AnimatableDreamer: Text-Guided Non-rigid 3D Model Generation and Reconstruction with Canonical Score Distillation

Xinzhou Wang[1,2,3], Yikai Wang[2†], Junliang Ye[2], Fuchun Sun[2†], Zhengyi Wang[2,3], Ling Wang[2,5], Pengkun Liu[2,4], Kai Sun[2], Xintong Wang[6], Wende Xie[7], Fangfu Liu[2], and Bin He[1]

[1] Tongji University [2] Tsinghua University [3] ShengShu [4] Fudan University
[5] Xi'an Research Institute of High-Tech [6] Zhejiang University [7] Didi

Abstract. Advances in 3D generation have facilitated sequential 3D model generation (a.k.a 4D generation), yet its application for animatable objects with large motion remains scarce. Our work proposes AnimatableDreamer, a text-to-4D generation framework capable of generating diverse categories of non-rigid objects on skeletons extracted from a monocular video. At its core, AnimatableDreamer is equipped with our novel optimization design dubbed Canonical Score Distillation (CSD), which lifts 2D diffusion for temporal consistent 4D generation. CSD, designed from a score gradient perspective, generates a canonical model with warp-robustness across different articulations. Notably, it also enhances the authenticity of bones and skinning by integrating inductive priors from a diffusion model. Furthermore, with multi-view distillation, CSD infers invisible regions, thereby improving the fidelity of monocular non-rigid reconstruction. Extensive experiments demonstrate the capability of our method in generating high-flexibility text-guided 3D models from the monocular video, while also showing improved reconstruction performance over existing non-rigid reconstruction methods.
Project page https://zz7379.github.io/AnimatableDreamer/.

Keywords: 4D generation · Diffusion model · Non-rigid reconstruction

1 Introduction

Automatically building animatable 3D models with non-rigid deformations and motions plays a crucial role in broad fields such as gaming, virtual reality, film special effects, etc. With the remarkable success of deep generative models, generating various 2D images through text prompts comes to reality [26,27,55], and this success is expanding beyond 2D generation. The application of Score Distillation Sampling (SDS) [24] has elevated 2D text-to-image diffusion models to generate high-quality 3D models. Numerous subsequent works have emerged in this domain [15,31,35,42]. However, generating deformable objects remains challenging due to their inherent unconstrained and ill-conditioned nature [21,22,38].

† Corresponding authors.

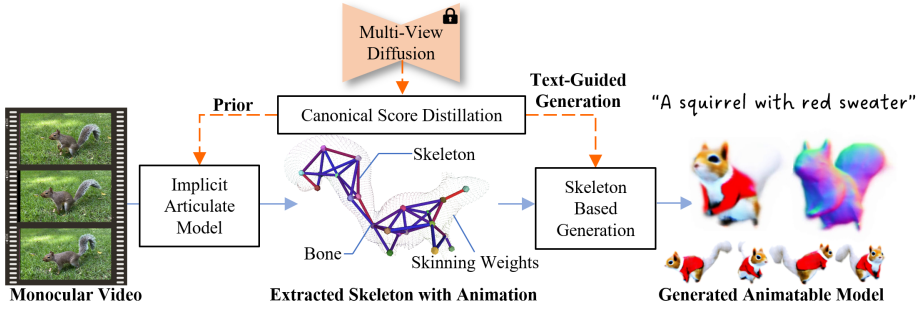

Fig. 1: Given a monocular generic category video, AnimatabaleDreamer initially extracts the skeleton with skinning and motions assisted byprior from a diffusion model. Subsequently, through our canonical score distillation, AnimatabaleDreamer generates novel animatable 3D models from the extracted skeleton and a text prompt with temporal consistency and warping robustness.

Intuitively, bones and motions extracted by implicit animatable models could serve as geometry constraints for deformable object generation. Recent efforts have been devoted to reconstructing animatable 3D models with pre-defined or learned skeletons [10, 33, 44, 50, 53]. Nevertheless, these methods are mostly category-specific with limited diversity or largely rest on the captured multi-view data [6, 34, 49]. We believe this problem can be greatly alleviated by distilling 2D priors from the diffusion model to hallucinate plausible geometry of invisible regions and avoid short-cut solutions [17].

To address these challenges, we propose **AnimatableDreamer**, a two-stage framework designed to extract skeletons from monocular videos and generate generic categories of non-rigid 3D models on these skeletons. Initially, the non-rigid object in the monocular video is disentangled into a canonical implicit field [4, 16] with a skeleton-based structure consisting of bones and neural skinning. AnimatableDreamer extracts bones and skinning from a monocular video, leveraging multi-view diffusion priors to refine the warping, geometry, and texture of unseen regions. Skeletons are generated based on the skinning weights of vertices and further constrain the pairwise relationship between bones of the generated model. Subsequently, under the constraint of the extracted skeleton and a specific text prompt, AnimatableDreamer generates 4D content with a diffusion model. Considering that directly employing SDS [24] or Variational Score Distillation (VSD) [42] will make the canonical model detach from the extracted motions and harm the plausibility of the warped model, we propose **Canonical Score Distillation (CSD)** to generate novel non-rigid 3D models. CSD is a novel distilling strategy designed to simultaneously generate a canonical model aligned with motions and refine the skeletons and skinning. CSD denoises multiple warped models through invertible warping functions while consistently optimizing a static canonical space shared by all animation frames. This novel approach simplifies 4D generation into a more manageable 3D process, yet maintains comprehensive supervision throughout the 4D space and ensures the morphological plausibility of the model under various object poses. Furthermore,

CSD refines the motions and skinning weights to ensure consistency with the canonical model.

To summarise, we make the following contributions:

- **AnimatableDreamer**: A novel framework that extracts skeletons with motions from a monocular video and generates generic categories of non-rigid 3D models based on these skeletons. This is the first implementation of generating text-guided non-rigid 4D content leveraging video-based skeletons.
- **Canonical Score Distillation**: A new distillation method enhances the generation and reconstruction of non-rigid 3D models. By back-propagating gradients from multiple camera spaces to a static canonical space, CSD ensures the morphological plausibility of models after warping. Besides, CSD refines bones and skinning weights through a specifically tailored gradient term. With these designs, CSD improves the reconstruction quality of unseen regions with diffusion prior and is capable of generating 4D models with time consistency and warping robustness.
- **Skeleton-based Generation**: An innovative approach for 4D generation taking skeletons, bones, skinning weights, and motions as prior. With constructed skeletons, a constraint with $SE(3)$ is utilized to guide the transformations of bone pairs, thereby preventing motion detaching and ensuring convergence. Furthermore, the density of Gaussian bones is considered as an indicator for the generation of surfaces with warping robustness.

2 Related Work

2.1 Neural Reconstruction for 3D Non-rigid Object

Neural Radiance Field (NeRF) has been a groundbreaking advancement in representing static scenes, enabling the generation of photorealistic novel views and detailed geometry reconstruction [1, 2, 13, 16]. Adapting NeRF for dynamic scenarios has involved augmenting the field into higher dimensions to accommodate objects with changing topologies [22]. An alternative strategy in dynamic object reconstruction employs an additional warping field to deform the NeRF [21, 25]. Nonetheless, these dynamic NeRF adaptations often encounter performance degradation, primarily due to the complexities introduced by the added temporal dimension. This could be alleviated by applying alternative representations including tesnors [29], Gaussian Splatting [43] and explicit representation [3, 9]. Despite these innovations, synthesizing space-time views from monocular perspectives remains a significant hurdle. Implementing spatio-temporal regularization methods, including depth and flow regularization, has shown potential in overcoming this issue [8, 46]. Furthermore, exploring category-specific or articulate priors offers promising avenues for reconstructing non-rigid objects [6, 23, 40, 49–51]. These approaches offer novel opportunities and insights for advancing the field of 3D non-rigid object reconstruction. However, they often overlook the application of generic priors, which could reduce the reliance on domain-specific priors and manually designed templates. Contrarily, our method leverages model training on extensive, generic datasets to distill such priors, thereby enhancing the reconstruction process.

2.2 Distillation-based 3D Generation from Diffusion Model

SDS [24] has gained prominence for its capability to elevate pre-trained 2D diffusion models to the realm of 3D generation. By distillate 2D prior learned from large-scale datasets and optimizing implicit field [16], SDS is able to generate high-quality 3D models based on text-prompt [5, 37, 42]. The integration of differentiable marching tetrahedra [30] further enhances the combination of explicit meshes and SDS [15]. However, semantic consistency challenges arise in distillation-based 3D generation methods due to their disconnection from the 3D dataset during training. Addressing this concern, MVDream [31] introduces a multi-view diffusion model for panoramas with homography-guided attention, improving semantic consistency by incorporating cross-view attention and camera conditions. Further, 2D diffusion model trained can be lifted to 4D via a temporal score distillation sampling [32], which integrates world knowledge into 3D temporal representations. In contrast, our method aims to produce a time-consistent and warp-robust 4D model by initially generating a non-rigid model based on a skeleton extracted from the video. By applying this model across various animations, we ensure morphological plausibility even when the skeletons exhibit differing articulations.

3 Method

Given a monocular video $V = \{(I_i, t_i)\}_{i=1}^{n}$, our objective is two-fold: first, to generate an object related to a specified prompt $\mathbf{y}$ on the skeletons and rigging extracted from the provided video; second, to reconstruct the original object with diffusion prior. The proposed framework, AnimatableDreamer, as illustrated in Fig. 2, comprises two distinct stages: skeletons extraction (Sec. 3.1) and skeletons-based generation (Sec. 3.2). Both stages employ CSD for content generation and warping refinement (Sec. 3.3). These workflows supervise the deformed model in **camera space** (articulated poses) and optimize the model in **canonical space** (rest pose) through differentiable warping.

3.1 Implicit Articulate Model

Canonical Model. We utilize the NeuS model [39] as our canonical representation to accurately reconstruct surface geometries, in conjunction with rendering an additional feature descriptor [48] to incorporate priors from off-the-shelf methods [14, 19] for self-supervised 3D registration. This approach facilitates the articulation extraction and 3D reconstruction of objects across various categories from monocular video. In our representation, each 3D point $\mathbf{X} \in \mathbb{R}^3$ on the canonical model is characterized by a color vector $\mathbf{c} \in \mathbb{R}^3$, a Signed Distance Field (SDF) value $\mathbf{d} \in \mathbb{R}$, and a feature descriptor $\psi \in \mathbb{R}^{16}$:

$$(\mathbf{c}, \mathbf{d}) = \mathbf{MLP}_{*}(\mathbf{X}, \mathbf{v}), \tag{1}$$

$$\psi = \mathbf{MLP}_{\psi}(\mathbf{X}), \tag{2}$$

where $\mathbf{v} \in SO(3)$ is view direction. Through the cumulative function of an unimodal distribution $\Gamma_\beta(\cdot)$, the Signed Distance Function (SDF) is transformed

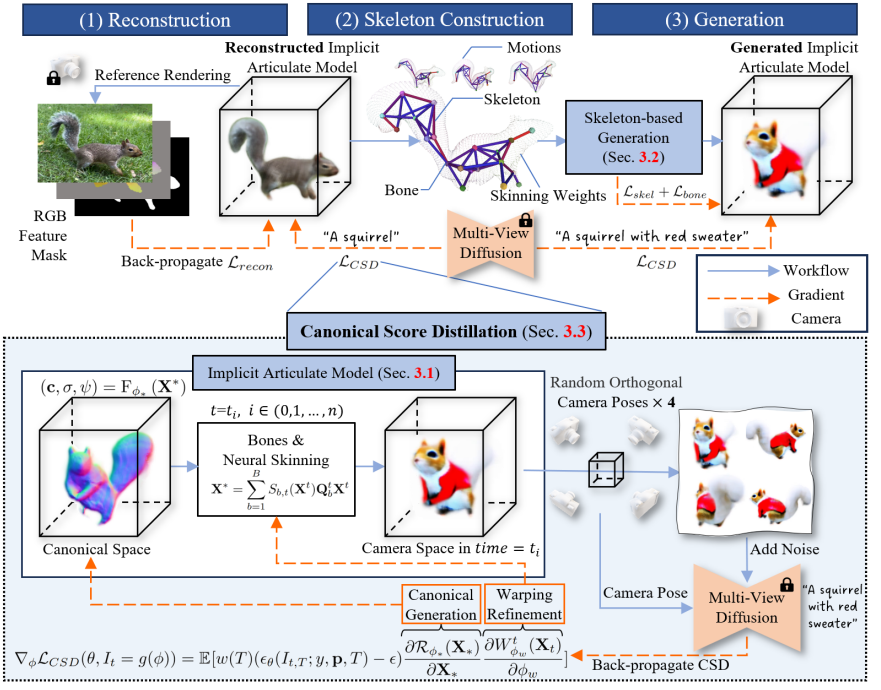

Fig. 2: Framework overview. Top: The AnimatableDreamer framework extracts the skeletons, skinning, and animation from monocular generic category videos with CSD to enhance the invisible regions. Subsequently, through our canonical distillation strategy, AnimatabaleDreamer generates text-guided novel animatable 3D models on the extracted skeleton. **Bottom:** We decompose the articulated model into a static neural field and time-varying neural skinning that transforms the object from canonical space to camera space. During training, we rendered a warped model on four random orthogonal views and optimized the model in camera space across different frames. The canonical generation term in CSD enhances the morphological plausibility of warped models, while the warping refinement term further refines the bone motions and skinning.

into a density representation [39, 41]. The feature descriptor ψ is learned based on 2D features extracted from the self-supervised vision model, DINOV2 [19]. Considering that CSD supervises a point along the time axis without strong texture continuity constraints, we employ a time-invariant canonical model to prevent texture flickering.

Warping Field. In contrast to dense motion fields [3, 21, 22], we disentangle non-rigid objects into a canonical model and a compact motion field [18, 49]. Such disentanglement mitigates the challenges associated with the ill-conditioned nature inherent to 4D generation and enables the application of distillation-based 3D generators. To warp the field from camera space to canonical space, we build the mapping between the 3D point in canonical space $\mathbf{X}^*$ and 3D point in camera space $\mathbf{X}^t$ through a blend skinning deformation defined on B rigid bones:

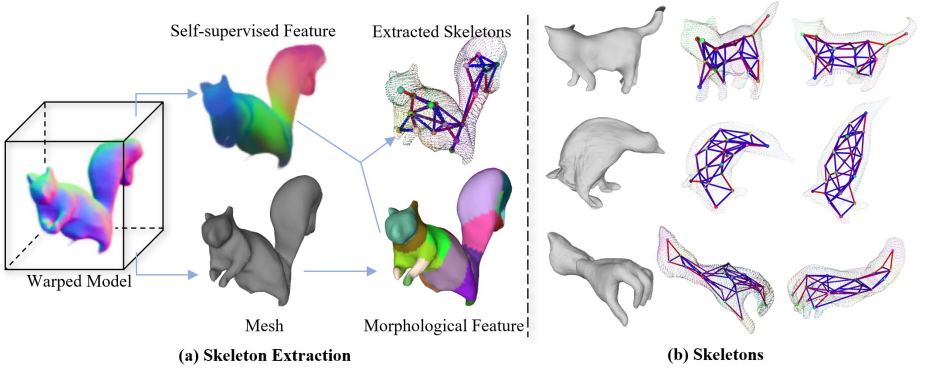

Fig. 3: Skeleton Construction. (a) Pipeline of skeleton construction. We extract mesh from canonical space and construct skeletons using both semantic correlation and morphological correlation. (b) Constructed skeletons for cat, penguin, and hand.

$$\mathbf{X}^* = W_{\phi_w}(\mathbf{X}^t) = \sum_{b=1}^{B} S_{b,t}(\mathbf{X}^t)\mathbf{Q}_b^t\mathbf{X}^t, \tag{3}$$

where $W(\cdot)$ is warping field, ϕ_w are warping-related parameters, and transformations of the b-th bone $\mathbf{Q}_b^t \in SE(3)$ is dual quaternion blend skinning (DQB) [11] learned from $\mathbf{MLP}_\mathbf{Q}(t)$. Here we applied the Fourier function for time embedding [16]. $S_{b,t}(X^t)$ is skinning weight of the b-th bone related to X^t:

$$S_{b,t}(\mathbf{X}^t) = \mathcal{M}(\mathbf{Q}_b^t, \sigma_b, \mathbf{X}^t) + \Delta S_{b,t}. \tag{4}$$

Here $\mathcal{M}$ is the Mahalanobis distance between $\mathbf{X}^t$ and the b-th bone at time t. Each Gaussian bone is assigned with a learnable scaling parameter σ_b. The term $\Delta S_{b,t}$ represents the delta skinning weights derived from $\mathbf{MLP}_\Delta$. It is important to note that the global motion of the object is integrated into the camera poses to enhance clarity. The warping-related parameters ϕ_w includes $\mathbf{MLP}_Q$, σ_b, and $\mathbf{MLP}_\Delta$. To render a pixel $\mathbf{c}$ in camera space for a given time t, we warp camera space sampling points $\mathbf{X_i^t} \in \mathbb{R}^3$ to canonical space with warping function and apply volume rendering:

$$\mathbf{c}\left(\mathbf{x}^t\right) = \mathcal{R}_{\phi_*}(\mathbf{X}^*(\mathbf{X}^t)), \tag{5}$$

where ϕ_* are parameters of $\mathbf{MLP}_*$ defined in Eq. (2). $\mathcal{R}(\cdot)$ is pixel-level volume rendering function [16].

3.2 Skeleton-based Generation

With a deformable model with the bone motions extracted from the reference video, a vanilla way to generate a new text-guided model is to directly modify the canonical model using SDS. However, this raises two problems: first, considering that SDS is an incremental process without a target function, brutally applying SDS to a well-reconstructed model will limit the generated model's diversity. Secondly, the distribution of rendered images from the reconstructed model is different from that of the diffusion model's training set. The reconstructed model

may hurt the performance instead. In this context, we proposed skeleton-based generation, which generates from a new start while being constrained by previously extracted motion.

Skeleton Construction. We pass model irrelevant parameters including the position embedding and time embedding from the reconstructed model to the generated model. Subsequently, we initialize the density field of canonical space of the generated model through bones and skeletons. To extract the skeletons from the reconstructed model, we first extract canonical mesh with edges $E = \{\mathbf{e}_i = \{\mathbf{v}_m, \mathbf{v}_n\}\}$ and vertices $V = \{\mathbf{v}_i \in \mathbb{R}^3\}$ using marching cube. Then we estimate the relation of each pair of bones based on the skinning weights and feature descriptor ψ of vertices. The feature descriptor of each bone b is defined as the weighted sum of related vertices in Eq. (6):

$$\psi_b = \frac{\sum_{i=1}^{N} S_{b,*}(\mathbf{v}_i)(\mathbf{MLP}_\psi(\mathbf{v}_i))}{\sum_{i=1}^{N} S_{b,*}(\mathbf{v}_i)}, \tag{6}$$

and the semantic correlation of two bones $\mathbf{G}_{j,k}$ is calculated as:

$$\mathbf{G}_{j,k} = \text{softmax}\left(\langle \psi_j, \psi_k \rangle\right), \tag{7}$$

where $\langle \cdot \rangle$ is the cosine similarity score. Though bones with similar features are intended to be connected, it may fail when multiple instances share one semantic feature (e.g. arms of a squirrel). To address this issue, we further explore the morphological correlation matrix $\mathbf{M} \in \mathbb{R}^{B \cdot B}$ of each bone pair. For each edge connecting a pair of vertices, we calculate the $\mathbf{M}$ as defined in Eq. (8):

$$\mathbf{M} = \frac{\sqrt{\sum_{e_i=\{\mathbf{v}_m,\mathbf{v}_n\}}^{L} \mathbf{S}_*(\mathbf{v}_m)\mathbf{S}_*(\mathbf{v}_n)^T}}{L}, \tag{8}$$

where L is the number of edges, $\mathbf{S}_*(\cdot) \in \mathbb{R}^B$ is the skinning weight matrix in canonical space. The higher the value of $\mathbf{M}$, the greater the degree to which two bones jointly influence the control over the same surface regions. We balance the semantic correlation and morphological correlation of a pair of bones and define the strength of the skeleton $\mathcal{T}_{j,k}$ as:

$$\mathcal{T}_{j,k} = \begin{cases} \mathbf{G}_{j,k} + \alpha \mathbf{M}_{j,k} & \mathbf{M}_{j,k} \geq \xi \\ 0 & \mathbf{M}_{j,k} < \xi \end{cases}, \tag{9}$$

where weighting scalar α and threshold ξ are learned in a hierarchical manner. The constructed skeletons are shown in Fig. 3.

Constrain with Skeletons and Bones. Given that we have retained only motion-related parameters and discarded the canonical space model, our goal is to ensure that generated objects remain aligned with the motion, and to prevent model collapse during subsequent optimization of the motion. To achieve this, we employ skeletons and bones as constraints in the generation process. We convert $\mathbf{Q}_b^t$ into a rotation quaternion $\mathbf{R}_b^t \in SO(3)$ and a translation vector $\mathbf{T}_b^t$. For j-th bone and k-th bone share a skeleton, we iterate over all time t, and compute the

range of relative position $\mathbf{T}_{jk}^t = \|\mathbf{T}_j^t - \mathbf{T}_k^t\|_2$ and the range of quaternion angle $\mathbf{A}_{jk}^t = \angle(\mathbf{R}_j^t, \mathbf{R}_k^t)$. With this setting, constraints can be applied to the motion of bones, thus preventing motion divergence when generative loss is applied:

$$\mathcal{L}_{skel} = \lambda_T \mathcal{L}_{skel,T} + \lambda_A \mathcal{A}_{skel,T}, \tag{10}$$

$$\mathcal{L}_{skel,T} = \sum\nolimits_{j,k} \mathcal{T}_{j,k} \max(\mathbf{T}_{jk} - \mathbf{T}_{max}, \mathbf{T}_{min} - \mathbf{T}_{jk}, 0)^2, \tag{11}$$

$$\mathcal{L}_{skel,A} = \sum\nolimits_{j,k} \mathcal{T}_{j,k} \max(\mathbf{A}_{jk} - \mathbf{A}_{max}, \mathbf{A}_{min} - \mathbf{A}_{jk}, 0)^2, \tag{12}$$

where $\mathbf{T}_{\max}$ and $\mathbf{T}_{\min}$ denote the maximum and minimum values of $\mathbf{T}_{jk}^t$ across time t, respectively, while $\mathbf{T}_{\max}$ and $\mathbf{T}_{\min}$ represent the maximum and minimum values of $\mathbf{A}_{jk}^t$ across time t, respectively.

Besides, we also use bones to constrain the generated surface as well as skinning weights:

$$\mathcal{L}_{bone} = \sum_{\mathbf{X}} \mathbf{H}\left(d(\mathbf{X}), d_g(\mathbf{X})\right) + \sum_{t,\mathbf{X}^t} \mathbf{S}_t(\mathbf{X}^t) \log\left(\frac{1}{\mathbf{S}_t(\mathbf{X}^t)}\right), \tag{13}$$

where $\mathbf{H}$ denotes the binary cross entropy, and d and d_g represent the Signed Distance Function (SDF) value of the canonical model and Gaussian bones, respectively. The first term aims to ensure that the surface is closely aligned with the Gaussian bones, taking into account their covariance matrix. It also seeks to maintain consistency with the neural skinning weights, thereby enhancing the convergence of generation. The second term is designed to encourage the sparsity of the skinning weights in order to mitigate potential degradation resulting from the first term as well as the generative loss.

3.3 Canonical Score Distillation

We distillate prior from the diffusion model for bot reconstruction and generation. This 2D supervision is sufficient for static object generation [24, 42]. However, in 4D reconstructing, supervising a single view at a single time point t becomes insufficient. As illustrated in Fig. 4(a), the supervision from the reference video forms a hyper-plane in the 4D space, resulting in lower quality for unobserved viewpoints. When denoising an image rendered from a viewpoint far away from reference with a large guidance scale, SDS tends to sample a new instance from its own distribution, due to the lack of sufficient information about the original instance. Here we incorporate multi-view consistent diffusion model MVDream [31], which is trained on a large 2D and 3D dataset [7, 28], to generate multi-view consistent images. Cross-view attention spreads the known information to unobserved views, as shown in Fig. 4(b).

Design of CSD. A straightforward way for temporal consistent articulate object generation is to supervise the model in canonical space. However, the canonical model is merely a "time-slice" of the 4D object, as depicted in Fig. 4(a). Roughly supervising the canonical model without considering articulations will result in the degradation of the morphological plausibility of the model in camera spaces. Additionally, unreachable points in canonical space will not be optimized

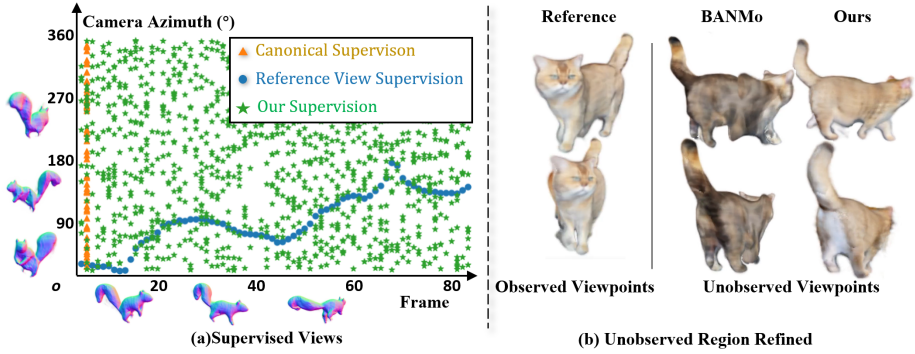

Fig. 4: CSD supervises the entire 4D space. (a) Our approach can traverse the entire XYZ-t space. In contrast, approaches that focus exclusively on the canonical space or solely on reference views are limited to supervising just a single hyperplane within this 4D space. (b) In unobserved regions, CSD achieves better texture consistency and geometry quality compared to BANMo.

(e.g., occluded body parts at canonical pose), even though they will be rendered in camera space.

To mitigate these issues, we employ a multi-view consistent diffusion model and propose a novel Canonical Score Distillation (CSD) to generate a 4D articulate model with both time consistency and warp robustness across different articulations. We elaborately design CSD to utilize the warping field as a bridge between diffusion prior and canonical model. Conversely, the warp is also refined by diffusion prior, especially in contexts of significant motion and in regions lacking ground-truth images. Distinct from SDS [24], CSD traverses all frames and replaces the image gradient with two terms: canonical generation and warping refinement, as defined in Eq. (14):

$$\nabla_{\phi}\mathcal{L}_{CSD} = \mathbb{E}[\underbrace{w(T)(\epsilon_{\theta}(I_{t,T}; y, \mathbf{p}, T) - \epsilon)}_{\text{Diffusion Prior}} \underbrace{\frac{\partial \mathcal{R}_{\phi_*}(\mathbf{X}_*)}{\partial \mathbf{X}_*}}_{\substack{\text{Canonical} \\ \text{Generation}}} \underbrace{\frac{\partial W_{\phi_w}(\mathbf{X}_t)}{\partial \phi_w}}_{\substack{\text{Warping} \\ \text{Refinement}}}]. \tag{14}$$

To distinguish from the previous time t of the articulate model, we refer to the time step of diffusion as T. $w(T)$ is hyper-parameter controlling the weight of T, y is text prompt, I_t are four images rendered in the camera space of time t from four orthogonal view-points $\mathbf{p}$, $I_{t,T}$ are sampled noisy images relative to frame time t and diffusion time step T, $\epsilon_{\theta}(I_{t,T}; y, \mathbf{p}, T)$ is noise predicted by diffusion model conditioned on prompt and camera poses.

Canonical Generation. The first term is the gradient of the canonical rendering with respect to the sampling point. It is designed to simplify the generation process from a 4D time-varying model to a series of static 3D models warped from a shared canonical model. Although we traverse all camera spaces, the distillation process is consistently conducted with respect to the canonical model parameters ϕ_*.

Warping Refinement. Regarding the second term, it signifies that the warping parameters ϕ_w are optimized to better collaborate with the canonical model in different poses. As depicted in Fig. 5(a), CSD corrects faulty skinning weights $S_{b,t}(\mathbf{X^t})$ and misplaced bone transformations $\mathbf{Q}_b^t$.

3.4 Optimization

We optimize all learnable parameters during the reconstruction phase. Following the extraction of skeletons from the reference video, we optimize a new implicit articulate model with $\mathbf{MLP}_\phi$ and camera-related parameters discarded. The total loss for skeleton extraction and model generation is defined as Eq. (15) and Eq. (16), respectively:

$$\mathcal{L}_{Ext} = \mathcal{L}_{recon} + \mathcal{L}_{CSD} + \mathcal{L}_{reg}, \tag{15}$$

$$\mathcal{L}_{Gen} = \mathcal{L}_{skel} + \mathcal{L}_{bone} + \mathcal{L}_{CSD} + \mathcal{L}_{reg}. \tag{16}$$

The reconstruction loss, denoted as $\mathcal{L}_{recon}$, comprises the photometric Mean-Square Error (MSE) $\mathcal{L}_{rgb}$, silhouette reconstruction MSE $\mathcal{L}_{sil}$, and flow loss $\mathcal{L}_{\text{OF}}$ [49], as defined in Eq. (17):

$$\mathcal{L}_{recon} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{geo}\left(\mathcal{L}_{sil} + \mathcal{L}_{\text{OF}}\right), \tag{17}$$

where silhouette and optical flow are pre-computed with an off-the-shelf model [52]. λ_{rgb} and λ_{geo} are balancing weights. The registration loss $\mathcal{L}_{reg}$ is defined in Eq. (18):

$$\mathcal{L}_{reg} = \mathcal{L}_{match} + \mathcal{L}_{2D-cyc} + \mathcal{L}_{3D-cyc}, \tag{18}$$

where $\mathcal{L}_{\text{match}}$, $\mathcal{L}_{\text{2D-cyc}}$, and $\mathcal{L}_{\text{3D-cyc}}$ represent the loss functions for 3D point feature matching [49], 2D cycle consistency [48], and 3D cycle consistency [14], respectively.

Here, we have designed a two-stage schedule with balanced weights for generation. During the articulation extraction stage, the model is primarily supervised by images, with pre-computed mask, flow, and features. Concurrently, the weight of $\mathcal{L}_{CSD}$ is configured to be low for the complementation of unseen regions, thereby ensuring a gentle adjustment without overpowering the original data. For the generation stage, the position embedding bandwidth is set to max value to guarantee the detail of the generated model.

4 Experiments

We conducted experiments on generation (Sec. 4.2) and reconstruction (Sec. 4.3) tasks using the Casual Videos dataset and Animated Objects dataset [49]. The experiments conducted spanned a diverse array of species, such as squirrels, cats, finches, eagles, humans, hands, and manipulators, among others, to convincingly showcase the capability of our method across generic categories. In the context of the generation task, our method excels in creating spatiotemporally consistent, animatable 3D models with text prompts and a template video. With the proposed CSD method, the generated 4D model exhibits superior performance

Method	Prompt	Category	Generation	3D Model	Motion	Articulation
ProlificDreamer [42]	Text	Generic	Shape+Texture	NeRF	-	-
Text2Video-Zero [12]	Text	Generic	-	-	Learned	-
BANMo [49]	-	Generic	-	NeRF	Learned	Learned
Farm3D [10]	Image	Specific	Texture	Mesh	Manual	Pre-defined
Ours	**Video+Text**	**Generic**	**Shape+Texture**	**NeRF**	**Learned**	**Learned**

Table 1: Related work overview on non-rigid 3D model reconstruction and generation. Distinguishing from previous work, our method, AnimatableDreamer, generates text-guided animatable models across generic categories without the need for pre-defined templates. This attribute establishes AnimatableDreamer as a versatile and user-friendly 4D generation tool.

over existing distillation strategies in terms of temporal consistency and warp robustness, as demonstrated in Table 2. For the reconstruction task, our approach significantly outperforms previous methods, particularly when the number of viewpoints and videos is limited, as shown in Table 3.

4.1 Technical Details

To enlarge the Casual Videos dataset [49], we collect videos containing a single complete instance with large kinesis from the internet. We utilize off-the-shelf models [19, 47, 52]to extract mask, optical flow, and features. We modify the camera distance and near-far plane calculation to avoid the articulated model out of frustum or obstructing the camera. Considering that viewpoints are fixed for reconstruction and randomly selected for generation, we alternate the loss calculation of reconstruction and generation in practice. On a single Nvidia A800 GPU, we sampled 128 pixels from 32 images in reconstruction and rendered four 200×200 images for generation. The complete training takes 5 hours for 12000 iterations. Here we adopted a gradient cache technology for saving memories [54].

4.2 Animatable 3D Model Generation

Qualitative Comparisons. We present our generated results alongside the input videos in Fig. 6. By disentangling the deformation and canonical model, our generated models demonstrate time consistency, even in cases where the video duration is extensive. Through optimization across all frames, our approach effectively eliminates issues such as disconnected shapes, flickering, and shape inconsistency. Notably, our method is capable of generating various species including quadruped, squirrel, eagle, bird, penguin and so on, and goes beyond mere texture generation with modifying the model's geometry. We employ VSD [42] and MVDream [31] as baselines and qualitative comparisons are shown in Fig. 5(b). Also, we conducted a comparative analysis with several notable 3D generation and 4D reconstruction methods: ProlificDreamer [42], 3D reconstructor BANMo [49], texture-swap articulated representation Farm3D [10] and text to video generator Text2Video-Zero [12] and summarize our strengths in Tab. 1.

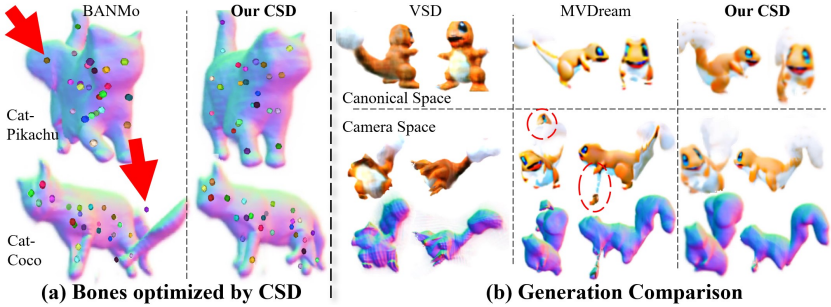

Fig. 5: Canonical score distillation results. (a) Through the warping refinement term, bones are optimized by the diffusion prior and warped surfaces. (b) Compared with VSD and MVDream, our method generates a more detailed result with plausible motions.

Methods	CLIP↑	CLIP-T↑	R-Precision@10↑	GPT Eval3D elo ↑
ProlificDreamer [42]	33.1	95.9	56.3	959
MVDream [31]	34.8	94.4	31.2	979
w/o $\mathcal{L}_{bone/skel}$	27.1	94.2	35.6	954
w/o $\mathcal{L}_{skel}$	28.4	94.0	40.1	960
w/o $\mathcal{L}_{bone}$	37.8	96.1	81.7	1070
Our CSD	**38.2**	**96.6**	**87.5**	**1098**

Table 2: Quantitative results for generation. We use CLIP ViT-B/32 for evaluation. CLIP-T is the average CLIP between adjacent frames, CLIP R-Precision@10 is a metric to evaluate the text-image consistency. GPT Eval3D [45] is an evaluator for Text-to-3D generation. Our method outperforms previous works, and we find that without skeleton restriction, the generation tends to diverge.

Quantitative Comparisons. We perform a quantitative evaluation of the generation quality and the consistency between text and images by utilizing the CLIP-score [20] and the GPT-4 Eval3D [45] methodologies. A video is rendered employing the camera trajectory specified in Eval3D [45], which navigates around the scene at a constant elevation angle while varying the azimuth. Each video frame is then assessed using the CLIP ViT-B/32 model, and the scores are aggregated across all frames and text prompts to compute the overall CLIP score. Additionally, we assess the temporal consistency of CLIP-T by calculating the CLIP similarity between consecutive frames. We employ VSD [42] and MVDream [31] as baselines, as presented in Tab. 2.

4.3 Animatable 3D Model Reconstruction

Qualitative Comparisons. As depicted in Fig. 7, we outperform the existing animatable object reconstruction method on Casual Videos dataset and Animation Objects dataset (Sec. 4). A notable observation is that our approach goes beyond merely refining the canonical model. It also encompasses modifications to the bones and warping fields, as stated in Eq. (14). This comprehensive modification is especially effective in scenarios where the initial bone structures are

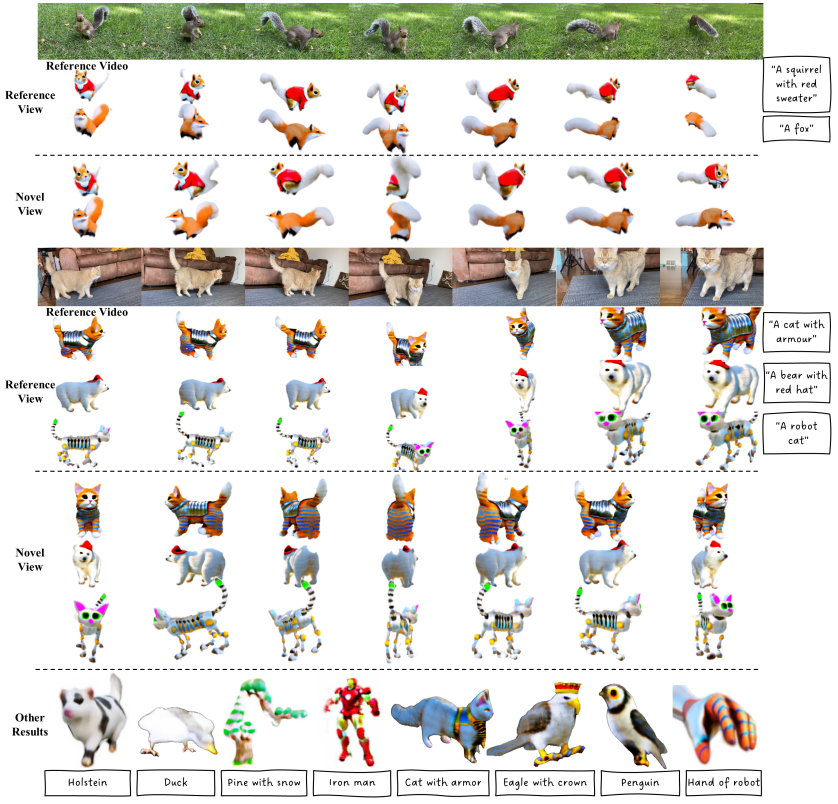

Fig. 6: Text-to-4D Generation. We generate animatable 3D models by leveraging a text prompt and a template video, achieving diverse outcomes that encompass both texture and geometry while maintaining temporal consistency and morphological plausibility across different poses. See the Appendix for more results.

Methods	Number of videos	Cat-Coco		Cat-Pikachu		Penguin		Shiba	
		CD↓	F@%2↑	CD↓	F@%2↑	CD↓	F@%2↑	CD↓	F@%2↑
BANMo [49]	1	10.7	15.3	3.71	57.3	6.47	43.9	6.81	36.6
BANMo [49]	4	4.66	51.6	4.51	52.7	3.75	60.3	4.66	51.9
RAC [50]	1	6.25	42.2	3.60	60.2	4.68	53.7	7.94	30.1
RAC [50]	4	4.48	55.8	3.39	68.1	8.77	26.9	6.86	41.9
w/o $\mathcal{L}_{CSD}$	1	8.34	32.6	3.88	59.6	6.94	42.3	5.83	35.2
Ours	**1**	**3.65**	**63.3**	**2.0**	**88.9**	**3.7**	**64.0**	**4.54**	**53.9**

Table 3: Quantitative results of monocular reconstruction on Casual Videos and Animated Objects. We calculate the Chamfer distance (cm, ↓) and F-score (%, ↑), averaging the results over all frames and videos. Leveraging prior from diffusion, our method outperforms existing methods, despite the absence of multiple videos or templates.

implausible. In such cases, our method adeptly repositions the bones, accompanied by corresponding adjustments to the warping function, which is clearly

Fig. 7: Monocular reconstruction result on Casual videos dataset. Our method is visibly superior to BANMo, particularly in frames with large motion (squirrel) or in regions not present in the reference image (cat and penguin). Each experiment is performed based on a **single monocular video**. See the Appendix for more results.

evidenced in Fig. 5(a). It is evident that while BANMo delivers good results at the input views, its performance significantly diminishes in unobserved spaces. This discrepancy in quality can be attributed to the model's tendency to overfit on the reference views, especially when there is a lack of supervision from other viewpoints, as we discussed in Fig. 4.

Quantitative Comparisons. Our evaluation utilizes both Chamfer distances [36] and F-scores using a threshold set to 2% of the bounding box size. Considering that Casual Videos has no ground truth, we employ BANMo on multiple video sequences and extract meshes to serve as pseudo ground truth. The result in Tab. 3 suggests AnimatableDreamer significantly outperforms BANMo, even when the pseudo ground truth is derived from BANMo itself. This indicates that, with CSD, our framework effectively supplements the missing information, surpassing the need for additional video data.

4.4 Ablation Study

For the generation process, we conduct ablation studies on $\mathcal{L}_{bone}$ and $\mathcal{L}_{skel}$ as detailed in Tab. 2. Our findings indicate that the absence of skeletal constraints $\mathcal{L}_{skel}$ leads to divergence in generation or results in motions becoming disconnected from the model. Additionally, it is observed that incorporating $\mathcal{L}_{bone}$ enhances the surface quality of the generated models. In the context of reconstruction, the ablation of $\mathcal{L}_{CSD}$ reveals a significant enhancement in performance, for it refines the texture and geometry of unobserved regions. Please refer to the Appendix for more results.

5 Conclusion

In this work, we present AnimatableDreamer, a pioneering framework for the generation and reconstruction of generic-category non-rigid 3D models. With the proposed Canonical Score Distillation (CSD), AnimatableDreamer addresses the challenges of unconstrained deformable object generation by simplifying the 4D generation problem into 3D space. Our method excels in generating diverse spatial-temporally consistent non-rigid 3D models based on textual prompts. With articulations extracted from monocular video, users can manipulate and animate these models by controlling the rigid transformations of bones. We demonstrate improved performance compared with existing monocular non-rigid body reconstruction methods, especially in scenarios with limited viewpoints and substantial motion.

Limitations. Our method requires large VRAM to render high-resolution images for CSD during training, due to the long gradient chain from camera space to canonical space, posing constraints on the optimization process. The requirement to feed four images into MVDream simultaneously poses a computational burden further. Addressing these issues could enhance the overall performance and versatility.

References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 3
2. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 3
3. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 3, 5
4. Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.: Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 2
5. Chen, Z., Wang, G., Liu, Z.: SceneDreamer: Unbounded 3D Scene Generation from 2D Image Collections (Apr 2023), http://arxiv.org/abs/2302.01330, arXiv:2302.01330 [cs] 4
6. Cheng, W., Chen, R., Fan, S., Yin, W., Chen, K., Cai, Z., Wang, J., Gao, Y., Yu, Z., Lin, Z., et al.: Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023) 2, 3
7. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13142–13153 (2023) 8

8. Du, Y., Zhang, Y., Yu, H.X., Tenenbaum, J.B., Wu, J.: Neural radiance flow for 4d view synthesis and video processing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14304–14314. IEEE Computer Society (2021) 3
9. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12479–12488 (2023) 3
10. Jakab, T., Li, R., Wu, S., Rupprecht, C., Vedaldi, A.: Farm3d: Learning articulated 3d animals by distilling 2d diffusion. arXiv preprint arXiv:2304.10535 (2023) 2, 11
11. Kavan, L., Collins, S., Žára, J., O'Sullivan, C.: Skinning with dual quaternions. In: Proceedings of the 2007 symposium on Interactive 3D graphics and games. pp. 39–46 (2007) 6
12. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439 (2023) 11
13. Li, Z., Müller, T., Evans, A., Taylor, R.H., Unberath, M., Liu, M.Y., Lin, C.H.: Neuralangelo: High-fidelity neural surface reconstruction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 3
14. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4, 10
15. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 1, 4
16. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 2, 3, 4, 6
17. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 2
18. Noguchi, A., Iqbal, U., Tremblay, J., Harada, T., Gallo, O.: 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. pp. 3677–3687 (2022) 5
19. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023) 4, 5, 11
20. Park, D.H., Azadi, S., Liu, X., Darrell, T., Rohrbach, A.: Benchmark for compositional text-to-image synthesis. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021) 12
21. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5865–5874 (2021) 1, 3, 5
22. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topo-

logically varying neural radiance fields. arXiv preprint arXiv:2106.13228 (2021) 1, 3, 5
23. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 3
24. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022) 1, 2, 4, 8, 9
25. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: Neural Radiance Fields for Dynamic Scenes. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10313–10322. IEEE, Nashville, TN, USA (Jun 2021). https://doi.org/10.1109/CVPR46437.2021.01018, https://ieeexplore.ieee.org/document/9578753/ 3
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 1
27. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems (NeurIPS) (2022) 1
28. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems (NeurIPS) **35**, 25278–25294 (2022) 8
29. Shao, R., Zheng, Z., Tu, H., Liu, B., Zhang, H., Liu, Y.: Tensor4D: Efficient Neural 4D Decomposition for High-Fidelity Dynamic Reconstruction and Rendering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16632–16642. IEEE, Vancouver, BC, Canada (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.01596, https://ieeexplore.ieee.org/document/10204587/ 3
30. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2021) 4
31. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023) 1, 4, 8, 11, 12
32. Singer, U., Sheynin, S., Polyak, A., Ashual, O., Makarov, I., Kokkinos, F., Goyal, N., Vedaldi, A., Parikh, D., Johnson, J., Taigman, Y.: Text-To-4D Dynamic Scene Generation (Jan 2023), http://arxiv.org/abs/2301.11280, arXiv:2301.11280 [cs] 4
33. Stathopoulos, A., Pavlakos, G., Han, L., Metaxas, D.N.: Learning articulated shape with keypoint pseudo-labels from web images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13092–13101 (2023) 2
34. Su, S.Y., Yu, F., Zollhöfer, M., Rhodin, H.: A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In: Advances in Neural Information Processing Systems (NeurIPS) (2021) 2
35. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023) 1
36. Tatarchenko, M., Richter, S.R., Ranftl, R., Li, Z., Koltun, V., Brox, T.: What do single-view 3d reconstruction networks learn? In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 14

37. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation (Dec 2022), http://arxiv.org/abs/2212.00774, arXiv:2212.00774 [cs] 4
38. Wang, L., Zhang, J., Liu, X., Zhao, F., Zhang, Y., Zhang, Y., Wu, M., Yu, J., Xu, L.: Fourier plenoctrees for dynamic radiance field rendering in real-time. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 1
39. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. Advances in Neural Information Processing Systems (NeurIPS) (2021) 4, 5
40. Wang, Y., Dong, Y., Sun, F., Yang, X.: Root pose decomposition towards generic non-rigid 3d reconstruction with monocular videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023) 3
41. Wang, Y., Han, Q., Habermann, M., Daniilidis, K., Theobalt, C., Liu, L.: Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023) 5
42. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. arXiv preprint arXiv:2305.16213 (2023) 1, 2, 4, 8, 11, 12
43. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering (Oct 2023), http://arxiv.org/abs/2310.08528, arXiv:2310.08528 [cs] 3
44. Wu, S., Li, R., Jakab, T., Rupprecht, C., Vedaldi, A.: Magicpony: Learning articulated 3d animals in the wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 2
45. Wu, T., Yang, G., Li, Z., Zhang, K., Liu, Z., Guibas, L., Lin, D., Wetzstein, G.: Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. arXiv preprint arXiv:2401.04092 (2024) 12
46. Xian, W., Huang, J.B., Kopf, J., Kim, C.: Space-time neural irradiance fields for free-viewpoint video. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9421–9431 (2021) 3
47. Yang, G., Ramanan, D.: Learning to segment rigid motions from two frames. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 11
48. Yang, G., Sun, D., Jampani, V., Vlasic, D., Cole, F., Liu, C., Ramanan, D.: Viser: Video-specific surface embeddings for articulated 3d shape reconstruction. Advances in Neural Information Processing Systems (NeurIPS) (2021) 4, 10
49. Yang, G., Vo, M., Neverova, N., Ramanan, D., Vedaldi, A., Joo, H.: Banmo: Building animatable 3d neural models from many casual videos. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 2, 3, 5, 10, 11, 13
50. Yang, G., Wang, C., Reddy, N.D., Ramanan, D.: Reconstructing animatable categories from videos. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 2, 3, 13
51. Yang, G., Yang, S., Zhang, J.Z., Manchester, Z., Ramanan, D.: Ppr: Physically plausible reconstruction from monocular videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3914–3924 (October 2023) 3
52. Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., Zheng, F.: Track anything: Segment anything meets videos (2023) 10, 11

53. Yao, C.H., Raj, A., Hung, W.C., Li, Y., Rubinstein, M., Yang, M.H., Jampani, V.: Artic3d: Learning robust articulated 3d shapes from noisy web image collections. arXiv preprint arXiv:2306.04619 (2023) 2
54. Zhang, K., Kolkin, N., Bi, S., Luan, F., Xu, Z., Shechtman, E., Snavely, N.: Arf: Artistic radiance fields. In: European Conference on Computer Vision. pp. 717–733. Springer (2022) 11
55. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023) 1