This supplementary file provides additional ablation analysis of our design components (Sec. 5) and weight of various loss terms (Appendix C), content and feature visualization(Appendix B), implementation details (Appendix D), Evaluation Metric (Appendix E), cross-style and homo-style interpolation results (Appendix G), user study introduction (Appendix F) and failure cases (Appendix H).

Video. We also provide several supplementary videos, which contains dynamic animations of our stylization results, visual comparisons, interpolation, stylized text2motion and failure cases. We strongly encourage our audience to watch these videos. It will be much helpful to understand our work. The videos are submitted along with supplementary files, or accessible online (1080P): https://drive.google.com/drive/folders/1UeGuE1qCceLFQJa3vpYoOHC2MoLdBifK?usp=sharing

Code and Model. The code of our approach and implemented baselines are also submitted for reference. Code and trained model will be publicly available upon acceptance.

A ABLATION ANALYSIS

S / U	Method	(Aberman et al., 2020)			(Xia et al., 2015)			
		Style Acc↑	Style FID ↓	Geo Dis↓	Style Acc↑	Content Acc↑	Content FID↓	Geo Dis ↓
S	Ours (A)	$\mathbf{0.945}^{\pm 007}$	$\mathbf{0.020}^{\pm 002}$	$\mathbf{0.344}^{\pm 002}$	$\mathbf{0.926}^{\pm 008}$	$\mathbf{0.674}^{\pm 011}$	$\mathbf{0.189}^{\pm 005}$	$\mathbf{0.680}^{\pm 002}$
	w/o latent	$\underline{0.932}^{\pm 008}$	$0.022^{\pm 002}$	$0.463^{\pm 003}$	$0.851^{\pm 012}$	$\underline{0.654}^{\pm 012}$	$0.258^{\pm 007}$	$0.707^{\pm 003}$
	w/o prob-style	$0.913^{\pm 007}$	$0.022^{\pm 002}$	$0.509^{\pm 004}$	$0.870^{\pm 010}$	$0.524^{\pm 015}$	$0.249^{\pm 008}$	$0.767^{\pm 004}$
	w/o homo-style	$0.883^{\pm 012}$	$0.032^{\pm 004}$	$0.507^{\pm 003}$	$0.851^{\pm 012}$	$0.537^{\pm 016}$	$0.232^{\pm 006}$	$0.760^{\pm 004}$
	w/o autoencoding	$0.900^{\pm 010}$	$0.026^{\pm 002}$	$0.427^{\pm 003}$	$\underline{0.879}^{\pm 010}$	$0.634^{\pm 011}$	$\underline{0.198}^{\pm 005}$	$0.720^{\pm 004}$
	w/o cycle-recon	$0.917^{\pm 009}$	$\underline{0.021}^{\pm 002}$	$\underline{0.385}^{\pm 003}$	$0.872^{\pm 006}$	$0.627^{\pm 011}$	$0.208^{\pm 004}$	$\underline{0.699}^{\pm 002}$
U	Ours (A)	$\mathbf{0.804}^{\pm 011}$	$\mathbf{0.040}^{\pm 003}$	$\mathbf{0.441}^{\pm 003}$	$\mathbf{0.814}^{\pm 011}$	$\underline{0.588}^{\pm 010}$	$\mathbf{0.217}^{\pm 006}$	$0.735^{\pm 003}$
	w/o latent	$\underline{0.780}^{\pm 014}$	$0.048^{\pm 003}$	$0.466^{\pm 004}$	$0.734^{\pm 014}$	$0.584^{\pm 011}$	$0.272^{\pm 008}$	$\underline{0.721}^{\pm 003}$
	w/o prob-style	$0.734^{\pm 018}$	$0.058^{\pm 004}$	$0.461^{\pm 003}$	$0.666^{\pm 016}$	$\mathbf{0.597}^{\pm 015}$	$0.270^{\pm 010}$	$\mathbf{0.718}^{\pm 003}$
	w/o homo-style	$0.753^{\pm 016}$	$0.050^{\pm 002}$	$0.513^{\pm 003}$	$0.730^{\pm 009}$	$0.526^{\pm 013}$	$0.250^{\pm 005}$	$0.803^{\pm 002}$
	w/o autoencoding	$0.777^{\pm 012}$	$0.049^{\pm 004}$	$0.493^{\pm 004}$	$\underline{0.811}^{\pm 011}$	$0.491^{\pm 015}$	$\underline{0.230}^{\pm 007}$	$0.759^{\pm 005}$
	w/o cycle-recon	$0.765^{\pm 011}$	$\underline{0.043}^{\pm 004}$	$\underline{0.560}^{\pm 005}$	$0.756^{\pm 017}$	$0.479^{\pm 013}$	$0.233^{\pm 007}$	$0.869^{\pm 002}$

Table 6: Ablation study on **different components of our model design**. ± indicates 95% confidence interval. **Bold** face indicates the best result, while underscore refers to the second best. (S) and (U) denote *supervised* and *unsupervised* setting. Motion-based stylization is presented for both settings. *Prob-style* refers to probabilistic style space.

Table 6 presents the results of ablation experiments investigating various components of our latent stylization models. These components include stylization on the latent space (*latent*), the use of a probabilistic style space (*prob-style*), homo-style alignment (*homo-style*), autoencoding, and cycle reconstruction. The experiments are conducted within the framework of Ours (A) and are focused on the task of motion-based stylization. Results are reported on two datasets (Aberman et al., 2020) and (Xia et al., 2015). It's important to note that the dataset of (Xia et al., 2015) is exclusively used for testing the generalization ability of our models and has not been used during training.

Overall, we observe a notable performance improvement by incorporating different modules into our framework. For instance, our key designs—latent stylization and the use of a probabilistic style space—significantly enhance performance on the unseen (Xia et al., 2015) dataset, resulting in a 7% increase in stylization accuracy in the supervised setting. Additionally, homo-style alignment, despite its simplicity, provides a substantial performance boost across all metrics. Notably, content accuracy sees a remarkable improvement of 13% and 6% in supervised and unsupervised settings, respectively, underscoring the effectiveness of homo-style alignment in preserving semantic information.

In the subsequent sections, we delve into a detailed discussion of three other critical choices in our model architecture and learning scheme: probabilistic (or deterministic) space for content and style features, separate (or end-to-end) training of latent extractor and stylization model, and the incorporation of a global motion predictor.

Content Space	Style Space	(Aberman et al., 2020)			(Xia et al., 2015)			
		Style Acc↑	Style FID ↓	Geo Dis↓	Style Acc↑	Content Acc↑	Content FID↓	Geo Dis ↓
D	D	$0.913^{\pm 007}$	$0.022^{\pm 002}$	$0.509^{\pm 004}$	$0.870^{\pm 010}$	$0.524^{\pm 015}$	$0.249^{\pm 008}$	$0.767^{\pm 004}$
D	P	$0.945^{\pm 007}$	$0.020^{\pm 002}$	$\mathbf{0.344}^{\pm 002}$	$\mathbf{0.926}^{\pm 008}$	$\mathbf{0.674}^{\pm 011}$	$\mathbf{0.189}^{\pm 005}$	$\mathbf{0.680}^{\pm 002}$
P	P	$\mathbf{0.947}^{\pm 001}$	$\mathbf{0.017}^{\pm 001}$	$0.489^{\pm 003}$	$0.891^{\pm 003}$	$0.417^{\pm 012}$	$0.322^{\pm 011}$	$0.758^{\pm 003}$

Table 7: Ablation study on the **choice of probabilistic (P) or deterministic (D) space** for content and style, in supervised setting. ± indicates 95% confidence interval. **Bold** face indicates the best result, while <u>underscore</u> refers to the second best. Motion-based stylization is presented.

Probabilistic Modeling of Style and Content Spaces. Table 7 presents a comparison between deterministic and probabilistic modeling approaches for both style and content spaces. In our study, the introduction of a probabilistic style space not only provides remarkable flexibility during inference, enabling diverse stylization and multiple applications, but it also consistently enhances performance and generalization capabilities. An intriguing aspect to explore is the impact of modeling the content space non-deterministically. As highlighted in Tab. 7, we observe that a probabilistic content space achieves superior stylization accuracy on in-domain datasets (Aberman et al., 2020). However, it exhibits sub-optimal generalization performance on out-domain cases (Xia et al., 2015).

Training Strategy	(Aberman et al., 2020)			(Xia et al., 2015)			
	Style Acc↑	Style FID ↓	Geo Dis↓	Style Acc↑	Content Acc↑	Content FID↓	Geo Dis ↓
Separately	$\mathbf{0.945}^{\pm 007}$	$\mathbf{0.020}^{\pm 002}$	$\mathbf{0.344}^{\pm 002}$	$\mathbf{0.926}^{\pm 008}$	$\mathbf{0.674}^{\pm 011}$	$\mathbf{0.189}^{\pm 005}$	$\mathbf{0.680}^{\pm 002}$
End-to-end	$0.125^{\pm 010}$	$1.521^{\pm 024}$	$0.577^{\pm 001}$	$0.174^{\pm 014}$	$0.293^{\pm 002}$	$1.417^{\pm 009}$	$0.700^{\pm 001}$

Table 8: Ablation study on **separately or end-to-end training** the latent model and stylization model, in supervised setting. ± indicates 95% confidence interval. **Bold** face indicates the best result, while <u>underscore</u> refers to the second best. (S) and (U) denote *supervised* and *unsupervised* setting. Motion-based stylization is presented.

Separate / End-to-end Training. Our two-stage framework can alternatively be trained in an end-to-end fashion. We also conduct ablation analysis to evaluate the impact of such choice of training strategy. The results are presented in Table 8. In practice, we observed that end-to-end training posed significant challenges. The model struggled to simultaneously learn meaningful latent motion representation and effectively transfer style traits between stages. Experimental results align with this observation, revealing that stylization accuracy is merely around 15% on both datasets in the end-to-end training scenario, in contrast to the accuracy of 92% achieved by stage-by-stage training.

Method	(Aberman et al., 2020)		CMU Mocap (CMU)		(Xia et al., 2015)	
	Style Acc↑	Foot Skating↓	Style Acc↑	Foot Skating↓	Style Acc↑	Foot Skating↓
Ours (S)	$\mathbf{0.945}^{\pm 007}$	$\mathbf{0.130}^{\pm 001}$	$0.918^{\pm 007}$	$\mathbf{0.140}^{\pm 001}$	$\mathbf{0.926}^{\pm 008}$	$\mathbf{0.263}^{\pm 003}$
Ours *w/o* GMP (S)	$0.942^{\pm 003}$	$0.141^{\pm 001}$	$\mathbf{0.920}^{\pm 006}$	$0.160^{\pm 001}$	$0.882^{\pm 008}$	$0.331^{\pm 002}$
Ours (U)	$\mathbf{0.840}^{\pm 010}$	$\mathbf{0.102}^{\pm 001}$	$\mathbf{0.828}^{\pm 010}$	$\mathbf{0.099}^{\pm 001}$	$\mathbf{0.860}^{\pm 010}$	$\mathbf{0.179}^{\pm 002}$
Ours *w/o* GMP (U)	$0.817^{\pm 013}$	$0.116^{\pm 001}$	$0.820^{\pm 009}$	$0.122^{\pm 001}$	$0.777^{\pm 018}$	$0.307^{\pm 002}$

Table 9: Ablation study on **global motion prediction** (GMP, see Sec. 3.2.3). The symbol ± indicates the 95% confidence interval. **Bold** indicates the best result. (S) and (U) denote *supervised* and *unsupervised* settings, respectively. Results of motion-based stylization are presented. *Foot skating* is measured by the average velocity of foot joints on the XZ-plane during foot contact.

Global Motion Prediction (GMP). The primary objective of our global motion prediction is to facilitate adaptive pacing for diverse motion contents and styles. As illustrated in Tab. 10, we quantify the mean square error of GMP in predicting root positions across three test sets, measured in millimeters. Notably, even on the previously unseen dataset Xia et al. (2015), the lightweight GMP performs admirably, with an error of 57.7 mm.

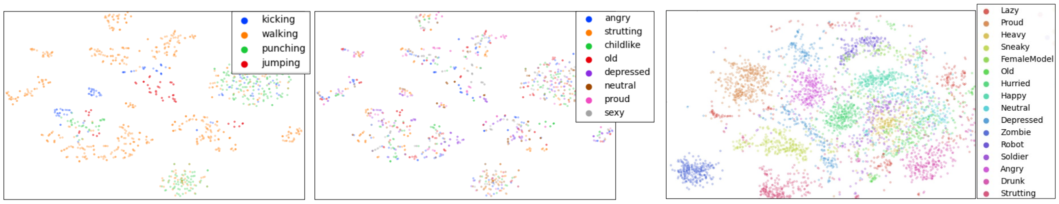
To assess the impact of GMP on stylization performance, we compare against a contrast setting (Ours *w/o* GMP), where global motions are directly obtained from the source content input, akin to

(Aberman et al., 2020)	CMU Mocap (CMU)	(Xia et al., 2015)
46.2	48.7	57.7

Table 10: **Mean Square Error of Root Position Prediction.** The metric is measured in millimeters. Note the dataset of (Xia et al., 2015) is untouched during the training of the global motion predictor.

previous approaches. Additionally, we introduce a *foot skating* metric to gauge foot sliding artifacts, calculated by the average velocity of foot joints on the XZ-plane during foot contact.Table 9 showcases motion-based results on (Aberman et al., 2020; CMU; Xia et al., 2015) test sets. Across all comparisons, our proposed GMP effectively mitigates foot skating issues. Although 2-dimensional global motion features constitute only a small fraction of the entire 260-dimensional pose vectors, it makes considerable difference on the dataset of (Xia et al., 2015), improving the stylization accuracy by around 9%. In our 3rd and 4th supplementary videos, we also illustrate how our GMP enables adaptive pacing in different stylization outcomes (label-based and motion-based) for the same content.

B FEATURE VISUALIZATION



(a) Content Codes Colored by Content Label (left), and Style Label (right).

(b) Style Codes Colored by Style Label.

Figure 8: **Latent Visualization.** Panel (a) displays the projection of the **identical** set of content codes onto a 2D space using t-SNE, colored according to content labels (left) and style labels (right). This visualization suggests that content codes faithfully capture content traits, while style information has been effectively removed. In panel (b), style codes are projected onto a 2D space using t-SNE and colored by their corresponding style labels. Notably, clear style clusters emerge unsupervisedly, aligning with style labels.

Given that our content encoder accommodates motions of arbitrary length, we extract content codes from the Xia et al. (2015) dataset. This dataset, unseen by our models, provides annotations for both style and content labels. Notably, the motions in this dataset are usually short, typically within 3s, which is insufficient to our style encoder. Therefore, for style codes, we takes the motions from dataset (Aberman et al., 2020) for visualization. The models are learned in unsupervised setting, using VAE as latent model.

Content Code Visualization. Figure 8 (a) visually presents 2D projections of our content codes. The content codes are colored by their content labels on the left and by their style labels on the right. To generate these projections, the temporal content codes are aggregated along the temporal dimension and then mapped to 2D space using t-SNE. When the content codes are colored by content label (e.g., *walking, kicking*), distinct clusters aligned with the corresponding labels become apparent. However, when the same set of content codes is colored by their style label, these labels are evenly distributed within these clusters. This observation suggests that the content code adeptly captures the characteristics of various contents while effectively erasing style information.

Style Code Visualization. Figure 8 (b) visualizes the style codes in a 2D space, color-coded by their style labels. Notably, these style labels were never used during model training. In contrast to the content code visualization in Fig. 8 (a), the projected style codes exhibit a strong connection with the external style label annotations. This observation underscores the effectiveness of our style encoder in extracting style features from the motion corpus.

C LOSS WEIGHT ANALYSIS

Tab. 11 presents more quantitative results of our models on (Aberman et al., 2020) and (Xia et al., 2015) test sets. Specifically, we provide the ablation evaluations in both supervised (S) and unsupervised setting (U). For supervised setting, we conduct experiments on label-based stylization which also compares the diversity; and for unsupervised setting we adopt motion-based stylization. Note the base models are not necessarily our final models, here they are set only for reference.

S / U	λ_{cyc}	λ_{kl}	λ_{hsa}	(Aberman et al., 2020)			(Xia et al., 2015)		
				Style Acc↑	Geo Dis↓	Div↑	Style Acc↑	Content Acc↑	Content FID↓
S (base)	0.1	0.01	0.1	$0.937^{\pm 008}$	$0.415^{\pm 003}$	$0.153^{\pm 016}$	$0.913^{\pm 008}$	$\underline{0.669}^{\pm 013}$	$0.202^{\pm 006}$
			0.5	$0.936^{\pm 008}$	$\underline{0.369}^{\pm 003}$	$0.091^{\pm 011}$	$0.924^{\pm 007}$	$\mathbf{0.706}^{\pm 010}$	$\mathbf{0.197}^{\pm 007}$
		0.001		$\mathbf{0.962}^{\pm 006}$	$0.429^{\pm 004}$	$0.125^{\pm 016}$	$\underline{0.933}^{\pm 009}$	$0.619^{\pm 014}$	$\underline{0.197}^{\pm 005}$
		0.1		$0.940^{\pm 008}$	$0.414^{\pm 004}$	$0.141^{\pm 015}$	$0.914^{\pm 009}$	$0.634^{\pm 013}$	$0.209^{\pm 005}$
	0.01			$\underline{0.955}^{\pm 006}$	$0.419^{\pm 003}$	$0.107^{\pm 011}$	$\mathbf{0.957}^{\pm 007}$	$0.609^{\pm 011}$	$0.207^{\pm 006}$
	1			$0.880^{\pm 011}$	$0.423^{\pm 003}$	$\underline{0.302}^{\pm 026}$	$0.833^{\pm 011}$	$0.625^{\pm 013}$	$0.236^{\pm 006}$
U (base)	1	0.01	0.1	$\mathbf{0.804}^{\pm 011}$	$0.441^{\pm 003}$	-	$\mathbf{0.814}^{\pm 014}$	$0.588^{\pm 010}$	$0.217^{\pm 006}$
			0.01	$\underline{0.790}^{\pm 015}$	$0.489^{\pm 004}$	-	$0.761^{\pm 012}$	$0.567^{\pm 016}$	$0.224^{\pm 007}$
		0.1		$0.659^{\pm 018}$	$0.430^{\pm 004}$	-	$0.701^{\pm 014}$	$0.619^{\pm 013}$	$\mathbf{0.190}^{\pm 005}$
	0.01			$0.669^{\pm 013}$	$\mathbf{0.388}^{\pm 003}$	-	$0.671^{\pm 015}$	$\mathbf{0.641}^{\pm 012}$	$\underline{0.206}^{\pm 006}$
	0.1			$0.739^{\pm 015}$	$\underline{0.420}^{\pm 004}$	-	$\underline{0.762}^{\pm 016}$	$\underline{0.619}^{\pm 014}$	$0.214^{\pm 007}$

Table 11: Effect of hyper-parameters of *ours (A)* on the (Aberman et al., 2020) and (Xia et al., 2015) test sets. ± indicates 95% confidence interval. **Bold** face indicates the best result, while underscore refers to the second best. (S) and (U) denote *supervised* and *unsupervised* setting. For (S), we present results of label-based stylization; and for (U), we present motion-based stylization.

Effect of λ_{hsa}. Homo-style alignment ensures the style space of the sub-clips from one motion sequence to be close to each other; it is an important self-supervised signal in our approach. Increasing the weight of homo-style commonly helps style modeling (style accuracy) and content preservation (content accuracy, FID), which however also comes with lower diversity. A common observation is that the performance on style and content always contradicts with the diversity. It could be possibly attributed to the inherently limited diversity in our training dataset (Aberman et al., 2020), which is collected by one person performing several styles.

Effect of λ_{kl}. λ_{kl} weighs how much the overall style space aligns with the prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Smaller λ_{kl} usually increases the capacity of the model exploiting styles, which on the other hand deteriorate the performance on content maintenance and diversity.

Effect of λ_{cyc}. Cycle reconstruction constraint plays an important role in unsupervised setting. In supervised setting, strong cycle reconstruction constraint is detrimental to style modeling. In contrast, while learning unsupervisedly, strengthening the cycle constraint enhances the performance on style transferring, and at the same time compromises the preservation of content.

λ_{l1}	λ_{sms}	(Aberman et al., 2020)			(Xia et al., 2015)			
		MPJPE (Recon)↓	Style Acc↑	Style FID↓	MPJPE (Recon)↓	Style Acc↑	Content Acc↑	Content FID↓
0.001	0.001	**39.4**	$\mathbf{0.945}^{\pm 007}$	$\mathbf{0.020}^{\pm 002}$	**62.5**	$\mathbf{0.926}^{\pm 008}$	$\mathbf{0.674}^{\pm 011}$	$\mathbf{0.189}^{\pm 005}$
0.1	0.1	360.1	$0.862^{\pm 010}$	$0.041^{\pm 004}$	431.8	$0.804^{\pm 011}$	$0.589^{\pm 012}$	$0.276^{\pm 007}$
0.01	0.01	180.4	$0.873^{\pm 010}$	$0.041^{\pm 004}$	250.5	$0.830^{\pm 009}$	$0.656^{\pm 012}$	$0.244^{\pm 007}$
0.0001	0.0001	77.6	$0.857^{\pm 010}$	$0.042^{\pm 003}$	130.9	$0.901^{\pm 011}$	$0.661^{\pm 013}$	$0.239^{\pm 007}$

Table 12: Effect of hyper-parameters of autoencoder on the (Aberman et al., 2020) and (Xia et al., 2015) test sets. ± indicates 95% confidence interval. **Bold** face indicates the best result, while underscore refers to the second best. Results of motion-based stylization in supervised setting are presented. MPJPE is measured in millimeter.

Effect of Autoencoder Hyper-Parameters. In Tab. 12, we investigate the impact of autoencoder hyper-parameters (λ_{l1} and λ_{sms}) on both motion reconstruction and stylization performance. Specifically, λ_{l1} encourages sparsity in latent features, while λ_{sms} enforces the smoothness of temporal features. Through experimentation, we identify an optimal set of hyper-parameters with

$\lambda_{l1} = 0.001$ and $\lambda_{sms} = 0.001$, which yields optimal performance in both reconstruction and stylization tasks. Notably, imposing excessive penalties on smoothness and sparsity proves detrimental to the model's capabilities, resulting in lower reconstruction quality. Additionally, we observe a substantial correlation between reconstruction and stylization performance, indicating that better reconstruction often translates to improved stylization.

D IMPLEMENTATION DETAILS

Our models are implemented by Pytorch. Motion encoder $\mathcal{E}$ and decoder $\mathcal{D}$ consists of 2 1-D convolution layers; global motion regressor is a 3-layer 1D convolution network. The content encoder E_c and style encoder E_s are also downsampling convolutional networks, where style encoder contains a average pooling layer before the output dense layer. The spatial dimensions of content and style code are both 512. Detailed model architecture is provided in Figs. 9 and 10. The values of λ_{kld}^{l}, λ_{l1} and λ_{sms} are all set to 0.001, and dimension D_z of $\mathbf{z}$ is 512. During training our latent stylization network, the value of λ_{hsa}, λ_{cyc} and λ_{kl} are (1, 0.1, 0.1) and (0.1, 1, 0.01) in supervised setting and unsupervised setting, respectively.

D.1 MODEL STRUCTURE

The detailed architectures of our motion latent auto-encoder and motion latent stylization model are illustrated in Figure 9 and Figure 10 respectively, where "w/o N", "IN" and "AdaIN" refer to without-Normalization, Instance Normalization and Adaptive Instance Normalization operations (Huang & Belongie, 2017). Dropout and Activation layer are omitted for simplicity.

D.2 DATA PROCESSING

We mostly adopt the pose processing procedure in (Guo et al., 2022a). In short, a single pose is represented by a tuple of root angular velocity, root linear velocity, root height, local joint positions, velocities, 6D rotations (Zhou et al., 2019) and foot contact labels, resulting in 260-D pose representation. Meanwhile, all data is downsampled to 30 FPS, augmented by mirroring, and applied with Z-nomalization.

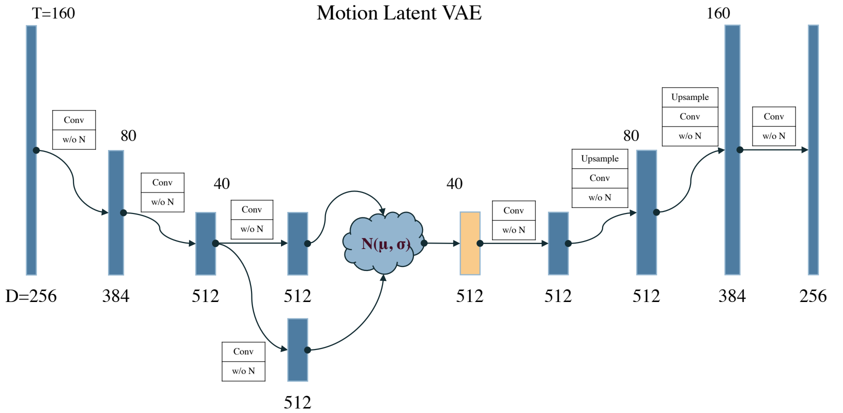


Figure 9: Detailed architecture of our VAE based motion latent model. The AE based latent model keeps only one convolution branch before the latent space. All convolutions, except the last layer of encoder, decoder and generator, use kernel size of 3.

D.3 BASELINE IMPLEMENTATION

For a fair comparison, we adapt the baseline models with minimal changes from their official implementations, training them on the same data splits. More specifically, without violating their design of input representation and networks, all the re-implemented baseline methods strictly load the same preprocessed data for training.

(Aberman et al., 2020). Due to the intentional dual representations for style and content inputs in (Aberman et al., 2020), we make some modifications in the dataloader. We first recover the raw 21-

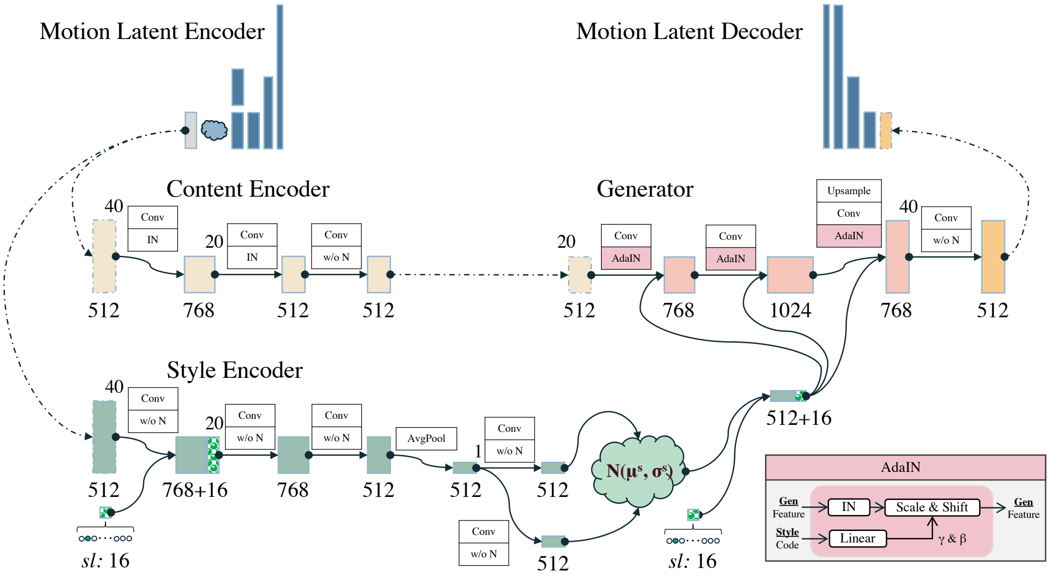
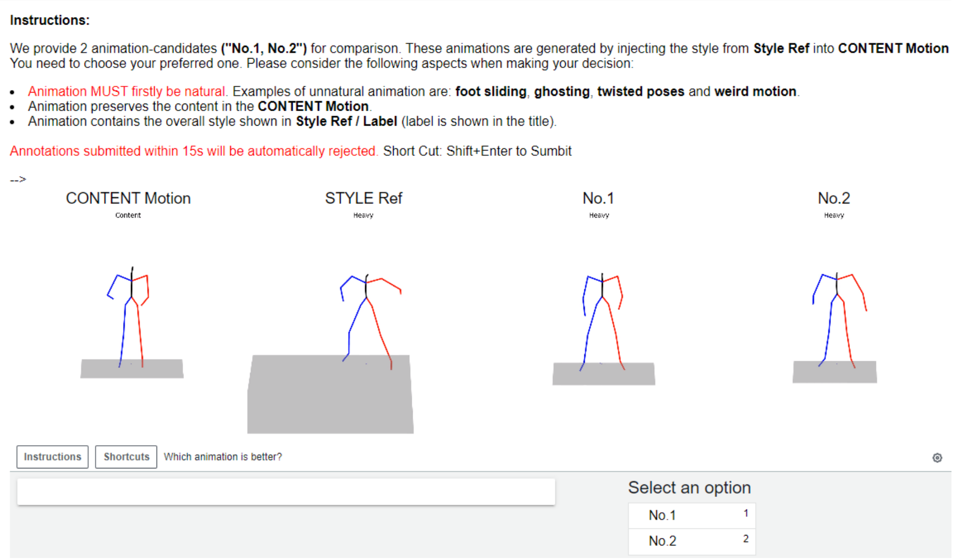


Figure 10: Detailed architecture of our motion latent stylization model in supervised setting. In unsupervised setting, the style label input is dropped. All convolutions, except the last layer of encoders and generator, use kernel size of 3.

joints structural motion data from the preprocessed data, and convert them into 84-D rotation-based content feature along with 4-D global motion, and 63-D position-based style feature, using their motion parsing function. In addition, we modified the channels of network input/output layers to fit the adapted data. Since our experiments solely consider style from 3-D motions, we disable the 2-D branch as well as the related loss functions. However, we suffer from extremely unstable training process and poor results using the same hyper-parameters. It may result from the length extension of motion sequence (now 160 vs original 32) and inherent flaws of GANs (Zhu et al., 2017; Karras et al., 2019). Thus, empirically, we lower the coefficient for adversarial loss α_{adv} from 1 to 0.5, and update the frequency of discriminator training from 1 per-iteration to 0.2 per-iteration.


Instructions:

We provide 2 animation-candidates **("No.1, No.2")** for comparison. These animations are generated by injecting the style from **Style Ref** into **CONTENT Motion**
You need to choose your preferred one. Please consider the following aspects when making your decision:

- Animation MUST firstly be natural. Examples of unnatural animation are: **foot sliding**, **ghosting**, **twisted poses** and **weird motion**.
- Animation preserves the content in the **CONTENT Motion**.
- Animation contains the overall style shown in **Style Ref / Label** (label is shown in the title).

Annotations submitted within 15s will be automatically rejected. Short Cut: Shift+Enter to Sumbit

-->

Which animation is better?

Figure 11: User study interface on Amazon Mechanical Turk.

(Park et al., 2021). We extract 63-D joint position feature and 126-D joint rotation feature from our preprocessed data, catering for the designated dataloader in (Park et al., 2021). Meanwhile, we replace the original 4-D quaternion with the equally functional 6-D rotation (Zhou et al., 2019) without any loss of capability. Their model design is limited to fixed motion length, due to the unscalable linear layer. Therefore, during evaluation on (Xia et al., 2015) test set, we duplicate the sequence to meet the 160-length setting and then extract the corresponding result from the output.

(Jang et al., 2022) takes the motion representation building from per-joint's 6-D position-based feature (i.e. position, velocity) and 6-D rotation-based feature (i.e. upward direction, forward direction) which is almost coherent with our preprocessed data. Thus, we directly re-organize our data to serve the baseline (Jang et al., 2022), keeping everything else unchanged.

E EVALUATION METRIC

Why certain metrics are not used across all datasets? Given our latent stylization models are trained on (Aberman et al., 2020), CMU Mocap (CMU) and (Xia et al., 2015) aims to emphasize zero-shot performance on Style Precision and Content Preservation respectively. Style classifier is trained on (Aberman et al., 2020), where all style motions come from. Compared to (Aberman et al., 2020) and CMU Mocap, (Xia et al., 2015) is quite small (570 clips), comprising variable-length short motions (< 3s). Style FID isn't computed for (Xia et al., 2015) due to substantial length differences between the style motion (from (Aberman et al., 2020), 5.3s) and output motion (< 3s). Content classifier is trained only on (Xia et al., 2015) to evaluate the Content Preservation as content labels are only available on this dataset. Since there is no evidence that this content classifier can generalize to other datasets, we only use it for (Xia et al., 2015).

F USER STUDY

The interface of the user study on Amazon Mechanical Turk for our experiments is shown in Figure 11. Since motion style is not as obvious as other qualitative attributes for common users, to simplify the study, we only compare one baseline result with ours each time. Moreover, for introduction, we briefly explain the concept of motion stylization, presenting the content motion as well as style motion for reference. Users are instructed to choose their preferred results over two generated stylization results based on judgement on *naturalism*, *content preservation* and *style visibility*. This study only involves users that are recognized as **master** by AMT.

G INTERPOLATION

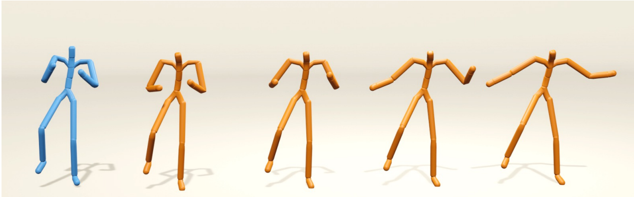
We present the results of interpolation in the respective style spaces learned unsupervisedly Fig. 12(a) and supervisedly Fig. 12(b). We are able to interpolate between styles from different labels in unsupervised setting. Specifically, two style codes are extracted from *sneaky* motion and *heavy* motion respectively. Then we mix these two style codes through linear interpolation, and apply them to stylize the given content motion. In supervised setting, the generator is conditioned on a specific style label. Here, we interpolate styles between two random style codes sampled from the prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Stylization results are produced conditioned a common style label, *heavy*. From Figure 12, we can observe the smooth transitions along the interpolation trajectory of two different style codes. Please refer to our supplementary video for better visualization.

H LIMITATIONS AND FAILURE CASES

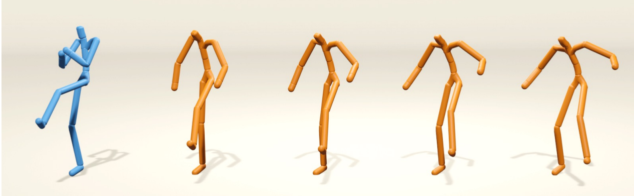
Firstly, our model may encounter difficulties when the input motion substantially deviates from our training data. Figure 13 presents two failed stylization results on rare content actions, *i.e.,* breaking dance and push-up. Given that our model has only seen standing motions during training, it commonly fails to reserve the lower-body movements in these two cases. Interestingly, our model can still retain the general motions of upper-body.

Secondly, the underlying reason for different performance of ours(V) and ours(A) on for example, diversity, style and content accuracy, remains unclear.

Lastly, certain styles are inherently linked to specific content characteristics, particularly within the datasets of (Aberman et al., 2020; Xia et al., 2015). For instance, styles like old, depressed and lazy



(a) Cross-style Interpolation



(b) Homo-style Interpolation

Figure 12: **Style Interpolation. (a)** Cross-style interpolation in unsupervisedly learned style space. Styles are interpolated between style codes of *sneaky* (left) and *heavy* (right) motions. **(b)** Homo-style interpolation in supervisedly learned style space. With style label *heavy* as condition input, styles are interpolated between two style codes that randomly sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. One key pose for each motion is displayed.

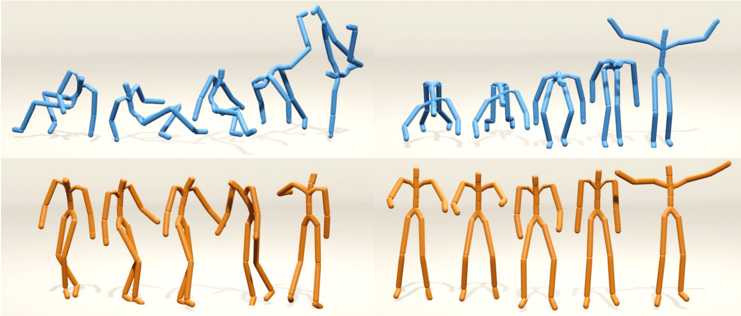


Figure 13: Failure cases. Top row shows content motion; bottom row shows our corresponding results. Stylization results of breaking dance motion (left) and push-up motion (right) using *happy* style label are displayed.

typically relate to slow motions, while happy, hurried, angry motions tend to be fast. As our stylization process aims to preserve content information, including speed, there could be contradictions with these style attributes. For instance, stylizing an slow motion with a hurried style might not yield an outcome resembling a hurried motion. We acknowledge this aspect for potential exploration in future studies.