Text2Performer: Text-Driven Human Video Generation Supplementary File

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

In this supplementary file, we will explain our design of the proposed continuous VQ-diffsuer for handling interpolation and sampling within one model in Section A. Then we will introduce implementation details in Section B. In Section C, we will give more detailed explanations on evaluation metrics. In Section D, we provide more analysis on motion-aware masking strategy. In Section E, we provide more qualitative and quantitative comparisons with baseline methods. Then we provide more visual results in Section F. In Section G, we will discuss the limitations of this work. Finally, we will show a supplementary video.

A. Interpolation and Sampling within One Model in Continuous Diffuser

During the training of the continuous VQ-diffuser S_θ, video clips are normalized to a fixed number of video frames. The normalization guarantees that the first frame and the last frame are the exact initial and ending poses corresponding to the given motion text $\boldsymbol{t}_m$. However, videos generated in this way are not temporally smooth since the normalization operation extracts unconsecutive frames from the original video clips.

To further improve the temporal consistency and support scalable frame rates, we train S_θ to perform video interpolation as well. Specifically, for the interpolation task, we set the text description $\boldsymbol{t}_m$ as a predefined text "empty" and randomly mask the pose embeddings of intermediate frames. In this mode, the pose embeddings at the two ends (*i.e.*, the first frame and the last frame) are always unmasked as the predictions of the intermediate frames are supposed to be conditioned on the neighbouring unmasked frames. The interpolation task is trained at a lower probability than the sampling task since the interpolation task is simpler compared to the generation task.

The training video clips are mixed with normalized video clips and original video clips. When normalized video clips are fed into the framework, the corresponding text descriptions are $\boldsymbol{t}_m$. If the original video clips are provided, the text description $\boldsymbol{t}_m$ will be replaced with "empty" in our setting.

B. Implementation Details

Our main experiments are conducted on our proposed Fashion-Text2Video dataset. The Fashion-Text2Video dataset contains 600 videos in total. We exclude 5 videos from training. The excluded videos are used for visualizing the performance of trained decomposed VQVAE and VQ-Sampler on unseen data. Models are trained with 4 Tesla V100 GPUs. Our main experiments are conducted on videos with resolution of 256×128. The decomposed VQVAE is trained with batch size of 64 for $145,000$ iterations. To improve the generation quality of decomposed VQVAE, we also include the data from SHHQ [1] to train the VQVAE. The sampler for exemplar pose and appearance is trained with batch size of 4. The texts for each exemplar image are generated on-the-fly for generalizability. The continuous VQ-Diffuser is trained with batch size of 4. The probability of training continuous VQ-Diffuser with interpolation mode is set to 0.2. The probability of masking all temporal frames is set to 0.375. The diffusion steps for sampling the frames at the two ends are set to 6 at the spatial dimension.

C. Further Explanations on Evaluation Metrics

Diversity and Quality. As described in Section 5.2 of our main paper, we adopt FID [3], FVD and KVD [8] as the metrics to indicate the diversity and quality of synthesized videos. We extract $2,048$ video clips from the original dataset as the real videos. Each video clip has 20 frames. We then use the text descriptions corresponding to the extracted $2,048$ video clips to

(a) w/o masking strategy (b) w/ masking strategy

Figure A1: **Ablation Studies on Motion-aware Masking Strategy.**

Table A1: **Further Analysis on Masking Strategy**

Method	FID ↓	FVD ↓	KVD ↓	Face ↓	ReID ↑
w/o masking	14.31	**122.93**	22.50	0.7617	0.9494
Full Model	**14.19**	125.38	**22.27**	**0.7562**	**0.9522**

synthesize videos. To calculate FID, we use the all frames as the images, *i.e.*, we calculate the FID values on $40,960$ images. For FVD and KVD, we directly compute the values on $2,048$ video clips.

Identity Preservation. For face scores [6], we first detect the faces and then extract the features for the cropped faces. Since there are frames without detectable faces (the target person rotates to the back view), we compute the distances between the frames with detectable faces and the first frame with detectable faces in one video clip. We then use the average of these distances as the final score. For ReID scores, features are extracted on the original generated frames [6], and the scores are computed on the extracted features.

D. Further Analysis on Motion-Aware Masking Strategy

We conduct further analysis on the motion-aware masking strategy. As shown in Table A1, after adopting the motion-aware masking strategy, the performance improves slightly. Using motion-aware masking strategy achieves superior FID, KVD, Face scores and ReID scores. We further demonstrate the effectiveness of motion-aware masking strategy in Fig. A1 as the improvements cannot be well reflected in the quantitative results. As shown in Fig. A1, the use of motion-aware masking strategy enhances of the completeness of the body structures.

E. More Comparisons

In Table A2, we present more quantitative comparisons. Apart from our proposed Fashion Text2Video dataset, we also compare with MMVID [2] on iPer dataset [5]. We use the original implementation of MMVID and retrain the method as no pretrained model is provided for this dataset. As shown in Table A2, our method achieves a better FID score as we synthesize frames with higher quality. We also achieve a comparable FVD score with MMVID. We attribute the value gap of FVD to the nature of FVD metrics. We find that the FVD score focuses more on the motion than the appearance consistency on this dataset. Since the resolution is relatively low on this dataset, some appearance flickering artifacts generated by MMVID are deemed as motions, and thus result in a slightly lower FVD score. Qualitative comparisons can be found in Fig. A2.

F. More Qualitative Results

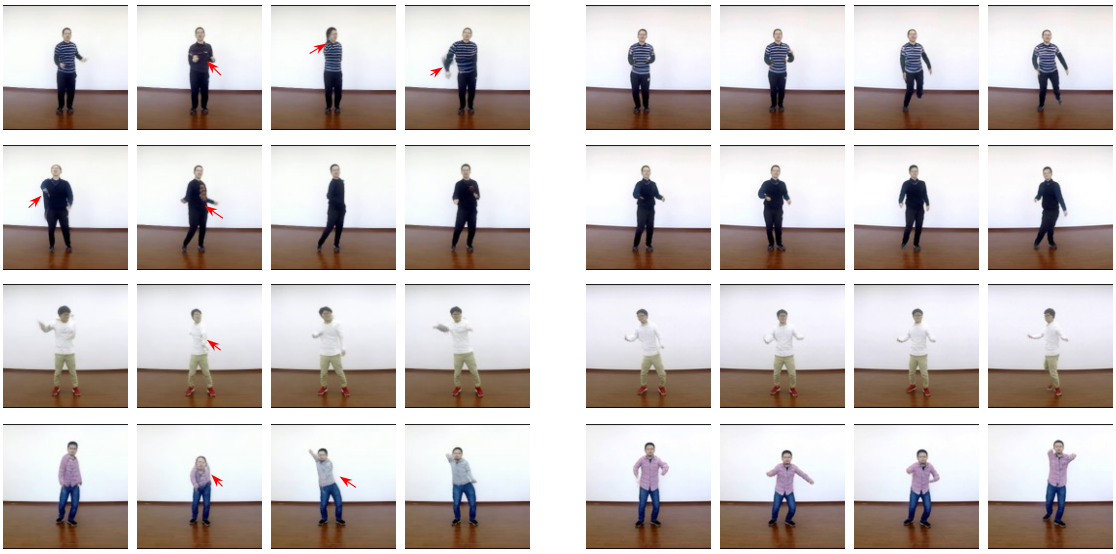
We show more results in Figs. A3-A5. Figure A3 shows the video clips with resolution of 256×128. Video clips in Figs. A4-A5 are with the resolution of 512×256. We also show the results on the iPerDataset in Fig. A6, where we follow the same setting as MMVID [2] to train and generate videos.

G. Limitations

Text2Performer is trained on videos with relatively clean background. In future works, more designs should be introduced to handle the complex background. In addition, the synthesized human videos are biased toward generating females with dresses. This is because the original FashionDataset [9] only contains videos of females with dresses. In future works, more data can be involved in the training to alleviate the issue caused by the dataset bias.

Table A2: **More Quantitative Comparisons.**

Dataset	Method	FID ↓	FVD ↓
Fashion-Text2Video	StyleGAN-V [7]	29.68	219.63
	CogVideo-v1 [4]	39.47	645.03
	CogVideo-v2 [4]	51.76	799.80
	MMVID [2]	11.85	303.02
	Text2Performer	**9.60**	**124.78**
iPer [5]	MMVID [2]	30.80	**256.72**
	Text2Performer	**15.47**	289.69



MMVID Ours

Figure A2: **Qualitative Comparisons with MMVID [2] on iPer Dataset [5].**

H. Supplementary Video

The supplementary video can be found in https://youtu.be/YwhaJUk_qo0.

References

[1] 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. 1

[2] 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. 2, 3

[3] 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. 1

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

[5] 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. 2, 3, 6

[6] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 2

[7] 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

[8] 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. 1

[9] 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

This female wears a dress, with graphic pattern. It has medium sleeves and it is of medium length. The person moves over to the right.

This woman is wearing a dress, with pure color pattern. It has long sleeves and it is of short length. This woman turns right from the side to the front.

This woman is wearing a dress. It has no sleeves and it is of medium length. The texture of it is pure color. The female shifts to the right.

The lady is wearing a long-sleeve dress. It is of long length. Its pattern is solid color. The lady is moving in the direction of the center while approaching from the right.

This lady is wearing a tank dress. It is of medium length. Its pattern is pure color. This female turns right from the side to the back.

This woman is wearing a dress. It has long sleeves and it is of long length. The texture of it is graphic. This person turns her head to the right.

Figure A3: **Our generated videos with the resolution of** 256×128**.**

This female wears a dress, with solid color pattern. It has sleeves cut off and it is of short length. The lady swings to the left.

This female is wearing a dress. It has long sleeves and it is of medium length. The texture of it is floral. The person turns right from the front to the side.

This female wears a dress, with graphic pattern. It has medium sleeves and it is of medium length. The person moves over to the right.

Figure A4: **Our generated videos with the resolution of** 512×256**.**

The female wears a dress, with floral pattern. It has no sleeves and it is of short length. The female makes a right turn from the side to the front.

This woman is wearing a long-sleeve dress. It is of short length. The texture of it is pure color. This woman is turning right from the side to the back.

Figure A5: **Our generated videos with the resolution of** 512×256**.**

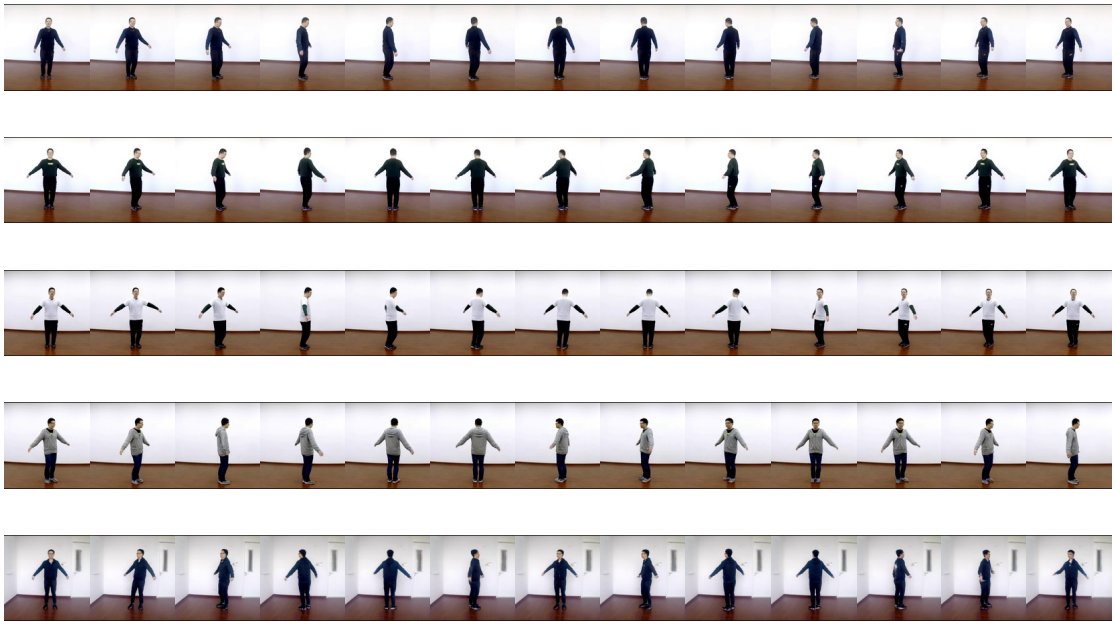


Figure A6: **Our results on iPer Dataset [5].**