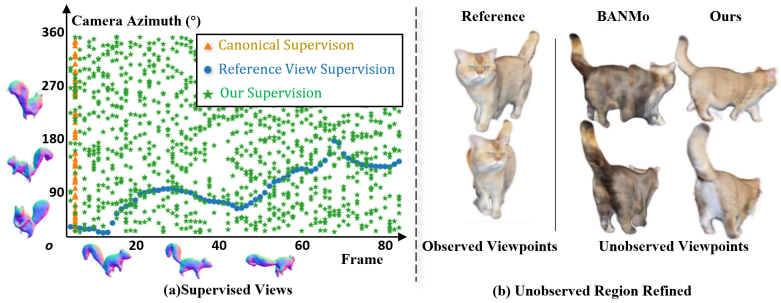

Fig. 1: CSD supervises the entire 4D space.

A Project Website

For more detailed and intractable results, please visit our Project Website at `https://AnimatableDreamer.github.io/`.

B CSD supervises the entire 4D space

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. In unobserved regions, CSD achieves better texture consistency and geometry quality compared to BANMo.

C Ablation Results of Generation

For the generation process, we conduct ablation studies on $\mathcal{L}_{bone}$ and $\mathcal{L}_{skel}$ as detailed in main paper. 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. As depicted in Figure 2, an ablation study was conducted to evaluate the proposed techniques. Specifically, the inclusion of $\mathcal{L}_{skel}$ enhances the coherence between motion and the generated model, ensuring a tighter bond. The model is prone to collapse in the absence of $\mathcal{L}_{\text{skel}}$. Concurrently, $\mathcal{L}_{bone}$ plays a critical role in aligning the generated surface meticulously with the underlying skeletal structure.

We present visualizations of the skeletal structures in the collapsed case above without $\mathcal{L}_{\text{skel}}$, attributing the collapse to the divergence in bone transformations. To substantiate the efficacy of $\mathcal{L}_{\text{skel}}$, a comparative visualization with $\mathcal{L}_{\text{skel}}$ applied is provided, as shown in Figure 3.

Fig. 2: Comparison and Ablation study. The model is prone to collapse in the absence of $\mathcal{L}_{\text{skel}}$. Concurrently, $\mathcal{L}_{\text{bone}}$ plays an indispensable role in precisely aligning the generated surface with the underlying skeletal structure. ProlificDreamer fails to generate meaningful results because of the lack of multi-view information. MVDream generates relatively reasonable result, but degrades after warping.

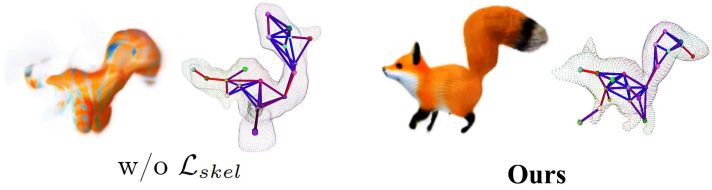

Fig. 3: Bone transformations diverge without skeleton restriction. The bones go into implausible positions without skeleton restriction.

D Additional Generation Results

We present additional results of AnimatableDreamer for non-rigid generation in Figure 4. AnimatableDreamer exhibits robust generalization performance across

various types of non-rigid objects. In addition to the generation of texture, the geometry is also produced.

Fig. 4: Results of high-quality animatable models from AnimatableDreamer. In addition to the generation of texture, geometry is also produced. Here we set the background to black in some cases for better generation.

E Additional Reconstruction Results

We present additional results of AnimatableDreamer for non-rigid reconstruction in Figure 5. In contrast to existing methods, our framework effectively completes the unobserved regions on the 3D model, leveraging the inductive priors and instance information provided by the multi-view diffusion model. We also visualize the extract skeletons with animations in Figure 6.

F Pseudo-code for AnimatableDreamer

A more detailed pseudo-code for AnimatableDreamer is presented in Algorithm 1.

Algorithm 1 Pseudo-code for AnimatableDreamer

Require: $\{I\}$, $\{t\}$, $\{\mathbf{p}\}$, $\mathbf{y}$
Ensure: ϕ_{Gen}
Initialize ϕ_{Recon}
for each $i \in [1, N]$ **do**
    Set position embedding bandwidth
    if i mod 2 == 0 **then**
        $I, t, \mathbf{p} \leftarrow$ Sampler($\{I\}$, $\{t\}$, $\mathbf{p}$)
        $I_r \leftarrow$ Render(ϕ_{Recon}, t, $\mathbf{p}$)
        $\mathcal{L} \leftarrow (\mathcal{L}_{\text{Recon}}(I_r; I) + \mathcal{L}_{\text{Reg}})$
        Optimize ϕ_{Recon}
    else
        $t, \mathbf{p}, T \leftarrow$ Sampler($\{t\}$, random pose, schedule)
        $I_r \leftarrow$ Render(ϕ, t, $\mathbf{p}$)
        $\mathcal{L} \leftarrow (\mathcal{L}_{\text{CSD}}(I_r; t, T, \mathbf{p}, \mathbf{y}) + \mathcal{L}_{\text{Reg}})$
        Freeze CameraMLP, Embedder
        Optimize ϕ_{Recon}
        Unfreeze CameraMLP, Embedder
Generate Skeletons S
Initialize ϕ_{Gen} with S
for each $i \in [1, N]$ **do**
    Set position embedding bandwidth
    Freeze CameraMLP, Embedder
    $t, \mathbf{p}, T \leftarrow$ Sampler($\{t\}$, random pose, schedule)
    $I_r \leftarrow$ Render(ϕ, t, $\mathbf{p}$)
    $\mathcal{L} \leftarrow (\mathcal{L}_{\text{CSD}}(I_r; t, T, \mathbf{p}, \mathbf{y}) + \mathcal{L}_{\text{Reg}} + \mathcal{L}_{\text{Skel}})$
    Optimize ϕ
return ϕ_{Gen}

This pseudo-code provides an overview of the CSD process, including initialization, sampling, rendering, loss calculation, and optimization steps for training

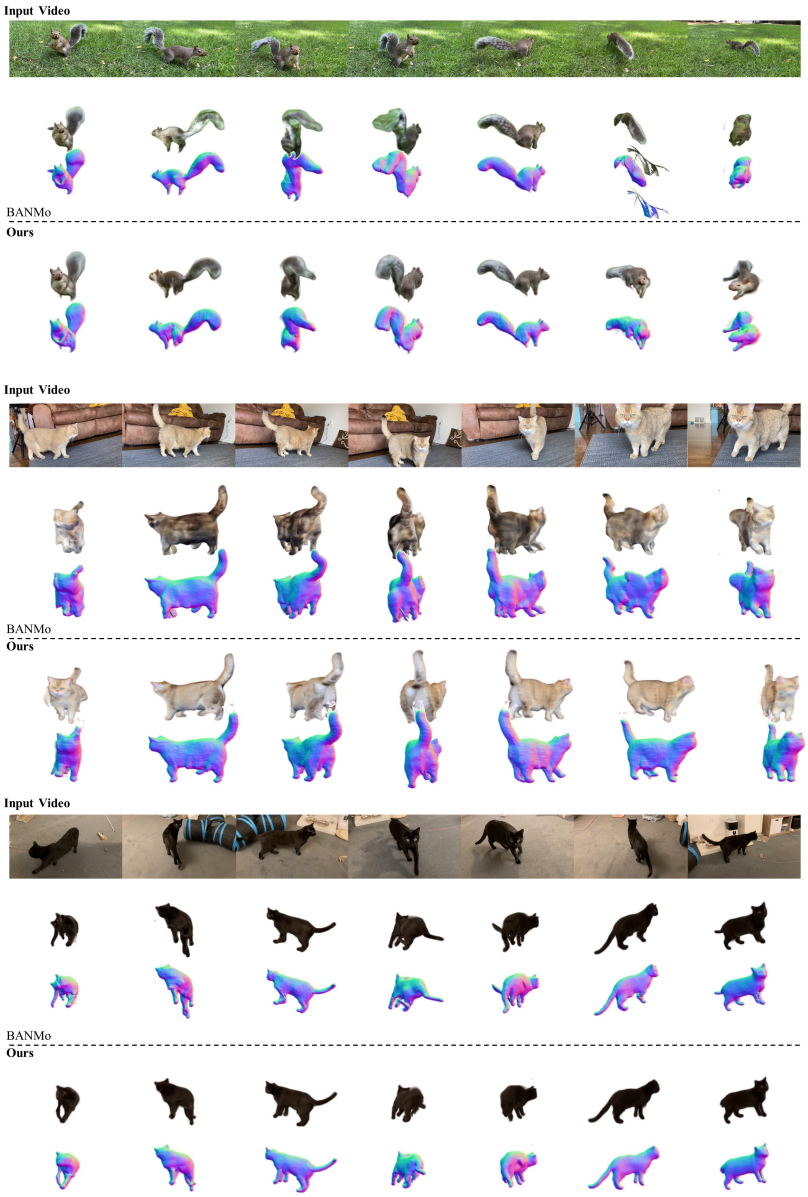

Fig. 5: More reconstruction result. Our approach enhances the fidelity of the reconstructed models, particularly in regions not previously observed.

the model. The Sampler function is responsible for extracting relevant informa-

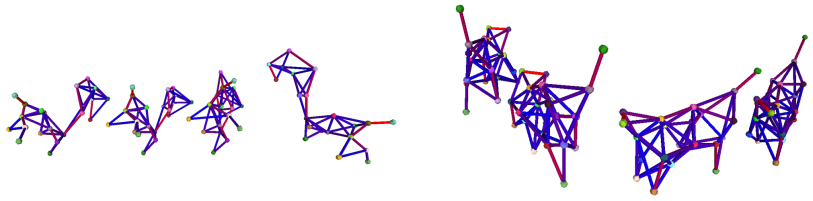

Fig. 6: The extracted skeletons with animations. We extract skeletons associated with extended animations.

tion from input datasets, and the Render function renders images based on the current model parameters. The loss terms include reconstruction loss $\mathcal{L}_{\text{Recon}}$, CSD loss $\mathcal{L}_{\text{CSD}}$, and regularization loss $\mathcal{L}_{\text{Reg}}$. Optimization of the reconstructed articulated model, denoted as θ_{Recon}, is conducted in an alternating manner between the generation and the reconstruction loss functions. Subsequently, the extracted skeletons are utilized to steer the 4D generation process of θ_{Gen}. It is noteworthy that both the CameraMLP and Embedder components are maintained in a frozen state when the optimization is driven by the generation loss $\mathcal{L}_{\text{Gen}}$.

G Weighting scalar α and threshold ξ

In this study, the weighting scalar α and the threshold ξ are determined through a hierarchical approach. The parameter α is optimized to ensure that geometrically distinct parts possessing similar semantic features are disconnected. Conversely, the threshold ξ is established to facilitate the connection of skeletons with the top 80% of scores. Subsequent to these adjustments, bones lacking a corresponding skeleton are eliminated.

H Additional Details

Hyper-parameters. During the generation stage, the weight of $\mathcal{L}_{CSD}$ is progressively increased from 0 to 0.0001, while the weight of $\mathcal{L}_{reg}$ is adjusted according to a logarithmic function, ranging from 0.01 to 1. Throughout the skeleton extraction stage, the weight assigned to $\mathcal{L}_{CSD}$ is methodically reduced from 0.001 to 0.00001. Furthermore, given the disparate magnitudes of $\mathcal{L}_{skel}$ and $\mathcal{L}_{bone}$ in comparison to other loss terms, a balancing factor related to the dimensions of the canonical mesh is employed to maintain equilibrium. An AdamW optimization algorithm is utilized, configured with a learning rate of 5×10^{-4}.

CSD is more CFG friendly. For multi-view diffusion, the classifier-free guidance (CFG) weight is set to 50. In our experiments, we observe that setting CFG to around 30 yields the most beneficial results, as illustrated in Figure 8. This observation is attributed to the fact that the reconstruction loss aids in controlling the consistency of the generated content. However, for the reconstruction

Fig. 7: The effect of the SDS time step. The broader the range of the time step, the more significant the modifications to the model will be. An excessively large time step can compromise the integrity of the skeleton's structure, leading to erroneous results. Conversely, a time step that is too small may not induce sufficient changes in the model. Here we assign different time steps for reconstruction and generation.

Fig. 8: Influence of guidance scale. Setting CFG to 100 enhances the mode-seeking feature and almost bypasses the generation loss, while setting CFG to 10 results in the collapse of the 3D model. This behavior may be attributed to the strong constraints imposed by the articulation extracted from the template video, rendering high CFG unnecessary for ensuring the consistency of the diffusion model.

task, we have observed that setting CFG to a relatively large value is beneficial, as shown in Table 1. Therefore, we set CFG to 100 for the reconstruction process.

CDS time step schedule. Given that our framework is a combination of both a generator and a reconstructor, we have carefully designed the SDS time step schedule based on a series of experiments. For prompts that induce significant changes in the object's geometry, we implement an annealing time step schedule ranging from 0.8 to 0.5. In cases where the prompt primarily affects texture or involves minimal geometry changes, we sample the time step in the range of 0 to 0.5. The impact of the SDS time step for generation is illustrated in Figure 7. For reconstruction, we fix the time step to 0.5 after experiments (Table 2).

Near-far planes. Given that we render from a free viewpoint rather than fixing on the reference, it is essential to compute the near-far plane of each frame dynamically. Therefore, our near-far planes are calculated on the fly to encompass all points of the proxy geometry with a considerable margin.

Issue about MVDream. MVDream is trained in a controlled environment where the object always faces the "forward" direction and is maintained at a proper distance from the camera. In contrast, our framework defines the "forward" direction based on the input video. To align the canonical model with the "forward" direction, we introduce an azimuth offset.

Guidance scale	CD	F@%2
10	5.52	44.8
30	5.33	43.7
50	5.45	45.5
100	4.6	52.1

Table 1: Ablation of the guidance scale in reconstruction. A large guidance scale can ensure the consistency of the reconstructed model.

Maximum T	CD	F@%2
0.5	4.6	52.1
0.8	5.06	47.8
0.98	4.68	49.1

Table 2: Ablation of the time step schedule in reconstruction. Setting the maximum T to an excessively large value can alter the original content and reduce the accuracy of reconstruction

The prompt impact result. Our findings indicate that the incorporation of certain fixed negative prompts significantly benefits our task. Examples of these prompts include *ugly, bad anatomy, blurry, pixelated*, and *obscure*. Additionally, the inclusion of descriptors pertaining to the background proves advantageous, particularly in scenarios where the model's color closely resembles that of the background, thereby mitigating the risk of model disappearance.

Time-invariant RGB. Given the generally time-invariant appearance of the template object, we configure RGB as time-invariant to enhance the model's temporal consistency, especially when employing the CSD loss for reconstruction assistance.