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

Johanna Karras¹, Aleksander Holynski²,³, Ting-Chun Wang⁴, Ira Kemelmacher-Shlizerman¹

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

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 pre-trained 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 [50]. 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: https://grail.cs.washington.edu/projects/dreampose

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 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-
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 fine-tune 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.

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 [13, 17, 19, 40, 49]. 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 that achieves photorealistic results on a diverse range of garment patterns and shapes, (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 net-
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.

works, such as for predicting the background [39, 47, 54], motion representation [20, 37, 38, 39, 44, 54], occlusion maps [38, 39, 47, 54], or depth maps [3]. However, multiple networks require separate training for each stage and potentially unavailable or imperfect ground-truth intermediate data, such as motion or depth. Especially with large and complex motion, these ground-truth estimates are harder to derive and more error-prone. 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

Many prior pose-guided fashion image synthesis methods are generative adversarial network (GAN)-based and rely on optical flow to align image features to pose [1, 8, 23, 33, 51, 55]. 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 [7, 25, 33].


**2.4. Diffusion Models for Video Synthesis**

Many text-to-video diffusion models rely on adapting text-to-image diffusion models for video synthesis [13, 14, 17, 19, 40, 49]. 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, 14]. Concurrently, Tune-A-Video fine-tunes a text-to-image pretrained diffusion model for text-and-image conditioned video generation [48]. 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 is popular among image 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, 10, 27, 34, 46]. 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 [48, 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 condition-
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 architectural modifications in Section 4.2. Then, we describe the two-stage finetuning 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 \( x'_i = D(z'_i) \).

4.2. Architecture

The DreamPose model is a pose- and image-conditioned image generation model that modifies and fine-tunes 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. Full implementation details are provided in the supplementary material.

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 finetuning 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 finetuning, 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 finetuning schemes [5, 52] 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 \( A \) 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:

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

### 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 [12]) 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 fine-tuned 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.

![Figure 5: Ablation of VAE Finetuning. We find that fine-tuning 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 fine-tuning the UNet alone.]

### 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 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:

$$\epsilon_\theta(z_t, c_I, c_p) = \epsilon_\theta(z_t, \emptyset, \emptyset) + s_I(\epsilon_\theta(z_t, c_I, \emptyset) - \epsilon_\theta(z_t, \emptyset, \emptyset)) + s_p(\epsilon_\theta(z_t, c_I, c_p) - \epsilon_\theta(z_t, c_I, \emptyset)) \tag{4}$$

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.

### 4.5. Dataset

We train and test our method on the UBC Fashion dataset [50]. 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. During training, we randomly sample pairs of frames from the training videos. We compute poses with DensePose [12].

### 5. 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 [50]. DreamPose handles diverse human and clothing appearances well, even from different viewpoints and loose garments. Moreover, for a given subject image, DreamPose
Figure 6: Qualitative comparisons of our method versus MRAA [39], TPSMM [54], and PIDM [4]. Our method produces more photorealistic details aligned with the input frame, such as garment folds, fine-grain patterns, and face identity. Our method can also better handle challenging cases, including pattern synthesis in occluded regions.

Table 1: Quantitative comparisons of our method with MRAA, TPSMM, PIDM, and our method. Bolded values indicate best scores in each column.
can be conditioned on driving poses derived from a different video, as shown in the supplementary Figure 11. We also show additional results on images from the DeepFashion dataset [11] in the supplementary materials.

5.1. 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) [54], and a concurrent diffusion-based pose transfer method, PIDM [4]. We train all methods from scratch on the UBC Fashion Dataset [50], using the provided training scripts and configurations. For evaluating MRAA [39] and TPSMM [54], we use the provided test scripts in the “AVD” mode.

5.1.1 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 [50]. For each video, we extract 50 frames for testing, where they are at least 50 frames away from the input frame. The full DreamPose model quantitatively outperforms all three methods in all metrics: L1, SSIM [45], VGG [22], LPIPS [53], FID [15], FVD [43], and Average Euclidean Distance (AED). Note that while PIDM produces realistic images, it struggles to preserve clothing and person identity, as indicated by the quantitative results. Moreover, there is poor temporal consistency frame-to-frame, resulting in a large FVD [43] score.

5.1.2 Qualitative Analysis

We qualitatively compare our method to MRAA, TPSMM, and PIDM in Figure 3. With these other methods, 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 or blurry limbs. Compared to PIDM, DreamPose produces higher-fidelity results, in terms of both face identity and clothing patterns. While 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. As such, PIDM does not work well as-is for video synthesis. We provide additional comparisons to PIDM on the DeepFashion dataset [11] in the supplementary materials. Furthermore, we conduct a user study and provide the results in the supplementary material.

5.2. Ablation Studies

We perform comparisons of ablated versions of our method to verify our design choices. Namely, we compare five variants: (1) OursCLIP: We use a pretrained CLIP image encoder, instead of our dual CLIP-VAE encoder, (2) OursNo-VAE-FT: We do subject-specific finetuning of the UNet only, not the VAE decoder, (3) Ours1-pose: We concatenate only one target pose, instead of 5 consecutive poses, to the noise. (4) OursSmooth: Like ablation Ours1-pose, but apply temporal smoothing to the output frames. (5) OursFull: 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 [50]. Shown in Table 2, our full model outperforms the ablated versions in all four metrics.

<table>
<thead>
<tr>
<th></th>
<th>L1 ↓</th>
<th>SSIM ↑</th>
<th>VGG ↓</th>
<th>LPIPS ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>OursCLIP</td>
<td>0.025</td>
<td>0.882</td>
<td>0.247</td>
<td>0.070</td>
</tr>
<tr>
<td>OursNo-VAE-FT</td>
<td>0.025</td>
<td>0.897</td>
<td>0.210</td>
<td>0.057</td>
</tr>
<tr>
<td>Ours1-pose</td>
<td>0.019</td>
<td>0.899</td>
<td>0.208</td>
<td>0.056</td>
</tr>
<tr>
<td>OursSmooth</td>
<td>0.767</td>
<td>0.758</td>
<td>0.502</td>
<td>0.202</td>
</tr>
<tr>
<td>OursFull</td>
<td>0.019</td>
<td>0.900</td>
<td>0.207</td>
<td>0.056</td>
</tr>
</tbody>
</table>

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,
Figure 8: Qualitative Ablation of Image Conditioning. We compare results of the original text-to-image Stable Diffusion model, our model with CLIP-only image embeddings without finetuning, our model with CLIP-V AE image encoder embeddings without V AE finetuning, and our full model.

our method with only a CLIP image encoder, our method with CLIP-V AE encoder, and our full method with CLIP-V AE encoder and subject-specific V AE 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.

6. Limitations & Future Work

In Figure 9, 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.

Figure 9: 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).

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

Acknowledgments

This work was supported by NVIDIA and the UW Reality Lab, Meta, Google, OPPO, and Amazon.
References


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


