedding or both, and it steers the process, often with a classifier-free guidance [22].

3.2. Semantic Diffusion Features

Intermediate activations of image diffusion networks have been shown to encode semantic features and provide robust correspondences across image samples [46, 80, 97]. We adopt the methodology of Tang et al. [80], where f_θ is parameterized by a U-Net. The semantic feature maps $\mathbf{A}_u \in \mathbb{R}^{H_u \times W_u \times A_u}$ are extracted from the intermediate activations of a U-Net layer u with height, width and feature size H_u, W_u, and A_u.

Given a pair of images with feature maps $\mathbf{A}_u, \mathbf{B}_u$ and a chosen spatial location $\phi^A \in \mathbb{R}^2$ in the first image, we find a semantically corresponding spatial location $\phi^B \in \mathbb{R}^2$ in the other image as $\phi^B = \arg\max_{\phi^B} \kappa(\mathbf{A}_u[\phi^A], \mathbf{B}_u[\phi^B])$, where

$$\kappa(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{b}}{||\mathbf{a}||_2 ||\mathbf{b}||_2} \tag{1}$$

is a cosine similarity $\kappa : \mathbb{R}^{A_u} \times \mathbb{R}^{A_u} \to \mathbb{R}$ and $\mathbf{x}[\phi]$ denotes spatial sampling of a map $\mathbf{x}$ at location ϕ, which we implement as a bilinear interpolation. For video, we treat each frame as an image with its own feature map, and we optimize semantic correspondences of reposed meshes using Eq. 1.

4. Method

Our methods accepts an unseen 2-manifold 3D mesh in an arbitrary pose and uses a pre-trained VDM to generate a temporal sequence of animation parameters (see Fig. 2 for an overview). We first describe our method for a general VDM and animation model before discussing specific realizations in Sec. 4.4.

Definitions We define the input mesh $\mathcal{M}$ as a tuple of N vertices and M triangular faces $\mathcal{M} := (\{\mathbf{u}_n \in \mathbb{R}^3 | n = 0, ..., N-1\}, \{\mathbf{f}_m \in \mathbb{N}^3 | m = 0, ..., M-1\})$. Next, we define $\tau : (\mathcal{M}, \mathbf{p}) \to \mathcal{M}'$ as a function transforming vertices to produce a mesh $\mathcal{M}' := (\{\mathbf{u}'_n\}, \{\mathbf{f}_m\})$ with a novel pose described by animation parameters $\mathbf{p} \in \mathbb{R}^P$. We refer to $\mathbf{p}_{\text{init}}$ as the input pose where $\tau(\mathcal{M}, \mathbf{p}_{\text{init}}) \equiv \mathcal{M}$ and, without a loss of generality, we assume it matches the first frame. Finally, $r_{rgb} : (\mathcal{M}, \mathcal{C}, \mathcal{T}, \mathbf{B}) \to \mathbf{I}_{rgb}$ is a rendering function producing an RGB image $\mathbf{I}_{rgb} \in \mathbb{R}^{H \times W \times 3}$ of the mesh $\mathcal{M}$ for a manually defined canonical camera $\mathcal{C}$, surface texture $\mathcal{T}$, and a background image $\mathbf{B} \in \mathbb{R}^{H \times W \times 3}$.

4.1. Single-View Texturing

While the visual datasets used to train existing VDMs are very large, they favor natural looking textured images with backgrounds (see Appendix D.1 for examples). We reduce the domain gap for our rendered image by automatically generating an RGB texture $\mathcal{T}$ and a semantically fitting background image $\mathbf{B}$. First, we render a depth map and a foreground mask ψ for a single fixed viewpoint of $\mathcal{M}$. Next, we style-transfer the depth map using a pre-trained ControlNet diffusion model [98] conditioned by a user-provided textual description to obtain a textured RGB image $\mathbf{S}$. Then, we crop the foreground texture $\mathcal{T} = \textit{unproject}(\mathbf{S} \odot \psi)$ and apply it to the mesh $\mathcal{M}_0$ using projective texturing [92]. Importantly, we do not strive for a complete texture of the entire mesh, but merely for a stylization of the single-view VDM input image. Finally, we obtain the background image $\mathbf{B}$ by inpainting the remainder of $\mathbf{S}$ outside of the foreground bounding box using Stable Diffusion XL [62]. See Appendix B.2 for prompt details.

4.2. Motion Generation

The motion produced by our method originates from a VDM conditioned by our rendered mesh image $\mathbf{I}^0_{rgb} = r_{rgb}(\mathcal{M}^0, \mathcal{C}, \mathcal{T}, \mathbf{B})$ and an embedding of the intended motion text description. We sample the generator in a multi-step diffusion process over T steps denoted as $t \in [0, ..., T-1]$ with scheduling details specific to each VDM. Because the temporally incoherent visual artifacts in RGB video outputs make motion tracking difficult (see Fig. 1), we extract semantically meaningful U-Net features $\mathbf{A}^t_u$ at time step $t = \hat{t}$ and U-Net layer $u = \hat{u}$ as explained in Sec. 3.2, and we

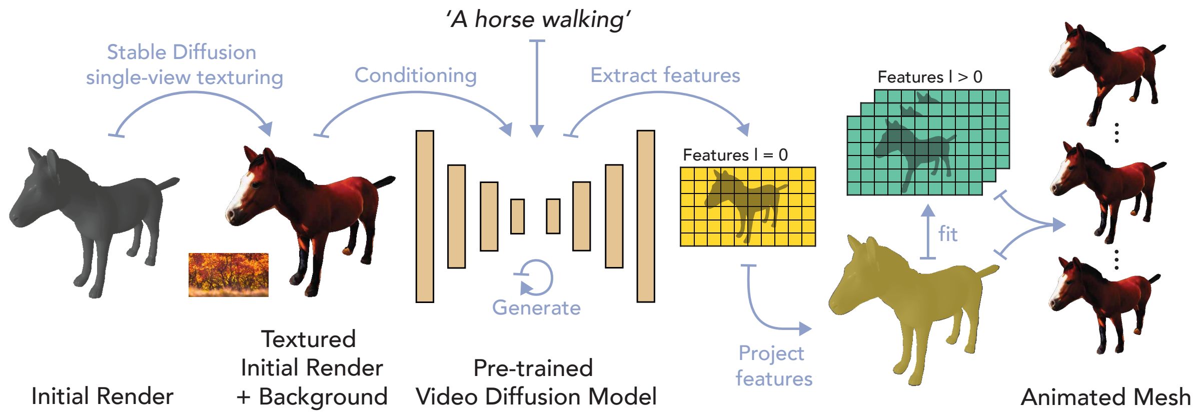

Figure 2. A diagram of our method. First, we automatically texture the input mesh $\mathcal{M}$ to reduce the domain gap to the VDM prior (Sec. 4.1). Second, we condition the VDM by a rendered image $\mathbf{I}_{rgb}$ to produce a video with motion and to extract features $\hat{\mathbf{A}}$ for all L frames from its internal U-Net (Sec. 4.2). Finally, we reproject the input frame features $\hat{\mathbf{A}}^0$ on the mesh surface and we optimize mesh animation parameters $\mathbf{p}$ to match the reposed mesh features to the video (Sec. 4.3). produce

show that this improves the fitting accuracy. We motivate our choice of $\hat{t}$ and $\hat{u}$ in Sec. 5.3, and will omit the suffixes from now on for brevity, such that $\hat{\mathbf{A}} \in \mathbb{R}^{L\times\hat{H}\times\hat{W}\times\hat{A}} \equiv \mathbf{A}^{\hat{t}}_{\hat{u}}$ and $\hat{\mathbf{A}}^l$ selects the video frame l of L. We further assume $\hat{\mathbf{A}}^0$ corresponds to the input image $\mathbf{I}^0_{rgb}$ (see Appendix A.1 for a discussion).

4.3. Motion fitting

Given the known correspondence of the mesh $\mathcal{M}$, initial pose $\mathbf{p}_{\text{init}}$, image $\mathbf{I}^0_{rgb}$ and features $\hat{\mathbf{A}}^0$ for the input frame $l = 0$, we aim to recover all animation parameters $\mathbf{p}^l$ for $l \in [0, ..., L-1]$.We achieve this by optimizing $\mathbf{p}$ to match reprojections of the input $\hat{\mathbf{A}}^0$ to $\hat{\mathbf{A}}^l$ extracted from the video. To this goal, we first reproject $\hat{\mathbf{A}}^0$ to new poses $\mathbf{p}^l$ and optimize these poses using a gradient descent.

Feature Reprojection Our mesh pose fitting is based on reprojection of $\hat{\mathbf{A}}^0$ to any new pose $\mathbf{p}^l$. First, we use projective texturing to map $\hat{\mathbf{A}}^0$ to $\mathcal{M}$. We obtain per-vertex features $\{\mathbf{a}_n\}$ by mapping each mesh vertex $\mathbf{u}_n$ to the image plane of the camera $\mathcal{C}$ and sampling $\hat{\mathbf{A}}^0$ as $\mathbf{a}_n = \hat{\mathbf{A}}^0[P(\mathbf{u}_n, \mathcal{C})]$, where $P(.)$ is a world space to image plane projection function and $[.]$ is a bilinear sampler. Finally, we transform $\mathcal{M}$ to $\mathcal{M}^l = \tau(\mathcal{M}, \mathbf{p}^l)$ for a given novel pose $\mathbf{p}^l$ and we render a feature image

$$\mathbf{I}^l_{\mathbf{A}} = r_{\mathbf{A}}(\mathcal{M}^l, \mathcal{C}, \{\mathbf{a}_n\}, \mathbf{B}_{\mathbf{A}}) \tag{2}$$

where $r_{\mathbf{A}}$ is a rasterization function interpolating the vertex attributes $\{\mathbf{a}_n\}$ and $\mathbf{B}_{\mathbf{A}}$ is a background feature map produced by inpainting the background $\hat{\mathbf{A}}^0 \odot (1-\psi)$ with a mean of valid features. Notice that Eq. 2 implies an approximate identity $\mathbf{I}^0_{\mathbf{A}} \approx \hat{\mathbf{A}}^0$, and we optimize $\mathbf{p}$ to improve this match for the full animation.

Mesh Pose Optimization We observe that direct optimization of each $\mathbf{p}^l$ independently is prone to local minima. Instead, we exploit the implicit bias of Multi-Layer-Perceptrons (MLPs) towards smooth functions, and regress $\mathbf{p}^l$ as a frame-dependent offset from an initial pose $\mathbf{p}_{\text{init}}$ such that $\mathbf{p}^l = \alpha m_\omega(\gamma(l)) + \mathbf{p}_{\text{init}}$, where $\alpha = 0.01$ is a scaling constant, γ is a frequency encoding [51], and $m(.)$ is an MLP with learnable parameters ω. We optimize ω by gradient descent to enforce semantic correspondences between the animated mesh and the video, i.e. $\mathbf{I}^l_{feat} \approx \hat{\mathbf{A}}^l$, using the rendering loss:

$$\mathcal{L}_r = 1 - \frac{1}{L\hat{H}\hat{W}} \sum_{l=0}^{L-1} \sum_{i\in\Omega_{\mathbf{A}}} \kappa(\mathbf{I}^l_{feat}[i], \hat{\mathbf{A}}^l[i]), \tag{3}$$

where $\kappa()$ is the cosine similarity (Eq. 1), $\Omega_{\mathbf{A}}$ is the spatial domain of $\hat{\mathbf{A}}$ and $[i]$ a spatial sampler.

Regularization losses First, our monocular video provides only a limited supervision for motion-in-depth. We discourage the optimization from explaining spatial deformation artifacts in the input video via motion-in-depth by per-vertex regularization loss

$$\mathcal{L}_d = \frac{1}{LN} \sum_{l=0}^{L-1} \sum_{n=0}^{N-1} ||(\bar{d}^0 - d^0_n) - (\bar{d}^l - d^l_n)||_1, \tag{4}$$

where d^l_n is the projected depth of vertex u_n in frame l, and $\bar{d}^l = 1/N \sum_{n=0}^{N-1} d^l_n$. Second, we enforce temporal smoothness beyond the MLP's implicit bias to further reduce jitter using the smoothness loss $\mathcal{L}_s = 1/((L-1)N) \sum_{l=0}^{L-2} ||\mathbf{p}^l - \mathbf{p}^{l+1}||_1$. Lastly, we penalize propagation of local spatial distortions from video by suppressing large deformations using the fidelity loss $\mathcal{L}_f = 1/(LN)\sum_{l=0}^{L-1}||\mathbf{p}^l||_1$. Consequently, our complete optimization objective is $\mathcal{L} = w_r\mathcal{L}_r + w_d\mathcal{L}_d + w_s\mathcal{L}_s + w_f\mathcal{L}_f$ with $w_r = 5$, $w_d = 0.01$, $w_s = 0.1$, $w_f = 0.01$.

4.4. Implementations Details

We implement our method in PyTorch [58] with PyTorch3D [67] mesh rasterizer, and we optimize the poses with the Adam optimizer [33] for 1 000 steps. We discuss further details in Appendix A.

Animation Models We experiment with four animation models for poses $\mathbf{p}$. For domain specific shapes, we use low-dimensional articulated models SMPL [45] (for humans), SMAL [103] (animals) and FLAME [38] (faces), where $\mathbf{p}^l$ are the joint angles and the other shape parameters are fixed. For other meshes, we use Neural Jacobian Fields (NJF) [1], which encodes the pose $\mathbf{p}^l$ by surface Jacobians, in combination with a single global translation and rotation - see Appendix A.2 for details and for an additional rigidity regularizer $\mathcal{L}_j$ applied for NJF.

VDMs We evaluate 2 VDMs: VideoComposer [91] (VC) and DynamiCrafter [93] (DC) with $\hat{\mathbf{A}}$ resolution of $(160, 88)$ and $(128, 72)$ respectively ($1/8$ of their outputs). We use their recommended inference schedulers with $T = 50$ steps. Our assumption of $\hat{\mathbf{A}}^0 \sim \mathbf{I}^0_{rgb}$ is satisfied by design for VC, and we present a solution for DC in Appendix A.1. We empirically find VC performs better for images with the background $\mathbf{B}$, while DC performs well even with a uniform white background. Additionally, we assessed another VDM, Stable Video Diffusion [5], but we discarded it due to its low motion quality (see Appendix D.1, D.5).

5. Experiments

We compare our zero-shot motion generation to other methods in a user study. Further, we quantitatively evaluate our pose fitting algorithm on a synthetic human motion dataset and measure the contribution of the individual components in an ablation study.

5.1. User Study

We compare our method to two other approaches for zero-shot 3D motion synthesis. First, we compare to DG4D [68] as an end-to-end shape-and-motion generative method based on image and video diffusion. Second, in absence of a class-agnotic method, we compare to a human motion diffusion model (MDM) [83] combined with motion retargeting (MT) [41]. We provide an additional qualitative comparison to a contemporary end-to-end generative method Consistent4D [28] in Appendix D.2. We run our method with both VC and DC backbones and use the same generated videos as inputs for DG4D (see Appendix B.1 for details). We use 9 meshes and a total of 12 prompts combined to obtain 2 human stimuli (using the SMPL mesh), 2 horses (SMAL), 2 faces (FLAME) and 4 other stimuli each with a unique mesh (NJF). See Fig. 1, Fig. 3, and the supplemental video for visual examples, Appendix B.2 for a complete list.

12 participants aged 24–41, naïve to the purpose of the experiment and with a normal or corrected-to-normal vision, participated in a $\sim$ 20 min low-risk IRB approved study, after signing an informed consent without any compensation. 16 frame (1 second) long video pairs from different methods were presented side-by-side in random order. Each displayed the same untextured animated shape from two viewpoints to clearly display the motion. Videos were looped until a binary answer was entered using a keyboard. The same stimuli were used for three different questions in three blocks. See Appendix B.3 for details.

In Fig. 4 (Left), we observe a statistically significant preference for our method compared to both DG4D and MDM-MT in terms of having “more natural motion”, “fewer visual artifacts” and “capturing the prompt better” ($p < 0.001$, binomial test). We provide a break-down for individual shapes in Appendix B.4. As expected, the human-specific MDM-MT approach excels for human stimuli but fails for morphologically distinct shapes, where correspondences are difficult to establish, which results in semantically incorrect and visually distracting motion (see Fig. 3 “Bunny”). In contrast, the other class-agnostic model, DG4D, struggles to accurately represent the video motion sequences leading to noisy reconstructions (see Fig. 3 “Raptor”). Moreover, the motion optimization in DG4D (Stage 2) takes 233 ± 5 seconds on an NVIDIA RTX 3090, while our method leverages fast rasterization and performs pose optimization in only 148 ± 39 seconds. Fig. 1, Fig. 3, Appendix B.4, and our video show more examples.

Input Accuracy After the main study, we asked each participant to additionally compare our textured output to the full DG4D color rendering and to the unprocessed VDM videos for overall preference (see the last bars Fig. 4 Left). The participants strongly prefer our method to DG4D, likely due to the more accurate geometry (Fig. 3). There was no effect when comparing to the VDMs, suggesting that our method closely preserves characters of the generated videos and should, therefore, benefit from future VDMs with a more accurate motion depiction.

5.2. Pose Optimization

We observe that the limiting factor of our method is the VDM motion quality. To remove this influence, we quantitatively evaluate performance of our pose fitting component using a captured human dancing motion dataset AIST++ [37] with known poses. First, we randomly select 20 test sequences

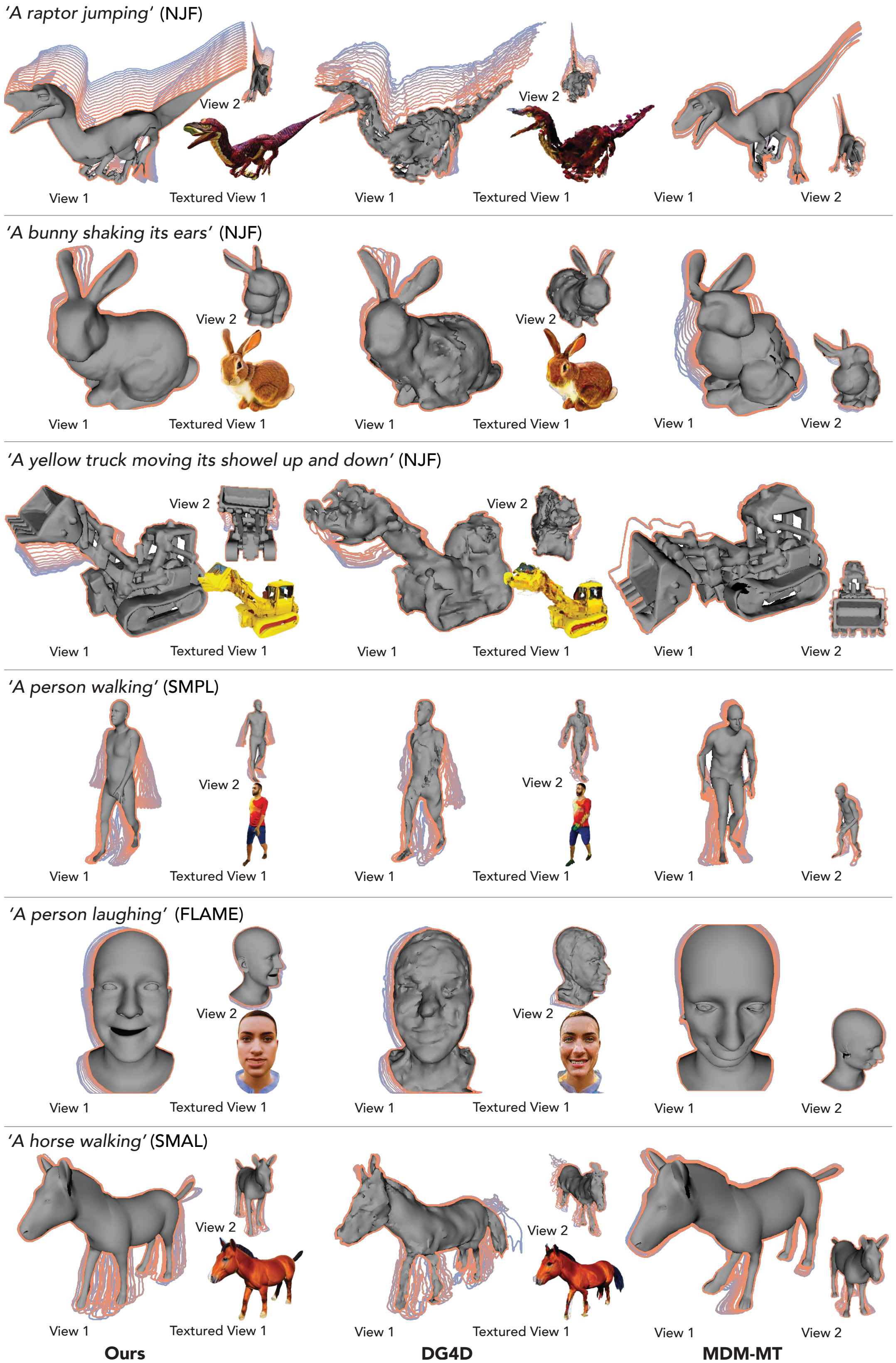

Figure 3. A qualitative comparison of our method to DG4D and MDM-MT for the prompts and the shapes used in our study. We display 2 untextured views of the last frame with one one additional textured image for reference. The contours convey the motion trajectory.

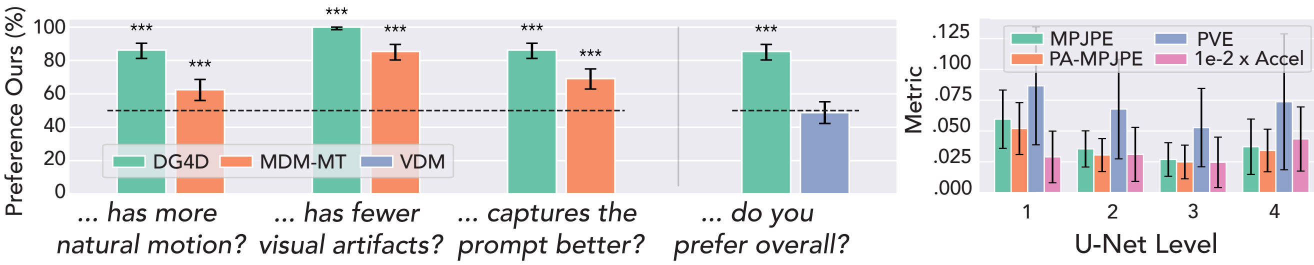

Figure 4. Left: Results of the user study, asking the question: *"Which video... ?"* For the first three questions we compare our method against untextured renders of DG4D and MDM-MT. For the last question we compare against the full RGB outputs of DG4D and the VDM output. *** denotes significance at $p < 0.001$ (bars show 95% confidence intervals). Right: Pose fitting errors for $\mathbf{A}_u^{\hat{t}}$ extracted across U-Net layers u with bars showing standard deviations.

and re-render the first 20 frames from each using the available SMPL mesh to simulate a perfect VDM. Then, we use VC to extract $\hat{\mathbf{A}}$ from the rendered video following Tang et al. [80] and optimize $\mathbf{p}$ for the SMPL model (Sec. 4.3) before evaluating the common metrics [75]: the Mean Per Joint Position Error (MPJPE), the Procrustes-aligned MPJPE (PA-MPJPE), the Per Vertex Error (PVE), and finally the Acceleration error (Accel) for smoothness.

We conduct three comparisons. First, we compare our single-view texturing (*textured*, Sec. 4.1) to a uniform gray shading (*untextured*). Second, we compare our semantic features $\hat{\mathbf{A}}$ (*Ours*) to RGB features (*RGB*) extracted directly from the input videos. Finally, we additionally test a state-of-the-art human pose estimation method WHAM [75] as a domain-specific reference. Since our method always starts with known $\mathbf{p}_{\text{init}}$, we emulate the same for WHAM by measuring its first-frame per-joint error and transform all predictions accordingly. This empirically improves WHAM scores relative to the unprocessed outputs. Appendix C provides details and alternative alignment strategies.

Results As summarized in Tbl. 1, *Ours* consistently achieves better results with *textured* inputs than with *untextured* inputs, which motivates our Single-View Texturing (Sec. 4.1). Furthermore, *Ours (full)* with semantic features $\hat{\mathbf{A}}$ achieves lower errors than the variant with *RGB* features, which documents the utility of these features for our task. Moreover, *Ours (full)* compares favorably even to the WHAM pose estimator despite the lack of human-specific training. This might be explained by the artificial appearance of our input videos which differ from common human pose estimation datasets. We do not claim general supremacy of our method for human pose estimation. This is showcased in Fig. 8 (right), where our method struggles to avoid physiologically implausible poses. Finally, in Fig. 6 we compare both features qualitatively in our full generative method and confirm that our semantic features lead to a better motion fit with fewer artifacts. See Appendix D.3 for more examples.

	MPJPE	PA-MPJPE	PVE	Accel
***Textured* (default)**				
WHAM	.059 ± .029	.042 ± .016	.075 ± .036	7.9 ± 9.0
RGB	.044 ± .051	.044 ± .042	.077 ± .059	7.5 ± 16.7
Ours (full)	**.041** ± .036	**.039** ± .035	**.063** ± .057	**5.0** ± 7.2
Untextured				
WHAM	.057 ± .028	**.039** ± .015	**.070** ± .035	7.4 ± 9.1
RGB	.146 ± .056	.126 ± .043	.203 ± .074	**3.2** ± 3.0
Ours	**.051** ± .037	.044 ± .034	.073 ± .054	4.7 ± 5.7

Table 1. The pose fitting performance of WHAM [75] and variants of our method for re-rendered AIST++ human body sequences [37]. Less is better for all metrics (see Sec. 5.2).

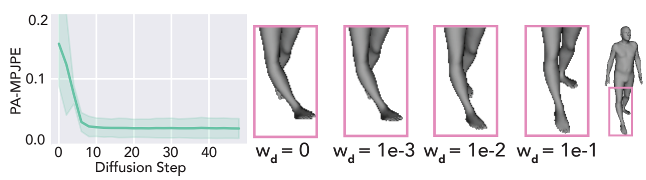

Figure 5. Left: PA-MPJPE ↓ with a standard deviation range for features $\mathbf{A}_{\hat{u}}^{t}$ extracted for different diffusion steps t. Right: Depth regularization prevents undesirable motion-in-depth explanations.

	MPJPE	PA-MPJPE	PVE	Accel
no $\mathcal{L}_s$	.027 ± .016	.025 ± .016	.053 ± .036	**2.72 ± 2.19**
no $\mathcal{L}_f$	.026 ± .017	.025 ± .018	**.103 ± .044**	2.52 ± 2.28
no $\mathcal{L}_d$	.025 ± .006	.024 ± .008	.046 ± .016	2.46 ± 2.04
Full	.027 ± .016	.025 ± .016	.053 ± .036	2.50 ± 2.36

Table 2. Performance of our ablated method variants in pose fitting. Notable performance impacts highlighted in red.

5.3. Ablations

We reuse the pose optimization experiment to validate our design choices. To this end, we follow the same procedure for 6 of the same AIST++ sequences [37]. First, we analyze the choice of $\hat{u}$ (Fig. 4 right) and $\hat{t}$ (Fig. 5 left) for extraction of $\hat{\mathbf{A}}$ using PA-MPJPE. We observe the best performance for $\hat{u} = 3$, which we consequently use for both VDMs in our other experiments. We further find our method is not

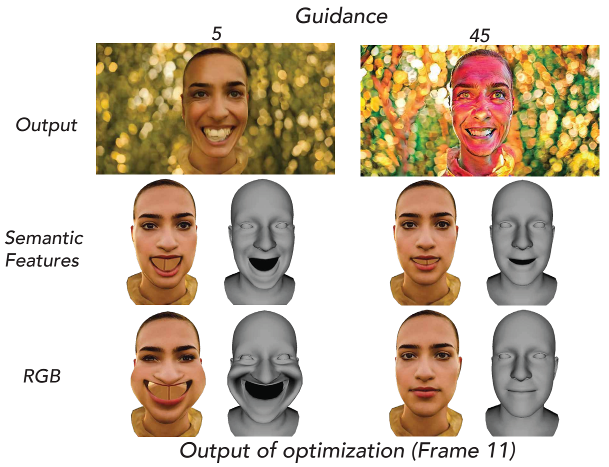

Figure 6. Effect of features on optimization under different guidance scales. Higher guidance scales are usually accompanied by stronger artifacts in RGB space (right column). The top row shows frame 11 of the VDM output, i.e. the target expression. The consecutive rows show semantic features vs RGB used for optimization.

sensitive to the choice of $\hat{t}$ above $t \approx 15$. Therefore, we select $\hat{t} = 20$ for VC and $\hat{t} = 40$ for DC.

Next, we ablate our regularization losses (Tbl. 2). As expected, the smoothness of $\mathcal{L}_s$ reduces the Acceleration error, while $\mathcal{L}_f$ reduces shape distortions recorded by the Per-Vertex Error. In contrast, the depth regularization of $\mathcal{L}_d$ does not lead to an improvement in performance metrics, but we observe that it discourages perceptually-objectionable depth errors (Fig. 5 right).

Finally, in Fig. 7 we ablate the mesh resolution in our method with NJF. We find that the output quality degrades gracefully and predictably with reduced vertex count. See Appendix D.4 for an extended discussion.

6. Discussion

Limitations and Future Work Single-view motion supervision struggles to resolve motion-in-depth or occlusions, which we mitigate using regularization at a risk of overall motion reduction (see Fig. 5 right). We acknowledge this as a limitation and a motivation for further research which could offer an improvement through multi-view supervision at the cost of additional training data [30]. We demonstrate a zero-shot method supporting a range of animation models, but we acknowledge that the high degree-of-freedom in NJF permits undesired distortions (see Fig. 8a). These could be potentially remedied though a static shape supervision inspired by 3D generative models [35] with a possible diversity reduction stemming from SDS [40]. On top of this, the motion produced by the current VDMs might not adhere to the prompt or might contain physically impossible transitions (see Fig. 8b). To counter this, the fast run-time of our

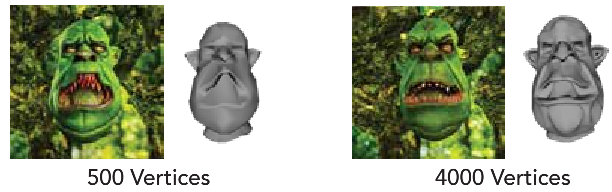

Figure 7. Effect of number of vertices on our method. Model source: Jaka Ardian 3D art / model from Indonesia.

Figure 8. Failure cases showing frames of the VC VDM output and our fitted motion. (a) The VDM produces fast motion accompanied by ear disappearance that our model explains as an undesired head deformation. (b) The VDM suddenly flips body orientation by 180 degrees which confuses our tracking and leads to self-intersections.

method could be combined with a suitable rejection heuristic. Furthermore, we expect to benefit from future VDM improvements [7, 9]. This will also allow for generating longer sequences, necessitating memory off-loading, which is currently absent in our implementation. Finally, an interesting future direction is to constrain the VDM generation with our simultaneously optimized 3D animation model in order to prevent any distortions from emerging.

Conclusion We presented a novel generative method for zero-shot 3D animation. Despite its limitations stemming from the single-view supervision, we demonstrated that it produces visually preferable motions across diverse unseen 3D shapes at computation cost lower than end-to-end 4D generative methods. We see our method as a capable tool for analysis of motion spaces in VDMs, and for affordable re-animation of static 3D assets in virtual environments.

Ethical Considerations Our method produces novel poses for 3D objects including human bodies and faces, but we do not focus on realistic appearance modeling. The biases in backbone VDMs can influence our method and are a priority research interest to the community.

Acknowledgments This work was partially supported by the Convergence AI Immersive Tech Lab at TU Delft.

References

[1] Noam Aigerman, Kunal Gupta, Vladimir G. Kim, Siddhartha Chaudhuri, Jun Saito, and Thibault Groueix. Neural jacobian fields: learning intrinsic mappings of arbitrary meshes. *ACM Trans. Graph.*, 41(4), 2022. 1, 2, 5, 14

[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. *arXiv preprint arXiv:2311.17984*, 2023. 1, 2

[3] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out? 3d animal reconstruction with expectation maximization in the loop. In *European Conference on Computer Vision*, pages 195–211. Springer, 2020. 13

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *ACM SIGGRAPH*, page 187–194, 1999. 2

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1, 2, 5, 13, 20

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 2

[7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2, 8

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017. 2

[9] Google DeepMind. Veo, 2024. 2, 8

[10] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 1

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[12] Niladri Shekhar Dutt, Sanjeev Muralikrishnan, and Niloy J Mitra. Diffusion 3d features (diff3f): Decorating untextured shapes with distilled semantic features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4494–4504, 2024. 1

[13] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 20

[14] William Gao, Noam Aigerman, Thibault Groueix, Vova Kim, and Rana Hanocka. Textdeformer: Geometry manipulation using text guidance. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 14

[15] Michael Garland and Paul S Heckbert. Surface simplification using quadric error metrics. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 209–216, 1997. 14

[16] Michael Gleicher. Retargetting motion to new characters. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 33–42, 1998. 3

[17] Michael Gleicher, S Grassia, Zoran Popovic, S Rosenthal, and Jeffrey Thingvold. Motion editing, principles, practice and promise. In *SIGGRAPH 2000 Course Notes 26*, 2000. 1

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. 2014. 2

[19] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 2

[20] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 2

[21] 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. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 2

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3, 13

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2, 3

[24] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1, 2

[25] David Hogg. Model-based vision: a program to see a walking person. *Image and Vision computing*, 1(1):5–20, 1983. 2

[26] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2

[27] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language.

Advances in Neural Information Processing Systems, 36, 2024. 2

[28] Yanqin Jiang, Li Zhang, Jin Gao, Weiming Hu, and Yao Yao. Consistent4d: Consistent 360° dynamic object generation from monocular video. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 2, 5, 20

[29] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 2

[30] Roy Kapon, Guy Tevet, Daniel Cohen-Or, and Amit H Bermano. Mas: Multi-view ancestral sampling for 3d motion generation using 2d diffusion. *arXiv preprint arXiv:2310.14729*, 2023. 2, 8

[31] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 2

[32] Aliasghar Khani, Saeid Asgari, Aditya Sanghi, Ali Mahdavi Amiri, and Ghassan Hamarneh. Slime: Segment like me. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 2

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5, 13

[34] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[35] Hanhung Lee, Manolis Savva, and Angel X Chang. Text-to-3d shape generation. In *Computer Graphics Forum*, page e15061. Wiley Online Library, 2024. 1, 8

[36] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172, 2000. 2

[37] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 5, 7, 15

[38] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph. (SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 1, 2, 5, 13, 14

[39] Tianyu Li, Jungdam Won, Alexander Clegg, Jeonghwan Kim, Akshara Rai, and Sehoon Ha. Ace: Adversarial correspondence embedding for cross morphology motion retargeting from human to nonhuman characters. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 3

[40] 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. *arXiv preprint arXiv:2311.11284*, 2023. 1, 2, 8

[41] Zhouyingcheng Liao, Jimei Yang, Jun Saito, Gerard Pons-Moll, and Yang Zhou. Skeleton-free pose transfer for stylized 3d characters. In *European Conference on Computer Vision*, pages 640–656. Springer, 2022. 3, 5, 14

[42] 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. *arXiv preprint arXiv:2312.13763*, 2023. 1, 2

[43] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2022. 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 2, 23

[45] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 1, 2, 5, 13, 14

[46] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[47] Sebastian Lutz, Richard Blythman, Koustav Ghosal, Matthew Moynihan, Ciaran Simms, and Aljosa Smolic. Jointformer: Single-frame lifting transformer with error prediction and refinement for 3d human pose estimation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1156–1163. IEEE, 2022. 2

[48] Adrien Maglo, Guillaume Lavoué, Florent Dupont, and Céline Hudelot. 3d mesh compression: Survey, comparisons, and emerging trends. *ACM Computing Surveys (CSUR)*, 47 (3):1–41, 2015. 2

[49] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5442–5451, 2019. 2

[50] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 2

[51] 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. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 2, 4, 13, 15

[52] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 1, 2

[53] Atsuhiro Noguchi, Umar Iqbal, Jonathan Tremblay, Tatsuya Harada, and Orazio Gallo. Watch it move: Unsupervised discovery of 3d joints for re-posing of articulated objects.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3687, 2022. 2

[54] Joseph O'rourke and Norman I Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):522–536, 1980. 2

[55] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2

[56] Frederick I Parke. Computer generated animation of faces. In *Proceedings of the ACM annual conference-Volume 1*, pages 451–457, 1972. 2

[57] Frederic Ira Parke. *A parametric model for human faces.* The University of Utah, 1974. 2

[58] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5, 13

[59] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 2

[60] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. Articulated people detection and pose estimation: Reshaping the future. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3178–3185. IEEE, 2012. 2

[61] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, 2016. 2

[62] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3, 14

[63] Jose Luis Ponton, Haoran Yun, Andreas Aristidou, Carlos Andujar, and Nuria Pelechano. Sparseposer: Real-time full-body motion reconstruction from sparse data. *ACM Transactions on Graphics*, 43(1):1–14, 2023. 2

[64] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2

[65] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. *arXiv preprint arXiv:2011.13961*, 2020. 2

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

[67] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 5

[68] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 1, 2, 5, 14, 20

[69] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 1

[70] Helge Rhodin, James Tompkin, Kwang In Kim, Kiran Varanasi, Hans-Peter Seidel, and Christian Theobalt. Interactive motion mapping for real-time character control. In *Computer Graphics Forum*, pages 273–282. Wiley Online Library, 2014. 3

[71] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 3, 20

[72] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graph. (SIGGRAPH Asia)*, 36(6), 2017. 2

[73] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2

[74] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1, 2

[75] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. *arXiv preprint arXiv:2312.07531*, 2023. 2, 7, 20

[76] Leonid Sigal, Sidharth Bhatia, Stefan Roth, Michael J Black, and Michael Isard. Tracking loose-limbed people. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, pages I–I. IEEE, 2004. 2

[77] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1): 4–27, 2010. 2

[78] 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. In *Proceedings of the 40th International Conference on Machine Learning*, pages 31915–31929, 2023. 1, 2

[79] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1, 2, 3

[80] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 1, 2, 3, 7, 20

[81] Jochen Tautges, Arno Zinke, Björn Krüger, Jan Baumann, Andreas Weber, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bernd Eberhardt. Motion reconstruction using sparse accelerometer data. *ACM Transactions on Graphics (ToG)*, 30(3):1–12, 2011. 2

[82] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 2

[83] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 5, 14

[84] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Wang Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, pages 703–735. Wiley Online Library, 2022. 2

[85] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014. 2

[86] Lukas Uzolas, Elmar Eisemann, and Petr Kellnhofer. Template-free articulated neural point clouds for reposable view synthesis. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[87] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022. 13

[88] Jiashun Wang, Xueting Li, Sifei Liu, Shalini De Mello, Orazio Gallo, Xiaolong Wang, and Jan Kautz. Zero-shot pose transfer for unrigged stylized 3d characters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8704–8714, 2023. 3

[89] Peng-Shuai Wang, Yang Liu, and Xin Tong. Dual octree graph networks for learning adaptive volumetric shape representations. *ACM Transactions on Graphics (SIGGRAPH)*, 41(4), 2022. 14

[90] Xinzhou Wang, Yikai Wang, Junliang Ye, Zhengyi Wang, Fuchun Sun, Pengkun Liu, Ling Wang, Kai Sun, Xintong Wang, and Bin He. Animatabledreamer: Text-guided non-rigid 3d model generation and reconstruction with canonical score distillation. *arXiv preprint arXiv:2312.03795*, 2023. 1, 2

[91] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 5, 13, 20

[92] Lance Williams. Casting curved shadows on curved surfaces. In *Proceedings of the 5th annual conference on Computer graphics and interactive techniques*, pages 270–274, 1978. 3

[93] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023. 1, 2, 5, 13, 20

[94] Tianhan Xu and Tatsuya Harada. Deforming radiance fields with cages. In *European Conference on Computer Vision*, pages 159–175. Springer, 2022. 2

[95] Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Hi-lassie: High-fidelity articulated shape and skeleton discovery from sparse image ensemble. *arXiv preprint arXiv:2212.11042*, 2022. 2

[96] Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*, 2023. 1, 2

[97] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[98] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3, 14

[99] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. Rohm: Robust human motion reconstruction via diffusion. *arXiv preprint arXiv:2401.08570*, 2024. 2

[100] 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. 1, 2

[101] Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text-and image-guided 4d scene generation. *arXiv preprint arXiv:2311.16854*, 2023. 1, 2

[102] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[103] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *CVPR*, pages 6365–6373, 2017. 1, 2, 5, 13, 14

[104] Xinxin Zuo, Sen Wang, Jiangbin Zheng, Weiwei Yu, Minglun Gong, Ruigang Yang, and Li Cheng. Sparsefusion: Dynamic human avatar modeling from sparse rgbd images. *IEEE Transactions on Multimedia*, 23:1617–1629, 2021. 14