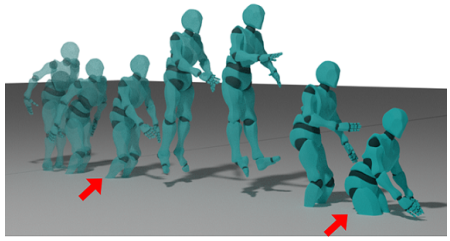
MoST: Motion Style Transformer between Diverse Action Contents

Supplementary Material

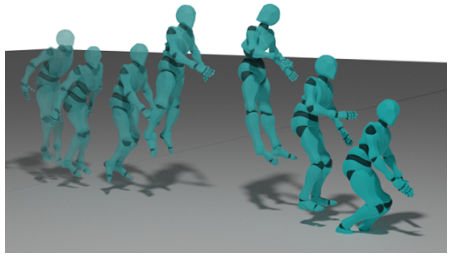
A. Additional Visualization of Limitations in Existing Methods

Fig.A illustrates the *correction of foot contacts* which is employed in existing methods[1, 14, 20] as a post-processing step. The raw outputs from these methods often suffer from issues such as floor penetration or foot skating, prompting the use of heuristic post-processing. In the *correction of foot contacts* scheme, they extract foot contact information from the *content motion*, use this data to adjust the positions of the feet, and apply inverse kinematics for output motion rectification. In contrast, our method generates plausible motion without using any heuristic post-processing.

Fig.B illustrates representative failure cases of existing methods, complementing Fig.1. In Fig.B (a), we observe the corruption of the generated motion due to a lack of clear disentanglement between style and content. The resulting motion fails to preserve the content and struggles to express style effectively. Furthermore, in Fig.B (b), we examine the result from Wen *et al*. [28], which encounters difficulties in expressing style. The method by Wen *et al*. [28] faces limitations, especially in handling short videos, attributed to its reliance on several front frames as a seed. As demonstrated in both Fig. B (a) and (b), our method consistently produces well-stylized and plausible outputs.

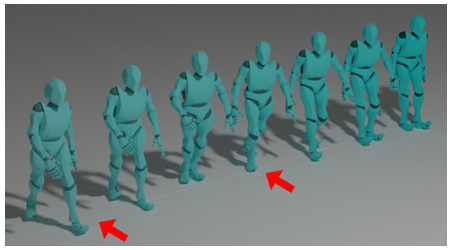


Before post-processing

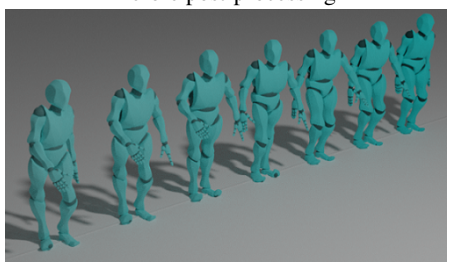


After post-processing

(a)



Before post-processing



After post-processing

(b)

Figure A. Post-processing in existing methods. (a) The 'Depressed Jump' motions generated by Park *et al*. [20]. In the output motion sequence of the model, the body penetrates the floor. Post-processing is applied to force correction and make contact with the ground. (b) The 'Angry Punch' motion generated by Motion-Puzzle [14] network contains feet movement. Post-processing is employed to enforce the fixity of the feet

B. Details of Motion Representation

In this section, we explain more details about the motion representation in Sec. 3.2. m_t^j represents j-th joint vector at t-th frame. The joint vector is represented by $m_t^j = [o_t^j; q_t^j]$, where $o_t^j \in \mathbb{R}^3$ indicates the 3-dimensional vector of each joint from the root joint and $q_t^j \in \mathbb{R}^4$ indicates joint rotation expressed by a unit quaternion. q_t^j is based on the world axis set with the anterior direction of the body. The global joint vector of t-th frame is represented as $m_t^{root} = [o_t^{root}; q_t^{root}]$, where o_t^{root} and q_t^{root} indicate the position of the root joint and the global rotation, respectively. A global velocity vector $v_t = [\acute{o}_t^{root}; a_t]$ is additionally used for the global motion following [20] and [28], where $\acute{o}_t^{root} \in \mathbb{R}^3$ and $a_t^{root} \in \mathbb{R}$ denotes a positional velocity and an angular velocity of the root joint, respectively.

C. Details of Adopted Loss

We provide details of adopted losses from existing methods in this section. Adversarial loss [1, 20] is written as

$$L_{adv} = \mathbb{E}_{M^S \sim \mathbb{M}} \left[log(\mathcal{D}(M^S)) \right] + \mathbb{E}_{M^C, M^S \sim \mathbb{M}} \left[log(1 - \mathcal{D}(M^G)) \right]. \quad (21)$$

Reconstruction loss [1, 20] is written as

$$L_{recon} = \mathbb{E}_{M^C \sim \mathbb{M}} \, ||\texttt{MoST}(M^C, M^C) - M^C||_2. \quad (22)$$

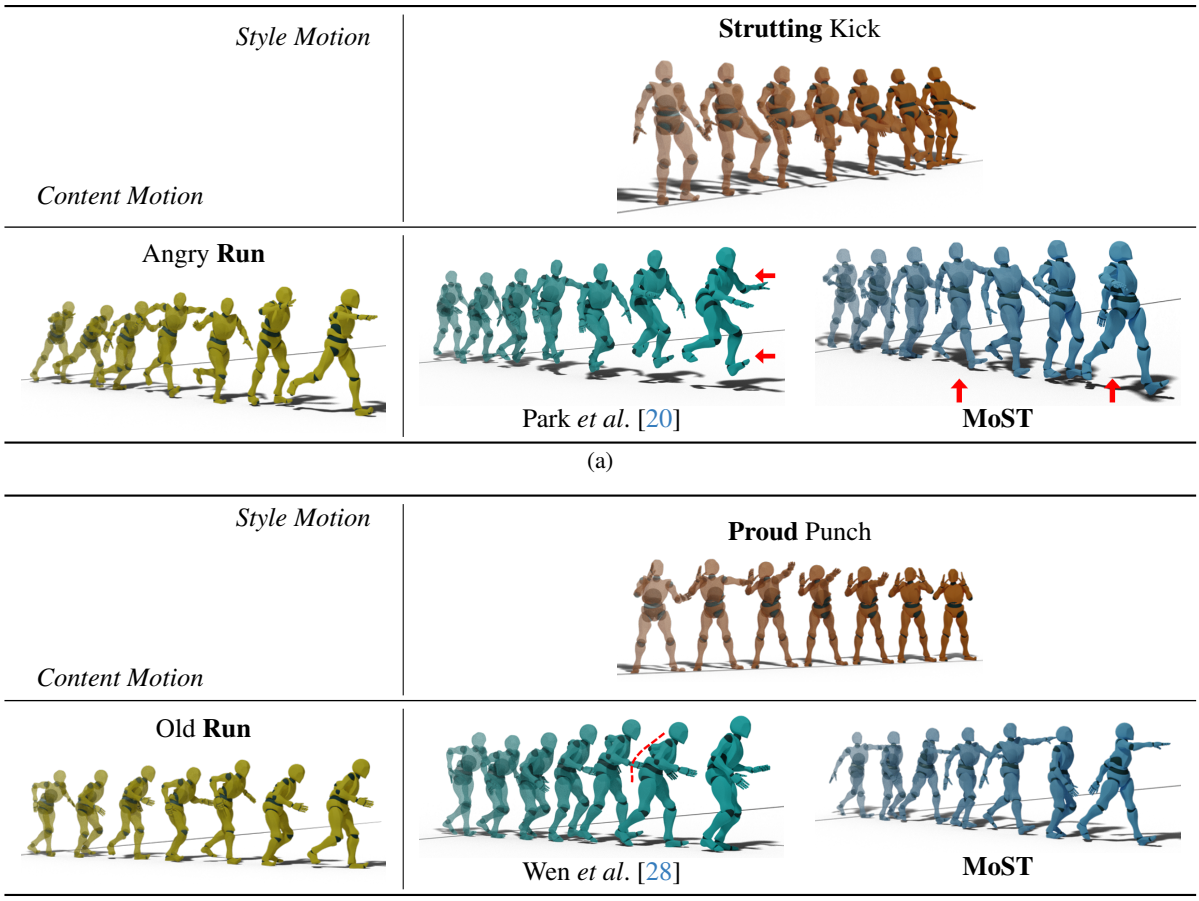


Figure B. Representative failure cases in existing methods: (a) A result from Park *et al*. [20] shows corrupted motion, with the generated animation depicting the leg swinging in mid-air and the arms losing movement. (b) A result from Wen *et al*. [28] exhibits issues in expressing style. For a clearer visualization, please refer to the attached video.

Cycle consistency loss [14] is written as

$$L_{cyc} = L_{cyc\text{-}s} + L_{cyc\text{-}c}, \tag{23}$$

$$L_{cyc\text{-}s} = \mathbb{E}_{M^C, M^S \sim \mathbb{M}} \, || \, \texttt{MoST}(M^S, M^G) - M^S ||_2, \tag{24}$$

$$L_{cyc\text{-}c} = \mathbb{E}_{M^C, M^S \sim \mathbb{M}} \, || \, \texttt{MoST}(M^G, M^C) - M^C ||_2, \tag{25}$$

$$M^G = \texttt{MoST}(M^C, M^S), \tag{26}$$

Inspired by [23], we introduce velocity and acceleration regularization in Eq.(18) for the generated motion M^G as

$$R_{vel} = \frac{1}{T-1} \sum_{t=1}^{T-1} \left(\sum_{j=1}^{J} ||\dot{\hat{m}}_t^j||_2 + ||\dot{\hat{m}}_t^{root}||_2 \right), \tag{27}$$

$$R_{acc} = \frac{1}{T-2} \sum_{t=1}^{T-2} \left(\sum_{j=1}^{J} ||\ddot{\hat{m}}_t^j||_2 + ||\hat{v}_{t+1} - \hat{v}_t||_2 \right),$$

where $\dot{\hat{m}}_t^j = \hat{m}_{t+1}^j - \hat{m}_t^j$, $\dot{\hat{m}}_t^{root} = \hat{m}_{t+1}^{root} - \hat{m}_t^{root}$ and $\ddot{\hat{m}}_t^j = \dot{\hat{m}}_{t+1}^j - \dot{\hat{m}}_t^j$.

D. Details of Proposed Loss

Fig. C illustrates the proposed *style disentanglement loss* (L_D). We generate motions from two different *style motions* with identical style labels but different content labels. L_D makes both generated motions similar. For more stable training, we halt back propagation for the path to $\texttt{MoST}(M^C, M_b^S)$.

E. Details of Evaluation Metrics

In this section, we provide equations of evaluation metrics in Sec. 4. We evaluate M^G which is generated from a motion pair of M^C and M^S drawn from the test dataset. l_{C^C} and l_{S^C} denote the content label and style label of M^C, respectively. l_{C^S} and l_{S^S} denote the content label and style label of M^S, respectively. l_{C^T} and l_{S^T} denote the content label and style label of a ground truth motion sequence M^T

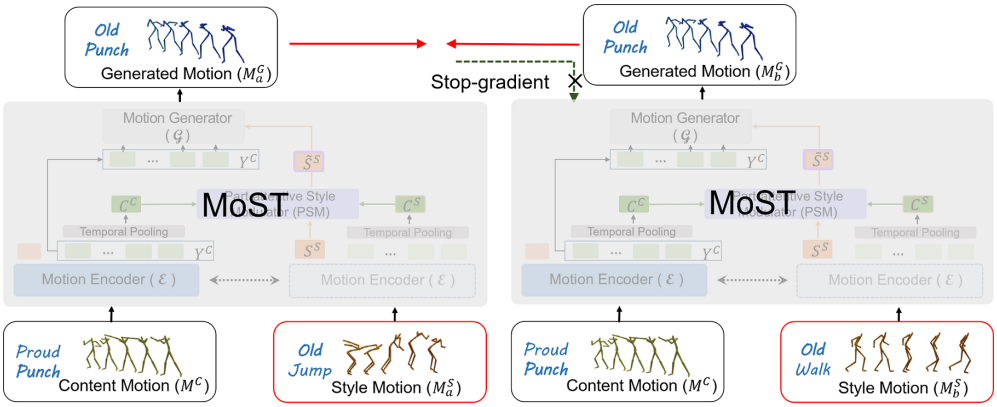


Figure C. Illustration explaining the proposed style disentanglement loss (L_D)

Table C. Ablation study of replacing $\mathcal{E}$, $\mathcal{G}$, and $\mathcal{D}$ with the different network architectures proposed by [1] and [20]. Applied components are labeled as ○. Note that the original method by Park *et al*. [20] utilizes a style label, whereas our method does not.

	Networks of $\mathcal{E}$, $\mathcal{G}$, $\mathcal{D}$	Style Label	Siamese Encoder	PSM	$\mathbf{L_D}$	CC ↓	SC^{++} ↓
Aberman *et al*. [1]	Temporal-CNN	×	×	×	×	46.0	189.7
OURS	Temporal-CNN	×	×	×	○	**39.6**	**81.2**
Park *et al*. [20]	GCN	○	×	×	×	38.4	**65.7**
OURS	GCN	×	×	○	○	**25.1**	68.5
OURS (MoST)	Transformer	×	○	○	○	**8.5**	**63.0**

in a training sample, respectively. Metrics are written as

$$\text{CC} = \mathbb{E}_{M^C, M^S \sim \{m_{test} | l_{SC} = l_{SS}\}} ||M^G - M^C||_2, \quad (28)$$

$$\text{SC} = \mathbb{E}_{M^C, M^S \sim \{m_{test} | l_{CC} = l_{CS}\}} ||M^G - M^S||_2, \quad (29)$$

$$\text{SC}^{++} = \mathbb{E}_{M^C, M^S \sim \{m_{test}\}} \left(\mathbb{E}_{M^T \sim \{m | l_{CT} = l_{CC},\, l_{ST} = l_{SS}\}} ||M^G - M^T||_2 \right),$$

where m_{test} denotes a random variable of motion data in the test set. m denotes a random variable of motion data in the training set. The global translation is excluded when calculating the metric for a fair comparison.

F. Ablation Study in Architectures

In Table C, we conduct an experiment by replacing the proposed transformer architecture in the encoder, generator, and discriminator with other architectures. Specifically, we employ temporal-CNN and GCN architectures, sourced from the methods of Aberman *et al*. [1] and Park *et al*. [20], respectively. In both architectures, our proposed components couldn't be entirely implemented due to their structural limitations. The temporal-CNN [1] and GCN [20] architectures lack the capability to extract both style and content features within a single network. Consequently, we constructed separate content encoders ($\mathcal{C}$) and style encoders ($\mathcal{S}$). Additionally, obtaining part-specific features from the temporal-CNN [1] was not feasible. As a result, PSM couldn't be utilized in this case. Nevertheless, both new losses were applied.

Each network benefited from our overall framework and loss functions, surpassing the individual performances of the original methods listed in Table 1. Notably, the temporal-CNN achieved significantly lower values in both CC and SC^{++}. Importantly, our proposed transformer architecture demonstrated superior performance compared to the other two networks.

G. Evaluation Across Motion Categories

Table D and Table E present the CC and SC^{++} in each content and style category. Regarding content, 'Run' and 'Kick' are proven to be challenging to retain. For style, 'Proud,' 'Angry,' and 'Childlike' are proven to be difficult to express.

H. Additional Results in BFA dataset

Fig. F illustrates the additional results on BFA dataset [1]. Our method effectively performs style transfer even when complex motions are mixed within a motion clip. The figure demonstrates a clear differentiation of styles in the generated motions when different styles are transferred. In addition, our method achieves robust transfer between two inputs with different contents.

Table D. Content consistency (CC) in each content category and in each style category on Xia dataset [30].

Methods	Content categories of content motion					Style categories of style motion								
	walk	run	jump	kick	punch	neutral	angry	childlike	depressed	old	proud	sexy	strutting	**average**
MotionPuzzle [14]	37.0	50.5	68.8	63.7	65.9	51.9	57.9	56.4	49.9	42.4	59.3	52.0	41.8	51.4
Aberman *et al*. [1]	29.3	38.5	40.3	66.2	55.7	42.2	51.9	59.3	34.4	34.1	67.2	39.8	39.2	46.0
Park *et al*. [20]	34.1	46.2	44.0	41.4	35.0	35.7	41.2	46.5	32.4	33.9	50.7	35.8	31.2	38.4
Wen *et al*. [28]	14.0	17.6	31.1	20.0	18.8	18.5	**9.6**	19.9	19.4	17.6	23.5	21.7	16.7	18.5
MoST	**8.1**	**9.6**	**8.2**	**9.7**	**7.9**	**6.9**	9.9	**7.5**	**7.5**	**10.1**	**10.4**	**7.6**	**8.5**	**8.5**

Table E. Style consistency^{++} (SC^{++}) in each content category and in each style category on Xia dataset [30].

Methods	Content categories of content motion					Style categories of style motion								
	walk	run	jump	kick	punch	neutral	angry	childlike	depressed	old	proud	sexy	strutting	**average**
MotionPuzzle [14]	69.6	78.2	89.4	77.9	77.5	71.5	84.2	79.3	72.4	65.2	92.3	74.8	68.1	76.0
Aberman *et al*. [1]	258.9	237.1	115.0	112.8	85.9	185.1	174.0	209.9	181.3	182.2	196.9	195.3	192.5	189.7
Park *et al*. [20]	65.2	77.9	64.8	68.3	53.3	59.2	**66.5**	70.5	59.8	62.0	81.1	**62.4**	64.2	65.7
Wen *et al*. [28]	72.7	87.5	97.4	97.5	65.0	71.1	79.5	84.5	77.2	83.8	87.2	78.6	84.5	80.8
MoST	**61.6**	**75.4**	**60.5**	**68.1**	**51.9**	**55.1**	67.5	**63.8**	**58.8**	**59.9**	**75.5**	62.8	**60.2**	**63.0**

I. Generation of Global Translation

Fig.D displays the global translation generated by our method. The figure illustrates the variation in motion speeds for different styles produced by our method. The upper motion is 'Childlike Walk,' while the lower motion is 'Old Walk.' The global translation of 'Old Walk' is generated at a slower speed. It is worth noting that Aberman *et al*. [1] utilized a heuristic post-processing technique called *global velocity warping* to diversify motion speeds between styles.

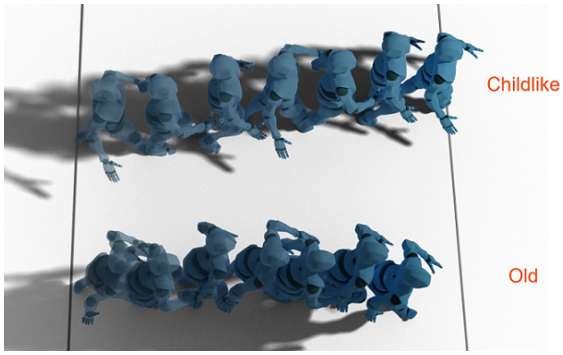
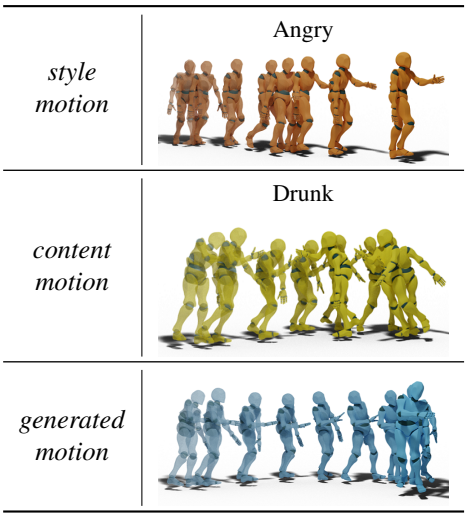


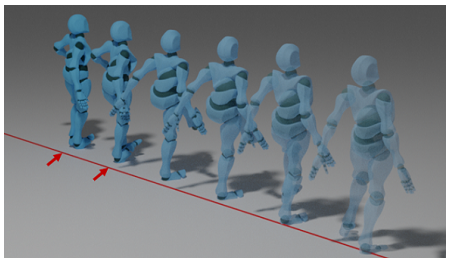
Figure D. Generated global translation reflecting styles. (Top) 'Childlike Walk' generated from 'Neutral Walk' (M^C) and 'Childlike Kick' (M^S). (Bottom) 'Old Walk' generated from 'Neutral Walk' (M^C) and 'Old Kick' (M^S).

J. Failure Cases of MoST

In Fig.E, we present the failure cases observed in our results. As shown in Fig.E (a), style transfer was relatively less successful for the *content motion* featuring intricate motion, such as 'Drunk.' In the generated motion, while the staggering appearance decreased, 'Angry' was not expressed perfectly. Fig. E (b) illustrates foot skating observed in the generated motion.

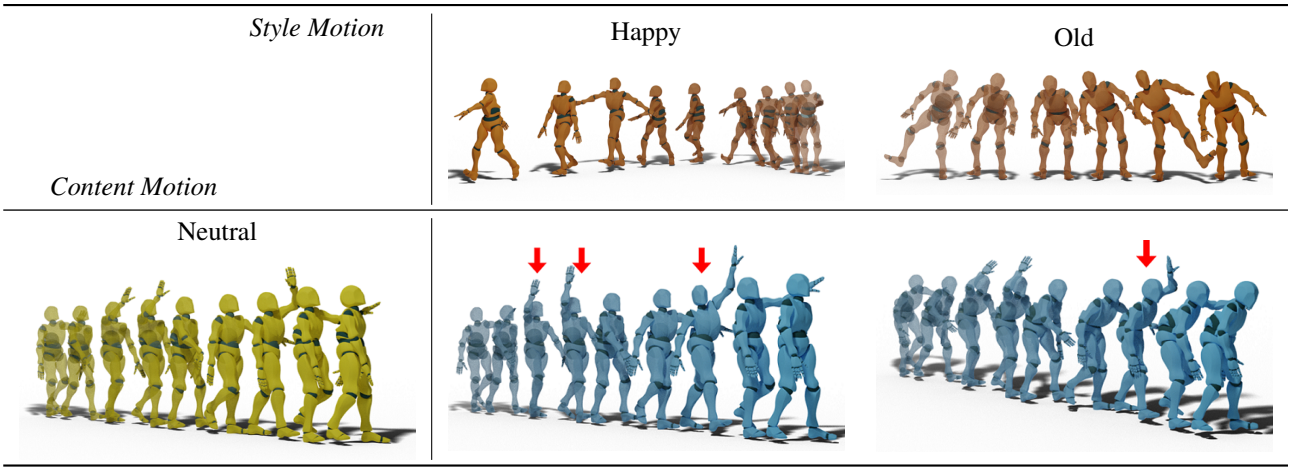


(a)

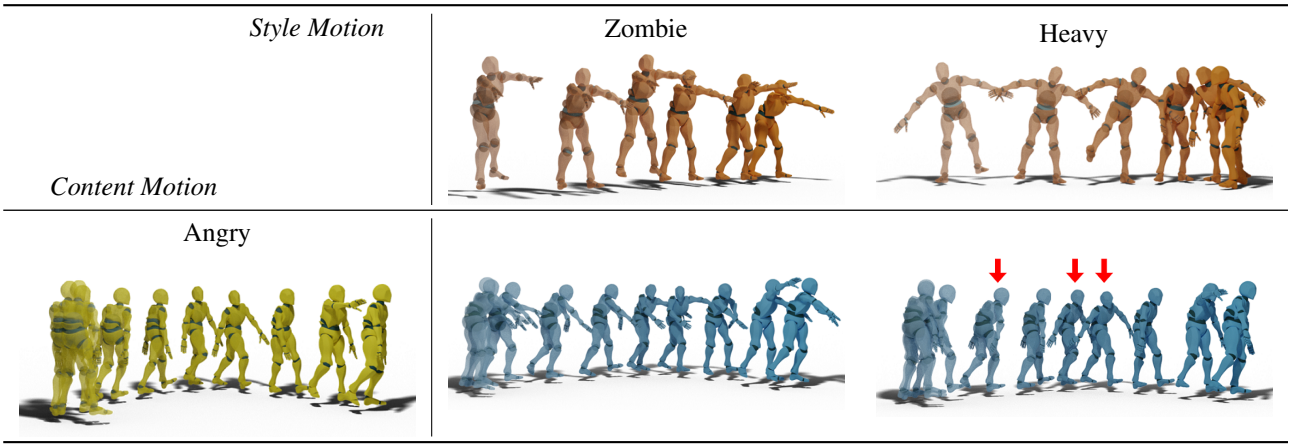


(b)

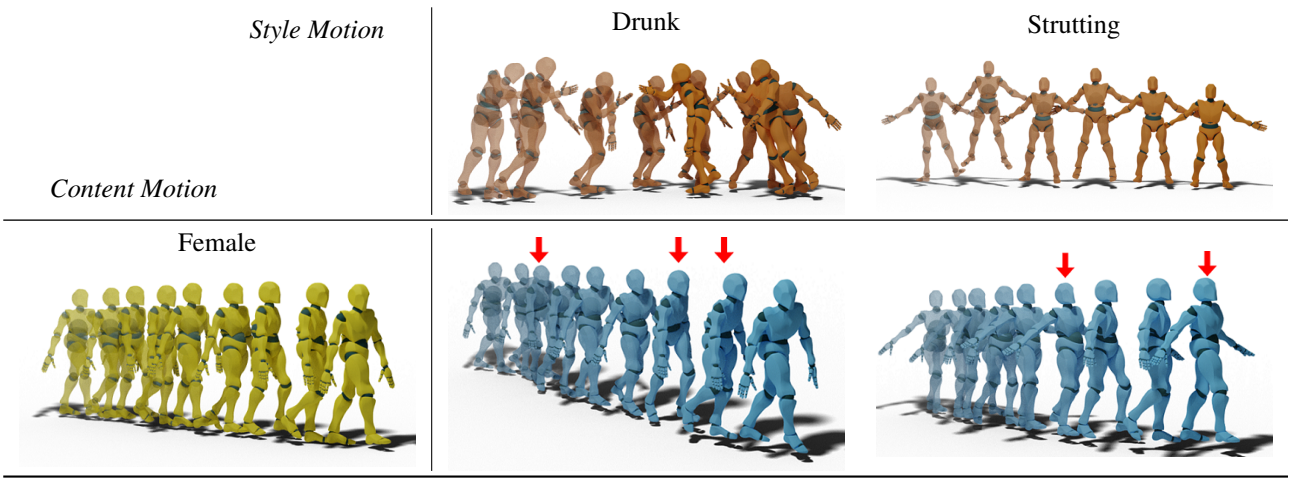
Figure E. Failure cases obtained in our method. (a) Less successful style transfer for the intricate *content motion*. (b) Foot skating observed in the generated motion of 'Proud Kick.'



(a)



(b)



(c)

Figure F. Qualitative results in the BFA [1] dataset. The BFA dataset comprises long motion sequences not segmented by contents. Therefore, we label only the style categories at the upper side of each motion clip. Each generated motion successfully reflects the desired styles, as highlighted in the motion segments indicated by the red arrows. (a) The character straightens its arms and opens its chest in the 'Happy' style, whereas it bends its back and arms in the 'Old' style. (b) The character walks staggering with arms extended forward in the 'Zombie' style, and the 'Heavy' style expresses the weight when pressing down on the ground. (c) The 'Drunk' style expresses staggering motion, while the character in the 'Strutting' style opens its chest and arms. For clearer visualization, please refer to the attached video.