Text2Performer: Text-Driven Human Video Generation

Yuming Jiang[1] Shuai Yang[1] Tong Liang Koh[1] Wayne Wu[2] Chen Change Loy[1] Ziwei Liu[1]✉
[1]S-Lab, Nanyang Technological University [2]Shanghai AI Laboratory
{yuming002, shuai.yang, koht0029, ccloy, ziwei.liu}@ntu.edu.sg wuwenyan0503@gmail.com

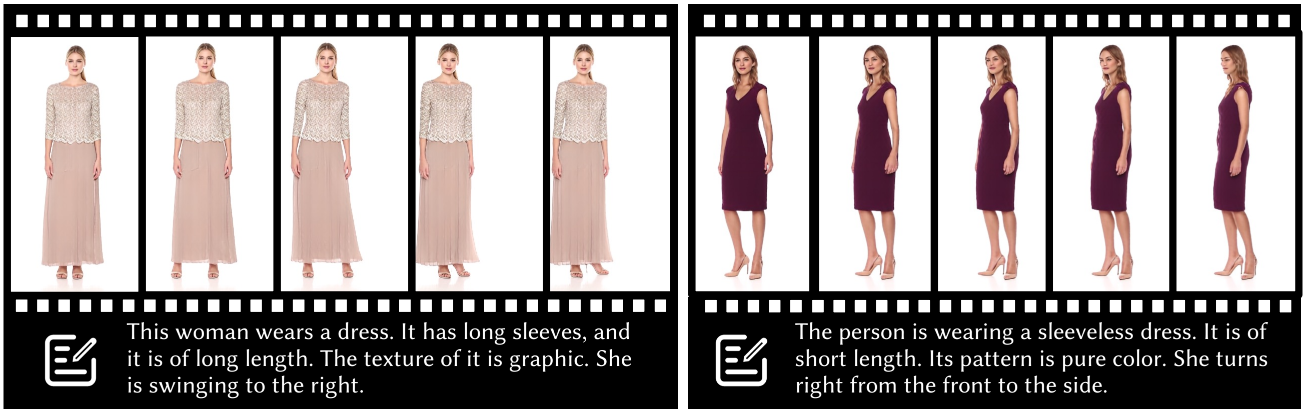

This woman wears a dress. It has long sleeves, and it is of long length. The texture of it is graphic. She is swinging to the right.
The person is wearing a sleeveless dress. It is of short length. Its pattern is pure color. She turns right from the front to the side.

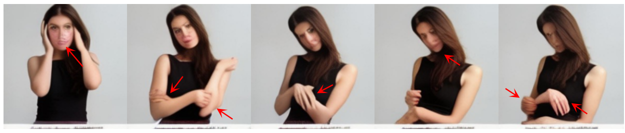
Figure 1: **High-resolution videos generated by *Text2Performer***. The videos are generated by taking the text descriptions as the only input. The generated videos contain diverse appearances and flexible motions. Identities are well maintained.

Abstract

Text-driven content creation has evolved to be a transformative technique that revolutionizes creativity. Here we study the task of text-driven human video generation, where a video sequence is synthesized from texts describing the appearance and motions of a target performer. Compared to general text-driven video generation, human-centric video generation requires maintaining the appearance of synthesized human while performing complex motions. In this work, we present ***Text2Performer*** *to generate vivid human videos with articulated motions from texts. Text2Performer has two novel designs:* ***1)*** *decomposed human representation and* ***2)*** *diffusion-based motion sampler. First, we decompose the VQVAE latent space into human appearance and pose representation in an unsupervised manner by utilizing the nature of human videos. In this way, the appearance is well maintained along the generated frames. Then, we propose* ***continuous VQ-diffuser*** *to sample a sequence of pose embeddings. Unlike existing VQ-based methods that operate in the discrete space, continuous VQ-diffuser directly outputs the continuous pose embeddings for better motion modeling. Finally, motion-aware masking strategy is designed to mask the pose embeddings spatial-temporally to enhance the temporal coherence. Moreover, to facilitate the task of text-driven human video generation, we contribute a Fashion-Text2Video dataset with manually annotated action labels and text descriptions. Extensive experiments demonstrate that Text2Performer generates high-quality human videos (up to 512 × 256 resolution) with diverse appearances and flexible motions. Our project page is* `https://yumingj.github.io/projects/Text2Performer.html`

1. Introduction

Since its emergence, text-guided image synthesis (*e.g.* DALLE [37, 38]) has attracted substantial attention. Recent works [9, 10, 14, 18, 40, 47] have demonstrated fascinating performance for the quality of synthesized images and their consistency with the texts. Beyond image generation, text-driven video generation [22, 24, 44, 53] is an advanced topic to explore. Existing works rely on large-scale datasets to drive large models. Although they have achieved surprising performance on general objects, when applied to the generation of some specific tasks, such as generating videos of garment presents for e-commerce websites, they fail to generate plausible results. Take the CogVideo [24] as an example. As shown in Fig. 2, the generated human objects contain incomplete human structures, and the temporal



The person is wearing a sleeveless dress. It is of short length. Its pattern is pure color. She turns right from the front to the side.

Figure 2: **Results of General Large Text-to-Video Models.** We use the pretrained general large Text-to-Video Models [24] to generate videos using the same texts as the right example of Fig. 1. The result fails to generate complete human structures and maintains temporal consistency.

consistency is poorly maintained. On the other hand, these methods need billions of training data, which hampers the application to those specific tasks (*e.g*. human video generation) without a large amount of paired data.

Therefore, it is worthwhile to explore text-to-video generation in human video generation, which has numerous applications [31, 45]. In this paper, we focus on the task of text-driven human video generation. Compared to general text-to-video generation, text-driven human video generation poses several unique challenges: **1)** The human structure is articulated. The joint movements of different body components form many complicated out-of-plane motions, *e.g*., rotations. **2)** When performing complicated motions, the appearance of the synthesized human should remain the same. For example, the appearance of a target human after turning around should be consistent with that at the first beginning. In sum, to achieve high-fidelity human video generation, consistent human representation and complicated human motions should be carefully modeled.

We propose a novel text-driven human video generation framework **Text2Performer** to handle consistent human representations and complex out-of-plane motions. As shown in Fig. 1, given texts describing appearance and motions, Text2Performer is able to generate temporally consistent human videos with complete human structures and unified human appearances. Text2Performer is built upon VQVAE-based frameworks [4, 11, 18, 52]. In Text2Performer, thanks to the specific nature of human videos which shares the same objects across the frames within one video, VQVAE latent space can be decomposed into appearance and pose representations. With the decomposed VQ-space, videos are generated by sampling appearance representation and a sequence of pose representations separately. This decomposition contributes to the maintenance of human identity. Besides, it makes the motion modeling more tractable, as the motion sampler does not need to take the appearance information into consideration.

To model complicated human motions, a novel continuous VQ-diffuser is proposed to sample a sequence of meaningful pose representations. The architecture of the continuous VQ-diffuser is transformer. The key difference to the previous transformer-based samplers [4, 11, 52] is that the continuous VQ-diffuser directly predicts the continuous pose embeddings rather than their indices in the codebook. After predicting continuous pose embeddings, we also make use of the rich embeddings stored in the codebook by retrieving the nearest embeddings of the predicted embeddings from the codebook. Predicting continuous embeddings alleviates the one-to-many prediction issue in previous discrete methods and the use of codebook constrains the prediction space. With this design, more temporally coherent human motions and more complete structures of human frames are sampled. In addition, we borrow the idea of diffusion models [4, 18, 23, 39, 26] to progressively predict the long sequence of the pose embeddings. We propose a motion-aware masking strategy to sample the pose embeddings of the first frame and last frame firstly. Then the pose embeddings of the intermediate frames are gradually diffused. The motion-aware masking strategy enhances the completeness of human structures and temporal coherence.

To facilitate the task of text-guided human video generation, we propose the Fashion-Text2Video Dataset. It is built upon the FashionDataset [56], which consists of 600 human videos performing the fashion show. We manually segment the whole video into clips and label the motion types. Each clip is performing one motion. With the manually labeled motion labels, we then pair them with text descriptions.

Our contributions are summarized as follows:

- We present and study the task of text-guided human video generation. Our proposed **Text2Performer** can be well trained and have generative abilities with only a small amount of data for training.
- We propose to decompose the VQ-space into appearance and pose representations. The decomposition is achieved by making use of the nature of human videos, *i.e*., the motions are performed under one identity (appearance) across the frames. The decomposition of VQ-space improves the appearance coherence among frames and eases motion modeling.
- We propose the **continuous VQ-diffuser** to predict continuous pose embeddings with the pose codebook. The final continuous pose embeddings are iteratively predicted with the guidance of the motion-aware masking scheme. These designs contribute to generated frames of high quality and temporal coherence.
- We construct the Fashion-Text2Video Dataset with human motion labels and text descriptions to facilitate the research on text-driven human video generation.

2. Related Work

Video Synthesis. Similar to image generation, GAN [5, 16, 29] and VQVAE [11, 52] are two common paradigms in the field of video synthesis [49]. MoCoGAN-HD [48] and

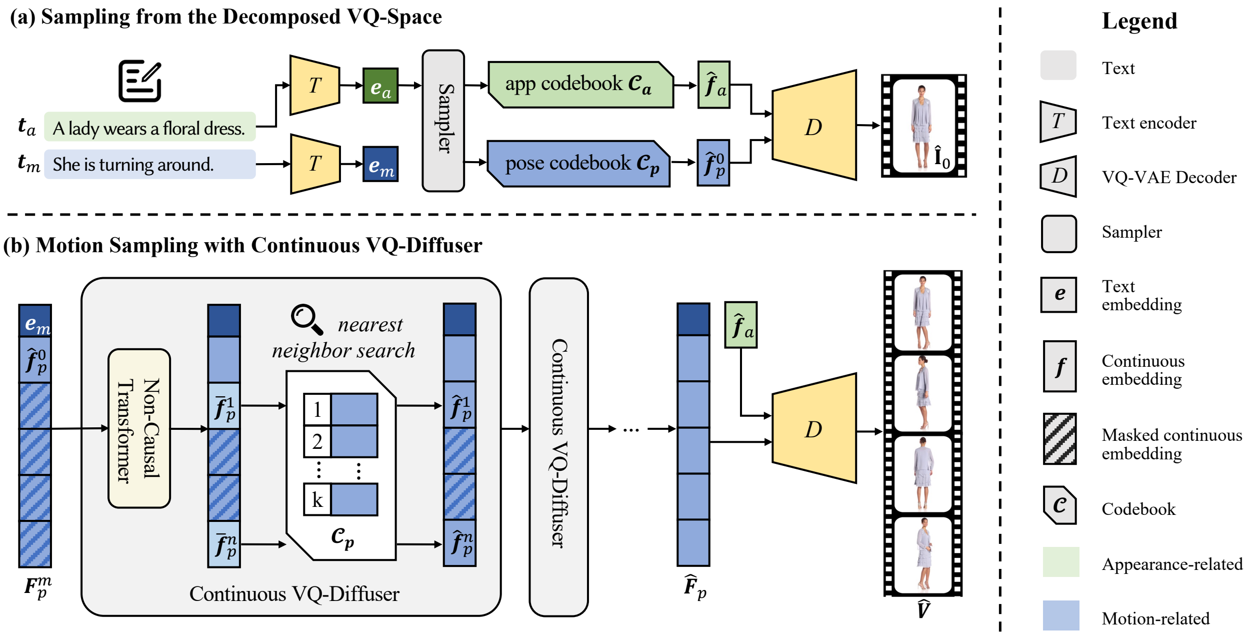


Figure 3: **Overview of Text2Performer** with (a) Sampling from the Decomposed VQ-Space and (b) Motion Sampling with Continuous VQ-Diffuser. Given a text, we first sample the target appearance features $\hat{\boldsymbol{f}}_a$ and exemplar pose features $\hat{\boldsymbol{f}}_p^0$ conditioned on the language features $\boldsymbol{e}_a$ extracted by a pretrained text model. The motion sequence $\hat{\mathbf{F}}_p$ is then sampled by our proposed continuous VQ-Diffuser. The continuous VQ-Diffuser takes the extracted language features $\boldsymbol{e}_m$ and $\hat{\boldsymbol{f}}_p^0$ as inputs. The prediction of continuous motion sequences $\hat{\mathbf{F}}_p$ starts with fully masked pose features and predicts continuous pose embeddings $\{\bar{\boldsymbol{f}}_p^1, ..., \bar{\boldsymbol{f}}_p^n\}$ with Non-Causal Transformer. The nearest neighbor pose embeddings $\{\hat{\boldsymbol{f}}_p^1, ..., \hat{\boldsymbol{f}}_p^n\}$ are then retrieved from the pose codebook $\mathcal{C}_p$. Guided by a motion-aware masking strategy, the continuous VQ-Diffuser is iteratively applied until the whole motion sequence is unmasked. The final videos are generated by feeding the continuous pose features and appearance features into the decoder of VQVAE.

StyleGAN-V [46] harness the powerful StyleGAN [29, 30] to generate videos. DI-GAN [55] models an implicit neural representation to generate videos. Brooks *et al.* [6] propose to regularize the temporal consistency via a temporal discriminator. These works focus on unconditional video generation. As for the VQVAE-based methods, VideoGPT [54] firstly extends the idea of VQVAE and autoregressive transformer to video generation. The following works introduce time-agnostic VQGAN with time-sensitive transformer [15] and masked sequence modeling [20] to further improve the performance. These VQVAE-based video synthesis frameworks accept conditions appended to the beginning of the token sequences. Our proposed Text2Performer is developed on VQVAE. The differences lie in that the designs of VQ-space decomposition and continuous sampling. Recently, there are some concurrent works [22, 24, 44, 53] for text-to-video generation. These methods are designed for general objects, while our Text2Performer has the design to separate human appearance and pose representations, which is specific to human video generation.

Human Content Generation and Manipulation. Existing works focus on human images, and they are mainly divided into two types: unconditional generation and conditional generation. StyleGAN-Human [12] employs StyleGAN to synthesize high-fidelity human images. TryOnGAN [32] and HumanGAN [41] generate human images conditioned on given human poses. Text2Human [27] proposes to generate human images conditioned on texts and poses. We focus on human video generation. For human content manipulation, pose transfer [1, 7, 33, 34, 35] is a popular topic. Pose transfer deals with video data. The task is to transfer the poses from the reference videos to the source video. Chan *et al.* [7] take human poses as inputs and the desired video is generated by an image-to-image translation network. Our task differs in that we only take texts as inputs and exemplar images are not compulsory. Motion synthesis works [19, 36] generate human motions in form of 3D human representations (*e.g.*, SMPL and skeleton space), while our method generates human motions at the image level.

3. Text2Performer

As shown in Fig. 3, our proposed Text2Performer synthesizes desired human videos by taking the texts as inputs ($\boldsymbol{t}_a$ for appearance and $\boldsymbol{t}_m$ for motions). To ensure consistent human representation, we propose to decompose the

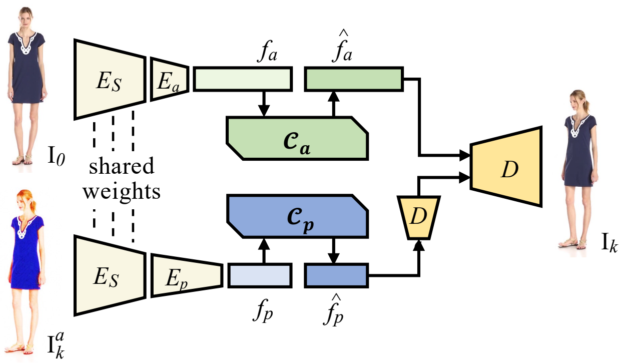


Figure 4: **Pipeline of Decomposed VQ-Space.** Two different frames I_0 and I_k of a video serve as the identity frame and the pose frame to provide appearance and pose features, respectively. We apply data augmentations to I_k to erase its appearance information to avoid information leakage. Two codebooks are built to store the pose features and appearance features. The quantized features are finally fed into the decoder to reconstruct the pose frame I_k.

VQ-space into appearance representation $\boldsymbol{f}_a$ and pose representation $\boldsymbol{f}_p$ as shown in Fig. 4. With the decomposed VQ-space, we sample the human appearance features $\hat{\boldsymbol{f}}_a$ and exemplar pose features $\hat{\boldsymbol{f}}_p^0$ according to $\boldsymbol{t}_a$. To model the motion dynamics, we propose continuous VQ-diffuser as the motion sampler. It progressively diffuses the masked sequence of pose features until the whole sequence is unmasked. The final video clip is generated by feeding the appearance and pose embeddings into the decoder.

3.1. Decomposed VQ-Space

Human video generation requires consistent human appearance, *i.e.*, the face and clothing of the target person should remain unchanged while performing the motion. Vanilla VQVAE [52] encodes images into a unified feature representation and builds a codebook on it. However, such design is prone to generate drifted identities along the video frames. To overcome this problem, we propose to decompose the VQ-space into the appearance representation $\boldsymbol{f}_a$ and pose representation $\boldsymbol{f}_p$ in an unsupervised manner. With the decomposed VQ-space, we can separately sample $\hat{\boldsymbol{f}}_a$ and a sequence of $\hat{\boldsymbol{f}}_p$ to generate the desired video.

In human video data, different frames in one video share the same human identity. We utilize this property to train the decomposed VQ-space [13, 25]. The pipeline is shown in Fig. 4. Given a video $\mathbf{V} = \{\mathbf{I}_0, \mathbf{I}_1, ..., \mathbf{I}_n\}$, we use its first frame $\mathbf{I}_0$ for appearance information and another randomly sampled frame $\mathbf{I}_k$ for pose information. $\mathbf{I}_0$ and $\mathbf{I}_k$ are fed into two encoders E_a and E_p, respectively. The two branches share the encoder E_s. To prevent $\mathbf{I}_k$ from leaking appearance information, data augmentations (*e.g.* color jittering and Gaussian blur) are applied to $\mathbf{I}_k$ before the pose branch. $\boldsymbol{f}_a$ and $\boldsymbol{f}_p$ are obtained as

$$\boldsymbol{f}_a = E_a(E_s(\mathbf{I}_0)), \boldsymbol{f}_p = E_p(E_s(\mathbf{I}_k^a))), \tag{1}$$

where $\mathbf{I}_k^a$ is the augmented $\mathbf{I}_k$. To make $\boldsymbol{f}_p$ learn compact and necessary pose information, we make the spatial size of $\boldsymbol{f}_p$ smaller. Given $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, we empirically find that the optimal spatial sizes for $\boldsymbol{f}_a$ and $\boldsymbol{f}_p$ are $H/16 \times W/16$ and $H/64 \times W/64$, respectively.

After obtaining $\boldsymbol{f}_a$ and $\boldsymbol{f}_p$, two codebooks $\mathcal{C}_a$ and $\mathcal{C}_p$ are built to store the appearance and pose embeddings. The quantized feature $\hat{\boldsymbol{f}}$ given codebook $\mathcal{C}$ is obtained by

$$\hat{\boldsymbol{f}} = Quant(\boldsymbol{f}) := \{\underset{c_k \in \mathcal{C}}{\text{argmin}} \left\| f_{ij} - c_k \right\|_2 \, | f_{ij} \in \boldsymbol{f}\}. \tag{2}$$

With the quantized $\hat{\boldsymbol{f}}_a$ and $\hat{\boldsymbol{f}}_p$, we then feed them into decoder D to reconstruct the target pose frame $\hat{\mathbf{I}}_k$:

$$\hat{\mathbf{I}}_k = D([\hat{\boldsymbol{f}}_a, D_p(\hat{\boldsymbol{f}}_p)]), \tag{3}$$

where $[\cdot]$ is the concatenation operation, and D_p upsamples $\hat{\boldsymbol{f}}_p$ to make it have the same resolution as $\hat{\boldsymbol{f}}_a$. The whole network (including the encoders, decoders, and codebooks) is trained using:

$$\begin{aligned} \mathcal{L} = \left\| \mathbf{I}_k - \hat{\mathbf{I}}_k \right\|_1 + \left\| \text{sg}(\hat{\boldsymbol{f}}_a) - \boldsymbol{f}_a \right\|_2^2 + \left\| \text{sg}(\boldsymbol{f}_a) - \hat{\boldsymbol{f}}_a \right\|_2^2 \\ + \left\| \text{sg}(\hat{\boldsymbol{f}}_p) - \boldsymbol{f}_p \right\|_2^2 + \left\| \text{sg}(\boldsymbol{f}_p) - \hat{\boldsymbol{f}}_p \right\|_2^2, \end{aligned} \tag{4}$$

where sg$(\cdot)$ is the stop-gradient operation.

With the decomposed VQ-space, an additional sampler is trained to sample $\hat{\boldsymbol{f}}_a$ and $\hat{\boldsymbol{f}}_p^0$. This sampler has the same design as previous methods [4, 27].

3.2. Continuous VQ-Diffuser

We adopt the absorbing diffusion transformer [2, 4, 8, 18] to sample the motion, a sequence of pose embeddings $\hat{\mathbf{F}}_p = \{\hat{\boldsymbol{f}}_p^1, \hat{\boldsymbol{f}}_p^2, ..., \hat{\boldsymbol{f}}_p^n\}$ from the learned pose codebook $\mathcal{C}_p$. Different from autoregressive models [11, 52] that make predictions in a fixed order, absorbing diffusion transformer predicts multiple codebook indices in parallel. The prediction of codebook indices starts from $\mathbf{F}_p^0$, *i.e.*, fully masked $\mathbf{F}_p^m$. The prediction at time step t is represented as follows:

$$\hat{\mathbf{F}}_p^t \sim q_\theta(\mathbf{F}_p^t | \mathbf{F}_p^{t-1}), \tag{5}$$

where θ denotes the parameters of the transformer sampler.

To model human motions, we propose **1)** continuous space sampling, and **2)** motion-aware masking strategy.

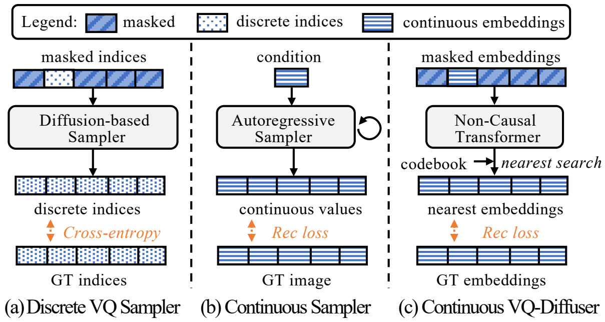


Figure 5: **Comparison with Image Samplers.** (a) Discrete VQ Sampler predicts discrete codebook indices, which are trained by cross-entropy loss. (b) Continuous Sampler operates on unconstrained continuous RGB spaces. (c) Our VQ-Diffuser predicts continuous embeddings but also utilizes rich embeddings stored in codebooks.

3.2.1 Continuous Space Sampling

As shown in Fig. 5(a), previous VQVAE-based methods [4, 8, 11, 18, 52] sample tokens in a discrete space. They predict codebook indices to form quantized features, which are then fed into the decoder to generate images. However, sampling at the discrete space is hard to converge because of redundant codebooks, which causes the one-to-many mapping prediction problem. This would lead to meaningless predicted pose sequences.

In continuous VQ-diffuser S_θ, we propose to train the non-causal transformer in the continuous embedding space as shown in Fig. 5(c). S_θ directly predicts continuous pose embeddings $\{\bar{\boldsymbol{f}}_p^k\}$. S_θ is trained to predict the full continuous $\bar{\mathbf{F}}_p$ from the masked pose sequences $\mathbf{F}_p^m$ conditioned on the motion description $\boldsymbol{t}_m$ and the initial pose feature $\hat{\boldsymbol{f}}_p^0$:

$$\bar{\mathbf{F}}_p = S_\theta([T(\boldsymbol{t}_m), \hat{\boldsymbol{f}}_p^0, \mathbf{F}_p^m]), \tag{6}$$

where $T(\cdot)$ is the pretrained text feature extractor.

To utilize rich embeddings stored in $\mathcal{C}_p$ to constrain the prediction space, we retrieve the nearest embedding of the predicted continuous $\bar{\mathbf{F}}_p$ from $\mathcal{C}_p$ to obtain the final $\hat{\mathbf{F}}_p$:

$$\hat{\mathbf{F}}_p = \text{Nearest}(\bar{\mathbf{F}}_p) := \{\underset{c_k \in \mathcal{C}_p}{\text{argmin}} \left\|\bar{\boldsymbol{f}}_p - c_k\right\|_2\}. \tag{7}$$

$\hat{\mathbf{F}}_p$ is then fed into D to reconstruct the final video $\hat{\mathbf{V}}$ following Eq. (3). Thanks to the continuous operation, we can add losses at both the image level and embedding level. The loss function to train our continuous VQ-diffuser is

$$\begin{aligned} \mathcal{L} = \left\|\bar{\mathbf{F}}_p - \mathbf{F}_p\right\|_1 + \left\|\text{sg}(\hat{\mathbf{F}}_p) - \bar{\mathbf{F}}_p\right\|_2^2 \\ + \left\|\text{sg}(\bar{\mathbf{F}}_p) - \hat{\mathbf{F}}_p\right\|_2^2 + L_{\text{rec}}(\hat{\mathbf{V}}, \mathbf{V}), \end{aligned} \tag{8}$$

where L_{rec} is composed of L_1 loss and perceptual loss [28]. It should be noted that stop-gradient operation is applied to make the nearest operation differentiable.

Compared to continuous samplers like PixelCNN [51] in Fig. 5(b), which directly predict the continuous RGB values, our continuous VQ-diffuser fully utilizes the rich contents stored in the codebook while restricting the prediction space. Therefore, it inherits the benefits of the discrete VQ sampler and continuous sampler at the same time. The pose sequences sampled by our continuous VQ-diffuser are temporally coherent and meaningful.

3.2.2 Motion-Aware Masking Strategy

To generate plausible videos, the sampled motion sequence $\hat{\mathbf{F}}_p$ should be both temporally and spatially reasonable. In order to make the continuous VQ-diffuser S_θ correctly condition on $\boldsymbol{t}_m$ as well as generate reasonable human poses for each frame, we design a motion-aware masking strategy to meticulously mask $\mathbf{F}_p$ temporally and spatially.

At the temporal dimension, S_θ first predicts pose embeddings of the first frame and the last frame based on $\boldsymbol{t}_m$ and $\hat{\boldsymbol{f}}_p^0$. Then the prediction diffuses to the intermediate frames according to the given conditions and previously unmasked frames. Therefore, during the training, if the first frame and the last frame are masked, we will mask all frames to prevent the intermediate frames from providing information to help frame predictions at two ends. A higher probability is assigned to masking all frames to help S_θ better learn prediction at the most challenging two ends.

At the spatial dimension, we propose to predict the first and the last frames with more than one diffusion step. Otherwise, the structure of the sampled human tends to be incomplete. To this end, during the training, we mask the first and last frames partially at the spatial dimension.

4. Fashion-Text2Video Dataset

Existing human video datasets are either of low resolution [3, 17, 49] or lack of text (action labels) [35, 43]. To support the study of text-driven human video generation, we propose a new dataset, Fashion-Text2Video dataset. Among existing human video datasets, we select videos in Fashion Dataset [56] as the source data for following reasons: 1) High resolution ($\geq 1024 \times 1024$); 2) Diverse appearance.

The Fashion Dataset contains a total of 600 videos. For each video, we manually annotate motion labels, clothing shapes (length of sleeves and length of clothing) and textures. Motion labels are classified into 22 classes, including standing, moving right-side, moving left-side, turning around, *etc*. The lengths of sleeves are classified into no-sleeve, three-point, medium and long sleeves. The length of clothing is the length of the upper part of the clothes.

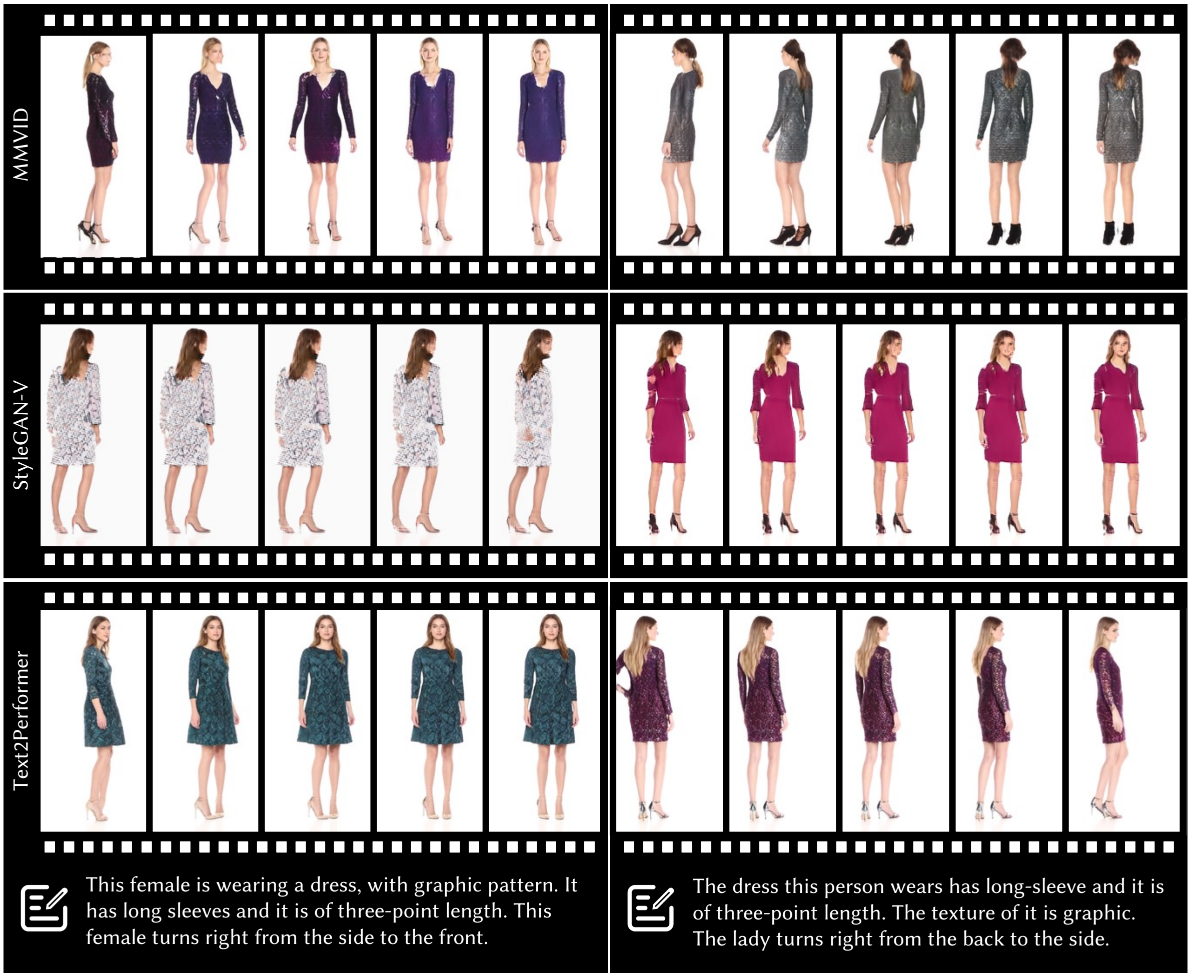

MMVID
StyleGAN-V
Text2Performer
This female is wearing a dress, with graphic pattern. It has long sleeves and it is of three-point length. This female turns right from the side to the front.
The dress this person wears has long-sleeve and it is of three-point length. The texture of it is graphic. The lady turns right from the back to the side.

Figure 6: **Qualitative Comparisons.** Text2Performer achieves superior generation qualities compared with baselines.

Clothing textures are mainly divided into pure color, floral, graphic, *etc*. Upon the above labels, we then generate texts for each video using some pre-defined templates. The texts include descriptions of the appearances of the dressed clothes and descriptions for each sub-motions. The proposed Fashion-Text2Video dataset can be applied to the research on Text2Video and Label2Video generation.

5. Experiments

5.1. Comparison Methods

MMVID [20] is a VQVAE-based method for multi-modal video synthesis. We follow the original setting to train MMVID on our dataset. **StyleGAN-V** [46] is a state-of-the-art video generation method built on StyleGAN. Action labels are used as conditions instead of texts, which we find significantly surpasses the performance of texts for this method. **CogVideo** [24] is a large-scale pretrained model for Text-to-Video generation. We implement two versions of CogVideo, *i.e*., CogVideo-v1 and CogVideo-v2. Since the original CogVideo utilized the pretrained Text2Image dataset, we train CogVideo-v1 to use the exemplar image as an additional input to compensate for the lack of large-scale pretrained Text2Image models for human images. CogVideo-v2 is with original designs.

5.2. Evaluation Metrics

Diversity and Quality. 1) We adopt Fréchet Inception Distance (FID) [21] to evaluate the quality of generated human frames. **2)** We evaluate the quality of generated videos using Fréchet Video Distance (FVD) and Kernel Video Distance (KVD) [50]. We generate $2,048$ video clips to compute the metrics, with 20 frames for each clip.

Identity Preservation. We use two metrics to evaluate the maintenance of identity along the generated frames. **1)** We extract face features of each generated frames using FaceNet [42]. For each generated video clip, we calculate the l_2 distances between features of the frames and the first frame. **2)** We extract features of each generated frames using an off-the-shelf ReID model, *i.e*., OSNet [42]. These

Table 1: **Quantitative Comparisons.**

Method	**FID ↓**	**FVD ↓**	**KVD ↓**	**Face ↓**	**ReID ↑**
MMVID [20]	11.85	303.02	78.67	0.9047	0.9096
StyleGAN-V [46]	29.68	219.63	18.77	0.8675	**0.9568**
CogVideo-v1 [24]	39.47	645.03	89.53	1.0564	0.8148
CogVideo-v2 [24]	51.76	799.80	112.24	1.0621	0.7960
Text2Performer	**9.60**	**124.78**	**17.96**	**0.8593**	0.9382

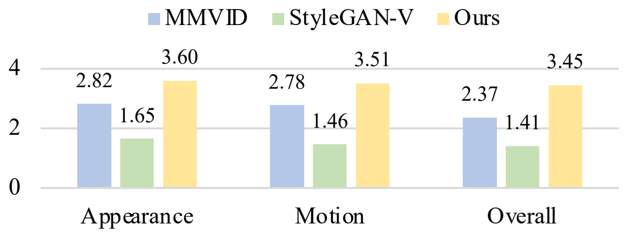


Figure 7: **User Study.** We achieve the highest scores.

features are related to identity features. We calculate the average cosine distances between extracted features of the frames and the first frame. The final metrics are obtained by averaging distances among $2,048$ generated video clips.

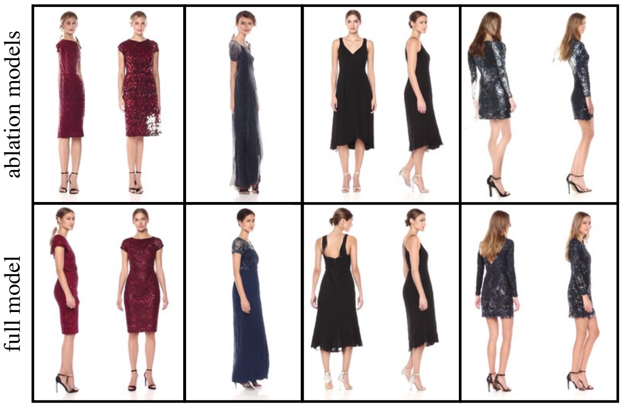
User Study. Users are asked to give three separate scores (the highest score is 4) at three dimensions: 1) The consistency with the text for appearance, 2) The consistency with the text for desired motion, and 3) The overall quality of the generated video in terms of temporal coherence and its realism. A total of 20 users participate in the user study. Each user is presented with 30 videos generated by different methods.

5.3. Qualitative Comparisons

We show two examples in Fig. 6. In the first example, MMVID and Text2Performer successfully generate desired human motions. However, frames generated by MMVID have drifted identities. StyleGAN-V fails to generate plausible frames as well as consistent motions. As for the second example, only Text2Performer generates corresponding motions. Artifacts appear in the video clips generated by MMVID and StyleGAN-V. The visual comparisons demonstrate the superiority of Text2Performer.

5.4. Quantitative Comparisons

The quantitative comparisons are shown in Table 1. In terms of FID, FVD and KVD, our proposed Text2Performer has significant improvements over baselines, which demonstrates the diversity and temporal coherence of the videos generated by our methods. Existing baselines generate videos from entangled human representations and thus suffer from temporal incoherence. As for the identity metrics, our method achieves the best performance on face distance and comparable performance on ReID distance with StyleGAN-V. Since StyleGAN-V is prone to generate videos with small motions, features tend to be similar, leading to high ReID scores. In our method, thanks to the decomposed VQ-space and continuous VQ-diffuser, temporal and identity consistency can be better achieved. The result of the user study is shown in Fig. 7. Text2Performer achieves the highest scores on all three dimensions. Specifically, the result of the assessment on overall quality is consistent with the other quantitative metrics, which further validates the effectiveness of Text2Performer.



(a) decomp. (b) res. (c) discrete (d) codebook

Figure 8: **Ablation Studies** on (a) Decomposed VQ-Space, (b) Spatial Resolution of Pose Branch, (c) Discrete VQ sampler, and (d) Usage of Codebook.

5.5. Ablation Study

Decomposed VQ-Space. The ablation model trains the continuous VQ-diffuser with a unified VQ-space, which fuses the identity and pose information within one feature. As shown in Fig. 8(a), compared to our results, sampling in the unified VQ-Space generates drifted identity and incomplete body structures. This leads to inferior quantitative metrics in Table 2. The unified VQ-Space requires the sampler to handle identity and pose at the same time, which poses burdens for the sampler. With the decomposed VQ-Space, the identity can be easily maintained by fixing the appearance features, which further helps learn motions.

Spatial Resolution of Pose Branch. In our design, the pose feature has 4× smaller spatial resolution than the appearance feature. In Fig. 8(b), we show the results generated by a decomposed VQVAE where pose and appearance features have the same resolution. Compared to results of the full model, body structures in the side view are incomplete. The quantitative metrics in Table 2 and Fig. 8(b) demonstrate that the smaller pose feature is necessary as the larger feature contains redundant information to harm valid feature disentanglement, thus adding challenges to the sampler.

Discrete VQ sampler in Decomposed VQ-Space. We train a variant of VQ-Diffuser to predict discrete codebook indices with the decomposed VQ-Space. In Fig. 8(c), the

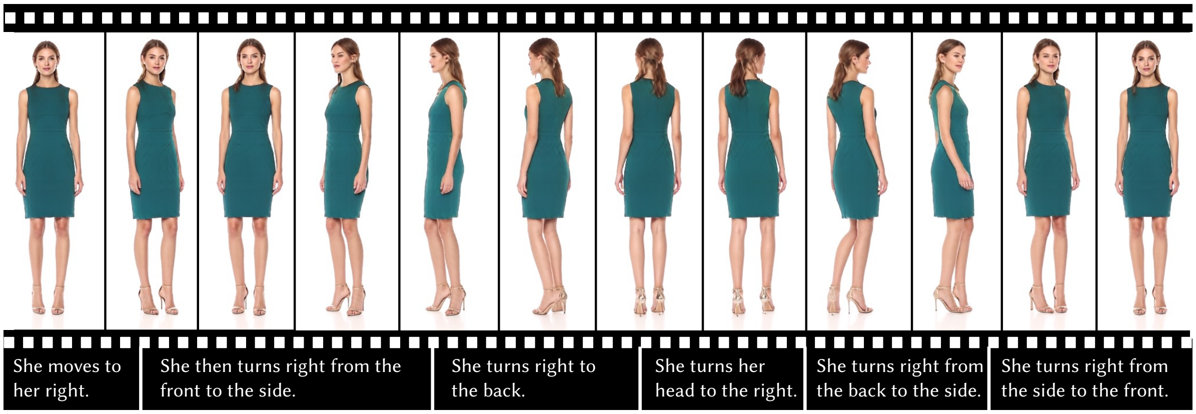

She moves to her right.
She then turns right from the front to the side.
She turns right to the back.
She turns her head to the right.
She turns right from the back to the side.
She turns right from the side to the front.

Figure 9: **High-Resolution Results.** We show one result generated by a long text sequence. The identity is well maintained.

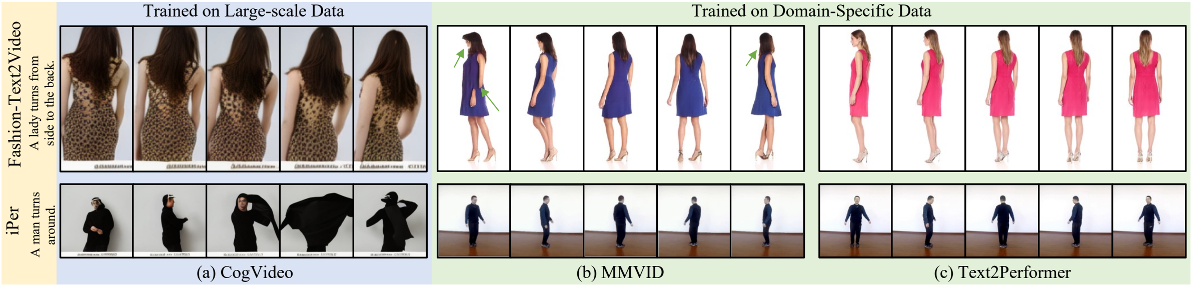

Trained on Large-scale Data
Trained on Domain-Specific Data
Fashion-Text2Video
A lady turns from side to the back.
iPer
A man turns around.
(a) CogVideo
(b) MMVID
(c) Text2Performer

Figure 10: **Cross-Dataset Performance.** We compare with models trained on large-scale data and domain-specific data.

Table 2: **Quantitative Results on Ablation Studies.**

Method	FID ↓	FVD ↓	KVD ↓	Face ↓	ReID ↑
w/o decomp.	20.52	266.97	35.52	0.8936	0.9213
same res.	15.69	202.06	32.96	0.8649	0.9399
disc. sampler	**10.37**	308.88	73.17	0.9106	0.9161
w/o codebook	11.31	154.93	22.80	0.9437	0.9109
Full Model	10.92	**134.43**	**20.31**	**0.8520**	**0.9421**

discrete VQ Sampler can generate complete body structures but fails to generate video clips consistent with the given action control, *i.e.*, turning right from the back to the side. Failure on conditioning on action controls does not harm the quality of each generated frame, and thus this ablation study achieves a comparable FID score as our method. However, our method significantly outperforms this variant on other metrics as shown in Table 2. In VQVAE, some codebook entries have similar contents but with different codebook indices. Therefore, treating the training of the sampler as a classification task makes the model hard to converge.

Usage of Codebook in Continuous VQ-Space. In our design, after the prediction of embeddings, the nearest neighbor of embeddings will be retrieved from the codebook. We train an auto-encoder without codebook in the pose branch. The sampler is trained to sample continuous embeddings from the latent space of auto-encoder without the retrieval from codebook. As shown in Fig. 8(d), without the codebook, the sampler results in a disrupted human image at the side view. The quality of generated videos deteriorates as reflected in Table 2. This demonstrates that predicting compact and meaningful continuous embeddings with codebooks enhances the quality of synthesized frames.

5.6. Additional Analysis

High-Resolution Results. Text2Performer can generate long videos with high resolutions (512 × 256) as shown in Fig. 1 and Fig. 9. Figure 9 shows an example of high-resolution videos for a long text sequence.

Cross-Dataset Performance. We further verify the performance of models trained on different datasets. The results are presented in Fig. 10. Compared with CogVideo [24] pretrained on large-scale dataset, our Text2Performer and MMVID succeed to generate plausible specific human content. As for models trained on small-scale domain-specific data, our Text2Performer is comparable to MMVID on the simple iPer dataset [35], and has superior performance on the more challenging Fashion-Text2Video dataset, especially on the completeness of body structures across frames.

Ability to Generate Novel Appearance. To verify that the generated humans are with novel appearances, we retrieve top-2 nearest neighbour images of generated frames using the perceptual distance [28]. As shown in Fig. 11, the generated query frames have different appearances from the top-2 nearest images retrieved from the dataset.

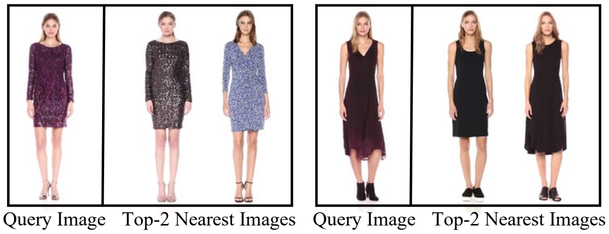


Figure 11: **Nearest Neighbour Image Search.** Top-2 nearest images of generated images are retrieved.

Figure 12: **Results on In-the-Wild Dataset.** In our current pipeline, We treat the background as a part of appearance information. The results contain some artifacts.

6. Discussions

We propose a novel Text2Performer for text-driven human video generation. We decompose the VQ-space into appearance and pose representations. Continuous VQ-diffuser is then introduced to sample motions conditioned on texts. These designs empower Text2Perfomer to synthesize high-quality and high-resolution human videos.

Limitations and Future Work. In this paper, we focus on the generation of human appearance and motion with clean backgrounds. The current pipeline is designed for generating human videos without textural backgrounds. It can be extended to generate plausible textual backgrounds by treating the background as a part of appearance information at the training time. The generated results are shown in Fig. 12. There are some artifacts in the generated videos. Generating videos with textual backgrounds is important in real-world applications. In future work, textural backgrounds can be segmented first. The results can be then improved by generating the foreground and textural background separately and fusing them harmoniously using an additional module. In addition, the synthesized human videos are biased toward generating females with dresses. This is because the Fashion Dataset only contains videos of females with dresses. In future applications, more data can be involved in the training to alleviate the bias. Besides, the control modality can also be extended to various types, such as audio and music.

Potential Negative Societal Impacts. The model may be applied to generate fake videos performing various actions. To alleviate the negative impacts, DeepFake detection methods can be applied to evaluate the realism of videos.

Acknowledgement. This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). This study is also supported by NTU NAP, MOE AcRF Tier 1 (2021-T1-001-088), and the Google PhD Fellowship.

References

[1] Badour Albahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Transactions on Graphics (TOG)*, 40(6):1–11, 2021. 3

[2] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *NeurIPS*, 34:17981–17993, 2021. 4

[3] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, pages 8340–8348, 2018. 5

[4] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P Breckon, and Chris G Willcocks. Unleashing transformers: parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *ECCV*, pages 170–188. Springer, 2022. 2, 4, 5

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *ICLR*, 2019. 2

[6] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *NeurIPS*, 35:31769–31781, 2022. 3

[7] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ICCV*, pages 5933–5942, 2019. 3

[8] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, pages 11315–11325, 2022. 4, 5

[9] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *NeurIPS*, 34:19822–19835, 2021. 1

[10] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *NeurIPS*, 35:16890–16902, 2022. 1

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 2, 4, 5

[12] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *ECCV*, pages 1–19. Springer, 2022. 3

[13] Aviv Gabbay and Yedid Hoshen. Demystifying inter-class disentanglement. *ICLR*, 2020. 4

[14] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *ECCV*, pages 89–106. Springer, 2022. 1

[15] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *ECCV*, pages 102–118. Springer, 2022. 3

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[17] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence*, 29(12):2247–2253, 2007. 5

[18] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, pages 10696–10706, 2022. 1, 2, 4, 5

[19] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 3

[20] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In *CVPR*, pages 3615–3625, 2022. 3, 6, 7

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017. 6

[22] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1, 3

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2

[24] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *ICLR*, 2023. 1, 2, 3, 6, 7, 8

[25] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. *NeurIPS*, 31, 2018. 4

[26] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *ICML*, pages 9902–9915. PMLR, 2022. 2

[27] Yuming Jiang, Shuai Yang, Haonan Qju, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. 3, 4

[28] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016. 5, 8

[29] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2, 3

[30] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR.*, pages 8110–8119, 2020. 3

[31] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. *NeurIPS*, 32, 2019. 2

[32] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG)*, 40(4):1–10, 2021. 3

[33] Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 05 2020. 3

[34] Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural rendering and reenactment of human actor videos. *ACM Transactions on Graphics (TOG)*, 38(5):1–14, 2019. 3

[35] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *ICCV*, pages 5904–5913, 2019. 3, 5, 8

[36] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *ICCV*, pages 10985–10995, 2021. 3

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831. PMLR, 2021. 1

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35:36479–36494, 2022. 1

[41] Kripasindhu Sarkar, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Humangan: A generative model of human images. In *2021 International Conference on 3D Vision (3DV)*, pages 258–267. IEEE, 2021. 3

[42] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 6

[43] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *NeurIPS*, 32, 2019. 5

[44] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *ICLR*, 2023. 1, 3

[45] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *CVPR*, pages 11050–11059, 2022. 2

[46] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *CVPR*, pages 3626–3636, 2022. 3, 6, 7

[47] Jianxin Sun, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun. Anyface: Free-style text-to-face synthesis and manipulation. In *CVPR*, pages 18687–18696, 2022. 1

[48] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. *ICLR*, 2021. 2

[49] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, pages 1526–1535, 2018. 2, 5

[50] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

[51] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *NeurIPS*, 29, 2016. 5

[52] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 30, 2017. 2, 4, 5

[53] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *ICLR*, 2023. 1, 3

[54] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 3

[55] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. *ICLR*, 2022. 3

[56] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019. 2, 5