Animating the Uncaptured: Humanoid Mesh Animation with Video Diffusion Models

Marc Benedí San Millán Angela Dai Matthias Nießner
Technical University of Munich

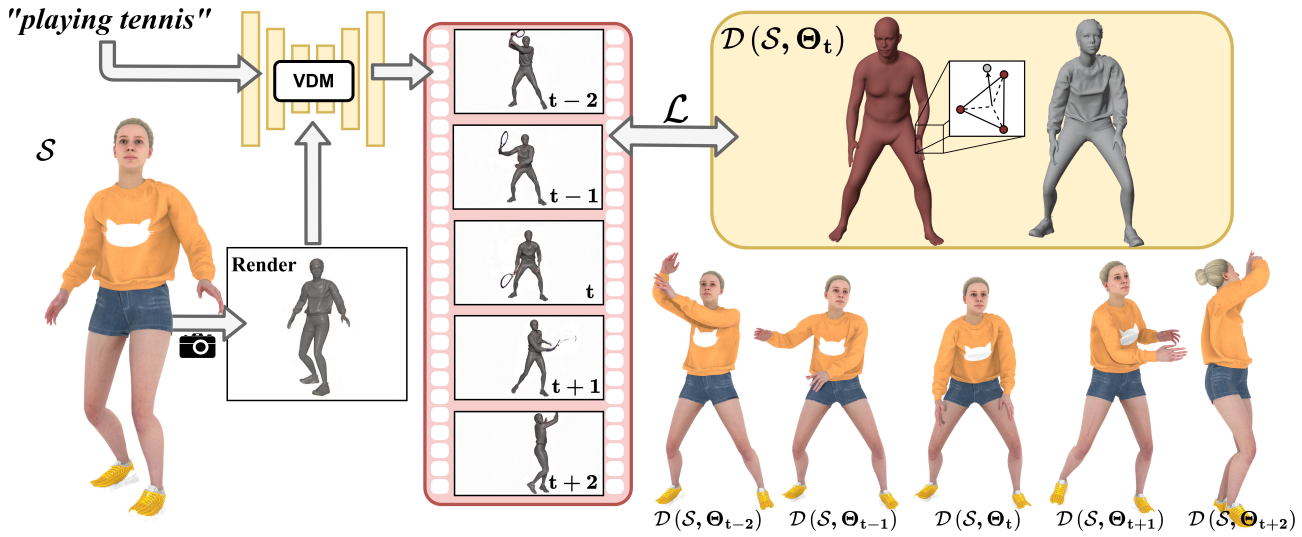

Figure 1. **Animating the Uncaptured**, a novel approach for animating 3D humanoid meshes from text prompts. Given an input mesh and a text prompt describing a motion, we use a video diffusion model to generate a video of the mesh performing the motion. We then transfer the motion to the mesh by sparse and dense tracking of the video.

Abstract

Animation of humanoid characters is essential in various graphics applications, but requires significant time and cost to create realistic animations. We propose an approach to synthesize 4D animated sequences of input static 3D humanoid meshes, leveraging strong generalized motion priors from generative video models – as such video models contain powerful motion information covering a wide variety of human motions. From an input static 3D humanoid mesh and a text prompt describing the desired animation, we synthesize a corresponding video conditioned on a rendered image of the 3D mesh. We then employ an underlying SMPL representation to animate the corresponding 3D mesh according to the video-generated motion, based on our motion optimization. This enables a cost-effective and accessible solution to enable the synthesis of diverse and realistic 4D animations.

Project Website: https://marcb.pro/atu

1. Introduction

Character animation is fundamental in computer graphics – enabling lifelike, expressive, and engaging virtual characters for applications such as movies, video games, mixed reality, robotics, and many more. Such characters portrayed with realistic motions help to drive storytelling and interactivity, making crafted content more engaging and immersive.

Traditionally, character animation requires significant manual labor from highly-trained artists, who manually craft character rigs, define keyframes for motions, and fine-tune detailed motion behavior. This is both costly and requires a significant amount of tedious effort from skilled artists. Thus, leveraging learned motion priors to inform character animation would enable much more efficient animation synthesis. However, ground-truth capture of human motion is very difficult and expensive to acquire, resulting in very limited data available [24, 29, 57] for training such motion priors, strongly limiting the diversity and generalization capability of methods fully supervised with such 4D data [41, 46].

Recently, video diffusion models [1, 14, 53], trained on large-scale datasets with abundant video data, have demonstrated the ability to generate diverse and realistic videos conditioned on textual prompts. This suggests that these models implicitly learn motion priors that capture how the world evolves over time. Building on this insight, we propose a general approach for animating a 3D humanoid mesh based on a text description of the intended motion. Instead of relying on small and constrained 4D human motion capture datasets, we leverage motion priors learned by video generative models, which possess strong representational capacity and can synthesize diverse, high-fidelity human motion sequences.

Given a text prompt describing the intended motion, we generate a synthetic video of a 3D humanoid mesh performing the specified motion using a text-to-video (T2V) diffusion model. This model is conditioned on both the rendering of the 3D mesh and the text prompt. To facilitate motion tracking, we employ the SMPL body model [25] as a deformation proxy for the input mesh, enabling us to estimate the SMPL parameters and track the 3D mesh throughout the generated video frames.

Our approach begins by registering the SMPL body model to the input mesh, leveraging estimated body joint locations from multiple views. We then reparameterize the input mesh's vertex coordinates into barycentric coordinates relative to the SMPL mesh faces. To transfer motion from the generated video to the input mesh, we extract 2D body landmarks, silhouettes, and dense DINOv2 [31] features from the video frames, which serve as tracking cues for accurate motion reconstruction. This provides an easy, accessible approach to generate a wide range of realistic 4D humanoid animations.

Our contributions are summarized as follows:

- We tackle the task of mesh animation by leveraging the strong generative capabilities of text-to-video (T2V) diffusion models.
- We propose a pipeline to robustly track the motion from the generated video by combining body landmarks, silhouettes, and dense features.

2. Related Work

Text-to-Human Motion Generation Human motion generation is a pivotal area in computer graphics, focusing on synthesizing realistic human movements for applications such as animation and virtual reality.

Although human motion can be captured using motion capture (MoCap) systems [15, 32, 49], this process requires specialized setups and actors which is expensive and time-consuming. This motivates the development of data-driven methods that can generate human motions from different input signals, such as actions [11], and audio [4, 5, 7]. enabling the creation of diverse and realistic human animations without the need to manually capture every new motion.

In particular, text-to-human motion generation aims to synthesize human motion sequences based on text prompts describing the desired action in natural language. [56] proposed the first text-conditioned diffusion model [13] for human motion generation. Such text-to-motion methods typically require paired text and body pose parameters for training, which are often obtained from motion capture data. However, such data is limited in quantity and diversity, and may not generalize well to unseen actions [11, 12, 37].

Similar to our approach, MotionDreamer [47] proposes to extract motions from video diffusion models to animate meshes. As proposed by [27, 45, 55], they extract semantic features from the intermediate activations of the diffusion model and perform feature matching between frames. This is an expensive process and limits their pipeline to use low resolution features. In contrast, we employ both sparse and dense features for mesh registration and tracking, and our focus on humanoid meshes enables us to use stronger body priors for regularization.

Human Pose and Shape Estimation from Images and Videos Monocular HPS estimation remains particularly challenging due to the inherent ambiguity in the 2D-to-3D mapping and the absence of depth information. To solve this, methods often rely on statistical body models [16, 25, 33, 35, 52] which provide a prior of shapes and poses and a low-dimensional parameterization of human body.

Two main approaches have been widely adopted to address this problem: optimization-based methods and regression-based methods. *Optimization-based* methods iteratively refine 3D pose and shape parameters by minimizing reprojection error between the 3D model and 2D image observations, such as silhouettes or body keypoints [3, 18, 26, 36]. [2] proposes to predict 2D joints location from an image and then optimize the SMPL body prameters to minimize the reprojection error. To seek a stronger body prior, [35] trains a variational autoencoder [19] to learn the representation of human poses. *Regression-based* methods leverage deep learning to directly estimate 3D pose and shape parameters from images or videos. Frameworks such as HMR [17] and SPIN [21] employ neural networks to predict SMPL parameters by learning from paired image data with pseudo-ground truth annotations. While these methods achieve real-time inference and improved robustness compared to optimization-based approaches, their performance is limited by the quality and diversity of the training datasets. As a result, regression-based models often struggle to generalize to out-of-distribution data, such as synthetic videos generated by text-to-video models.

Recent advancements have extended the image-based formulation to video-based. Crucially, these methods leverage temporal information to improve the consistency and realism

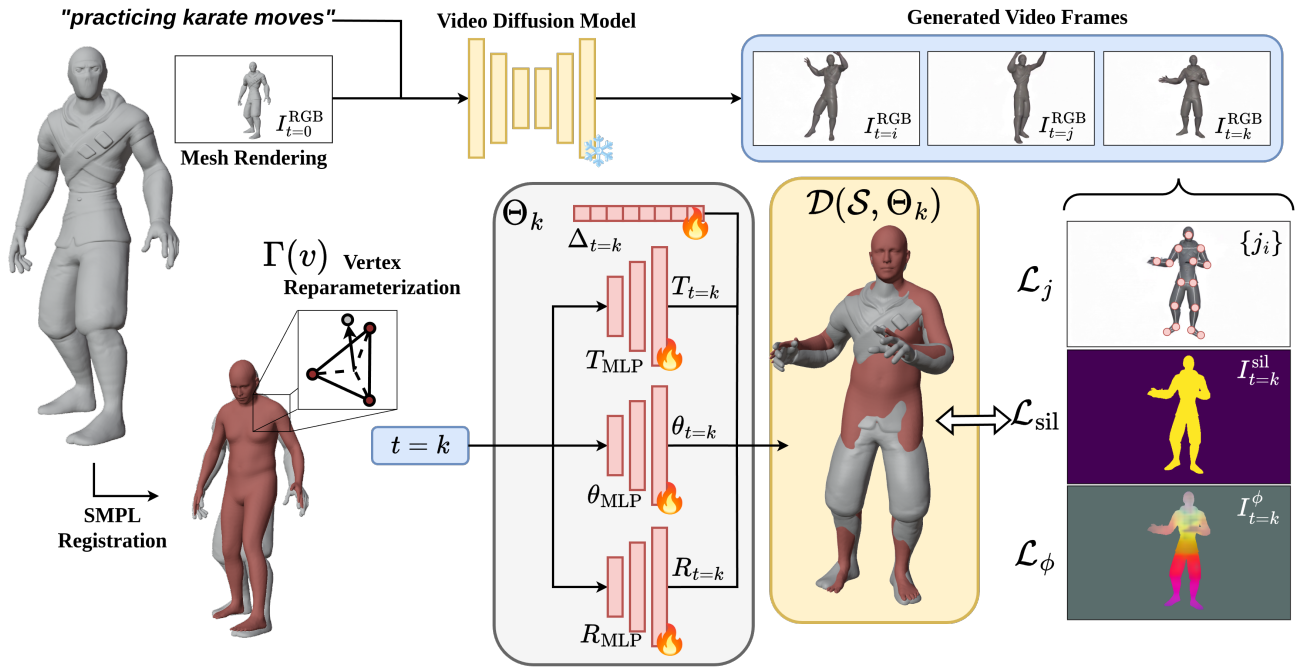

Figure 2. **Method Overview.** Given a mesh in an arbitrary pose and a text prompt describing the desired motion, we generate a video conditioned on the text prompt and the rendering of the mesh. We leverage the SMPL body model as a deformation proxy and to track the motion from the video and transfer it to the input mesh. For this, we fit SMPL to the input mesh and associate the vertices of the input mesh with the SMPL faces (Sec. 3.3.1). Finally, we optimize the SMPL paramters to match the motion in the video based on estimated body landmarks, silhouette and DINOv2 features from the frames (Sec. 3.3.2).

of the estimated human motions across frames [10, 42, 44].

Such human pose estimation from images and videos have also been used to synthesize 3D and 4D human-object interactions [22, 23]. In this work, we use an optimization-based approach to focus on humanoid animation. Our approach robustly handles the synthetic videos generated by a text-to-video model to effectively generate diverse, realistic motion for various humanoid meshes.

3. Method

Our method tackles the task of text-to-motion for humanoid meshes. Given a text prompt $(\mathcal{P})$ describing the desired motion and a humanoid mesh $(\mathcal{S})$ in an arbitrary pose, the method generates the deformation parameters (Θ) to animate the mesh over time. For this, we leverage the motion priors of Video Diffusion Models (VDMs) by generating a video conditioned on the prompt and the rendering of the mesh $\left(I^{\text{RGB}}_{t=0} \in \mathbb{R}^{H\times W\times 3}\right)$. We use the SMPL body model [25] as a deformation proxy to animate the input mesh through its parameters (Sec. 3.3.1). We then optimize the deformation proxy parameters to match the motion in the generated video (Sec. 3.3.2). The overview of our method is illustrated in Fig. 2.

3.1. Preliminaries

SMPL The Skinned Multi-Person Linear (SMPL) model [25] is a parametric body model that represents body shape and pose variations using a learned low-dimensional representation. It is defined by the shape parameters $\beta \in \mathbb{R}^{10}$ and the pose parameters $\theta \in \mathbb{R}^{23\times 3}$, where the shape components correspond to the principal components of the body shape and the pose parameters represent the rotations of 23 skeletal joints in axis-angle representation.

In this work we use the SMPL model as a deformation proxy to animate the input mesh. In particular, we use the encoding $Z \in \mathbb{R}^{32}$ of the variational autoencoder VPoser [35] as the pose representation for this work. Directly optimizing the latent representation ensures that the pose parameters lie within the manifold of valid human poses. To simplify the notation, we use the symbol θ to refer both to the pose parameters and the VPoser encoding.

3.2. Video Generation and Frame Features

We start by normalizing $\mathcal{S}$ to unit scale and centering it around the origin. We define $P : (\mathcal{S}, C) \rightarrow I \in \mathbb{R}^{H\times W\times 3}$ as the rendering function of the mesh $\mathcal{S}$ from camera C to obtain the frame I^{RGB}.

Video Generation We use a Video Diffusion Model (VDM) to generate a video $\left\{I_t^{\text{RGB}}\right\}_{t=0}^{F-1}$ with F frames depicting the mesh performing the motion described by the prompt. For this, we condition the VDM with the frame I_0^{RGB} and the prompt $\mathcal{P}$. Note that the first frame of the generated video is the same as the rendered image I_0^{RGB}.

Body Landmarks Estimation We use MediaPipe Pose Landmarker [26] to estimate the body pose landmarks for video. For each frame, MediaPipe provides 33 landmarks, (j_i, ω_i), where j_i are the normalized pixel coordinates and ω_i are the confidence scores. We rearrange them to be in the same order as the SMPL joints, and apply a smoothing filter to mitigate the effect of noisy predictions.

Dense Features Due to the jittery and sparse nature of landmark estimation, we additionally extract dense per-pixel features from the frames using a pre-trained DINOv2 [31] and obtain $\{I_t^{\phi}\}_{t=0}^{F-1}$. Following Dutt et al. [6], we annotate the vertices of the mesh $\mathcal{V} = \{v_i\}$ with the features $\mathcal{V}^{\phi} = \{v_i^{\phi}\}$, where v_i^{ϕ} is the feature vector of vertex v_i. To obtain the features, we render the mesh from 100 equally spaced cameras along the surface of a sphere centered on the mesh, extract the features of each view, and back-project and accumulate them on the mesh vertices. Finally, due to memory constraints, we perform PCA on the dense features and keep the first 64 components.

Silhouette We extract the silhouette I^{sil} of the mesh by thresholding the white background.

3.3. Optimization

We reduce the task of transferring the motion from the video to the mesh as a video tracking problem. First, we register the SMPL [25] model to the input mesh and reparameterize $\mathcal{V}$'s coordinates with respect to the SMPL faces $\mathcal{F}^{\text{SMPL}}$ (Sec. 3.3.1). Finally, we optimize the parameter Θ to deform the input mesh to match the motion in the video(Sec. 3.3.2).

Model Parameters The deformation function $\mathcal{D} : (\mathcal{S}, \Theta_t) \rightarrow \mathcal{S}_t$ is a function that deforms the input mesh $\mathcal{S}$ using the deformation parameters Θ_t for frame t. Note that for $\mathcal{D}(\mathcal{S}, \Theta_0) = \mathcal{S}$ due to the video generation process being conditioned on I_0^{RGB}. The parameters Θ are defined as $(s, \beta, \theta_t, T_t, R_t, \Delta_t)$, where s is the scale of the SMPL model, β is the shape parameters, and θ_t, T_t, R_t, and Δ_t are the pose, translation, rotation parameters and per-vertex offsets for each frame t. Note that the shape and scale parameters are shared across all frames. For simplicity, we use Θ_t to refer to the deformation parameters for frame t: $\Theta = \{\Theta_t\}_{t=0}^{F-1}$.

To mitigate the influence of noisy signals during optimization, such as the jitter in the estimated landmarks, we opt to use shallow Multi-Layer Perceptrons (MLPs) to parameterize θ_t, T_t, R_t. That is, we use θ_{MLP}, T_{MLP}, and R_{MLP} to predict their corresponding SMPL parameters for each frame. As input, they take the Sinusoidal Positional Encoding [48] of the frame index. For simplicity, we use the same notation to refer to the SMPL parameters produced by the MLPs. For instance, $\theta_t = \theta_{\text{MLP}}(PE(t))$ represents the pose parameters predicted by the MLP for time step t.

3.3.1. SMPL Registration

We choose to leverage the SMPL model [25] for the following reasons: (1) due to monocular tracking being inherently ambiguous, we benefit from the body prior in SMPL and VPoser [35], which ensures that the optimization remains within the plausible space of human poses; (2) the input mesh doesn't incorporate any deformation model such as skeleton or blend shapes.

We optimize for s, β, θ_0, T_0, R_0; and minimize both terms in Eq. (1) and Eq. (2). The first term ensures that the SMPL joints $\hat{J}$ are close to the estimated 3D joint locations J from the input mesh, while the second term minimizes the distance between the SMPL vertices and their nearest neighbors on the input mesh. To obtain J, we triangulate the 2D landmark predictions from MediaPipe [26] using views sampled around the mesh.

$$\mathcal{L}_J = \frac{1}{N} \sum_{i=1}^{N} \omega_i \left\| \hat{J}_i - J_i \right\|_2^2 \tag{1}$$

$$\mathcal{L}_{\text{p2p}}(\mathcal{V}) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \left\| v - \text{NN}(v, \mathcal{V}^{\text{SMPL}}) \right\|_2^2 \tag{2}$$

We also include simple L_2 priors for the shape and pose parameters, as defined in Eq. (3) and Eq. (4), respectively.

$$\mathcal{L}_{\beta}(\beta) = ||\beta|| \tag{3}$$

$$\mathcal{L}_{\theta}(\theta_t) = ||\theta_t|| \tag{4}$$

SMPL as a Deformation Proxy Finally, we reparametrize the input mesh vertices $\mathcal{V}$ with respect to the closest faces on the SMPL model $\mathcal{F}^{\text{SMPL}}$. We define the function Ψ that maps the input mesh vertices to their corresponding SMPL faces:

$$\Psi : v \rightarrow f^{\text{SMPL}} \tag{5}$$

We compute the barycentric coordinates of the corresponding SMPL face to obtain the 3D location of the

input mesh vertices in the SMPL space. We define this reparametrization as:

$$\Gamma(v) = \sum_{v_i^{\text{SMPL}} \in \Psi(v)} \gamma_i v_i^{\text{SMPL}} + d\mathbf{n} \tag{6}$$

where γ_i are the barycentric coordinates of the input mesh vertices v with respect to the SMPL face f^{SMPL}, d is the distance from the input mesh vertex to the SMPL face, and $\mathbf{n}$ is the normal of the SMPL face.

This ensures that the input mesh vertices are attached to the SMPL model, allowing us to optimize the SMPL parameters to match the motion in the video. To address potential outliers, we apply a robust filtering mechanism that identifies and excludes mismatches. Outliers are detected by combining multiple criteria: (1) absolute distance between the input mesh vertices and the SMPL model; (2) the angular deviation between vertex normals and face normals of the closest SMPL triangle, ensuring alignment within a threshold of 45 degrees; (3) neighborhood statistics, which analyze the mean and standard deviation of distances between vertices to identify points that deviate significantly from their neighbors.

where γ_i are the barycentric coordinates of the input mesh vertices v with respect to the SMPL face f^{SMPL}.

3.3.2. Video Tracking and Motion Transfer

We transfer the motion from the generated video to the input mesh by optimizing the following parameters of the deformation model Θ_t: T_t, R_t, θ_t, and Δ_t, keeping fixed the shape and scale parameters β and s. We formalize the deformation model as follows:

$$\mathcal{D}(\mathcal{S}, \Theta_t) = \Gamma\left(R_t \cdot \text{SMPL}(\beta, \theta_t) + T_t\right) + \Delta_t \tag{7}$$

Loss terms We optimize the parameters described above to minimize Eq. (8).

$$\mathcal{L}_{\text{total}} = \frac{1}{F} \sum_{t=1}^{F-1} (\mathcal{L}_j + \mathcal{L}_{\text{sil}} + \mathcal{L}_\phi) + \mathcal{L}_{\text{regs}} \tag{8}$$

The first data term, $\mathcal{L}_j$, shown in Eq. (9), minimizes the distance between the re-projected SMPL joints ($\hat{j}$), and the predicted 2D landmarks (j), where w is the confidence score of the predicted landmarks, ρ is the German-McClure loss function [8], and N is the number of landmarks.

$$\mathcal{L}_j = \frac{1}{N} \sum_{i=1}^{N} w_i \rho\left(\hat{j}_i - j_i\right) \tag{9}$$

The second data term, $\mathcal{L}_{\text{sil}}$, shown in Eq. (10), minimizes the binary cross-entropy loss between the rendered silhouette ($\hat{I^{\text{sil}}}$) and the silhouette extracted from the generated video (I^{sil}), where $N = HW$ is the number of pixels.

$$\mathcal{L}_{\text{sil}} = -\frac{1}{N} \sum_{i=1}^{N} \left(I_{t,i}^{\text{sil}} \log(\hat{I_{t,i}^{\text{sil}}}) + (1 - I_{t,i}^{\text{sil}}) \log(1 - \hat{I_{t,i}^{\text{sil}}}) \right), \tag{10}$$

The final data term, $\mathcal{L}_\phi$, shown in Eq. (11), minimizes the cosine similarity between the rendered features ($\hat{I^\phi}$) and the dense features extracted from the generated video (I^ϕ).

$$\mathcal{L}_\phi = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{\hat{I^\phi}_{t,i} \cdot I_{t,i}^\phi}{\|\hat{I^\phi}_{t,i}\|_2 \|I_{t,i}^\phi\|_2} \right), \tag{11}$$

The last term, $\mathcal{L}_{\text{regs}}$, includes all regularization terms: $\mathcal{L}_{\text{temp}}$, $\mathcal{L}_{\text{ex. ben.}}$, and $\mathcal{L}_\theta$, which are defined in Eq. (12), Eq. (13), and Eq. (4), respectively.

Temporal regularizers are used to ensure smooth motion across frames and mitigate the impact of landmark jitter. In particular, we penalize abrupt changes in translation, rotation, pose parameters, and 3D joint locations between consecutive frames. The temporal regularization terms are defined as follows:

$$\mathcal{L}_{\text{temp}}(x) = \sum_{t=1}^{T} \|x_t - x_{t-1}\|_2 \tag{12}$$

Inspired by [2], we also include a term to penalize extreme bending of the knees and elbows. See (13). This term is defined as the sum of the squared angles between the upper and lower limbs, ensuring that implausible poses with excessive bending are avoided.

$$\mathcal{L}_{\text{ex. ben.}}(\theta) = \sum_{i \in (\text{elbows, knees})} \exp\{\theta_i\} \tag{13}$$

Additionally, as described in Sec. 3.3.1, we penalize deviations from the manifold of valid human poses by using the VPoser regularization term defined in Eq. (4).

Finally, we employ As-Rigid-as-Possible [43] regularization to ensure that the resulting mesh deformation is smooth and preserves the mesh's intrinsic structure.

Feature Mapper Due to the appearance of the mesh $\mathcal{S}$ and the generated video $\{I_t^{\text{RGB}}\}_{t=0}^{F-1}$ diverging over time, we use optimize for a vector that learns to project the vertex features $\mathcal{V}^\phi$ to a space that is more similar to the features extracted from the video I^ϕ. Its parameters are optimized in a self-supervised manner.

4. Experiments

4.1. Implementation Details

In our implementation, we use PyTorch 2.0.1 [34] and CUDA 11.7 [30], and perform differentiable rendering with PyTorch3D 0.7.7 [40]. We use SMPL [25] and VPoser [35]

Figure 3. **Qualitative Results**. Visualization of generated mesh animations with our method. Each row shows: the text prompt, the input mesh, and intermediate frames of the generated motion. For all generations, we visualize the frames from the front and side views.

as body priors. For frame processing, we MediaPipe [26] and DINOv2 [31] to extract landmarks and dense features respectively. Our neural parameterization consists of shallow multilayer perceptrons (MLPs) with four layers and 128 hidden units per layer. We use a 64-dimensional positional encoding of the frame index as input. We used two different video diffusion models: RunwayML [9] and Kling AI [38]. Both models generate 5-second videos at a resolution of 768×1280 pixels, which we downsample to 384×640 pixels during optimization due to memory constraints. For SMPL registration, as described in Sec. 3.3.1, we set the batch size to 8 and optimize for $1,000$ iterations or until convergence, using a learning rate of 0.01 with the Adam optimizer [20]. Similarly, for video tracking, as described in Sec. 3.3.2, we use the same batch size and optimize for 4000 iterations or until convergence, with a learning rate of 0.001.

4.2. Video Tracking

We evaluate the tracking performance of our method using rendered videos of animated untextured meshes, as this enables evaluation with ground truth while also mimicking the output of a text-to-video model conditioned on untextured meshes.

Datasets For evaluation, we use the CAPE dataset [28, 39], which provides 4D sequences of clothed humans in motion. Each sequence includes a 3D human body mesh with the corresponding SMPL+D registration. We use these SMPL parameters as ground truth for evaluation. We evaluate a random set of 26 sequences from the CAPE dataset, totaling 7,000 frames, and render their untextured SMPL meshes.

Baselines We compare our method with a learning-based method, WHAM [42], and an optimization-based method, SMPLIFY-X [35]. Since SMPLIFY-X does not enforce temporal regularizations, we apply a smoothing filter to its results to improve the performance on video sequences. Finally, as our method uses the input mesh to get an initial alignment for the first frame of the sequence, we align the results of WHAM and SMPLIFY-X to the input mesh using Procrustes alignment as well for fair comparison.

Evaluation Metrics We evaluate the methods using three metrics: Mean-Per-Joint-Position-Error (MPJPE), Per-Vertex-Error (PVE), and acceleration error, which quantifies the inter-frame smoothness of the reconstruction. We follow established evaluation from [35], and include acceleration error as in [42] to measure the temporal consistency.

Results As shown in Tab. 1, our method outperforms SMPLIFY-X, SMPLIFY-X with smoothing, and WHAM in all metrics. We note that WHAM, as a learning-based method, is not trained for tracking untextured videos, which causes its performance to degrade from its original setting. Our approach is better suited for tracking videos generated by a text-to-video model conditioned on untextured meshes.

Method	MPJE	PVE	Accel
SMPLIFY-X	0.054	0.057	20.57
SMPLIFY-X*	0.053	0.056	02.18
WHAM	0.051	0.054	08.61
Ours	**0.036**	**0.041**	**01.49**

Table 1. Pose fitting performance comparison with SMPLIFY-X [35], it's smoothed version SMPLIFY-X*, WHAM [42], and our proposed method on untextured sequences from the CAPE dataset. Metrics include Mean-Per-Joint-Position-Error (MPJPE), Per-Vertex-Error (PVE), and acceleration error (Accel). Lower values indicate better performance across all metrics.

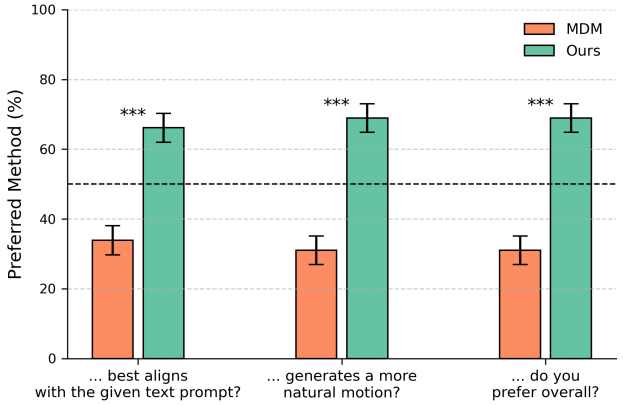

Figure 4. **User Study Results**. We ask the questions: *"Which video...?"*. We compare our method against MDM [46]. We observe a statistically significant preference for our method over MDM in all questions. *** denotes significance at $p \leq 0.001$.

4.3. Motion from Generated Videos

Perceptual Study We evaluate the quality of the motions generated by our method by conducting a perceptual study and comparing them with the motions generated by MDM [46]. A total of 30 participants took part in the study and were asked to answer the following questions:

- **Q1:** "Which motion looks more realistic?"
- **Q2:** "Which motion aligns better with the text prompt?"
- **Q3:** "Which motion do you prefer overall?"

For each test case, we generated the same motion using both our method and MDM [46]. Participants were shown 17 such motion pairs, each presented from two different views: front and side. To mitigate bias, the order of the motions was randomized for each participant. Note that the questions aim to not only evaluate prompt alignment but also the overall quality of the generated motion.

In Fig. 4, we observe a statistically significant preference, meaning $(p \leq 0.001, \text{binomial test})$, for our method over MDM [46] in all questions.

Method	MPJPE	PVE	Accel
w/o $\mathcal{L}_\theta$	0.0486	0.0556	1.7121
w/o $\mathcal{L}_{\text{ex. ben.}}$	0.0393	0.0453	1.5699
w/o $\mathcal{L}_\phi$	0.0392	0.0447	1.5895
w/o $\mathcal{L}_{\text{temp}}$	0.0403	0.0458	3.2005
Opt. Parameters	0.0453	0.0533	2.6494
Ours	**0.0362**	**0.0411**	**1.498**

Table 2. Ablation study on the effect of various components on performance. Metrics include MPJPE (Mean Per Joint Position Error), PVE (Per Vertex Error), and Accel (Acceleration Error). Lower is better for all metrics.

Qualitative Results In Fig. 5, we show a comparison of the generated motion using our method and MDM [46] for the same motion prompts. We observe that our method generates motions that align better with the text prompt and look more realistic. Specifically, by leveraging the wide prior of the VDM, we can generate more diverse and realistic motions.

In Fig. 3, we show additional qualitative results of our method. We generate motions given a text prompt and visualize the frames from the front and side views.

4.4. Ablations

We conduct different ablation studies to evaluate the impact of our various design choices. To evaluate a setting where ground truth is available, we run the same evaluation as for tracking (Sec. 4.2) and show ablation results in Tab. 2. We observe that the temporal loss $\mathcal{L}_{\text{temp}}$ clearly improves performance, particularly in acceleration error. *Opt. Parameters* refers to directly optimizing the SMPL parameters instead of using an MLP to predict them. This confirms that by using a neural parameterization, we can take advantage of the inductive bias of the MLP to improve the tracking performance, particularly, the smoothness. $\mathcal{L}_{\text{ex. ben.}}$ and $\mathcal{L}_\theta$ allow us to leverage the body prior, which helps to maintain anatomically plausible poses.

4.5. Limitations and Future Work

While video diffusion models have demonstrated significant potential in generating diverse and realistic human motions, they can generate artifacts such as morphing effects. However, with the rapid developemnt of VDMs, we anticipate that future iterations of VDMs will address these shortcomings and further enhance motion realism. Monocular tracking remains an inherently underconstrained problem, often leading to ambiguities and inaccuracies in the reconstructed motion. To mitigate these challenges, future work could explore the integration of depth predictors [18] or leverage multi-view diffusion models [50, 51, 54]. Our proposed approach opens up exciting possibilities for generating 4D datasets of human

Figure 5. **Qualitative Evaluation**. We compare the motions generated by MDM [46] and our method for some of the prompts used in the perceptual study. We show two views (front and side) of the generated motions for multiple frames.

motion. These datasets could serve as valuable resources for training and benchmarking models in human motion analysis. Moreover, our method has significant potential for practical applications, such as creating animations for virtual characters in video games, movies, and mixed-reality experiences.

5. Conclusion

In this work, we propose *Animating the Uncaptured*, a novel method for text-to-motion generation for humanoid meshes. By leveraging the strong priors of video diffusion models, our approach generates realistic and diverse human motions, which are transferred to 3D meshes. We use the SMPL model as a deformation proxy, anchoring the vertices of the input mesh to their closest SMPL face and optimizing the SMPL parameters to track the motion depicted in the generated video. This process is guided by extracting 2D body landmarks, silhouette information, and dense semantic features from the video frames. Experiments on the CAPE dataset demonstrate that our method quantitatively outperforms baseline approaches, particularly in tracking videos with untextured meshes. Finally, our user study highlights a strong preference for the motions produced by our method, both in terms of realism and alignment with text descriptions.

Acknowledgements

This work was supported by the ERC Consolidator Grant Gen3D (101171131) of Matthias Nießner and the ERC Starting Grant SpatialSem (101076253) of Angela Dai.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets, 2023. arXiv:2311.15127 [cs]. 2

[2] 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, 2016. 1344 citations (Semantic Scholar/arXiv) [2024-03-17] arXiv:1607.08128 [cs]. 2, 5

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[4] Luka Crnkovic-Friis and Louise Crnkovic-Friis. Generative choreography using deep learning. *arXiv preprint arXiv:1605.06921*, 2016. 2

[5] Xiang Deng, Youxin Pang, Xiaochen Zhao, Chao Xu, Lizhen Wang, Hongjiang Xiao, Shi Yan, Hongwen Zhang, and Yebin Liu. Stereo-talker: Audio-driven 3d human synthesis with prior-guided mixture-of-experts. *arXiv preprint arXiv:2410.23836*, 2024. 2

[6] Niladri Shekhar Dutt, Sanjeev Muralikrishnan, and Niloy J. Mitra. Diffusion 3D Features (Diff3F): Decorating Untextured Shapes with Distilled Semantic Features, 2024. 2 citations (Semantic Scholar/arXiv) [2024-07-14] arXiv:2311.17024 [cs]. 4

[7] Satoru Fukayama and Masataka Goto. Music content driven automated choreography with beat-wise motion connectivity constraints. *Proceedings of SMC*, pages 177–183, 2015. 2

[8] Stuart Geman. Statistical methods for tomographic image restoration. *Bull. Internat. Statist. Inst.*, 52:5–21, 1987. 5

[9] Anastasis Germanidis. Introducing gen-3 alpha: A new frontier for video generation, 2024. 6

[10] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and Tracking Humans with Transformers, 2023. 32 citations (Semantic Scholar/arXiv) [2024-03-17] arXiv:2305.20091 [cs]. 3

[11] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 2

[12] 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

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models, 2020. 6477 citations (Semantic Scholar/arXiv) [2024-03-17] arXiv:2006.11239 [cs, stat]. 2

[14] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video Diffusion Models, 2022. 840 citations (Semantic Scholar/arXiv) [2024-07-11] arXiv:2204.03458 [cs]. 2

[15] 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, 2014. 2

[16] Yujiao Jiang, Qingmin Liao, Zhaolong Wang, Xiangru Lin, Zongqing Lu, Yuxi Zhao, Hanqing Wei, Jingrui Ye, Yu Zhang, and Zhijing Shao. SMPLX-Lite: A Realistic and Drivable Avatar Benchmark with Rich Geometry and Texture Annotations, 2024. 0 citations (Semantic Scholar/arXiv) [2024-06-01] arXiv:2405.19609 [cs]. 2

[17] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end Recovery of Human Shape and Pose, 2018. 1499 citations (Semantic Scholar/arXiv) [2024-03-17] arXiv:1712.06584 [cs]. 2

[18] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for Human Vision Models, 2024. arXiv:2408.12569 [cs]. 2, 7

[19] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 6

[21] Nikos Kolotouros, Georgios Pavlakos, Michael Black, and Kostas Daniilidis. Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2252–2261, Seoul, Korea (South), 2019. IEEE. 2

[22] Hongjie Li, Hong-Xing Yu, Jiaman Li, and Jiajun Wu. Zerohsi: Zero-shot 4d human-scene interaction by video generation. *arXiv preprint arXiv:2412.18600*, 2024. 3

[23] Lei Li and Angela Dai. Genzi: Zero-shot 3d human-scene interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20465–20474, 2024. 3

[24] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10): 2684–2701, 2020. 1

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6): 248:1–248:16, 2015. 2368 citations (Crossref/DOI) [2024-10-11] 88 citations (Semantic Scholar/DOI) [2024-03-17] 1976 citations (Crossref) [2024-03-17]. 2, 3, 4, 5

[26] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Medi-

aPipe: A Framework for Building Perception Pipelines, 2019. arXiv:1906.08172 [cs]. 2, 4, 6

[27] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence, 2024. 48 citations (Semantic Scholar/arXiv) [2024-07-02] arXiv:2305.14334 [cs]. 2

[28] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to Dress 3D People in Generative Clothing, 2020. 11 citations (Semantic Scholar/arXiv) [2024-03-17] arXiv:1907.13615 [cs]. 6

[29] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of Motion Capture as Surface Shapes, 2019. 752 citations (Semantic Scholar/arXiv) [2024-03-17] arXiv:1904.03278 [cs]. 1

[30] John Nickolls, Ian Buck, Michael Garland, and Kevin Skadron. Scalable parallel programming with cuda. In *ACM SIGGRAPH 2008 Classes*, New York, NY, USA, 2008. Association for Computing Machinery. 5

[31] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, 2024. 1120 citations (Semantic Scholar/arXiv) [2024-07-17] arXiv:2304.07193 [cs]. 2, 4, 6

[32] 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*, PAMI-2(6):522–536, 1980. 2

[33] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. STAR: Sparse Trained Articulated Human Body Regressor. pages 598–613. 2020. 211 citations (Semantic Scholar/arXiv) [2024-03-17] arXiv:2008.08535 [cs]. 2

[34] 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. In *NIPS-W*, 2017. 5

[35] 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, 2019. 1103 citations (Semantic Scholar/arXiv) [2024-03-17] arXiv:1904.05866 [cs]. 2, 3, 4, 5, 6, 7

[36] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016. 2

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

[38] Kling AI: AI-Powered Video Generation Platform. Kling ai, 2024. 6

[39] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J. Black. ClothCap: seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics*, 36(4):1–15, 2017. 294 citations (Crossref/DOI) [2024-10-11] 350 citations (Semantic Scholar/DOI) [2024-03-18]. 6

[40] 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

[41] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human Motion Diffusion as a Generative Prior, 2023. 50 citations (Semantic Scholar/arXiv) [2024-03-17] 50 citations (Semantic Scholar/DOI) [2024-03-17] arXiv:2303.01418 [cs]. 1

[42] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing World-grounded Humans with Accurate 3D Motion, 2024. arXiv:2312.07531 [cs]. 3, 6, 7

[43] Olga Sorkine and Marc Alexa. *As-Rigid-As-Possible Surface Modeling*. The Eurographics Association, 2007. Accepted: 2014-01-29T09:43:10Z ISSN: 1727-8384. 5

[44] Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi Sing Leung, Ziwei Liu, Lei Yang, and Zhongang Cai. AiOS: All-in-One-Stage Expressive Human Pose and Shape Estimation, 2024. 0 citations (Semantic Scholar/arXiv) [2024-03-27] arXiv:2403.17934 [cs]. 3

[45] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent Correspondence from Image Diffusion, 2023. 95 citations (Semantic Scholar/arXiv) [2024-07-02] arXiv:2306.03881 [cs]. 2

[46] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human Motion Diffusion Model, 2022. 294 citations (Semantic Scholar/arXiv) [2024-03-17] 294 citations (Semantic Scholar/DOI) [2024-03-17] arXiv:2209.14916 [cs]. 1, 7, 8

[47] Lukas Uzolas, Elmar Eisemann, and Petr Kellnhofer. MotionDreamer: Zero-Shot 3D Mesh Animation from Video Diffusion Models, 2024. 0 citations (Semantic Scholar/arXiv) [2024-06-01] arXiv:2405.20155 [cs]. 2

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4

[49] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 2

[50] 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, 2024. arXiv:2411.18613 [cs]. 7

[51] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. SV4D: Dynamic 3D Content Generation with Multi-Frame and Multi-View Consistency, 2024. arXiv:2407.17470 [cs]. 7

[52] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6183–6192, Seattle, WA, USA, 2020. IEEE. 241 citations (Semantic Scholar/DOI) [2024-03-17]. 2

[53] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer, 2024. arXiv:2408.06072 [cs]. 2

[54] 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. 7

[55] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa F Polanía, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence. 2

[56] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 2

[57] Shihao Zou, Xinxin Zuo, Yiming Qian, Sen Wang, Chuan Guo, Chi Xu, Minglun Gong, and Li Cheng. Polarization human shape and pose dataset. *arXiv preprint arXiv:2004.14899*, 2020. 1