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_hp7iWZmvtxXLPnOrh19qW), 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=1ueQHNx56MWoqL9i1Hj2uBZourq4VrbKc); ii) “Multiple control-points generation” in Fig. 9 (more examples at https://drive.google.com/open?id=1OUOd2LjmKwHwVpRwldUE1gvzfpWucYjt); iii) “Loop generation” in Fig. 10 (more examples at https://drive.google.com/open?id=1kb8PCnTRZ1kE1JS6N1wyglxK1ChSBsB). 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 poses 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...
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 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.
We provide the training details and network architecture in this section.

D. Implementation Details

We implement our model in PyTorch. For SM-MNIST and Weizmann Action the input and output image size is $64 \times 64$, and for Human3.6M the input comprises the joint positions of size $17 \times 3$. 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.
pared under the same number of input frames. For the reconstruction loss in $L^\text{full}_{\theta,\phi,\psi}$, 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 $L^\text{full}_{\theta,\phi,\psi}$, we set $\beta = 10^{-4}$, $\alpha_{\text{cpc}} = 100$, $\alpha_{\text{align}} = 0.5$, $p_{\text{skip}} = 0.5$. And the length of training sequences is $12 \pm 3$.

**Weizmann Action:** For the weights in $L^\text{full}_{\theta,\phi,\psi}$, 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 $15 \pm 3$ 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 $L^\text{full}_{\theta,\phi,\psi}$: $\beta = 10^{-5}$, $\alpha_{\text{cpc}} = 10^5$, $\alpha_{\text{align}} = 1.0$, $p_{\text{skip}} = 0.3$. The length of training sequences for Human3.6M is $27 \pm 3$. 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 $q_\phi$, ii) prior $p_\psi$, and iii) generator $p_\theta$. The encoder is shared by $q_\phi$, $p_\psi$ 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, $q_\phi$, $p_\psi$ and $p_\theta$ 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 $q_\phi$, one-layer, 256 hidden units for $p_\phi$, two-layer, 256 hidden units for $p_\theta$.

**Weizmann Action:** We use $|h_t| = 512$, $|z_t| = 64$; one-layer, 1024 hidden units for $q_\phi$, one-layer, 1024 hidden units for $p_\phi$, two-layer, 1024 hidden units for $p_\theta$.

**Human3.6M:** The networks have $|h_t| = 512$, $|z_t| = 32$; one-layer, 1024 hidden units for $q_\phi$, one-layer, 1024 hidden units for $p_\phi$, two-layer, 1024 hidden units for $p_\theta$. 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.
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


