(Supplementary Material) Point-to-Point Video Generation

A. Overview

The supplementary material is organized as follows: First, we provide an overview video (video link: https://drive.google.com/open?id=1kS9f2oNGFPO_hp7iWZmvtXLPnOrhl9qW), which briefly summarizes our work. Second, we provide more quantitative results on all datasets: SM-MNIST, Weizmann Human Action, and Human3.6M in Sec. B. Furthermore in Sec. C, we present more qualitative evaluations with respect to *i*) "Generation with various length" in Figs. 6-8 (more examples at https://drive.google.com/open?id=1ueQHNx56MWoqL9ilHjZuBZourg4VrbKc); *ii*) "Multiple control-points generation" in Fig. 9 (more examples at https://drive.google.com/open?id=1OUOd2LjmKwHwVpRwldUEIgvzfpWucYjt); *iii*) "Loop generation" in Fig. 10 (more examples at https://drive.google.com/open?id=1kb8PCIR2_lkE1JS6NlwyglxKlChSBSbF). Finally, the implementation details are described in Sec. D.

B. Quantitative Results

B.1. Performance Under Various Length

In this section, we investigate control point consistency, generation quality and diversity under generation of different lengths on SM-MNIST, Weizmann Action, and Human3.6M dataset (refer to Sec. 4.4 in the main paper).

Control Point Consistency (S-CPC): In Fig. 1, we show the performance of CPC on the three datasets, where for SM-MNIST and Weizmann (the first and the second column), the higher (SSIM) the better, and for Human3.6M (the last column), the lower (MSE) the better. Our method (red line) significantly outperforms other baselines on all datasets, while different components of our method including CPC on prior, latent space alignment, and skip-frame training all introduce performance gain.

Quality (S-Best): In Fig. 2, we demonstrate that our method is able to sustain the generation quality on the three datasets, with the higher (SSIM) the better for SM-MNIST and Weizmann (the first and the second column), and the lower (MSE) the better for Human3.6M (the last column). Our method (red line) achieves superior quality on Human3.6M since its data contain 3D skeletons with highly diverse actions and imposing a targeted end-frame largely confines the S-Best error (more details mentioned in Sec. 4.5 in the main paper). On the other hand, for SM-MNIST and Weizmann, our method only suffers from marginal performance drop in comparison with other baselines. We point out that the generation quality in SM-MNIST gradually declines with increasing generation length since the two digits are prone to overlapping with each other in a longer sequence, resulting in blurry generation after the encounter. This can be potentially solved by representation disentanglement [34, 4, 33, 11, 38], which is out of scope of this paper and left to future work. Overall, we establish that our method attains comparable generation quality while achieving CPC.

Diversity (S-Div): Finally, we show the generation diversity on the three datasets in Fig. 3, where for all columns, the higher (SSIM or MSE) the better. We can observe that our method (red line) reaches superb and comparable performance on Human3.6M and SM-MNIST dataset respectively. On the contrary, Weizmann dataset involves video sequences with steady and fixed-speed action and hence tremendously reduces the possibility of generation if posing constraint at the end-frame (red line in the middle column). All in all, regardless of the limitation of dataset itself, our method is capable of generating diverse sequences and simultaneously achieving CPC.

B.2. Performance Through Time

In this section, we perform a more detailed analysis on generation quality and diversity through time (refer to Sec. 4.5 in the main paper).

Quality (S-Best): In Fig. 4, we show the generation quality at each timestep on the three datasets, with the higher (SSIM) the better for SM-MNIST and Weizmann (the first and second columns), and the lower (MSE) the better for Human3.6M (the last column). We can observe a consistent trend across methods and datasets that the quality progressively decreases as the timestep grows. This is expectable since the generated sequences will step-by-step deviate from the ground truth and induce compounding error

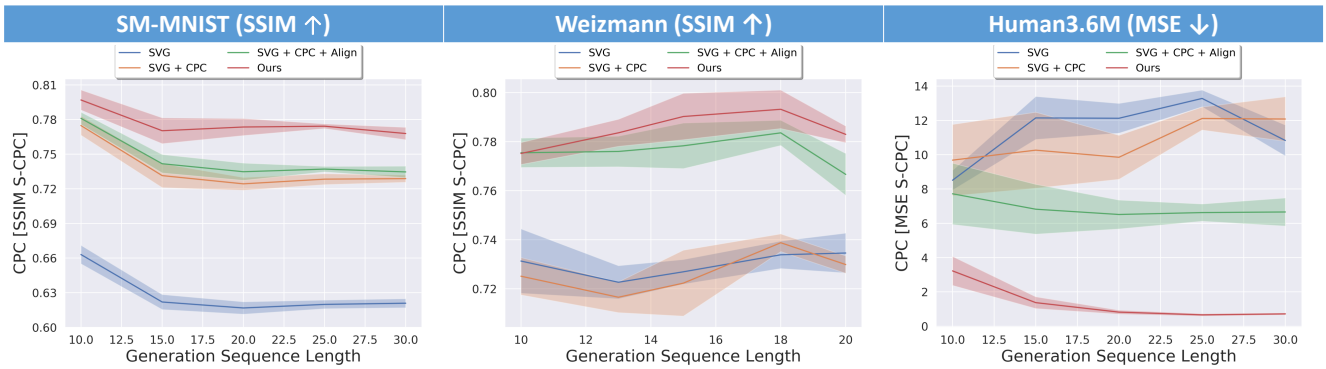

Figure 1: (Better view in color) Control point consistency in generation with various length. Our model significantly outperforms other baselines on all datasets under various lengths.

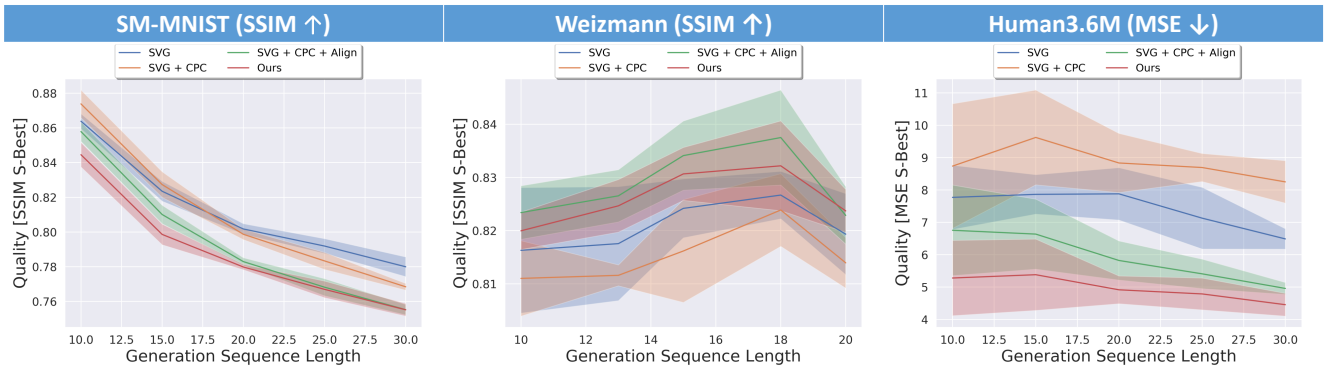

Figure 2: (Better view in color) Quality in generation with various length. Our model sustains the generation quality on the three datasets while achieving CPC.

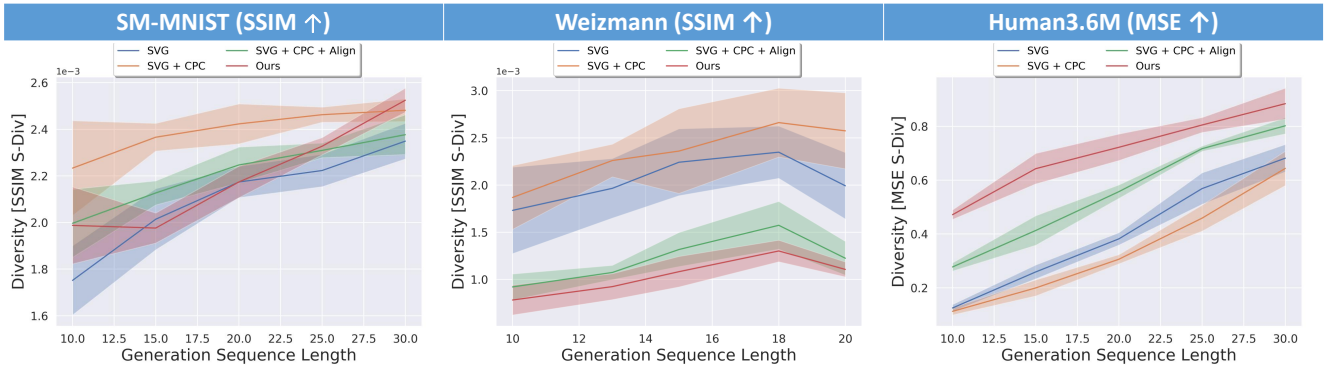

Figure 3: (Better view in color) Diversity in generation with various length. Ours achieve better or comparable diversity on SM-MNIST and Human3.6M while achieving CPC.

as the generation is gradually further from the given start-frame. Remarkably, for all methods taking CPC into consideration (orange, green, and red lines), there is a strong comeback on the generation quality at the end of the sequence since achieving CPC ensures that the generated end-frame converges to the targeted end-frame, thus leading to the results with better S-Best at the last timestep. Finally, the quality boost at the end-frame is lower in Weizmann dataset (the middle column) since unlike the other two (the first and the last columns), its data are captured in noisy background, posing more challenges to CPC and consequently causing lower quality at the end frame.

Diversity (S-Div): In Fig. 5, we demonstrate the generation diversity through time on the three datasets, with the higher (SSIM or MSE) the better in all columns. A consistent trend is shared across all datasets (all columns) in our method (red line) where the diversity is high in the inter-

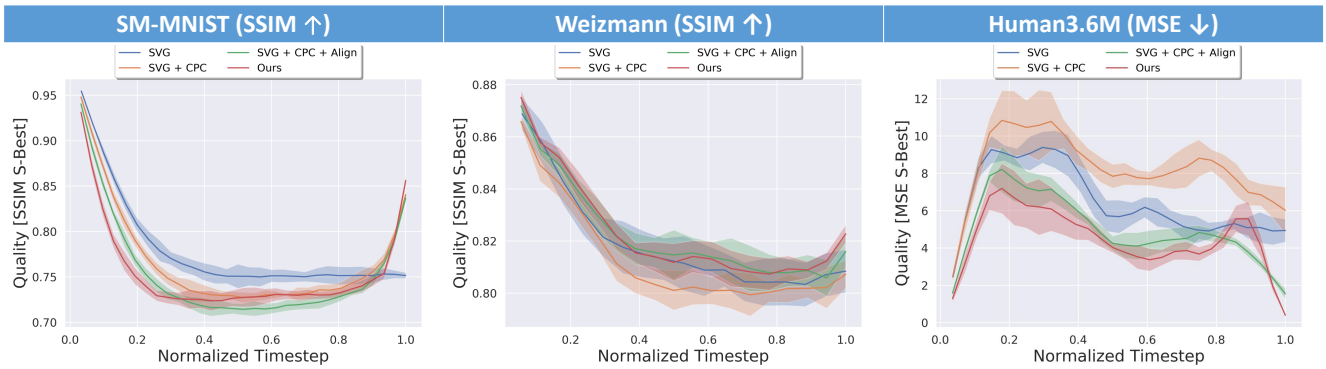

Figure 4: (Better view in color) Quality through time. The generation quality of our model is comparable on SM-MNIST, Weizmann and better on Human3.6M while achiving control-point consistency.

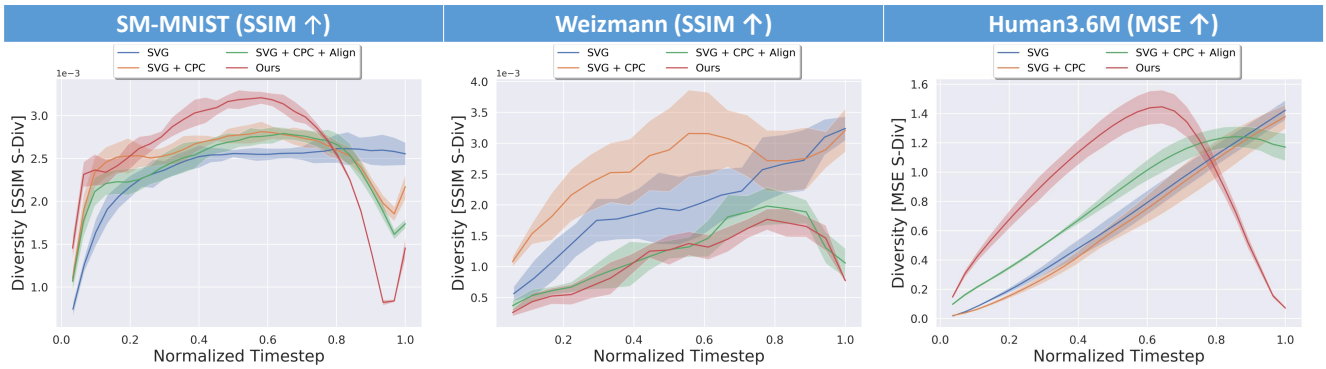

Figure 5: (Better view in color) Diversity through time. The diversity is high in the intermediate frames but reaches zero at the two control points—the targeted start- and end-frames.

mediate frames but reaches zero at the two control points—the targeted start- and end-frames. This suggests that our method is able to plan ahead, generate high-diversity frames at the timestep far from the end, and finally converge to the targeted end-frame with zero-approaching diversity. In addition, we point out that the diversity curve of Weizmann dataset (the middle column) indicates a slightly worse performance in comparison to the results on the other two datasets (the first and third columns) since Weizmann data is featured by unvarying actions, *e.g.*, walking in a fixed speed, that immensely reduces the potential diversity at the intermediate frames.

C. Qualitative Results

Generation with various length. In Fig. 6, Fig. 7, and Fig. 8, we demonstrate the generation results with various lengths on SM-MNIST, Weizmann Action, and Human3.6M datasets. For more generated examples, please see https://drive.google.com/open?id=1ueQHNx56MWoqL9ilHjZuBZourg4VrbKc.

Multiple control-points generation. In Fig. 9, given multiple targeted start- and end-frames, we show our model's ability to merge multiple generated clips into a longer video. For more generated examples, please see https://drive.google.com/open?id=1OUOd2LjmKwHwVpRwldUEIgvzfpWucYjt.

Loop generation. In Fig. 10, by setting the targeted start- and end-frame to be the same, we can achieve loop generation. For more generated examples, please see https://drive.google.com/open?id=1kb8PCIR2_lkE1JS6NlwyglxKlChSBSbF.

D. Implementation Details

We provide the training details and network architecture in this section.

D.1. Training Details

We implement our model in PyTorch. For SM-MNIST and Weizmann Action the input and output image size is 64×64, and for Human3.6M the input comprises the joint positions of size 17×3. Note that while our p2p generation models are fed with the targeted end-frames, the baseline method SVG [3], which is not CPC-aware, is introduced with one additional frame such that all methods are com-

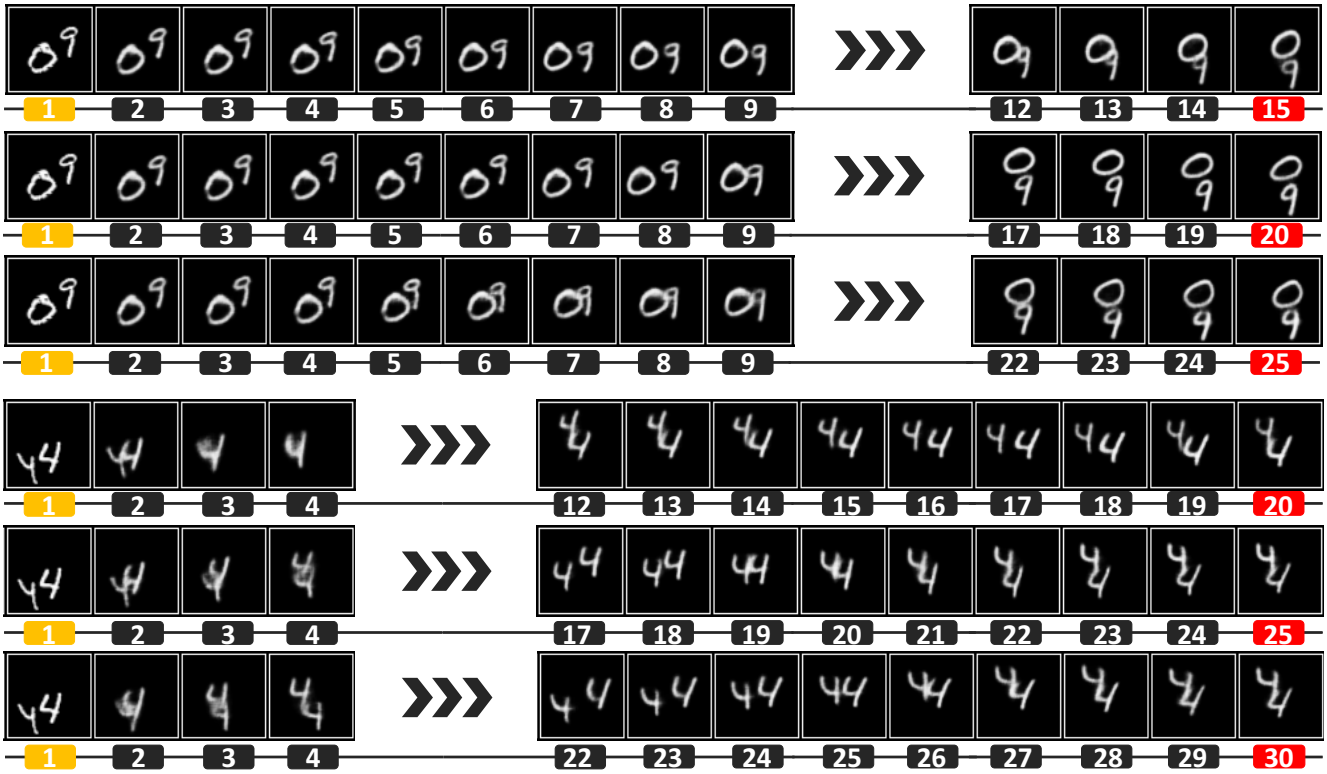

Figure 6: Generation with various length on SM-MNIST. Given the targeted (orange) start- and (red) end-frames, we show the generation results with various lengths on SM-MNIST.

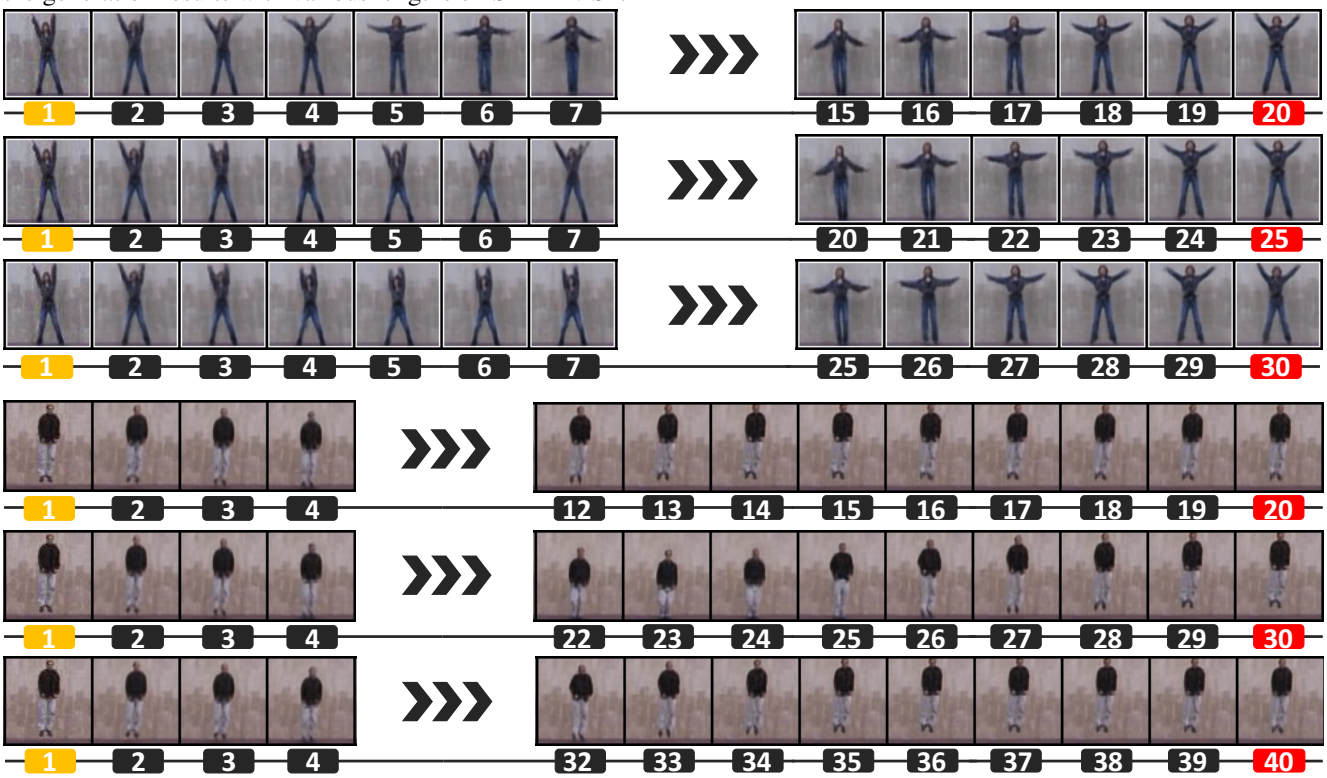

Figure 7: Generation with various length on Weizmann. Given the targeted (orange) start- and (red) end-frames, we show the generation results with various lengths on Weizmann.

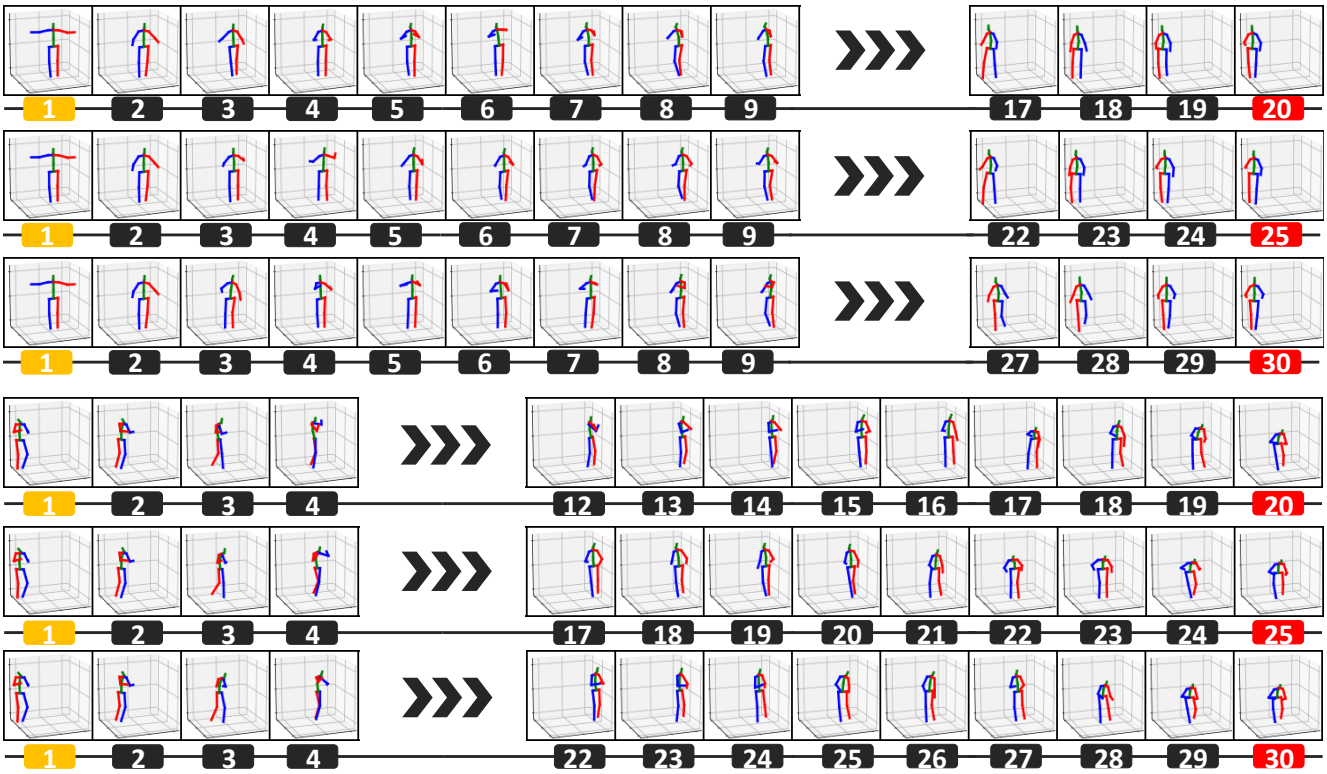

Figure 8: Generation with various length on Human3.6M. Given the targeted (orange) start- and (red) end-frames, we show the generation results with various lengths on Human3.6M.

pared under the same number of input frames. For the reconstruction loss in $\mathcal{L}^{\text{full}}_{\theta,\phi,\psi}$, we use L_2-loss. All models are trained with Adam optimizer, learning rate of 0.002, and batch size of $100, 64, 128$ for SM-MNIST, Weizmann Action and Human3.6M respectively. The weights in the full objective function and other details regarding each dataset are summarized as follows:

SM-MNIST: For the weights in $\mathcal{L}^{\text{full}}_{\theta,\phi,\psi}$, we set $\beta = 10^{-4}$, $\alpha_{\text{cpc}} = 100$, $\alpha_{\text{align}} = 0.5$, $p_{skip} = 0.5$. And the length of training sequences is 12 ± 3.

Weizmann Action: For the weights in $\mathcal{L}^{\text{full}}_{\theta,\phi,\psi}$, we set $\beta = 10^{-5}$, $\alpha_{\text{cpc}} = 10^5$, $\alpha_{\text{align}} = 0.1$, $p_{\text{skip}} = 0.3$. The length of training sequences is 15 ± 3 for Weizmann Action and we augment the dataset by flipping each sequence so that our model can learn to generate action sequences that proceed toward both directions.

Human3.6M: For the weights in the objective function $\mathcal{L}^{\text{full}}_{\theta,\phi,\psi}$: $\beta = 10^{-5}$, $\alpha_{\text{cpc}} = 10^5$, $\alpha_{\text{align}} = 1.0$, $p_{\text{skip}} = 0.3$. The length of training sequences for Human3.6M is 27 ± 3. Besides, we speed up the training sequences to $6\times$ since the adjacent frames in the original sequences are often too similar to each other, which may prevent the model from learning diverse actions.

D.2. Network Architecture

The networks for three datasets all contain the following main components: *i*) posterior q_ϕ, *ii*) prior p_ψ, and *iii*) generator p_θ. The encoder is shared by q_ϕ, p_ψ and the global descriptor. We choose DCGAN [**?**] as the backbone of our encoder and decoder for SM-MNIST and Weizmann Action, and choose multilayer perceptron (MLP) for Human3.6M. The hyper-parameters for the decoder, encoder, q_ϕ, p_ψ and p_θ for each dataset are listed below:

SM-MNIST: For the networks we set $|h_t| = 128$, $|z_t| = 10$; one-layer, 256 hidden units for q_ϕ, one-layer, 256 hidden units for p_ϕ, two-layer, 256 hidden units for p_θ.

Weizmann Action: We use $|h_t| = 512$, $|z_t| = 64$; one-layer, 1024 hidden units for q_ϕ, one-layer, 1024 hidden units for p_ϕ, two-layer, 1024 hidden units for p_θ.

Human3.6M: The networks have $|h_t| = 512$, $|z_t| = 32$; one-layer, 1024 hidden units for q_ϕ, one-layer, 1024 hidden units for p_ϕ, two-layer, 1024 hidden units for p_θ. The encoder MLP consists of 2 residual layers with hidden size of

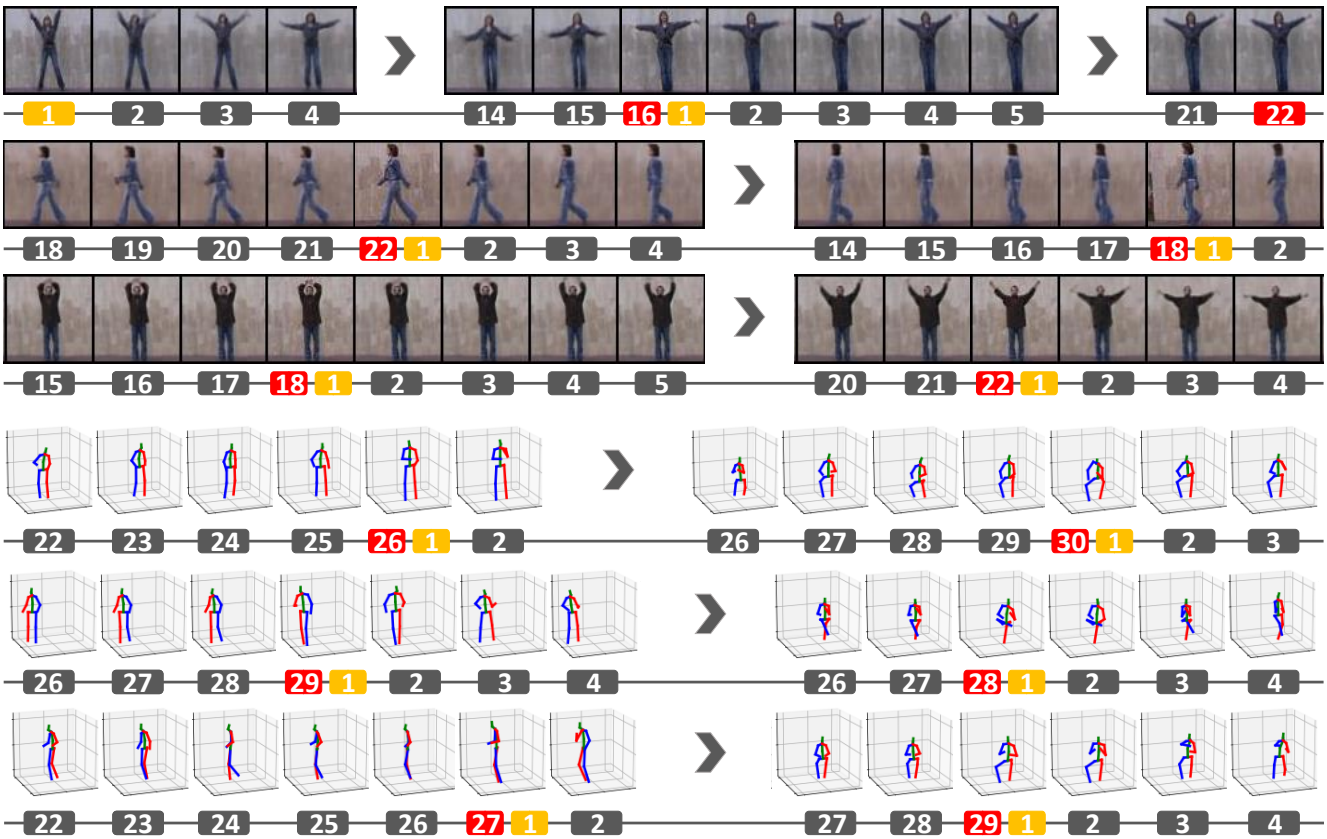

Figure 9: Multiple control points generation. Given multiple targeted (orange) start- and (red) end-frames, we can merge multiple generated clips into a longer video.

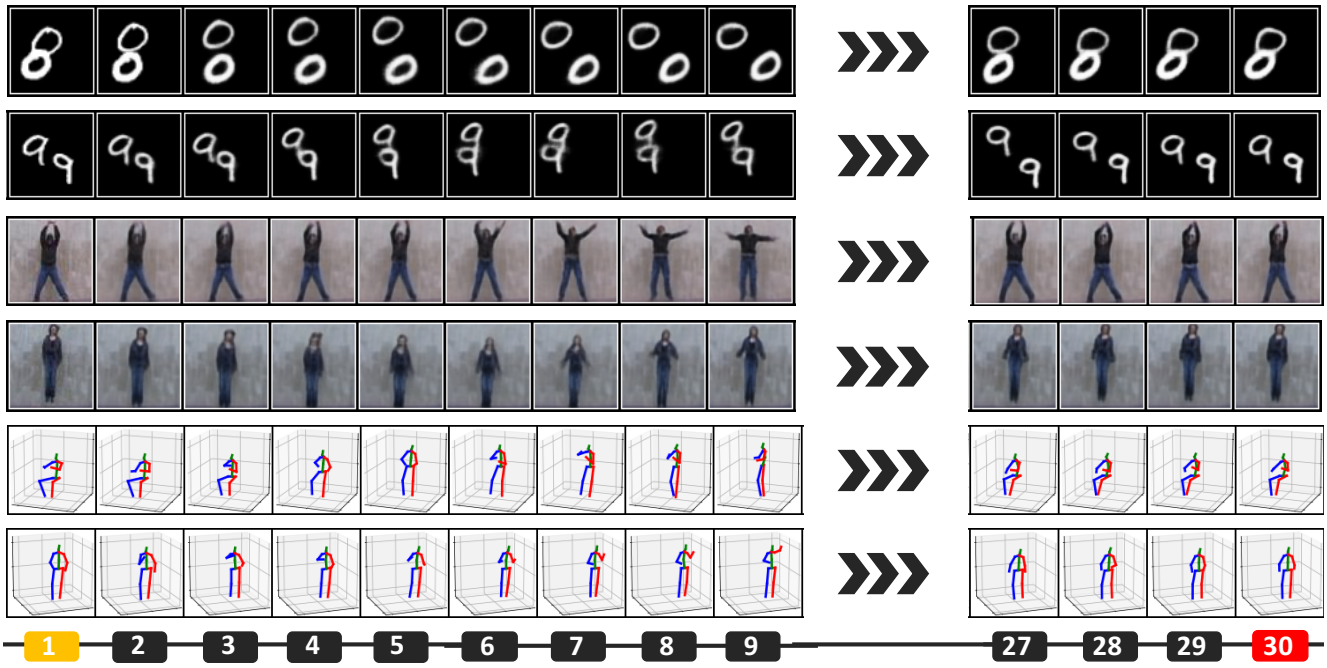

Figure 10: Loop generation. We set the targeted (orange) start- and end-frame with the same frame to achieve loop generation.

512, followed by one fully-connected layer and activated by $\tanh$ function; the decoder MLP is the mirrored version of the encoder but without $\tanh$ in the output layer.

References

[1] Sandra Aigner and Marco Körner. Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing autoencoder gans. *arXiv preprint arXiv:1810.01325*, 2018.

[2] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. In *Proceedings of the International Conference on Learning Representations*, 2018.

[3] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *Proceedings of the International Conference on Machine Learning*, 2018. 3

[4] Emily L. Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, pages 4414–4423, 2017. 1

[5] Frederik Ebert, Chelsea Finn, Alex Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. In *CoRL*, 2017.

[6] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems*, pages 64–72, 2016.

[7] Katerina Fragkiadaki, Jonathan Huang, Alex Alemi, Sudheendra Vijayanarasimhan, Susanna Ricco, and Rahul Sukthankar. Motion prediction under multimodality with conditional stochastic networks. *arXiv preprint arXiv:1705.02082*, 2017.

[8] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007.

[9] Zekun Hao, Xun Huang, and Serge Belongie. Controllable video generation with sparse trajectories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7854–7863, 2018.

[10] Jiawei He, Andreas Lehrmann, Joseph Marino, Greg Mori, and Leonid Sigal. Probabilistic video generation using holistic attribute control. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018.

[11] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*, pages 515–524, 2018. 1

[12] Qiyang Hu, Adrian Waelchli, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Video synthesis from a single image and motion stroke. *arXiv preprint arXiv:1812.01874*, 2018.

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[14] Dinesh Jayaraman, Frederik Ebert, Alexei A. Efros, and Sergey Levine. Time-agnostic prediction: Predicting predictable video frames. In *Proceedings of the International Conference on Learning Representations*, 2019.

[15] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *Advances in Neural Information Processing Systems*, pages 667–675, 2016.

[16] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: high quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[17] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1219–1228, 2018.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations*, 2014.

[19] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.

[20] Yitong Li, Martin Renqiang Min, Dinghan Shen, David E. Carlson, and Lawrence Carin. Video generation from text. *arXiv preprint arXiv:1710.00421*, 2017.

[21] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P Xing. Dual motion gan for future-flow embedded video prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1744–1752, 2017.

[22] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4463–4471, 2017.

[23] William Lotter, Gabriel Kreiman, and David Cox. Unsupervised learning of visual structure using predictive generative networks. In *Workshop Track of International Conference on Learning Representations*, 2016.

[24] Tanya Marwah, Gaurav Mittal, and Vineeth N Balasubramanian. Attentive semantic video generation using captions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1426–1434, 2017.

[25] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *Proceedings of the International Conference on Learning Representations*, 2016.

[26] Simone Meyer, Abdelaziz Djelouah, Brian McWilliams, Alexander Sorkine-Hornung, Markus H. Gross, and Christopher Schroers. Phasenet for video frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[27] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[28] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[29] MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014.

[30] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2830–2839, 2017.

[31] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015.

[32] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms. In *Proceedings of the International Conference on Machine Learning*, pages 843–852, 2015.

[33] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018. 1

[34] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *Proceedings of the International Conference on Learning Representations*, 2017. 1

[35] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances in Neural Information Processing Systems*, pages 613–621, 2016.

[36] Carl Vondrick and Antonio Torralba. Generating the future with adversarial transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1020–1028, 2017.

[37] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3332–3341, 2017.

[38] Olivia Wiles, A Koepke, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. In *Proceedings of the British Machine Vision Conference*, 2018. 1

[39] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018.

[40] Tianfan Xue, Jiajun Wu, Katherine Bouman, and Bill Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2016.

[41] Shohei Yamamoto, Antonio Tejero-de Pablos, Yoshitaka Ushiku, and Tatsuya Harada. Conditional video generation using action-appearance captions. *arXiv preprint arXiv:1812.01261*, 2018.

[42] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 265–281, 2018.

[43] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016.

[44] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. In *Proceedings of the International Conference on Learning Representations*, 2019.

[45] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

[46] Xiaoou Tang Yiming Liu Ziwei Liu, Raymond Yeh and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.