DreamPose: Fashion Image-to-Video Synthesis via Stable Diffusion

Johanna Karras[1], Aleksander Holynski[2,3], Ting-Chun Wang[4], Ira Kemelmacher-Shlizerman[1]

[1]University of Washington, [2]UC Berkeley, [3]Google Research, [4]NVIDIA

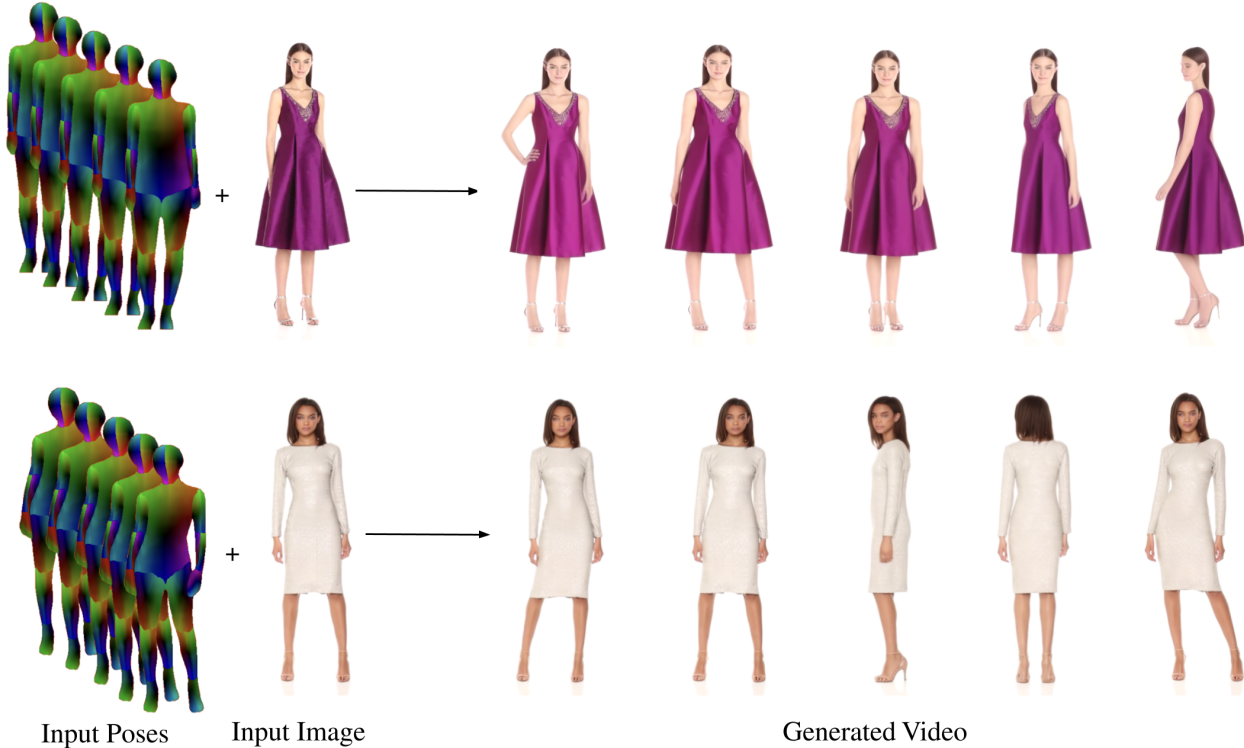Input Poses　　Input Image　　　　　　　　　　　　　　　　Generated Video

Figure 1: Given an image of a person and a sequence of body poses, DreamPose synthesizes a photorealistic video.

Abstract

We present DreamPose, a diffusion-based method for generating animated fashion videos from still images. Given an image and a sequence of human body poses, our method synthesizes a video containing both human and fabric motion. To achieve this, we transform a pretrained text-to-image model (Stable Diffusion [16]) into a pose-and-image guided video synthesis model, using a novel finetuning strategy, a set of architectural changes to support the added conditioning signals, and techniques to encourage temporal consistency. We fine-tune on a collection of fashion videos from the UBC Fashion dataset [49]. We evaluate our method on a variety of clothing styles and poses, and demonstrate that our method produces state-of-the-art results on fashion video animation. Video results are available on our project page.

1. Introduction

Fashion photography is incredibly prevalent online, from social media platforms to online retail sites. Unfortunately, these still photographs are limited in the information they convey, and fail to capture many of the crucial nuances of a garment, such as how it drapes and flows when worn. Fashion *videos*, on the other hand, do showcase all these details, and for this reason are highly informative for consumer decision-making. Despite this obvious benefit, however, these videos are a relatively rare commodity.

In this paper, we introduce DreamPose, a method that turns fashion photographs into realistic, animated videos, using a driving pose sequence. Our method is a diffusion video synthesis model based upon Stable Diffusion [16]. Given one or more images of a human and a pose sequence, DreamPose generates a high-quality video of the input sub-
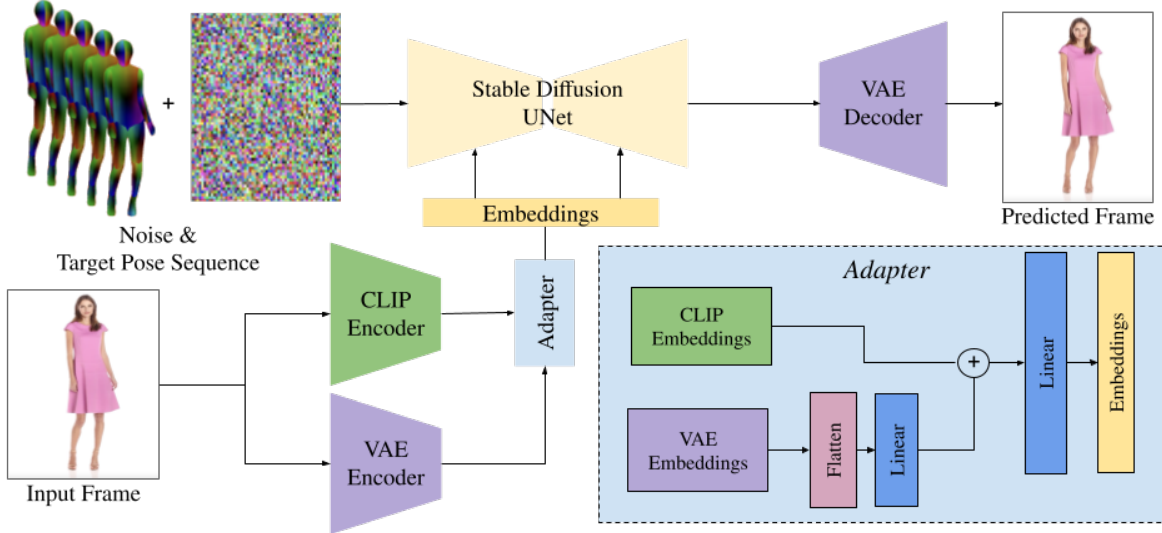
1

Figure 2: Architecture Overview. We modify the original Stable Diffusion architecture in order to enable image and pose conditioning. First, we replace the CLIP text encoder with a dual CLIP-VAE image encoder and adapter module (shown in the blue box). The adapter module jointly models and reshapes the pretrained CLIP and VAE input image embeddings. Then, we concatenate the target pose representation, consisting of 5 consecutive poses surrounding the target pose, to the input noise. During training, we finetune the denoising UNet and our Adapter module on the full dataset and further perform subject-specific finetuning of the UNet, Adapter, and VAE decoder on a single input image.

ject following the pose sequence (Figure 1).

This is a challenging task in several ways. While image diffusion models have shown impressive, high-quality results [16, 31, 35], video diffusion models have yet to achieve the same quality of results and are often limited to "textural" motion or cartoon-like appearance [14, 17, 19, 40, 48]. Moreover, existing video diffusion models suffer from poor temporal consistency, motion jitter, lack of realism, and the inability to control the motion or detailed object appearance in the target video. This is partly because existing models are primarily conditioned on text, as opposed to other conditioning signals (e.g., motion) which may offer more fine-grained control. In contrast, our image-and-pose conditioning scheme allows for greater appearance fidelity and frame-to-frame consistency.

Our model is fine-tuned from an existing pretrained image diffusion model, which already effectively models the distribution of natural images. When using such a model, the task of image animation can effectively be simplified to finding the subspace of natural images consistent with the conditioning signals. To accomplish this, we redesign the encoder and conditioning mechanisms of the Stable Diffusion [16] architecture, in order to enable aligned-image and unaligned-pose conditioning. Further, we propose a two-stage finetuning scheme that consists of finetuning both UNet and VAE from one or more input images.

To summarize, our contributions include: (1) *Dream-Pose*: an image-and-pose conditioned diffusion method for still fashion image animation, (2) a simple, yet effective,

pose conditioning approach that greatly improves temporal consistency across frames, (3) a split CLIP-VAE encoder that increases the output fidelity to the conditioning image, (4) a finetuning strategy that effectively balances image fidelity and generalization to new poses.

2. Related Work

2.1. Diffusion models

Diffusion models have recently demonstrated impressive results in text-conditioned image synthesis [16, 31, 35], video synthesis [17, 19, 27], and 3D generation tasks [28, 41]. However, training these models from scratch is computationally expensive and data intensive. Latent Diffusion Models (as in Stable Diffusion [16]) perform diffusion and denoising in the latent space, thereby drastically reducing the computational requirements and training time with only marginal reductions to quality. Since its release, Stable Diffusion and its pretrained checkpoints have been used by many for various image generation tasks [2, 5, 34]. Like these methods, our work leverages a pretrained Stable Diffusion model with subject-specific finetuning.

2.2. Still Image Animation

Still image animation refers to the task of generating a video from one or more input images. Existing, non-diffusion approaches often consist of multiple separate networks, such as for predicting the background [39, 46, 53], motion representation [20, 37, 38, 39, 43, 53], occlusion

maps [38, 39, 46, 53], or depth maps [3]. The downside of multiple networks is that each stage requires separate training and potentially unavailable or imperfect ground-truth intermediate data, such as ground-truth motion or depth. Especially when the motion is large and complex, these ground-truth estimates are harder to derive and more prone to errors. Several more recent papers explore end-to-end single-network approaches, such as by merging optical flow and warping [32], replacing motion estimation networks entirely with a cross-attention modules [26], or generating animatable 3D humans using a NeRF representation [21].

2.3. Fashion Image Synthesis

Prior pose-guided fashion image synthesis methods are typically generative adversarial network (GAN)-based and rely on optical flow to align image features to pose [1, 9, 23, 33, 50, 54]. However, GAN-based approaches often struggle with large pose changes, synthesizing occluded regions, and preserving garment style. More recent approaches rely on attention-based mechanisms, where self- and cross-attention are used to warp image features to the target frame [8, 25, 33].

Relatively few works exist for diffusion-based fashion image and video synthesis. DiffFashion [6] aims to edit a clothing item by transferring the style of a reference image. Concurrent work PIDM [4] also generates pose-conditioned human images, although unlike our method does not optimize for temporal consistency. We compare DreamPose and PIDM in Figures 9 and 14.

2.4. Diffusion Models for Video Synthesis

Many text-to-video diffusion models rely on adapting text-to-image diffusion models for video synthesis [14, 15, 17, 19, 40, 48]. While the results are promising, these methods still struggle to match the realism that text-to-image models do. Quality is largely hindered due to the new challenges introduced by video synthesis, such as maintaining temporal consistency across frames and generating realistic motion. Some video diffusion methods are instead trained from scratch, requiring expensive computational resources, huge training datasets, and extensive training time [17, 19, 28, 31, 27, 15]. Concurrently, Tune-A-Video finetunes a text-to-image pretrained diffusion model for text-and-image conditioned video generation [47]. However, like earlier video diffusion methods, Tune-A-Video's results exhibit textural flickering and structural inconsistencies. Our work aims to address these issues in order to synthesize realistic human and fabric motion.

2.5. Conditioning Mechanisms for Diffusion Models

Text-conditioning has been popular among image generation diffusion models [16, 30, 36]. While effective at controlling high-level details, text conditioning fails to provide

rich, detailed information about the *exact* identity or pose of a person and garment.

Several works tackle the challenge of image conditioning for a pretrained text-to-image Stable Diffusion model [2, 5, 11, 27, 34, 45]. These often incorporate text embeddings of some kind. For example, DreamBooth, the first method to perform subject-specific finetuning of Stable Diffusion on a set of images, learns a unique text token to represent the subject in the text encoder [34]. Others incorporate text to edit the appearance of existing images [5] and videos [47, 27]. PIDM [4] encodes image textures using a separate textural encoder and concatenates target pose with an input noisy image. DreamPose allows the user to not only control the appearance of subjects in video, but also the structure and motion. Similar to PIDM, our image conditioning approach directly incorporates image embeddings in the cross-attention layers of the UNet, but these image embeddings come from a mixture of two pretrained encoders: CLIP and VAE. Moreover, with our method, we achieve smooth, temporally consistent motion using a multi-pose input representation concatenated to the input noise.

3. Background

Diffusion models are a recent class of generative models that have surpassed GANs at synthesis tasks in terms of quality, diversity, and training stability [10]. A standard image diffusion model learns to iteratively recover an image from normally distributed random noise [42]. A *latent* diffusion model, e.g., Stable Diffusion [16], operates in the encoded latent space of an autoencoder, thereby saving computational complexity, while sacrificing minimal perceptual quality. Stable Diffusion is composed of two models: a variational autoencoder and a denoising UNet. The autoencoder consists of an encoder $\mathcal{E}$ that distills a frame x into a compact latent representation, $z = \mathcal{E}(x)$, and a decoder $\mathcal{D}$ that reconstructs the image from its latent representation, $x' = \mathcal{D}(z)$. During training, the latent features z are diffused in T timesteps by a deterministic Gaussian process to produce noisy features $\tilde{z}_T$, indistinguishable from random noise. In order to recover the original image, a time-conditioned UNet is trained to iteratively predict the noise of the latent features corresponding to each timestep $t \in \{1, ..., T\}$. The UNet ϵ_θ objective function is:

$$L_{DM} = \mathbb{E}_{z, \epsilon \in \mathcal{N}(0,1)}[||\epsilon - \epsilon_\theta(\tilde{z}_t, t, c)||_2^2] \qquad (1)$$

where c represents the embeddings of conditional information, such as text, image, segmentation mask, etc. In the case of text-to-image Stable Diffusion, c is obtained using a CLIP text encoder [29]. Finally, the predicted denoised latents z' are decoded to recover the predicted image $x' = \mathcal{D}(x')$.

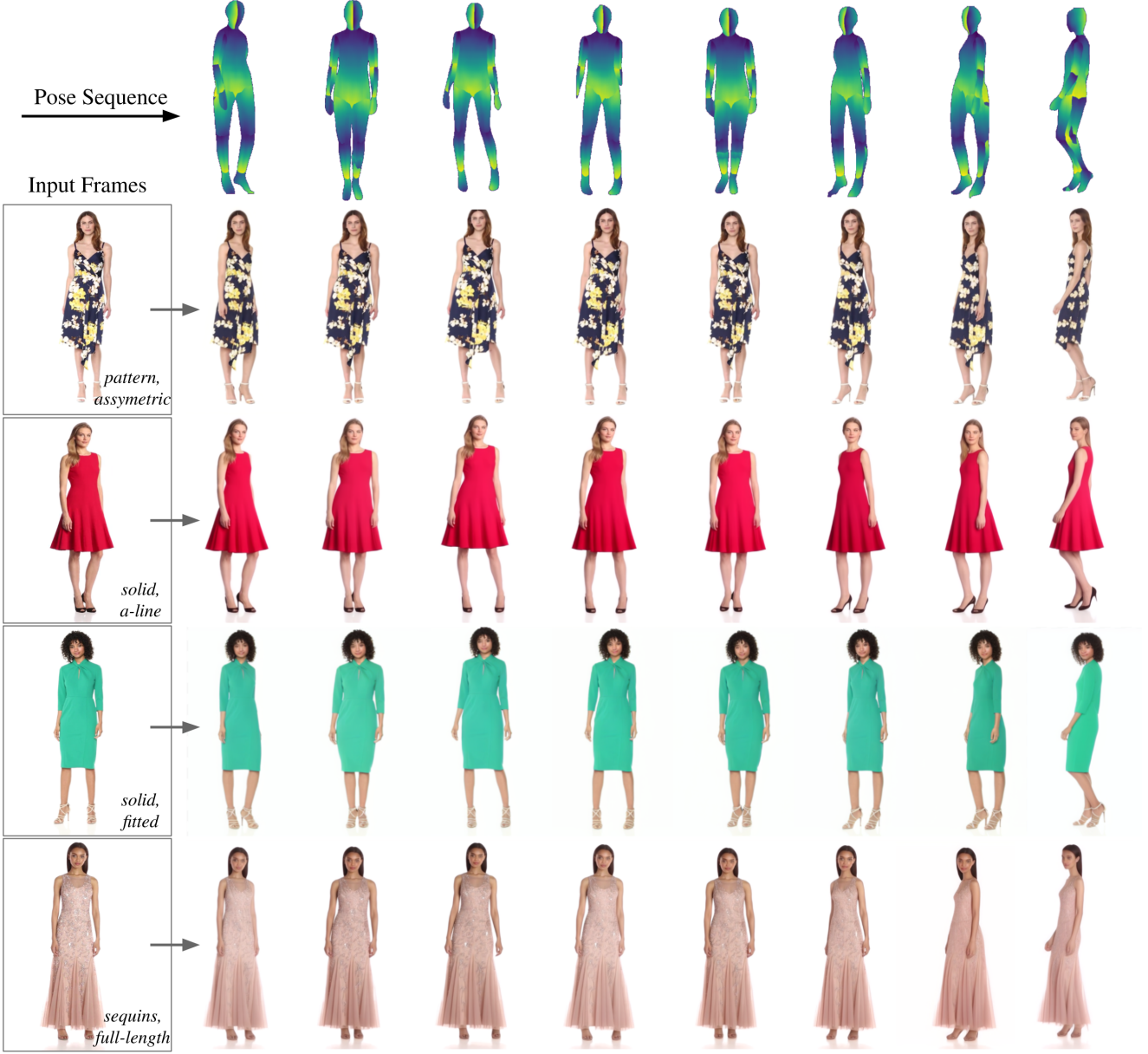*Classifier-free guidance* is a mechanism in sampling that

Figure 3: Qualitative Results. We showcase the results of our method on a variety of input frames and poses. DreamPose is capable of synthesizing photorealistic video frames consistent with a diverse range of patterns, fabric types, person identities, clothing shapes, and viewpoints.

pushes the distribution of predicted noise towards the conditional distribution via an implicit classifier [18]. This is practically achieved by dropout, a training scheme that, with a random probability, replaces real conditioning inputs with null inputs ($\varnothing$). During inference, the conditional prediction is used to guide the unconditional prediction towards the conditional, using a guidance scalar weight s:

$$\epsilon_\theta = \epsilon_\theta(\tilde{z}_t, t, \varnothing) + s \cdot (\epsilon_\theta(\tilde{z}_t, t, c) - \epsilon_\theta(\tilde{z}_t, t, \varnothing)) \quad (2)$$

4. Method

Our method aims to produce photorealistic animated videos from a single image and a pose sequence. To achieve this, we fine-tune a pretrained Stable Diffusion model on a collection of fashion videos. This involves adapting the architecture of Stable Diffusion (which is a text-to-image model) to accept additional conditioning signals (image and pose), and to output temporally consistent content that can be viewed as a video.

In the coming section, we begin by describing the ar-

4

chitectural modifications in Section 4.2. Then, we describe the two-stage fine-tuning strategy in Section 4.3. Finally, in Section 4.4, we describe the inference process of generating an animated video from a still image, which involves a novel formulation of classifier-free guidance.

4.1. Overview

Given input image x_0 and poses $\{p_1, ..., p_N\}$, our method generates a video $\{x'_1, ..., x'_N\}$, where x'_i is the i-th predicted frame corresponding to input pose p_i. Our method relies on a pretrained latent diffusion model [16], which is conditioned on an input image and a sequence of poses. At inference time, we generate each frame independently through a standard diffusion sampling procedure: starting with uniformly distributed Gaussian noise, the diffusion model is repeatedly queried with both conditioning signals to gradually denoise the noisy latent to a plausible estimate. Finally, the predicted denoised latent z'_i is decoded to produce the predicted video frame $x'_i = \mathcal{D}(z'_i)$.

4.2. Architecture

The DreamPose model is a pose- and image-conditioned image generation model that modifies and finetunes the original text-to-image Stable Diffusion model for the purpose of image animation. The objectives of image animation include: (1) faithfulness to the provided input image, (2) visual quality, and (3) temporal stability across generated frames. As such, DreamPose requires an image conditioning mechanism that captures the global structure, person identity, and fine-grained details of the garment, as well as a method to effectively condition the output image on target pose while also enabling temporal consistency between independently sampled output frames. We describe our approach to achieving these goals in the sections below. A diagram of our architecture can be found in Figure 2.
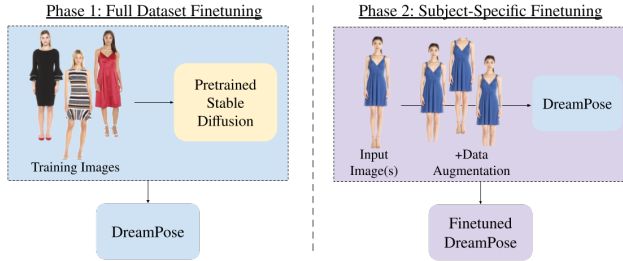

Figure 4: Two-phase finetuning scheme. In the first phase, our method finetunes the modified Stable Diffusion model on the full dataset. In the second phase, the model is further finetuned on a single subject image.

4.2.1 Split CLIP-VAE Encoder

In many prior works, such as InstructPix2Pix [5], image conditioning signals are often concatenated with the input

noise to the denoising U-Net. While this is effective for conditioning signals that are spatially aligned with the desired output image, in our case, our network aims specifically to produce images which are *not* spatially aligned with the input image. As such, we explore alternative approaches for image conditioning. In particular, we implement image conditioning by replacing the CLIP text encoder with a custom conditioning adapter that combines the encoded information from pretrained CLIP image and VAE encoders.

A crucial objective when fine-tuning from a pretrained network is to make training gradients as meaningful as possible by making the input signals as similar as possible to those used in the original network training. This helps avoid regressions in network performance during fine-tuning, or loss of learned priors, which can come from noisy gradients (e.g., if the network does not know how to parse new forms of input signals). For this reason, most diffusion-based fine-tuning schemes [5, 51] will retain all original conditioning signals, and will initialize network weights that interact with new (previously unseen) conditioning signals to zero.

For our purposes, given that Stable Diffusion is conditioned on CLIP embeddings of text prompts, and CLIP encodes both text and images to a shared embedding space, it may seem natural to simply replace the CLIP conditioning with the embedding derived from the conditioning image. While this would in theory pose a very small change to the original architecture and allow for image conditioning with minimal finetuning, we find that in practice that CLIP image embeddings alone are insufficient for capturing fine-grained details in the conditioning image. So, we instead additionally input the encoded latent embeddings from Stable Diffusion's VAE. Adding these latent embeddings as conditioning has the added benefit of coinciding with the output domain of the diffusion model.

Since the architecture does not support VAE latents as a conditioning signal by default, we add an adapter module $\mathcal{A}$ that combines the CLIP and VAE embeddings to produce one embedding that is used in the network's usual cross-attention operations. This adapter blends both the signals together and transforms the output into the typical shape expected by the cross-attention modules of the denoising U-Net. Initially, the weights corresponding to the VAE embeddings are set to zero, such that the network begins training with only the CLIP embeddings (as mentioned before, to mitigate network "shock" in training). We define the final image conditioning signal c_I as:

$$c_I = \mathcal{A}(c_{\text{CLIP}}, c_{\text{VAE}}) \qquad (3)$$

4.2.2 Modified UNet

Unlike the image conditioning, the pose conditioning *is* image-aligned. As such, we concatenate the noisy latents

$\tilde{z}_i$ with a target pose representation c_p. To account for noise in the poses (which are estimated from real videos using an off-the-shelf network [13]) and to maximize temporal consistency in the generated frames, we set c_p to consist of five consecutive pose frames: $c_p = \{p_{i-2}, p_{i-1}, p_i, p_{i+1}, p_{i+2}\}$. We observe that individual poses are prone to frame-to-frame jitter, but training the network with a set of consecutive poses increases the overall motion smoothness and temporal consistency. Architecturally, we modify the UNet input layer to take in 10 extra input channels, initialized to zero, while the original channels corresponding to the noisy latents are unmodified from the pretrained weights.

4.3. Finetuning

For initialization, the unmodified Stable Diffusion layers are initialized from a pretrained text-to-image Stable Diffusion checkpoint, except for the CLIP image encoder which is loaded from a separate pretrained checkpoint [16, 29]. As mentioned previously, the novel layers are initialized such that initially the new conditioning signals do not contribute to the network output.

Following initialization, DreamPose is finetuned in two stages (shown in Figure 4). The first phase fine-tunes the UNet and adapter module on the full training dataset in order to synthesize frames consistent with an input image and pose. The second phase refines the base model by finetuning the UNet and adapter module, then the VAE decoder, on one or more subject-specific input image(s) to create a subject-specific custom model used for inference.

Similar to other image-conditional diffusion methods [34, 28, 27], we find that sample-specific finetuning is essential to preserving the identity of the input image's person and garment, as well as maintaining a consistent appearance across frames. However, simply training on a single frame and pose pair quickly leads to artifacts in the output videos, such as texture-sticking. To prevent this, we augment the image-and-pose pair at each step, such as by adding random cropping.

We also find that finetuning the VAE decoder is crucial for recovering sharper, more photorealistic details in the synthesized output frames. Refer to Figure 5 and the supplementary videos for an ablated comparison. Furthermore, we show in Figure 7 that even single image finetuning of the decoder allows increased pose guidance, without sacrificing the person identity or appearance.

4.4. Pose and Image Classifier-Free Guidance

At inference time, we generate a video frame-by-frame from a single input image and a sequence of poses using the subject-specific model. We modulate the strength of image conditioning c_I and pose conditioning c_p during inference using dual classifier-free guidance [5]. The dual classifier-free guidance equation is modified from Equation 3 to be

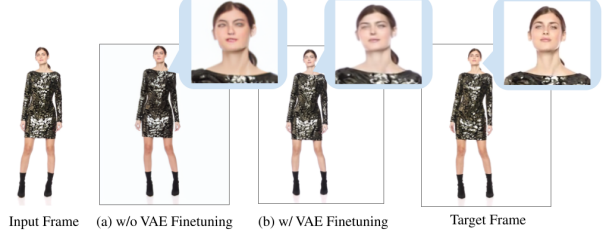Input Frame (a) w/o VAE Finetuning (b) w/ VAE Finetuning Target Frame

Figure 5: Ablation of VAE Finetuning. We find that finetuning the VAE decoder, in addition to the UNet, during the subject-specific finetuning phase, yields more photorealistic details and reduces high-frequency noise, compared to finetuning the UNet alone.

controlled by two guidance weights, s_I and s_p, which rule how similar the output image is to the input image c_I and input pose c_p, respectively:

$$
\begin{aligned}
\epsilon_\theta(z_t, c_I, c_p) = &\, \epsilon_\theta(z_t, \varnothing, \varnothing) \\
&+ s_I(\epsilon_\theta(z_t, c_I, \varnothing) - \epsilon_\theta(z_t, \varnothing, \varnothing)) \\
&+ s_p(\epsilon_\theta(z_t, c_I, c_p) - \epsilon_\theta(z_t, c_I, \varnothing))
\end{aligned}
\tag{4}
$$

In Figure 7, we show the effect of varying the classifier free guidance weights (s_I, s_p). A large s_I ensures high appearance fidelity to the input image, while a large s_p ensures alignment to the input pose. In addition to strengthening our pose and image guidance, the decoupled classifier-free guidance prevents overfitting to the one input pose after subject-specific finetuning.

5. Experiments

5.1. Implementation Details

Our experiments are trained on two NVIDIA A100 GPU's with resolution 512x512. In our first phase of training, we finetune our base model UNet on the full training dataset for a total of 5 epochs at a learning rate of $5e$-6. We use an effective batch size of 16 (through 4 gradient accumulation steps). We implement a dropout scheme where null values replace the pose input 5% of the time, the input image 5% of the time, and both input pose and input image 5% of the time during training. We further finetune the UNet on a specific sample frame for another 500 steps with a learning rate of $1e$-5 and no dropout. Lastly, we finetune the VAE decoder only for 1500 steps with a learning rate of $5e$-5. During inference, we use a PNDM sampler for 100 denoising steps [24].

5.2. Dataset

We train and test our method on the UBC Fashion dataset [49]. We follow the provided train/test split of 339 training and 100 test videos. Each video has a frame rate of 30 frames/second and is approximately 12 seconds long. Dur-

| Input Frame | MRAA | TPSMM | Ours | Ground Truth |

Figure 6: a) Qualitative comparisons of our method versus MRAA [39] and TPSMM [53]. Our method produces more photorealistic details aligned with the input frame, such as garment folds (top row), fine-grain patterns (second row), and face identity (fourth row). Our method can also better handle challenging cases, including pattern synthesis in occluded regions (third row) and large limb motion (bottom row).
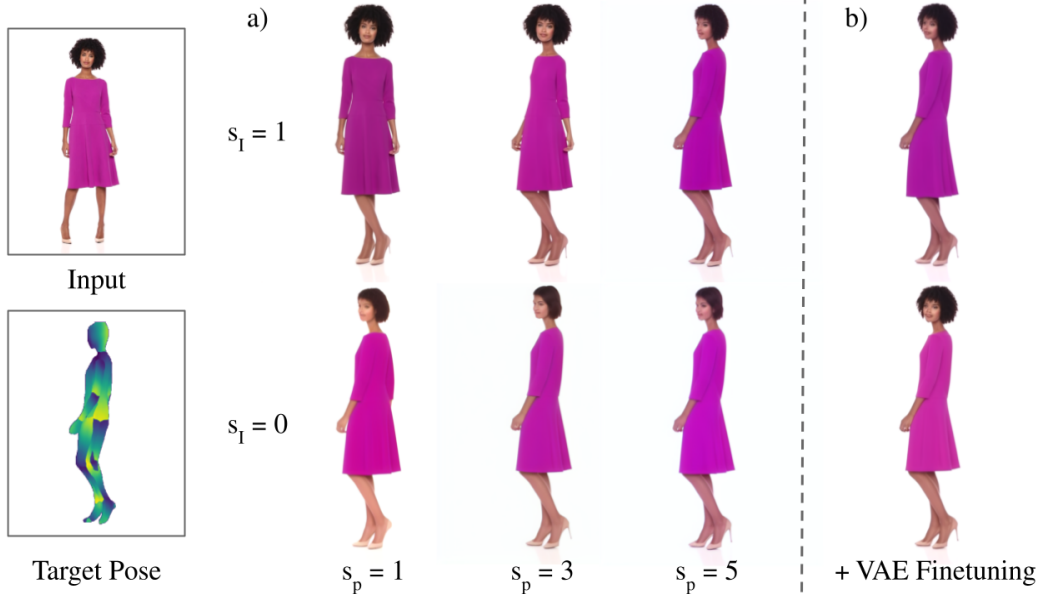
7

Figure 7: a) Pose and Image Classifier-Free Guidance. We demonstrate the effect of the relative weight between image and pose guidance weights, s_I and s_P. Results shown are after subject-specific finetuning. b) Finetuning the decoder improves the appearance and person identity, even with larger relative pose guidance.

ing training, we randomly sample pairs of frames from the training videos. We compute poses with DensePose [13].

6. Results

DreamPose is capable of generating state-of-the-art fashion videos from still images. In Figure 3, we showcase frames synthesized by DreamPose from a variety of input images and poses from the UBC Fashion dataset [49]. DreamPose handles diverse human and clothing appearances well, even from different viewpoints. We show additional results on images from the DeepFashion dataset [12] in the supplementary materials.

6.1. Comparisons

We compare DreamPose quantitatively and qualitatively to two publicly available state-of-the-art conditional video synthesis methods, Motion Representations for Articulated Animation (MRAA) [39] and Thin-Plate Spline Motion Model (TPSMM) [53]. We train both methods from scratch on the UBC Fashion Dataset [49], using the provided training scripts and recommended number of epochs. For evaluation, we use the provided test scripts in the "AVD" mode.

We also perform a qualitative comparison to PIDM [4], a concurrent pose-transfer diffusion model work. For PIDM, since training scripts are not yet available, we use the provided model checkpoint trained on the DeepFashion dataset for our comparisons. We run PIDM and our method with 100 denoising steps.

	L1 ↓	SSIM ↑	VGG ↓	LPIPS ↓
MRAA [39]	0.0857	0.749	0.534	0.212
TPSMM [53]	0.0858	0.746	0.547	0.213
Ours	**0.0256**	**0.885**	**0.235**	**0.068**

Table 1: Quantitative comparisons of our method with MRAA, TPSMM, and our method. Bolded values indicate best scores in each column.

6.1.1 Quantitative Analysis

We present our quantitative analysis in Table 1. We test all models on the UBC Fashion test set, consisting of 100 unique fashion videos, at 256px resolution [49]. For each video, we extract 50 frames for testing, where they are at least 50 frames away from the input frame. Note that both MRAA and TPSMM rely on extracted features from a driving video, whereas DreamPose only relies on driving UV-pose sequence. Even so, the full DreamPose model quantitatively outperforms both methods in all four quantitative metrics: L1, SSIM [44], VGG [22], and LPIPS [52].

6.1.2 Qualitative Analysis

We qualitatively compare our method to MRAA and TPSMM in Figure 3. With MRAA and TPSMM, note that the person identity, fabric folds, and fine patterns are lost in new poses, whereas DreamPose accurately retains those details. Plus, during large pose changes, MRAA may produce disjointed limbs. In Figure 9, we compare our method to PIDM. DreamPose produces higher-fidelity results, in terms of both face identity and clothing patterns. While
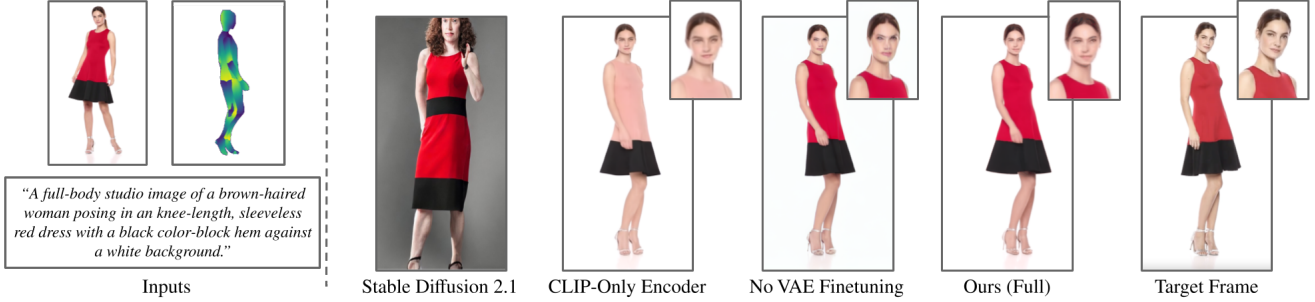
Figure 8: Qualitative Ablation of Image Conditioning. We compare results of text-conditioning alone with the original text-to-image Stable Diffusion model, our model with CLIP-only embeddings without finetuning, our model with CLIP-VAE encoder embeddings without VAE finetuning, and our full model.

PIDM synthesizes realistic faces, they do not necessarily align with the identity of the source person. Moreover, we find that both the identity and the dress appearance vary frame-to-frame, indicating PIDM would not work well as-is for video synthesis. We provide additional comparisons to PIDM on the DeepFashion dataset [12] in the supplementary materials.
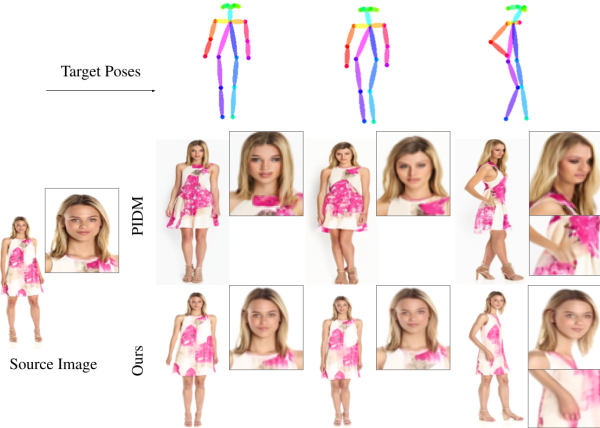

Figure 9: Qualitative comparisons of our method to PIDM [4]. Our method produces more high-fidelity results with respect to person identity and garment pattern. Note that PIDM is conditioned on OpenPose estimates [7] (top row).

6.2. Ablation Studies

We perform a quantitative and qualitative comparison of ablated versions of our method to verify our design choices. Namely, we compare four variants: **(1) Ours$_{CLIP}$:** We use a pretrained CLIP image encoder, instead of our dual CLIP-VAE encoder, **(2) Ours$_{No\text{-}VAE\text{-}FT}$:** We do subject-specific finetuning of the UNet only, not the VAE decoder, **(3) Ours$_{1\text{-}pose}$:** We concatenate only one target pose, instead of 5 consecutive poses, to the noise **(4) Ours$_{full}$:** Our full model, including subject-specific VAE finetuning, CLIP-VAE encoder, and 5-pose input.

Quantitative Comparison. For each ablated version, we compute the L1, SSIM, VGG, and LPIPS for 100 predicted video frames selected from each of the 100 test videos of the UBC Fashion dataset [49]. Shown in Table 2, our full model outperforms the ablated versions on all four metrics.

	L1 ↓	SSIM ↑	VGG ↓	LPIPS ↓
Ours$_{CLIP}$	0.025	0.882	0.247	0.070
Ours$_{No\text{-}VAE\text{-}FT}$	0.025	0.897	0.210	0.057
Ours$_{1\text{-}pose}$	0.019	0.899	0.208	0.056
Ours$_{full}$	**0.019**	**0.900**	**0.207**	**0.056**

Table 2: Quantitative comparison of ablated versions of our method. Note that the single-pose version of our method, although achieving similar numerical results, fails to achieve the motion smoothness as the full method. Please refer to our project page for a video comparison.

Qualitative Comparison. We visually show the effectiveness of our full method in Figure 8. We compare results from the original text-conditioned Stable Diffusion model, our method with only a CLIP image encoder, our method with CLIP-VAE encoder, and our full method with CLIP-VAE encoder and subject-specific VAE finetuning.

The original Stable Diffusion model with text-only conditioning via CLIP text encoder is unable to preserve rich details of the garment or person identity. Simply replacing the text encoder with a CLIP image encoder helps capture most image details, but there is still information loss about the appearance. Subject-specific finetuning of the UNet, similar to DreamBooth [34], is critical to preserving photorealistic details in the face and garment. Furthermore, we find that also finetuning the VAE decoder on the input image greatly improves the sharpness of these details and does not lead to overfitting to the input pose.

Lastly, with only a single input pose, there is noticeable flickering of the subject's shape, especially around the feet and hair. Please refer to video qualitative comparisons of each ablated version on our project page.

9

Figure 10: Results after training with 1, 3, 5, and 7 input images. Increasing the number of input frames improves fidelity of pose, facial identity, and color.

6.3. Multiple Input Images

While DreamPose demonstrates high-quality results with only a single input image, DreamPose can also be fine-tuned with an arbitrary number of input images of a subject. We showcase the results of training with multiple input images in Figure 10. We find that additional input images of a subject increase the quality and viewpoint consistency.

7. Limitations & Future Work

In Figure 11, we show failure cases of our method. On rare occasions, we observe limbs disappearing into the fabric, hallucinated dress features, and directional misalignment when the target pose is facing backwards. We suspect that some of these failures could be alleviated with improved pose estimation, a larger dataset, or a segmentation mask. Additionally, while our method produces realistic results on most plain and simple-patterned fabrics, some of our results present minor flickering behavior on large and complex patterns. Achieving better temporal consistency on such patterns, ideally without subject-specific finetuning, is left to future work. Lastly, similar to other diffusion models, our finetuning and inference times are slow compared to GAN or VAE methods. Finetuning the model on a specific subject takes approximately 10 minutes for the UNet and 20 minutes for the VAE decoder, in addition to an 18 second per-frame rendering time.

8. Conclusion

In this paper, we presented DreamPose, a novel diffusion-based method for still fashion image animation. Given a single image and pose sequence, we demonstrate how our method generates photorealistic fashion videos

Figure 11: Examples of failure cases of our method. Our method may merge limbs into underlying fabric textures (left), hallucinate features (middle), or predict front-facing person instead of a back-facing person (right).

from only a single image – animating a diverse range of fabrics, patterns, and person identities.

Acknowledgments

References

[1] Badour AlBahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan, 2021.

[2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models, 2023.

[3] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation, 2019.

[4] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via de-noising diffusion model, 2022.

[5] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2022.

[6] Shidong Cao, Wenhao Chai, Shengyu Hao, Yanting Zhang, Hangyue Chen, and Gaoang Wang. Difffashion: Reference-based fashion design with structure-aware transfer by diffusion models, 2023.

[7] 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.

[8] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization, 2021.

[9] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing, 2021.

[10] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.

[11] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models, 2023.

[12] Yuying Ge, Ruimao Zhang, Lingyun Wu, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images, 2019.

[13] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild, 2018.

[14] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos, 2022.

[15] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths, 2022.

[16] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022.

[17] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022.

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022.

[20] Aleksander Holynski, Brian Curless, Steven M. Seitz, and Richard Szeliski. Animating pictures with eulerian motion fields, 2020.

[21] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections, 2022.

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016.

[23] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation, 2021.

[24] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds, 2022.

[25] Songhua Liu, Jingwen Ye, Sucheng Ren, and Xinchao Wang. Dynast: Dynamic sparse transformer for exemplar-guided image generation, 2022.

[26] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit warping for animation with image sets, 2022.

[27] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors, 2023.

[28] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[31] Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. Dalle-2 is seeing double: Flaws in word-to-concept mapping in text2image models, 2022.

[32] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion, 2022.

[33] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H. Li, and Ge Li. Deep image spatial transformation for person image generation, 2020.

[34] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2022.

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[37] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer, 2018.

[38] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation, 2020.

[39] Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. 2021.

[40] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022.

[41] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea

Vedaldi, Devi Parikh, Justin Johnson, and Yaniv Taigman. Text-to-4d dynamic scene generation, 2023.

[42] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

[43] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation, 2022.

[44] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[45] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models, 2022.

[46] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo, 2018.

[47] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation, 2022.

[48] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation, 2022.

[49] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation, 2019.

[50] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan, 2021.

[51] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[52] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

[53] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation, 2022.

[54] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation, 2019.

Supplementary Material

A. User Studies

We conducted two user studies involving 50 distinct Amazon Mechanical Turk workers to compare our method with state-of-the art image animation approaches [39] [53] and evaluate the quality of our videos. In both surveys, workers evaluated results corresponding to 50 unique input images from the test set of the UBC Fashion dataset [49].

In the first user study, workers were asked their pair-wise preferences between our method and one of the other methods. For each input image, the workers were shown two videos: one containing the input image, our resulting video, and the MRAA resulting video and the other containing the input image, our resulting video, and the TPSMM resulting video. The ordering of our video and other video (MRAA or TPSMM) was randomized for each question. For each videos, workers selected their preference between the videos. The results are shown in Table 3. Overall, the workers had a preference for our method over MRAA and TPSMM.

In the second user study, workers were asked to rate our videos and TPSMM videos on a scale of 0 to 5, where 0 corresponds a video that does not match the input image at all and 5 corresponds to a realistic animation of the input image. During training, workers were shown a video of a different dress for as an example of a "0" rating and a ground-truth video of the input image as an example of the "5" rating. The results are shown in Figure 12. Our videos achieved higher scores for image similarity and quality than TPSMM and 85% of users rated the results of our method a 3 or higher.

	# Responses	Total Responses	(%)
Ours > MRAA [39]	1637	2500	(65%)
Ours > TPSMM [53]	1417	2500	(57%)

Table 3: Results of User Study #1: Workers choose between pairs of videos corresponding to input images, either our result vs. MRAA result or our result vs. TPSMM result. Overall, participants preferred our method over both MRAA and TPSMM in terms of quality and similarity to the input image.
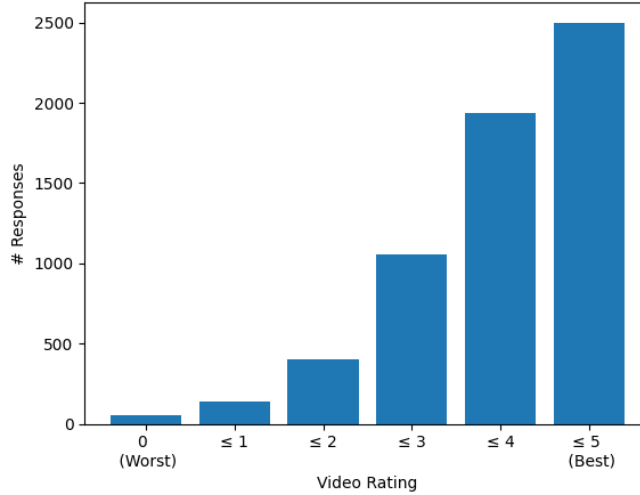

Figure 12: Results of User Study #2: Amazon Mechanical Turk worker ratings of our videos from 0 (video does not match input image) to 5 (video is a realistic animation of the input image). Overall, 85% of workers rated our method a 3 or higher.

B. Deep Fashion Results

We demonstrate the effectiveness of our method on a popular dataset, DeepFashion, in Figure 13 [4, 12]. Although trained exclusively on the UBC Fashion video dataset, DreamPose performs well on unseen retail images, even to new backgrounds, model identities, accessories, and patterns.

C. Application to Pose Transfer

While adapted for image-to-video synthesis, DreamPose is also an effective pose transfer tool. In Figure 14, we compare DreamPose to two state-of-the-art pose transfer models: DynaST [25] and PIDM [4]. Our method is better able to preserve

Target Pose

Input Image   **Predicted Video**

Figure 13: DreamPose results on unseen samples from the DeepFashion dataset [12]. Despite being trained exclusively on the UBC Fashion Dataset, our method generalizes to new garments and model identities after subject-specific finetuning of the base model.

fine-details, such as shoe appearance, hemline, and face identity, than DynaST or PIDM.

| Input Image | Target Pose | | DynaST | PIDM | Ours |

Figure 14: Comparison of Pose Transfer Results. We compare our method to two state-of-the-art pose transfer methods, DynaST [25] and PIDM [4].