future internet

MDPI

Article

Dynamic Fashion Video Synthesis from Static Imagery

Tasin Islam, Alina Miron, Xiaohui Liu and Yongmin Li *

Department of Computer Science, Brunel University London, London UB8 3PH, UK; tasin.islam2@brunel.ac.uk (T.I.); alina.miron@brunel.ac.uk (A.M.); xiaohui.liu@brunel.ac.uk (X.L.)
* Correspondence: yongmin.li@brunel.ac.uk

Abstract: Online shopping for clothing has become increasingly popular among many people. However, this trend comes with its own set of challenges. For example, it can be difficult for customers to make informed purchase decisions without trying on the clothes to see how they move and flow. We address this issue by introducing a new image-to-video generator called FashionFlow to generate fashion videos to show how clothing products move and flow on a person. By utilising a latent diffusion model and various other components, we are able to synthesise a high-fidelity video conditioned by a fashion image. The components include the use of pseudo-3D convolution, VAE, CLIP, frame interpolator and attention to generate a smooth video efficiently while preserving vital characteristics from the conditioning image. The contribution of our work is the creation of a model that can synthesise videos from images. We show how we use a pre-trained VAE decoder to process the latent space and generate a video. We demonstrate the effectiveness of our local and global conditioners, which help preserve the maximum amount of detail from the conditioning image. Our model is unique because it produces spontaneous and believable motion using only one image, while other diffusion models are either text-to-video or image-to-video using pre-recorded pose sequences. Overall, our research demonstrates a successful synthesis of fashion videos featuring models posing from various angles, showcasing the movement of the garment. Our findings hold great promise for improving and enhancing the online fashion industry's shopping experience.

Keywords: diffusion models; fashion synthesis; generative AI; image-to-video synthesis

Citation: Islam, T.; Miron, A.; Liu, X.; Li, Y. Dynamic Fashion Video Synthesis from Static Imagery. *Future Internet* **2024**, *16*, 287. https://doi.org/10.3390/fi16080287

Academic Editor: Gianluigi Ferrari

Received: 18 July 2024
Revised: 6 August 2024
Accepted: 7 August 2024
Published: 8 August 2024

1. Introduction

The rise of online clothing shopping has brought about challenges for both customers and businesses. Customers often have to guess whether a product will look good on them or fit properly, leading to a higher risk of returns and lower satisfaction for both parties [1]. Additionally, customers are unable to physically experiment with fashion products, such as feeling and flow when worn, which can result in a less satisfying online shopping experience compared to in-person shopping.

Currently, there is active ongoing research on how deep learning can help fashion businesses prosper and allow customers to have a better shopping experience [2]. One area of focus is image-based virtual try-on, which uses a deep learning model to combine images of the customer and the desired clothing product to create a try-on image [3,4]. This helps customers visualise how the clothing item will look on them, making it easier for them to make a purchase decision. Not only can deep learning approaches improve customer shopping experience, but they can make business processes more effective. There are deep learning models that can transform rough fashion design sketches into real products [5], allowing businesses to quickly develop and create new products. This speeds up the experimentation process and enables businesses to bring their initial ideas to life in a timely manner.

While other diffusion models are either text-to-video or image-to-video using pre-recorded pose sequences, we have developed a new deep learning framework that can help businesses market their fashion products in a more engaging way. This model creates a short

video from a still image, showcasing how a piece of clothing moves and flows when worn by an actor. Using a latent diffusion model [6], as well as additional components such as a variational autoencoder (VAE) [7], contrastive language–image pre-training (CLIP) [8], a frame interpolator and attention [9], we are able to capture vital characteristics from the still image and generate a high-fidelity video. Our model stands out from others because it can generate natural and convincing movements with just a single image, unlike other diffusion models that rely on pre-recorded pose sequences or are limited to text data. In addition, we prioritise making diffusion models more efficient to speed up video synthesis and reduce computational demands. This makes it more affordable for businesses, as they do not need to invest in expensive AI accelerators, and reduces the inference time so that customers can see results quickly. Efficiency is often overlooked in the research community.

The contribution from our work is the following:

- Developed a generative model that produces spontaneous fashion videos with realistic movements from a still image. The model is a latent diffusion model, utilising a cross-attention mechanism to combine the noisy latent with the conditioning data.
- Utilised a pre-trained VAE decoder to process each frame from the denoised latent space to synthesise a complete fashion video.
- Demonstration of local and global conditioning enables the model to preserve the most detail from the conditioning image.

We share our source code and provide pre-trained models on our GitHub repository located at https:/github.com/1702609/FashionFlow (accessed on 6 August 2024).

2. Background

In this section, we will discuss the existing literature that shows how deep learning can benefit the fashion industry, including its functionality and potential business and customer benefits. We will be concentrating on generative models that can synthesise images and videos.

2.1. Generative Adversarial Networks

Generative Adversarial Networks (GANs) have emerged as cutting-edge technology for image synthesis and generation, with significant advancements made in recent years. Particularly, StyleGANs have showcased the potential of GANs in generating remarkably realistic photographic images [10,11]. GANs comprise two neural networks that operate in an adversarial manner, allowing the generator network to produce samples that mimic the underlying dataset while the discriminator network evaluates whether the sample is real or generated [12].

MoCoGAN generates unconditioned videos from noise vectors [13]. The model uses two principles to generate a video—the content vector and motion vector. The content vector specifies the object and appearance of the video, while the motion vector specifies the motion and movement of the video. The framework generates a video by mapping a sequence of random vectors to a sequence of video frames. Each random vector consists of a content part and a motion part. The content part remains fixed, while the motion part is a stochastic process. MoCoGAN has two discriminators, one that differentiates between real and fake images and the other that differentiates between real and fake video stacks. This ensures that the generated video is of good quality and has realistic dynamics.

Vondrick et al. utilised a spatiotemporal convolutional architecture to synthesise a video [14]. The model takes a low-dimensional latent code produced by Gaussian noise as its input. The model comprises two streams. The first stream uses spatiotemporal convolutions to upsample the latent code and generate high-dimensional videos with numerous frames. The second stream controls the background and foreground objects to create the impression of a stationary camera with only the object in motion.

The StyleGAN-V is built on the modified StyleGAN2 architecture [15] to synthesise videos [16]. This model utilises a similar concept as the MoCoGAN [13], where the noise vector is split into two vectors—the content vector and the motion vector. However,

the motion code in StyleGAN-V is continuous, and this enables the production of very long videos with consistent frames. Unlike previous approaches, where the features of the image and video are evaluated separately by multiple discriminators [13,17,18], StyleGAN-V employs a single discriminator that is conditioned on the time distances between the frames. This discriminator is more efficient than traditional video discriminators and provides more effective feedback to the generator. The experimental results have shown that the StyleGAN-V model can produce videos for as long as one hour without any issues, which was a significant challenge for its competitors.

There are video GANs that use conditioning data to produce desired videos. For instance, StoryGAN uses a paragraph of text to visualise a story by generating a sequence of images that complement the text [17]. The generator does this in two steps. First, a story encoder maps the text into a low-dimensional vector using a multi-layer perceptron (MLP). Second, a context encoder, which is a recurrent neural network (RNN), tracks the story flow and generates the next image to keep the story moving forward. There are two discriminators—an image discriminator and a story discriminator—that evaluate the genuineness of the story and image. The image discriminator measures whether the generated image matches the sentence, while the story discriminator ensures that the generated image sequence aligns with the story.

Mocycle-GAN can translate videos from one form to another; for example, converting videos into segmentation labels and vice versa [19]. It can even transform day-time videos into sunset or night-time environments. Mocycle-GAN maintains consistency in the video by using optical flow and temporal constraints. The model ensures that the video is smooth by maintaining a similar optical flow between adjacent frames throughout the video. Furthermore, the researchers utilise a technique to ensure that the motion in the video aligns with reality. This involves transferring motion information across different types of videos.

ImaGINator is a model that can generate a video using just one image and a class label to control facial expression and action [18]. The generator consists of an image encoder that encodes the conditioning image into a latent vector, which is then combined with noise and the class label. The decoder is a more complex component that utilises pseudo-3D convolutional layers. These layers separate the convolutional filters into temporal (1D) and spatial (2D) components. This approach is better because pseudo-3D convolutional layers are more efficient to optimise than the standard 3D convolutional [20]. Additionally, the decoder employs a fusion mechanism to maintain the appearance of the video throughout.

Dynamic GAN can generate sign language videos by processing images of skeletal poses and people [21]. The researchers of this study achieved this through employing a U-Net model that generated target frames using skeletal pose key point information. Additionally, they incorporated a VGG-19 model that allows for classification, intermediate frame generation, de-blurring, image alignment and other video-quality enhancement purposes to produce high-quality photorealistic sign gesture videos. H-DNA extends the functionality of Dynamic GAN by allowing for the recognition and generation of sign language videos using a hybrid approach of neural machine translation (NMT) and MediaPipe [22].

A conditional invertible neural network (cINN) can be utilised for synthesising videos from images [23]. This model employs the invertible domain transfer method. To accommodate the distinctions between images and videos, a probabilistic residual representation is introduced. The bijective mapping only captures information that complements the initial image. This probabilistic approach enables sampling and generating new video sequences starting from the same initial frame.

Poke learns from fine-grained object dynamics, enabling the model to generate video sequences with natural responses to pixel-level user interactions with images [24]. The model utilises a hierarchical recurrent model to capture complex, detailed object dynamics. Poke is trained solely on video sequences without pre-defined object interaction assumptions or ground-truth interactions.

Most GAN-based video synthesis models have not developed a method for producing high-fidelity videos based on a conditioning image. Instead, they often generate videos based on random noise or perform video-to-video translation. These models lack the necessary components to extract specific details from the conditioning image and incorporate them into the video synthesis process.

2.2. Diffusion Model

GANs and diffusion models are two different techniques used to generate data. While GANs use adversarial training to improve the generator, diffusion models gradually add noise to real data to create a series of distorted versions. The model then learns to reverse this process by removing the noise sequentially to generate a coherent image or other data types, starting from random noise.

Similar to GANs, diffusion models offer the capability to manipulate the generated images using textual descriptions [6] or input images [25]. This adaptability opens up a wide array of possibilities for their use in the fashion industry. Diffusion models have emerged as a promising approach to generating videos. Numerous research studies have investigated the potential use of these models in this domain. Recent works such as [26–33] have shown that diffusion models are currently the most active area of research in the development of generative models for videos. Notably, these models have demonstrated their ability to generate videos with the best quality and temporal consistency.

Many researchers have pointed out that training diffusion models with large-scale video datasets are computationally expensive [26–28,34]. Some techniques have tackled this problem by leveraging models that are trained on datasets consisting of pairs of text and images [26,27]. These models are designed to understand how the world looks based on textual input, and they are then trained on unsupervised video footage to learn how images move in the real world [26].

In the field of video generation using diffusion models, most models rely on a text-to-video approach, where textual inputs are used to create videos [26,27,33,35,36]. Only a few models have demonstrated synthesing videos from conditioning images [26,37,38].

The video diffusion model [39] is the first approach to use the diffusion model for generating unconditional videos. This model generates low-resolution video frames first and then leverages a super-resolution module to upsample them. The key feature of this model is that it is trained to approximate the complex distributions of raw videos, which is computationally challenging.

Make-A-Video [26] utilises existing text-to-image models and converts them into text-to-video models through the use of spatiotemporal attention and convolution. These techniques enable the model to preserve the video's quality and ensure it has temporal smoothness. Similar to ImaGINator [18], Make-A-Video uses pseudo-3D convolutional layers where the 2D convolution is stacked against a 1D convolution. They would use unsupervised learning on unlabeled video data to teach a model how an image would move. The decoder generates low-resolution, low-framerate video frames, which are then interpolated to a higher frame rate and resolution as data passes through the decoder layers.

MagicVideo is an efficient model for text-to-video synthesis [35]. It uses a latent diffusion model (LDM) [6], which is an efficient approach to denoise the latent space to a lower dimension, making the entire process faster. Compared to other video synthesis tools, such as Make-A-Video [26], MagicVideo employs a 2D convolution with temporal computation operators to model both the spatial and temporal features of the video. The temporal computation operator is a lightweight adapter that can effectively exploit the correlation between video frames.

Tune-a-video trains a text-to-video generator using a single text-video pair and a pre-trained text-to-image model [27]. The spatiotemporal attention queries relevant positions in previous frames to ensure consistency with the generated video's temporal dimension. During the inference stage, structure guidance from the source video is incorporated. This means that the latent noise of the source video is obtained from the diffusion model with

no textual condition. The noise serves as the starting point for DDIM sampling, which is guided by an edited prompt to preserve the video's motion and structure while basing the content on the prompt.

Like Mocycle-GAN [19], there are diffusion models that perform as video-to-video translation diffusion models [34,38]. For example, Esser et al. [34] encodes the input video to extract its structure, including shapes, object locations, and changes over time, as well as its content, like colours, styles, and lighting. They use this information to control and influence the synthesised video. The resulting video must have the same structure as the input video, but the content can be changed through cross-attention, which adjusts the colour and appearance of the video. This allows the input video to be translated into a different style while respecting its motion and structure.

Given their success in generating high-quality videos, our proposed model also utilises a diffusion model to create fashion videos. The contribution of our work is the design of the diffusion model that generates high-fidelity video from conditional images. Our model captures the appearance of conditional images and synthesises believable movements. It differs from previous video diffusion models as most focus on text-to-video synthesis [26,27,33,35,36].

2.3. Deep Learning for Fashion Application

There are several deep learning models that can be used to support the fashion industry, as mentioned in [2]. For instance, virtual try-ons [3,4,40–43] give customers the ability to merge images of clothing items with their own images, allowing them to see how the garment will look and fit on them realistically. Facial makeup transfer allows the model to transfer makeup from a reference image to a source image [44]. Additionally, pose transfer can show different viewing angles of fashion products by altering the posture of a person in the image [45,46].

Image-based fashion recommendation systems (FRSs) offer consumers a highly personalised shopping experience by leveraging their browsing history and previous purchase records to provide tailored recommendations. Among these FRSs, some employ deep learning techniques [47,48]. These advanced systems go beyond basic recommendations by predicting compatibility scores between clothing items, particularly in terms of their style, such as colours and patterns.

For instance, when a customer is interested in purchasing a t-shirt, these models can thoroughly analyse the chosen t-shirt and, based on its style attributes, suggest the most suitable trousers that complement the selected t-shirt. Essentially, they make informed decisions on behalf of the customer, ensuring that the clothing combinations harmonise seamlessly. This functionality greatly improves the shopping experience for customers, making it not only convenient but also enhancing their overall satisfaction with the process [49].

Video virtual try-on, as demonstrated in works like [50–53], take virtual try-on experiences to the next level by seamlessly integrating clothing onto a person in a video. This dynamic approach enables customers to witness clothing items in motion as they respond and adapt to the wearer's movements. Notably, these models capture the subtle nuances of how clothing products move and flow as the person walks, gestures, or performs various activities. This level of detail and realism is exceptionally informative, as it allows customers to gauge not just how a garment looks when stationary but also how it behaves during real-world activities. Consequently, a virtual video try-on elevates the online shopping experience by providing a holistic view of a clothing item's fit, comfort, and aesthetic appeal in motion.

Although the use of diffusion models in fashion-related tasks is a recent development, only a few studies have explored this approach thus far [38,54,55]. Some of these studies include DreamPose [38], a fashion video synthesis model that uses conditional poses to guide video synthesis, and DiffFashion [54], a model that combines texture from one image with a fashion product. These models demonstrate the applicability of diffusion models in the fashion industry and their potential to revolutionise various aspects of it, from cre-

ative content generation and virtual try-ons to personalised styling recommendations and trend forecasting.

Our work solves a similar problem presented in DreamPose [38], where we generate videos from a single image. However, we differ from DreamPose in terms of producing spontaneous yet believable motions, while DreamPose uses pre-recorded posture data to guide video synthesis. Additionally, we utilise cross-attention [9] mechanisms that simultaneously condition all video frames. In contrast, DreamPose synthesises each video frame separately and applies cross-attention individually, making their model much slower than ours.

Compared to previous work, our goal is to make generative AI more attractive to businesses by developing more computationally efficient models. This means that businesses will not need to invest in expensive AI accelerators. Furthermore, efficient models can generate videos more quickly, thus preventing the testing of their customers' patience. Our research is focused on developing diffusion models that can produce high-quality fashion videos from conditional images without the need for time-consuming fine-tuning or other activities. Implementing fast and efficient models will offer businesses an innovative and enjoyable way to promote their fashion products while enhancing the overall customer's shopping experience.

3. Method

In this section, we present an approach for generating short, spontaneous fashion videos using a single image as a conditioner, as shown in Figure 1. Our method harnesses the power of the diffusion model. The model produces a video showcasing an actor performing spontaneous yet fashion-relevant poses and movements while maintaining stylistic consistency.

Overall, the model is fed with a conditioning image, which is processed into a latent space by the VAE and CLIP encoder. This latent space then influences the latent diffusion U-Net through cross-attention. The noisy video latent space is passed through pseudo-3D convolution layers, cross-attention and spatiotemporal attention layers, and finally, a VAE decoder to synthesise a final video.

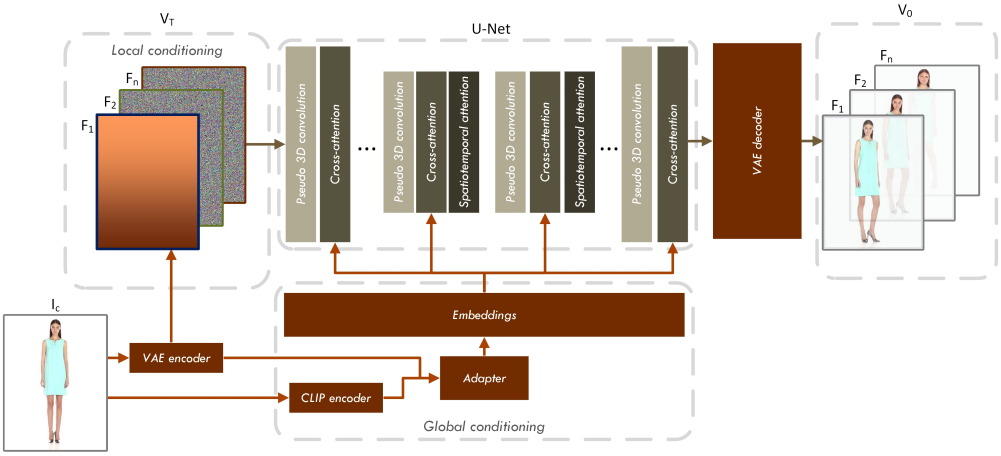


Figure 1. The architecture of our proposed image-to-video model. Our approach involves a latent diffusion model [6] to denoise the latent space of a video. Each frame of the latent space is then processed using a pre-trained VAE decoder to generate the final video. We condition the video in two ways: locally and globally. Local conditioning involves adding a VAE-encoded image as the first frame of the noisy latent, while global conditioning involves using cross-attention layers [9] to influence intermediate features with the conditioning image throughout the layers of the latent diffusion U-Net.

3.1. Diffusion Model

Using traditional diffusion models to denoise the pixels of videos would be extremely time-consuming and will require significant computational resources. We employ the latent diffusion model (LDM) [6] to create a video. LDM directly denoises the latent space, making them more efficient as they require fewer parameters and demand lesser computational resources, thus making them suitable for video synthesis. Generating videos involves producing multiple frames, which can be time-consuming. Therefore, efficiency is a crucial factor, and LDM's direct work on the latent space makes it an ideal option for video synthesis.

Our LDM is a U-Net architecture [56] that contains layers of pseudo-3D convolution, cross-attention, and spatiotemporal attention. These components can handle the latent space of the video and modify each frame of the latent space with the conditioning image. Our latent diffusion U-Net distinguishes itself from other models by employing cross-attention layers. We are the first to demonstrate the incorporation of cross-attention layers by combining latent videos with encoded conditioning images.

There are several ways in which diffusion models add noise to data [57], and we use the noise scheduler used by the diffusion probabilistic model (DDPM) [58]. The forward process can be described as follows:

$$x_t = \sqrt{1-\beta_t} \cdot x_{t-1} + \sqrt{\beta_t} \cdot \epsilon_t, \epsilon_t \sim \mathcal{N}(0, I) \tag{1}$$

where at each timestep t, the noisy data x_t is produced by adding noise to data from the previous timestep x_{t-1}. The noise level at timestep t is denoted by β_t. Gaussian noise is added to the data at each timestep, which is represented by ϵ_t. Similarly, the reverse process is presented as follows:

$$x_{t-1} = \mu_\theta(x_t, t) + \sqrt{\beta_t} \cdot \epsilon_t, \epsilon_t \sim \mathcal{N}(0, I) \tag{2}$$

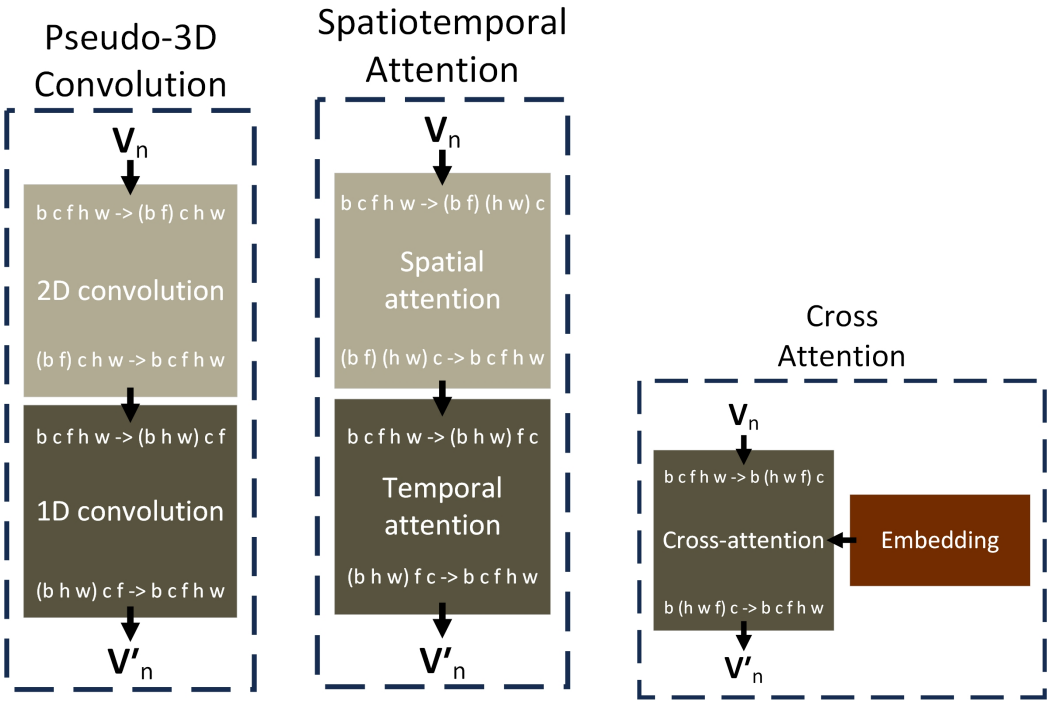
where a neural network $\mu_\theta(x_t, t)$ is used to predict the noise of x_t and subtract it to get x_{t-1}.

In Equations (1) and 2, the variable x is replaced with the latent video V, as shown in Figure 1. The video latent space V_T contains complete noise, except for the first frame, which includes a VAE-encoded condition image I_{VAE}. We train the latent diffusion U-Net not to modify the first frame by not adding any noise to it during the forward process. Additionally, we use global conditioning to ensure that the video retains high fidelity while preserving the details from I_c.

3.2. Pseudo-3D Convolution

Generating high-quality videos poses significant challenges, especially when it comes to ensuring the quality and fluidity of the video are exceptional. Two-dimensional convolution layers are the most practical approach to handle the spatial aspect of an image [59–61]. While it is possible to use 3D convolution layers to generate videos, they are more difficult to optimise and can be computationally expensive, as discussed in Section 2. We draw inspiration from the techniques proposed by [26,62], where they introduce an approach known as pseudo-3D convolution (shown in Figure 2a). This method involves stacking 1D convolutional layers alongside 2D convolutional layers to efficiently process video data. Adopting the pseudo-3D convolution technique offers a more efficient and effective way to handle both spatial and temporal dimensions in video processing tasks.

The pseudo-3D convolutional layer is considered better than the standard 3D convolutional layer because it can easily capture the nonlinear relationship between the spatial and temporal aspects of a video, allowing it to be more capable of understanding complex functions in the video data [20]. Even when both layers have the same parameters, the pseudo-3D convolutional layer processes the spatial and temporal aspects of the video separately, unlike the standard 3D convolution layer, which processes them together and leads to worse performance.



(**a**) (**b**) (**c**)

Figure 2. The architecture of the pseudo-3D convolutional and attention layers. b, c, f, h, w represent the number or value of batch, channel, frame, height and width, respectively. (**a**) The pseudo-3D convolutional layer eases optimisation and performs better than its standard counterpart. (**b**) The spatiotemporal attention layer helps the model generate high-quality video frames while maintaining smoothness and consistency. (**c**) The cross-attention layer allows the model to condition the synthesised video based on the input image.

In Figure 2a, we show the step-by-step transformation of V_n dimensions when passing it to a pseudo-3D convolutional layer. b, f, c, h and w represent the value of batch, frame, channel, height, and width, respectively. Firstly, we change the dimension V_n to $(b\,f)\;c\;h\;w$ for the 2D convolutional layer to process the spatial aspect. Then, we change the dimension again to $(b\;h\;w)\;c\;f$ for the 1D convolutional layer to process the temporal aspect.

3.3. Frame Interpolation

Our latent diffusion U-Net decoder has been specifically designed to perform frame interpolation for video smoothing purposes. Frame interpolation refers to a technique where additional frames are added between existing ones in order to make videos smoother and longer. This technique is helpful in situations where the original video has a low frame rate or is of low quality.

We achieve frame interpolation by masking. Masking involves hiding or masking half of the frames in the latent space. The decoder then synthesises how the masked frames would look based on the neighbouring visible frames. By doing this, the decoder is able to generate new frames that are similar in appearance to the original frames.

According to the implementation proposed by [63], we only apply masking to the frames when the latent space reaches the innermost layer of our latent diffusion U-Net. In our implementation, the latent space consists of 70 frames, and when it reaches the last encoder layer, 35 of those frames will be masked. The decoder then needs to synthesise what the masked frames might look like.

3.4. VAE and CLIP Encoder

We used the encoder of the variational autoencoder (VAE) to compress the conditional image I_c into a lower-dimensional latent representation, denoted as I_{vae}. We applied I_{vae} in

two ways to condition the latent space: local and global conditioning. Local conditioning involves using I_{vae} as the initial frame of the latent video noise. The LDM captures details from I_{vae} to ensure that the remaining noisy video latent frames preserve vital information such as colours and shapes.

For global conditioning, we take additional steps, which involves producing I_{clip} from I_c using contrastive language-image pre-training (CLIP) [8]. CLIP is a neural network that can generate relevant captions for an input image based on its training on various image and text pairs. We combine I_{vae} and I_{clip} using an adapter proposed by [38], and we use cross-attention [9] to condition the LDM right after the latent space has been processed using a pseudo-3D convolutional layer.

The stable diffusion model is dependent on the text prompt embeddings from the CLIP encoder [6]. We have a similar structure to stable diffusion, so we use the same CLIP embedding structure, but our embedding consists of image conditioning data. DreamPose has suggested using VAE and CLIP encoders to extract a more detailed image representation, which is more effective in conditioning the model [38]. By combining the embeddings obtained from both VAE and CLIP, we can have a more complete set of features that can better guide the model during the generation process.

3.5. Attention

Attention layers are very useful for generating images and videos. They can help models prioritise and extract relevant regions of data and disregard irrelevant parts of the data [64]. There are models that have used attention for synthesising videos [26,31]. The general equation for attention is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \tag{3}$$

where Q is the query matrix representing the encoded representation of the current input, K is the key matrix representing the encoded representation of the input sequence to attend to, V is the value matrix containing the information to be extracted for each input element, and d_k is the dimensionality of the key matrixes.

In each pseudo-3D convolutional block of our model, we integrate a cross-attention layer as shown in Figure 2c. This layer plays a crucial role in learning the significant connections between two sets of data [6]. In our case, these two sets of data are the noisy latent video V_n and the input try-on image I_c, which have been encoded by VAE and CLIP, I_{vae} and I_{clip}, respectively. The cross-attention layer calculates the relevance of each element in the noisy latent video with respect to each element in the input try-on image. By doing so, our model is able to focus on the most relevant parts from the input sources and generate the desired outcomes effectively.

In Figure 2c, we show that the dimension of V_n has to be changed to allow the cross-attention layer to perform its task. We change the dimension of V_n to $b\ (h\ w\ f)\ c$. The cross-attention layer processes and modifies the texture of V_n in the spatiotemporal dimension with the condition image at every channel.

The innermost layers of our latent diffusion U-Net use spatiotemporal attention layers that allow the model to understand complex relationships within video frames and improve quality. The spatiotemporal layers consist of spatial attention and temporal attention. Spatial attention evaluates the significance of individual elements (i.e., pixels) within a feature space relative to one another. This approach empowers the model to consider both local (neighbouring regions) and global (farther apart regions) dependencies in the data, thereby allowing it to capture intricate contextual information. Temporal attention mechanisms play a crucial role in unravelling temporal dependencies between consecutive frames within a video sequence. This enables our model to seamlessly synthesise fluid and coherent movements in video content.

The spatiotemporal attention shown in Figure 2b begins with performing spatial attention on V_n. The dimension of V_n is rearranged as $(b\ f)\ (h\ w)\ c$, which enables the

attention layer to process the spatial dimension at every channel. After this, the dimension is rearranged as $(b\ h\ w)\ f\ c$ to allow the temporal attention layer to operate and focus on the temporal dimension at every channel. Overall, the spatiotemporal attention layer helps the model to ensure that the video is high-quality and its movements look natural.

4. Experiments

In this section, we will provide a comprehensive overview of the training and evaluation process for FashionFlow. Our analysis encompasses both qualitative and quantitative comparisons, offering an in-depth examination of our results. Additionally, we delve into the distinct contributions made by each model component towards the overall performance. We further elaborate on the datasets utilized in our experiments, along with a detailed exposition of the network implementation details. This commitment to transparency and thorough documentation ensures the reproducibility of our work.

4.1. Dataset

We trained our model on the Fashion dataset [65]. This dataset features professional women models who pose at various angles to showcase their dresses. There is a vast range of clothing and textures available, offering a multitude of possible appearances.

This dataset includes 500 videos for training and 100 for testing. Each video consists of approximately 350 frames. The video resolution is set to 512 pixels in height and 400 pixels in width.

4.2. Implementation

We divide our proposed network into three sections: the VAE encoder, the latent diffusion U-Net, and the VAE decoder. The pre-trained VAE encoder and decoder are produced by [6]. The latent diffusion U-Net architecture consists of several blocks that contain pseudo-3D convolutional layers. The downsampler block has kernel sizes of 64, 128, 256, and 512. The middle section is made up of four pseudo-3D convolutional layers, all utilising 512 filters. The upsampler block has kernel sizes set to 512, 256, 128, and 64. All pseudo-3D convolutional layers use a kernel size of 3, a stride of 1, and padding set to 1.

After every block of pseudo-3D convolutional layers, we utilise cross-attention to enable the network to capture intrinsic detail from the conditioning image. The spatiotemporal attention is utilised in the innermost block of the latent diffusion U-Net.

We trained the latent diffusion U-Net for 2500 epochs with a denoising step of 1000. We chose DDPM [58] as our cosine noise scheduler as it adds noise at a slower rate than linear. This enhances the diffusion model's performance [66]. We utilised the AdamW optimiser [67] to train the latent diffusion U-Net. We set the learning rate hyperparameter to 0.0002 and the values of β_1 and β_2 to 0.5 and 0.999, respectively.

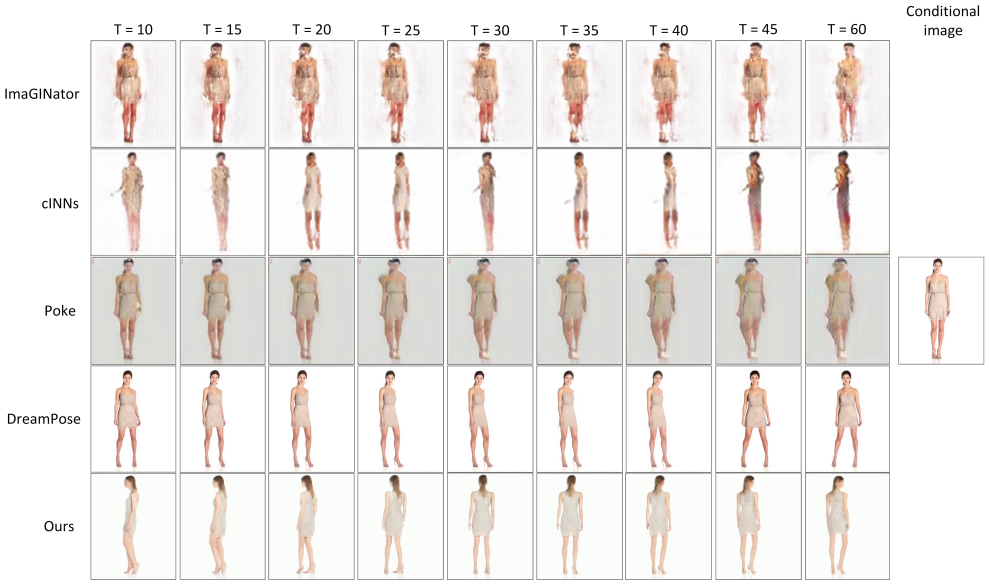
4.3. Qualitative Analysis

The videos presented in Figure 3 show a side-by-side comparison of our model with four other models, namely ImaGINator [18], cINNs [23], Poke [24], and DreamPose [38], whose VAE decoder has been fine-tuned. Our model outperforms others in terms of the range of motion it can perform, such as making a person turn significantly, and our result is much more temporally consistent.

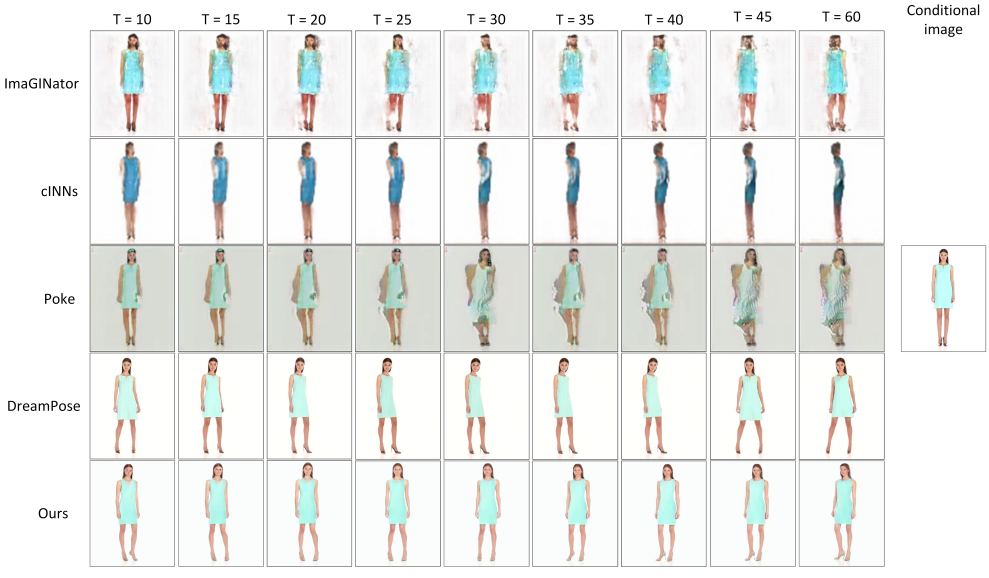
ImaGINator [18], cINNs [23], and Poke [24] use GANs [12] to synthesise videos. These models were initially designed to create short videos with a low number of frames (around 10 to 20) and a resolution of either 64×64 or 128×128. Because of this, the video quality of GAN-based models deteriorates beyond 20 frames. We used the Fashion dataset [65] to train ImaGINator and downloaded pre-trained models of cINNs and Poke, which were trained on the iPER dataset [68]. The iPER dataset showcases individuals performing tai chi moves, which is similar and relevant to the movements performed in the Fashion dataset. Our approach has demonstrated a better performance compared to previous work in terms of video quality, as our model generates videos with higher resolution and

smoother temporal consistency. Our model generates a longer video consisting of 70 frames with a resolution of 512 for height and 640 for width. This indicates that our model is more effective in informing customers about the product clearly and with greater detail. Regarding DreamPose, this model was able to retain the facial details of the person better than ours because they performed person-specific fine-tuning. We do not perform this because it is significantly time-consuming and inefficient. We discuss this in Section 4.5.

Apart from DreamPose, we encountered difficulties in accurately depicting the fine details of the face. Producing precise facial representations is a complex task. Other video diffusion models have not addressed the image-to-video problems, and there is a lack of research on how to capture intricate details such as facial features from an image. Based on the results of the text-to-video approaches, synthesising a person's face from a distance is challenging [28,39]. Additionally, there is a lack of experiments involving human faces in the existing literature, with most research focusing on animal movements and landscape transitions. Synthesising realistic faces in videos remains a formidable challenge, necessitating further research and development efforts.

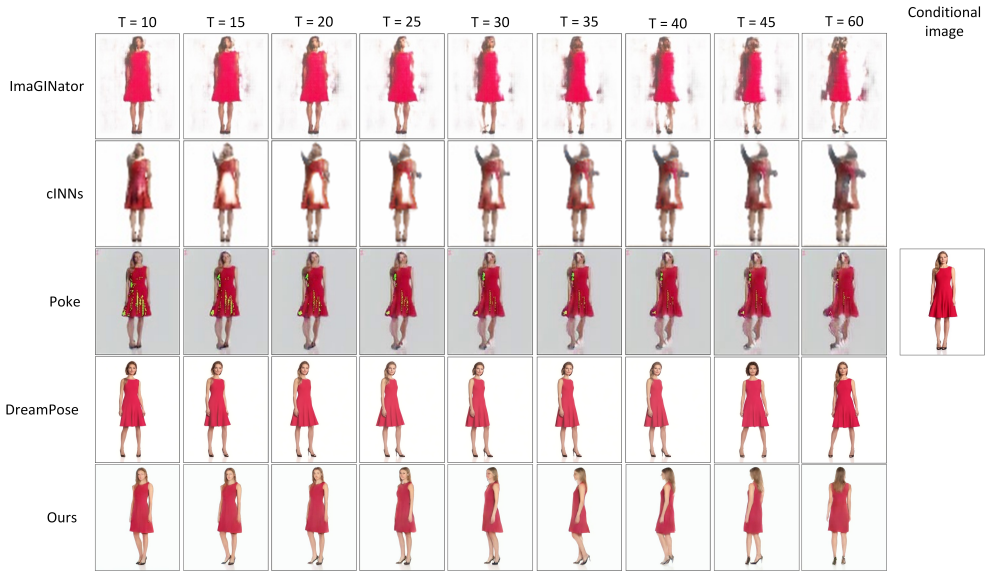


(**a**)



(**b**)

Figure 3. *Cont.*



(**c**)

Figure 3. Qualitative comparison of our method against ImaGINator [18], cINNs [23], Poke [24], and DreamPose [38]. Subfigures (**a**–**c**) show how our model and competitors handle conditional images with different coloured dresses and people. Our method executes a wider range of movements, showcasing the clothing from various angles. Our result is highly comparable to that of DreamPose in terms of both quality and temporal consistency. This means that our method is capable of producing high-quality results consistently over time, providing users with a reliable and efficient solution for their needs.

4.4. Quantitative Analysis

Video synthesis has received relatively less attention compared to image synthesis, resulting in a limited number of evaluation metrics available to assess the quality of synthesised videos. In this context, we have utilised evaluation metrics commonly employed for assessing the quality of fashion videos in the realm of video virtual try-on [42]. These metrics have also been adopted by other researchers working in the field of video synthesis, such as Poke [24] and cINNs [23]. Furthermore, the algorithms do not factor in the playback speed. Irrespective of whether the video is played at 30 or 60 frames per second, these algorithms assess each video frame independently and its complementarity with the others.

Frechet video distance (FVD) is an effective metric for evaluating the quality of a synthesised video [69,70]. It is influenced by the frechet inception distance (FID) [71], which is used for evaluating synthesised images. FVD introduces a feature representation that captures the temporal coherence of the content of a video, in addition to the quality of each frame. FVD consistently outperforms structural similarity (SSIM) and peak signal-to-noise ratio (PSNR) in terms of agreeing with human judgment [69]. FVD evaluates 16 frames of a video using the pre-trained inflated 3D convnet model (I3D) [72], which was designed to recognise the act performed in a video. The number of frames cannot be altered according to the implementation of [73]. However, the VFID I3D implementation by [74] is identical to FVD, except their I3D model can evaluate up to 60 frames.

The equation for FVD and VFID I3D is denoted as follows:

$$d(P_R, P_G) = |\mu_R + \mu_G|^2 + \mathrm{Tr}(\Sigma_R + \Sigma_G - 2(\Sigma_R \Sigma_G)^{1/2}) \quad (4)$$

where R represents the video from the dataset, G is the generated video, the variables μ_R and μ_G represent the mean, while Σ_R and Σ_G represent the covariance matrices of the recorded activation data from both the real P_R and generated P_G videos, respectively. P_R and P_G were obtained by passing the videos through a pre-trained I3D, which takes into account the visual content's temporal coherence across a sequence of frames.

The inception score (IS) serves as a tool to assess the performance of generative models [75]. Its purpose is to gauge both the variety and the aesthetic appeal of the images produced by these models. It achieves this by running the generated images through a classifier that has been trained beforehand and then determines a score based on the resulting probabilities. Specifically, the IS is derived from the exponential of the average KL divergence between the class distribution of the generated images and the class distribution of a large collection of real images. A higher IS suggests that the generated images possess greater diversity and visual appeal.

$$\mathrm{IS} = \exp\left(\mathbb{E}_{x \sim P_G}[D_{\mathrm{KL}}(P(y|x) \| P(y))]\right) \tag{5}$$

where P_G represents the distribution of the generated images, the notation $\mathbb{E}_{x \sim P_G}$ signifies that we are taking the average over samples x drawn from this distribution. The conditional distribution $P(y|x)$ indicates how labels y (values obtained from a pre-trained classifier) are distributed when we have a generated image x. The marginal distribution $P(y)$, on the other hand, shows the overall label distribution. The Kullback–Leibler divergence, denoted as D_{KL}, quantifies how dissimilar these two distributions are. The formula computes the expected value, denoted as $\mathbb{E}$, of the Kullback–Leibler divergence across all the generated images. Finally, the exponential function exp is applied to this expected Kullback–Leibler divergence to yield the IS.

Table 1 shows the results of our method in comparison to three other image-to-video models, namely ImaGINator [18], cINNs [23], and Poke [24]. The comparison was made on the Fashion dataset's testing subset [65]. We have downloaded pre-trained models of cINNs and Poke, which were trained on the iPER dataset [68] and Imagintor from the Weizmann Action dataset [76]. The movements performed are similar to those in the Fashion dataset [65], such as turning and stepping from side to side. This allows for a fair and meaningful comparison. Our model has outperformed the previous works in all metrics. The results were particularly impressive for the metrics FVD and VFID I3D. We have performed exceptionally well in the VFID I3D metric because the GAN-based models we compared were trained to generate only 20-30 frames. When they were made to produce more than that, the quality would seriously deteriorate. However, we have trained our model to generate more than 60 frames, which allows it to produce consistent frames and achieve a significantly better score.

Table 1. Quantitative comparison performed on the testing set of Fashion [65]. Our method has outperformed the other image-to-video models ImaGINator, cINNs, and Poke. Dreampose uses a video-to-video translation approach, which yields the best scores by processing individual video frames, but lacks the ability to generate spontaneous movements. Higher values indicate better results for IS, while lower values are desirable for FVD and VFID I3D.

Method	**IS ↑**	**FVD ↓**	**VFID I3D ↓**
ImaGINator [18]	2.003	3383.2	32.179
cINNs [23]	2.560	3326.0	79.386
Poke [24]	2.823	3514.5	25.929
Ours	2.965	1659.5	0.867
DreamPose [38]	3.016	1253.6	0.653

A quantitative evaluation was also conducted on DreamPose [38], and it scored slightly better than our model. However, it is important to note that DreamPose analyses five consecutive pose frames and adjusts the pose of the image accordingly, taking more of video-to-video translation approach. The metrics do not assess the ability of a model to synthesise spontaneous realistic movements, a feature which DreamPose does not provide.

4.5. Inference Time

We have observed that GAN models like ImaGINator, cINNs, and Poke can generate videos faster than diffusion models, including our own. The way GANs work is by synthesising images and videos instantly from input data, while diffusion models denoise the sample iteratively to obtain the final content. This means that it will take a significant amount of time for our model to generate a video. Additionally, the GAN models we compared our model with do not produce as many video frames. Although producing fewer frames may seem computationally less demanding and efficient, it does not effectively achieve the same objective as our model.

Evaluating the models on the same hardware, our model demonstrated the ability to synthesise a 70-frame video in 138 s, whereas DreamPose took over 1300 s or nearly 10 times the amount of time as ours, including the time required to fine-tune the VAE decoder to be person-specific. In real-world applications, DreamPose's longer processing time could lead to significant delays for customers, potentially causing them to lose interest. Our approach saves time by denoising the latent video instead of processing individual frames one by one, and we do not fine-tune the VAE decoder.

Generating videos is a complex task and requires a practical and efficient generative model. Fine-tuning a model for each individual is highly impractical as it takes a considerable amount of time and requires a large amount of additional storage space to save weight for each person. Our model is a more practical and faster alternative that does not require fine-tuning, nor does it consume additional storage space. Additionally, we believe that users would be highly frustrated if they had to wait for an AI model to fine-tune and then synthesise the video. In a business context, where multiple customers are being served, this would create a huge queue for those customers who want to use the model. Therefore, speed and efficiency are crucial factors when serving multiple people.

4.6. Ablation Study

In this section, we will be conducting an ablation study to analyse the impact of global and local conditioning. We will demonstrate the effects on the synthesised video when one of the conditioning methods is absent. Global conditioning involves using cross-attention mechanisms to influence the entire latent diffusion U-Net architecture. On the other hand, local conditioning entails inserting the encoded image as the first frame of our noise input.

In Figure 4, we can observe the effects of the conditioning image generation models based on local and global factors. The first row of the figure shows the results of using global conditioning. We can see that the model fails to capture small details, such as the white stripe on the dress. Instead, it captured the overall colour. The latent space generated by VAE and CLIP encoders was probably inadequate in compressing small details, potentially leading to the loss of such information during cross-attention operations on the U-Net. The second row shows the results of using only local conditioning. The model also struggled to capture intricate details, and as the video progresses, the texture undergoes significant alteration due to the model's inability to retain the appearance of the initial frame, resulting in a lack of temporal coherence across the frames. Finally, the third row shows the results of using both global and local conditioning, which yields the best result. In this case, the model does a better job of preserving specific details, such as the white stripe on the dress, while also maintaining the overall colour scheme. This is also supported by Table 2. Using both conditioners has quantitatively outperformed methods using a single variant across all metrics.

Figure 4. The effects of image conditioning. Global conditioning captures the overall colour of the garment, but it misses out on smaller details like the white stripes. Local conditioning darkened the skin colour too much and also failed to capture small clothing details. Using both local and global conditioning allows the model to capture the overall colour from the conditioning image, where it was able to pick up small details like stripes. This is because there is a greater information flow from the conditioning image to the noisy latent video.

Table 2. Ablation study on the conditioning methods. Our model performs best when both global and local conditioners are employed.

Method	IS ↑	FVD ↓	VFID I3D ↓
Global only	2.891	1812.5	0.881
Local only	2.801	2111.0	1.179
Global and local	2.965	1659.5	0.867

4.7. Extension of Virtual Try-On

Image-based virtual try-on is a deep learning model that can replace the clothing of a person with the desired item they want to wear [42]. FashionFlow has the potential to enhance virtual try-on models by bringing the virtual try-on image to life, showcasing clothing products from various angles, and making the experience more immersive and exciting.

In the current state of this paper, utilising FashionFlow in isolation proves to be advantageous for female customers, providing them with a visual understanding of how a dress flows and moves. However, the integration of FashionFlow with virtual try-on technology is expected to broaden the interest to a more diverse audience, as customers of any gender can engage with virtual try-on models. This approach holds the potential to maximise hedonic value, as it allows customers to wear clothing products virtually, encourages personalisation, and allows them to experiment with various poses, thereby enhancing the overall customer experience.

5. Limitations

Some GAN-based models, such as ImaGINator, cINNs, and Poke, have an advantage over our model in terms of the speed of video synthesis. This is because our model is a latent diffusion model that iteratively denoises latent space, which consumes more time than a GAN-based approach.

Currently, our model has a few areas that require improvement. The videos generated by our model sometimes have colours that look faded, facial identities that are poorly preserved, and clothing details that are lost as the video progresses. The second row of Figure 5 illustrates this. We believe that the global conditioning in the latent diffusion U-Net model is hindering its ability to capture intricate details from the conditioning image. This causes the latent diffusion U-Net to lose vital information about the conditioning image, leading to more hallucinations and a reduced degree of fidelity. We leave the improvement of information flow from the global conditioner for future research. The only model capable of preserving facial identity is DreamPose [38], as shown in Figure 3. However, this requires fine-tuning for every individual, which is significantly impractical and time-consuming.

Figure 5. Limitations of our model. In the first row, our model failed to preserve the skin colour of ethnic minorities accurately. We attribute this to the VAE decoder not being fine-tuned and the global conditioner being ineffective at capturing intricate details in images. Likewise, the second row depicts that the global conditioner was unable to preserve complex or patterned textures shown in the dress. For future work, we will look into improving the global conditioner's ability to pick up intricate details effectively. Additionally, we need to find ways to generate high-fidelity videos without the need for time-consuming tasks such as fine-tuning.

The Fashion dataset [65] has some limitations. One of the main limitations is that it is not representative of the entire population. This dataset only focuses on dresses and does not include models of men, children, or ethnic minorities. The first row of Figure 5 shows that the absence of representative data can result in a lower degree of fidelity in the generated video, where the skin colour is not preserved. Instead, it has been generated to match the most common skin tones present in the dataset. Due to this limitation, it may not be very useful for businesses that want to cater to a wider demographic. In order for our model to be more useful, a diverse dataset needs to be used. It can be argued that the dataset may not be the only issue. The main problem lies in our proposed model's ability to capture intricate details from the image as effectively as our competitor. Further research is needed to explore how global conditioning can capture more meaningful information from the image and what other techniques we can apply without increasing the inference time.

Maintaining consistency between clothing and characters is the most critical issue in the fashion field, and there is an urgent need to address this issue. If the display effect of the images of models wearing clothes is poor, it would make it difficult for users to imagine or match the effect of wearing clothes themselves. While the diffusion models have certainly improved the situation compared to earlier GAN-based models, the problem remains. Although our proposed model has demonstrated better performance over the other previous studies, maintaining fidelity and preserving details is still a challenging issue. Generating high-fidelity videos using diffusion models is still a relatively new field, and it is challenging to create convincing videos that exhibit consistency between the person and clothing. This research is only a beginning in understanding how the diffusion model

can use conditioning images to influence video synthesis. We leave it for future work to improve the model's ability to realistically simulate virtual clothing.

6. Conclusions

Our research introduces a novel architecture for diffusion models that is tailored specifically for synthesising high-fidelity fashion videos using conditional images. We propose a methodology that utilises pseudo-3D convolution, VAE, and CLIP encoder to condition synthesised videos both on a global and local scale. Our approach represents a significant advancement over previous efforts in this domain, such as synthesising videos at a faster pace and not requiring person-specific fine-tuning. Additionally, it produces videos that are temporally coherent and capture vital details from the conditioning image. We conducted a thorough comparison of our image-to-video model with various other models. We demonstrated that the video quality produced by our model is significantly better than the GAN-based methods. Our model generates videos with a higher resolution, allowing for a more detailed output. Additionally, we observed that our model produces videos that have better temporal consistency, meaning that the frames flow more seamlessly from one to the other.

Our diffusion model is unique compared to other diffusion models. Most of the previous works are text-to-video or image-to-video with pre-recorded pose sequences like DreamPose. Unlike these models, our diffusion model generates spontaneous movements that are realistic and distinctive without the need for pre-recorded data. This makes our model more practical and efficient, as it only requires one image as input.

We conducted an ablation study to evaluate the effectiveness of our approach. Our results show that conditioning the latent diffusion U-Net on local and global scales allows the model to preserve the most detail from the condition image. By doing so, our method can generate high-quality videos that demonstrate a high degree of fidelity and realism.

Overall, our work highlights the potential of combining deep learning techniques to synthesise high-quality videos with greater efficiency and precision. Our approach has significant implications for the fashion industry, where the ability to create high-quality videos is becoming increasingly important for marketing and showcasing products in a competitive setting.

The proposed image-to-video synthesis model also has a wide range of potential applications. At this stage, our model can be valuable for generating preliminary visualisations, concept art, and storyboarding elements for filmmakers and content creators, which can then be manually refined for final production. Video game developers might use it for early-stage animations, cutscenes, or prototyping visual elements before detailed finalisation. Additionally, artists and designers can leverage the model for experimental visual projects and initial drafts of visual artworks or animations that do not require extreme detail. In the future, when this model is improved, it can be used for higher-end applications.

Author Contributions: Conceptualisation, T.I., A.M. and Y.L.; methodology, T.I.; software, T.I.; validation, T.I.; formal analysis, T.I., A.M. and Y.L.; writing—original draft preparation, T.I.; writing—review and editing, T.I., A.M., X.L. and Y.L.; visualisation, T.I.; supervision, A.M. and Y.L.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Engineering and Physical Sciences Research Council (EPSRC) grant number EP/T518116/1.

Data Availability Statement: This paper did not generate any new data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pachoulakis, I.; Kapetanakiss, K. Augmented reality platforms for virtual fitting rooms. *Int. J. Multimed. Its Appl.* **2012**, *4*, 35. [CrossRef]
2. Cheng, W.H.; Song, S.; Chen, C.Y.; Hidayati, S.C.; Liu, J. Fashion meets computer vision: A survey. *ACM Comput. Surv. CSUR* **2021**, *54*, 1–41. [CrossRef]

3. Han, X.; Wu, Z.; Wu, Z.; Yu, R.; Davis, L.S. Viton: An image-based virtual try-on network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, UT, USA, 18–23 June 2018; pp. 7543–7552.
4. Wang, B.; Zheng, H.; Liang, X.; Chen, Y.; Lin, L.; Yang, M. Toward characteristic-preserving image-based virtual try-on network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 589–604.
5. Xian, W.; Sangkloy, P.; Agrawal, V.; Raj, A.; Lu, J.; Fang, C.; Yu, F.; Hays, J. Texturegan: Controlling deep image synthesis with texture patches. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, UT, USA, 18–23 June 2018; pp. 8456–8465.
6. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
7. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
8. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning. PMLR, Vienna, Austria, 18–24 July 2021; pp. 8748–8763.
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
10. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
11. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8110–8119.
12. Goodfellow, I. J,; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27* , 139–144. Available online: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf (accessed on 6 August 2024).
13. Tulyakov, S.; Liu, M.Y.; Yang, X.; Kautz, J. Mocogan: Decomposing motion and content for video generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1526–1535.
14. Vondrick, C.; Pirsiavash, H.; Torralba, A. Generating videos with scene dynamics. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 5–10 December 2016; Volume 29.
15. Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; Aila, T. Training generative adversarial networks with limited data. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12104–12114.
16. Skorokhodov, I.; Tulyakov, S.; Elhoseiny, M. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3626–3636.
17. Li, Y.; Gan, Z.; Shen, Y.; Liu, J.; Cheng, Y.; Wu, Y.; Carin, L.; Carlson, D.; Gao, J. Storygan: A sequential conditional gan for story visualization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6329–6338.
18. Wang, Y.; Bilinski, P.; Bremond, F.; Dantcheva, A. Imaginator: Conditional spatio-temporal gan for video generation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 1–5 March 2020; pp. 1160–1169.
19. Chen, Y.; Pan, Y.; Yao, T.; Tian, X.; Mei, T. Mocycle-gan: Unpaired video-to-video translation. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 647–655.
20. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6450–6459.
21. Natarajan, B.; Elakkiya, R. Dynamic GAN for high-quality sign language video generation from skeletal poses using generative adversarial networks. *Soft Comput.* **2022**, *26*, 13153–13175. [CrossRef]
22. Natarajan, B.; Rajalakshmi, E.; Elakkiya, R.; Kotecha, K.; Abraham, A.; Gabralla, L.A.; Subramaniyaswamy, V. Development of an end-to-end deep learning framework for sign language recognition, translation, and video generation. *IEEE Access* **2022**, *10*, 104358–104374. [CrossRef]
23. Dorkenwald, M.; Milbich, T.; Blattmann, A.; Rombach, R.; Derpanis, K.G.; Ommer, B. Stochastic image-to-video synthesis using cinns. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3742–3753.
24. Blattmann, A.; Milbich, T.; Dorkenwald, M.; Ommer, B. Understanding object dynamics for interactive image-to-video synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5171–5181.
25. Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; Norouzi, M. Palette: Image-to-image diffusion models. In Proceedings of the ACM SIGGRAPH 2022 Conference Proceedings, Vancouver, BC, Canada, 7–11 August 2022; pp. 1–10.

26. Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. Make-a-video: Text-to-video generation without text-video data. *arXiv* **2022**, arXiv:2209.14792.
27. Wu, J.Z.; Ge, Y.; Wang, X.; Lei, S.W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; Shou, M.Z. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 7623–7633.
28. Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D.P.; Poole, B.; Norouzi, M.; Fleet, D.J.; et al. Imagen video: High definition video generation with diffusion models. *arXiv* **2022**, arXiv:2210.02303.
29. Luo, Z.; Chen, D.; Zhang, Y.; Huang, Y.; Wang, L.; Shen, Y.; Zhao, D.; Zhou, J.; Tan, T. VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 10209–10218.
30. Villegas, R.; Babaeizadeh, M.; Kindermans, P.J.; Moraldo, H.; Zhang, H.; Saffar, M.T.; Castro, S.; Kunze, J.; Erhan, D. Phenaki: Variable length video generation from open domain textual description. *arXiv* **2022**, arXiv:2210.02399.
31. Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S.W.; Fidler, S.; Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 22563–22575.
32. Ge, S.; Nah, S.; Liu, G.; Poon, T.; Tao, A.; Catanzaro, B.; Jacobs, D.; Huang, J.B.; Liu, M.Y.; Balaji, Y. Preserve your own correlation: A noise prior for video diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 22930–22941.
33. Girdhar, R.; Singh, M.; Brown, A.; Duval, Q.; Azadi, S.; Rambhatla, S.S.; Shah, A.; Yin, X.; Parikh, D.; Misra, I. Emu Video: Factorizing Text-to-Video Generation by Explicit Image Conditioning. *arXiv* **2023**, arXiv:2311.10709.
34. Esser, P.; Chiu, J.; Atighehchian, P.; Granskog, J.; Germanidis, A. Structure and content-guided video synthesis with diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 7346–7356.
35. Zhou, D.; Wang, W.; Yan, H.; Lv, W.; Zhu, Y.; Feng, J. Magicvideo: Efficient video generation with latent diffusion models. *arXiv* **2022**, arXiv:2211.11018.
36. Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv* **2023**, arXiv:2311.15127.
37. Wang, W.; Liu, J.; Lin, Z.; Yan, J.; Chen, S.; Low, C.; Hoang, T.; Wu, J.; Liew, J.H.; Yan, H.; et al. MagicVideo-V2: Multi-Stage High-Aesthetic Video Generation. *arXiv* **2024**, arXiv:2401.04468.
38. Karras, J.; Holynski, A.; Wang, T.C.; Kemelmacher-Shlizerman, I. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv* **2023**, arXiv:2304.06025.
39. Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; Fleet, D.J. Video diffusion models. *arXiv* **2022**, arXiv:2204.03458.
40. Islam, T.; Miron, A.; Liu, X.; Li, Y. Svton: Simplified virtual try-on. In Proceedings of the 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), Nassau, Bahamas, 12–14 December 2022; pp. 369–374.
41. Islam, T.; Miron, A.; Liu, X.; Li, Y. StyleVTON: A multi-pose virtual try-on with identity and clothing detail preservation. *Neurocomputing* **2024**, *594*, 127887. [CrossRef]
42. Islam, T.; Miron, A.; Liu, X.; Li, Y. Deep Learning in Virtual Try-On: A Comprehensive Survey. *IEEE Access* **2024**, *12*, 29475–29502. [CrossRef]
43. Islam, T.; Miron, A.; Nandy, M.; Choudrie, J.; Liu, X.; Li, Y. Transforming Digital Marketing with Generative AI. *Computers* **2024**, *13*, 168. [CrossRef]
44. Chen, H.J.; Hui, K.M.; Wang, S.Y.; Tsao, L.W.; Shuai, H.H.; Cheng, W.H. Beautyglow: On-demand makeup transfer framework with reversible generative network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10042–10050.
45. Ma, L.; Jia, X.; Sun, Q.; Schiele, B.; Tuytelaars, T.; Van Gool, L. Pose guided person image generation. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
46. Dong, H.; Liang, X.; Shen, X.; Wang, B.; Lai, H.; Zhu, J.; Hu, Z.; Yin, J. Towards multi-pose guided virtual try-on network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9026–9035.
47. Polanía, L.F.; Gupte, S. Learning fashion compatibility across apparel categories for outfit recommendation. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 4489–4493.
48. McAuley, J.; Targett, C.; Shi, Q.; Van Den Hengel, A. Image-based recommendations on styles and substitutes. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 9–13 August 2015; pp. 43–52.
49. Chakraborty, S.; Hoque, M.S.; Rahman Jeem, N.; Biswas, M.C.; Bardhan, D.; Lobaton, E. Fashion recommendation systems, models and methods: A review. *Informatics* **2021**, *8*, 49. [CrossRef]
50. Dong, H.; Liang, X.; Shen, X.; Wu, B.; Chen, B.C.; Yin, J. Fw-gan: Flow-navigated warping gan for video virtual try-on. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1161–1170.

51. Kuppa, G.; Jong, A.; Liu, X.; Liu, Z.; Moh, T.S. ShineOn: Illuminating design choices for practical video-based virtual clothing try-on. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual Conference, 5–9 January 2021; pp. 191–200.
52. Zhong, X.; Wu, Z.; Tan, T.; Lin, G.; Wu, Q. Mv-ton: Memory-based video virtual try-on network. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 908–916.
53. Jiang, J.; Wang, T.; Yan, H.; Liu, J. Clothformer: Taming video virtual try-on in all module. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10799–10808.
54. Cao, S.; Chai, W.; Hao, S.; Zhang, Y.; Chen, H.; Wang, G. Difffashion: Reference-based fashion design with structure-aware transfer by diffusion models. *arXiv* **2023**, arXiv:2302.06826.
55. Bhunia, A.K.; Khan, S.; Cholakkal, H.; Anwer, R.M.; Laaksonen, J.; Shah, M.; Khan, F.S. Person image synthesis via denoising diffusion model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5968–5976.
56. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Toronto, ON, Canada, 18–22 September 2015; pp. 234–241.
57. Croitoru, F.A.; Hondru, V.; Ionescu, R.T.; Shah, M. Diffusion models in vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10850–10869. [CrossRef] [PubMed]
58. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
59. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
60. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*. [CrossRef]
61. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
62. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
63. Wang, P. GitHub—Lucidrains/Make-a-Video-Pytorch: Implementation of Make-A-Video, New SOTA Text to Video Generator from Meta AI, in Pytorch—github.com. 2022. Available online: https://github.com/lucidrains/make-a-video-pytorch (accessed on 17 March 2024).
64. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
65. Zablotskaia, P.; Siarohin, A.; Zhao, B.; Sigal, L. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv* **2019**, arXiv:1910.09139.
66. Nichol, A.Q.; Dhariwal, P. Improved denoising diffusion probabilistic models. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 19–24 July 2021; pp. 8162–8171.
67. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
68. Liu, W.; Piao, Z.; Min, J.; Luo, W.; Ma, L.; Gao, S. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5904–5913.
69. Unterthiner, T.; van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; Gelly, S. FVD: A New Metric for Video Generation. Available online: https://openreview.net/forum?id=rylgEULtdN (accessed on 6 August 2024)
70. Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; Gelly, S. Towards accurate generative models of video: A new metric & challenges. *arXiv* **2018**, arXiv:1812.01717.
71. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
72. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
73. Davtyan, A. GitHub—Araachie/Frechet_Video_Distance-Pytorch-: Frechet Video Distance Metric Implemented on PyTorch—github.com, 2020. Available online: https://github.com/Araachie/frechet_video_distance-pytorch- (accessed on 18 January 2024).
74. Chang, Y.L.; Liu, Z.Y.; Lee, K.Y.; Hsu, W. Learnable gated temporal shift module for deep video inpainting. *arXiv* **2019**, arXiv:1907.01131.

75. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 5–10 December 2016; Volume 29.
76. Gorelick, L.; Blank, M.; Shechtman, E.; Irani, M.; Basri, R. Actions as space-time shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 2247–2253. [CrossRef]