

# Supplement Material of MoCoDiff: A Controllable Autoregressive Diffusion Model for Expressive Motion Generation

Anonymous CVPR submission

Paper ID 30955

## 1. Overview

In this supplementary material, we first provide additional experiments to evaluate our Controllable Autoregressive Diffusion Model (MoCoDiff) in Sec. 2. Then, we provide a comprehensive explanation of our quantitative metrics in Sec. 3. We further introduce the details of our user study in Sec. 4. Lastly, we provide the details of baseline method modification in Sec. 6.

## 2. Additional Experiments

### 2.1. Importance of Temporal IMC in Diffusion

In this section, we compare the performance of our Temporal Injection Modulation Controller (TIMC) against a traditional fused-conditioning autoregressive (AR) setup, where feature fusion is typically implemented through adaptive normalization layers such as AdaIN [5]. Our TIMC is designed to directly modify the diffusion transition dynamics, formulated as:

$$x_{t-1} = f_{\theta}(x_t, t) + Ct(ht - 1). \quad (1)$$

Here,  $f_{\theta}(x_t, t)$  is the standard denoising operator, while  $Ct(ht - 1)$  provides a time-dependent modulation derived from the terminal state  $h_{t-1}$  of the previous segment. This design makes the denoising trajectory explicitly history-guided, offering a structural mechanism that enforces temporally smooth evolution.

In contrast, AdaIN-style feature injection treat the history feature  $F_H$  as an auxiliary signal concatenated or normalized together with other conditions. Unlike our step-wise modulation term  $Ct(ht - 1)$  that directly shapes the diffusion transition, such fused-conditioning merely augments the model’s input space, making it inadequate for controlling accumulated drift over long sequences.

Table 1 shows the quantitative comparison results. The significantly lower AUJ (From 1.58 to 2.31) and PJ (From 0.27 to 0.48) scores demonstrate that TIMC is superior at suppressing temporal drift and enforcing

Methods	FID↓	Div→	PJ→	AUJ↓
Ours w TIMC	<b>8.03</b> ±0.01	7.75±0.03	<b>0.27</b> ±0.00	<b>1.58</b> ±0.01
Ours w AdaIN	9.21±0.02	<b>7.96</b> ±0.01	0.48±0.00	2.31±0.02

Table 1. **Comparison of methods for adding historical information to diffusion models.** The results show that our MoCoDiff uses TIMC to add historical information constraints, which can better smooth the transition between motions.

smooth transitions compared to merely feeding the history feature into the fused-conditioning pathway. This validates our claim that modifying the diffusion transition dynamics is more effective than standard feature fusion for long-range coherence.

### 2.2. Effect of Dynamic Scheduling $\lambda_h(t)$ in TIMC

In this section, we evaluate the necessity of the dynamic scheduling factor  $\lambda_h(t)$  in TIMC. This factor gradually anneals the influence of historical conditions during the diffusion denoising process, aiming to mitigate the error accumulation commonly observed in autoregressive diffusion models. We compare four different history-conditioning strategies: our dynamic scheduling scheme, a fixed-strong setting ( $\lambda_h = 1.0$ ), a fixed-weak setting ( $\lambda_h = 0.5$ ), and a no-history variant ( $\lambda_h = 0.0$ ). The dynamic schedule follows a piecewise linear formulation: it maintains strong history alignment in the early diffusion steps ( $t \in [0, 10]$  with  $\lambda_h(t) = 1.0$ ), gradually decays within  $t \in [10, 20]$ , and eventually removes the history influence for  $t \geq 20$  ( $\lambda_h(t) = 0.0$ ). Table 2 show that our dynamic scheduling achieves the best performance across most metrics, including AUJ (1.58), PJ (0.27), and FID (8.03), significantly outperforming all constant baselines. In contrast, the fixed-strong variant applies excessive history constraints during the final denoising steps, conflicting with local detail refinement and leading to degraded AUJ and PJ, demonstrating the instability of aggressive AR rollouts. The no-history model performs the worst overall, highlighting the indispensable role of history conditioning in main-

Methods	FID↓	Div→	PJ→	AUJ↓
Fixed-Strong ( $\lambda_h = 1.0$ )	8.25 $\pm$ 0.02	7.50 $\pm$ 0.04	0.40 $\pm$ 0.00	2.05 $\pm$ 0.02
Fixed-Weak ( $\lambda_h = 0.5$ )	8.10 $\pm$ 0.02	7.65 $\pm$ 0.03	0.32 $\pm$ 0.00	1.70 $\pm$ 0.01
No History ( $\lambda_h = 0.0$ )	8.90 $\pm$ 0.03	<b>7.80</b> $\pm$ 0.02	1.79 $\pm$ 0.00	6.96 $\pm$ 0.13
<b>Ours (Dynamic Scheduling)</b>	<b>8.03</b> $\pm$ 0.01	7.75 $\pm$ 0.03	<b>0.27</b> $\pm$ 0.00	<b>1.58</b> $\pm$ 0.01

Table 2. **Evaluation of Dynamic Scheduling Factor  $\lambda_h(t)$  in TIMC.** The results demonstrate the necessity of annealing history influence for optimal temporal coherence and motion quality.

Methods	FID↓	R-Top3↑	SRA↑
Ours w STIMC	<b>5.95</b> $\pm$ 0.01	<b>0.564</b> $\pm$ 0.002	<b>26.37</b> $\pm$ 0.02
Ours w ControlNet	7.51 $\pm$ 0.02	0.431 $\pm$ 0.003	15.38 $\pm$ 0.02
Ours w AdaIN	9.12 $\pm$ 0.01	0.522 $\pm$ 0.004	14.93 $\pm$ 0.03

Table 3. **Comparison of methods for adding style information to diffusion models.** The results show that our MoCoDiff uses STIMC to add style information constraints, which can better express style and preserve content.

taining long-term temporal coherence. The fixed-weak setting improves stability compared to the no-history baseline but remains inferior to dynamic scheduling, indicating that a uniform influence level cannot capture the stage-dependent nature of history reliance in the denoising process. Overall, dynamic scheduling enables strong alignment in the early stages while avoiding over-conditioning in later steps, effectively reducing error accumulation and enhancing long-sequence generation quality.

### 2.3. Importance of Style IMC in Diffusion

In this section, we compare our Style Injection Modulation Controller (STIMC) with two representative conditioning paradigms: ControlNet [12] and AdaIN. STIMC targets high-frequency, pose-level stylistic cues such as rhythm and curvature, and performs style injection through a lightweight, modality-specific Cross-Attention module. The style features  $F_S$  modulate the backbone representation  $\mathcal{X}$  via dedicated key-value projections, enabling selective and content-aware control without interfering with semantic pathways.

ControlNet, in contrast, conditions the model by duplicating U-Net blocks and adding the outputs of a trainable auxiliary branch to the main network. This design is computationally heavy and often imposes overly strong constraints, making the generated motion rigid and disrupting natural temporal transitions. AdaIN performs style transfer by matching feature statistics between content and style, but its global perturbation mechanism lacks the spatial and temporal specificity required for fine-grained motion stylization.

Table 3 shows the quantitative comparison results. By combining residual style injection ( $\hat{X}_t = X_t +$

Classifier-free		FID↓	R-Top3↑	SRA↑
Dropout	Scale			
$p = 0.05$	$\lambda = 1.0$	6.20 $\pm$ 0.02	0.540 $\pm$ 0.003	27.85 $\pm$ 0.03
$p = 0.15$	$\lambda = 1.0$	6.05 $\pm$ 0.01	0.555 $\pm$ 0.002	<b>28.14</b> $\pm$ 0.02
$p = 0.25$	$\lambda = 1.0$	<b>5.95</b> $\pm$ 0.01	0.564 $\pm$ 0.002	26.37 $\pm$ 0.02
$p = 0.35$	$\lambda = 1.0$	6.78 $\pm$ 0.01	<b>0.579</b> $\pm$ 0.003	25.15 $\pm$ 0.02
$p = 0.45$	$\lambda = 1.0$	6.30 $\pm$ 0.02	0.568 $\pm$ 0.003	24.80 $\pm$ 0.03
$p = 0.25$	$\lambda = 0.5$	7.95 $\pm$ 0.02	<b>0.584</b> $\pm$ 0.002	21.63 $\pm$ 0.04
$p = 0.25$	$\lambda = 0.8$	6.50 $\pm$ 0.01	0.520 $\pm$ 0.003	23.50 $\pm$ 0.03
$p = 0.25$	$\lambda = 1.0$	<b>5.95</b> $\pm$ 0.01	0.564 $\pm$ 0.002	26.37 $\pm$ 0.02
$p = 0.25$	$\lambda = 1.5$	6.35 $\pm$ 0.01	0.546 $\pm$ 0.002	27.03 $\pm$ 0.02
$p = 0.25$	$\lambda = 2.0$	6.40 $\pm$ 0.02	0.531 $\pm$ 0.003	<b>27.54</b> $\pm$ 0.02

Table 4. **Evaluation of the classifier-free parameters dropout  $p$  and scale  $\lambda$ .** For balanced performance, we ultimately choose  $p=0.25$  and  $\lambda=1.0$ .

$O_{sty}^{(t)}$ ) with cross-attentive modulation, STIMC avoids the rigidity of ControlNet and the coarse global shifts of AdaIN. This results in more precise, disentangled, and stable style control, achieving higher stylistic fidelity without compromising structural or semantic coherence.

### 2.4. Classifier-free Parameters

In this section, we conduct experiments on the parameters of classifier-free diffusion guidance [6]. As shown in Table 4, we initially change the dropout ratio ( $p$ ) of the training none-style model from 0.05 to 0.45 and observe an upward trend in the FID and R-Top3 metrics, while the SRA metric shows a downward trend. We also experiment with adjusting the guidance scale ( $\lambda$ ) from 2.0 to 0.5 and observe the same trend. We further validate the control of style using classifier-free guidance through the selection of scales during inference, as shown in Fig. 1. The visualization supports the effectiveness of classifier-free guidance in achieving user-friendly style transfer. To maintain balanced performance, we ultimately choose  $p=0.25$  and  $\lambda=1.0$ .

### 2.5. Importance of the MotionCLIP as Our Style Encoder

In this section, we further investigate the reasons behind our choice of MotionCLIP [10] over a simple pre-trained AutoEncoder as our Style Extractor. The training process of MotionCLIP involves aligning the motion latent space with the text and image CLIP [9] spaces through the use of text loss and image loss. In the specific implementation process, this study first constructs an autoencoder model pre-trained on the HumanML3D dataset [4]. This model employs a multi-layer Transformer architecture and introduces a skip connection mechanism to enhance encoding capabilities. Compared with traditional autoencoders, this model combines an action reconstruction loss and a CLIP model alignment

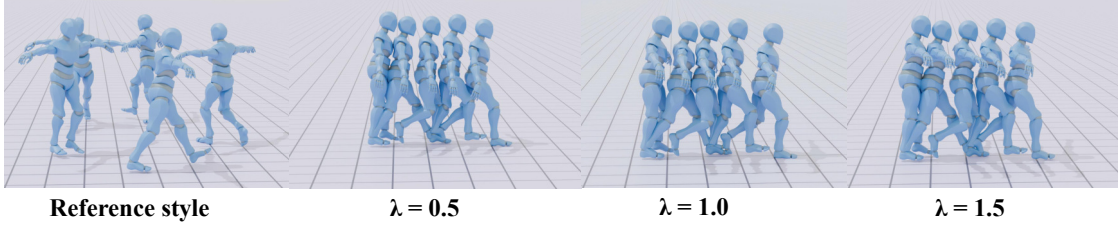


Figure 1. **Results of style control.** The results show that increasing the scale from 0.5 to 1.5 gradually strengthens the style, while further increases have minimal impact on style strength.

Methods	FID↓	R-Top3↑	SRA↑
Ours w AE	6.37±0.03	0.548±0.002	23.21±0.03
Ours w MotionCLIP	<b>5.95±0.01</b>	<b>0.564±0.002</b>	<b>26.37±0.02</b>

Table 5. **Comparison with variants of Style Encoder (AE, MotionCLIP).** The results show that using MotionCLIP as our Style Encoder enables better capturing of style features.

loss in its loss function design, significantly improving the accuracy of capturing action style features. In order to assess the importance of this alignment for our model, we retrain a MotionCLIP variant without text loss and image loss, which essentially functions as a basic AutoEncoder (AE). The results, as presented in Table 5, demonstrate a decrease in the SRA metric. This decrease indicates that MotionCLIP, which aligns with the Clip space of text and images, can more effectively extract style features, thereby enhancing the effectiveness of our method.

## 2.6. Experiments on Diverse Datasets

To further test the generalizability of our MoCoDiff, we conduct experiments on motion data in BVH format. Specifically, we train our MoCoDiff on the CMU [3] dataset and test it on the Xia [11] dataset. Since BVH format data does not have an existing MotionCLIP as our Style Encoder, we use a pretrained VAE encoder. We also compared our MoCoDiff with Motion Puzzle [7], which uses the same dataset.

As shown in Table 7, our MoCoDiff outperforms motion puzzle in terms of FID and R-Top3, which proves the high quality of our generated motion and the high degree of content preservation. Moreover, our AUJ metric is also higher than the Motion Puzzle, further validating the superiority of our MoCoDiff of treating trajectories as an additional learnable condition. This approach ensures more natural motion and avoids signif-

Methods	Content Preservation	Style Performance	Motion Smoothness
AutoMDM+PersonaBooth [8]	3.93±0.67	3.73±0.62	3.92±0.51
ControlNet+FlowMDM [2]	3.88±0.69	3.11±0.64	3.64±0.65
AutoMLD+SMooDi [13]	3.92±0.55	3.87±0.61	3.69±0.64
Ours	<b>4.27±0.49</b>	<b>4.49±0.21</b>	<b>4.38±0.30</b>

Table 6. **User study.** The results show that our MoCoDiff outperforms other methods in terms of content preservation, style performance and Motion Smoothness

Figure 2. **User interface in our user study.** Users need to rate each stylized motion on three metrics: Realism, Content Preservation, and Style Performance.

icant foot-sliding artifacts. Regarding the low performance on style metrics, we speculate that this is due to the VAE encoder’s inability to encode style features ef-

163  
164  
165

Methods	FID↓	R-Top3↑	SRA↑	AUJ↓
AdaIN+Motion Puzzle [7]	10.75±0.05	0.466±0.007	27.41±0.03	2.41±0.02
Ours	<b>8.93±0.05</b>	<b>0.512±0.006</b>	21.11±0.03	<b>0.69±0.03</b>

Table 7. **Evaluation on Xia [11] test set.** The results show that our MoCoDiff outperforms Motion Puzzle in terms of FID, R-Top3, and AUJ metrics. Our lower performance of the SRA metric might be the inappropriate style encoder.

fectively. It might be necessary to train a MotionCLIP equivalent with BVH format data, but this presents challenges with the current BVH datasets due to the lack of a large amount of annotated data.

### 3. Details of Quantitative Metrics

In this section, we provide the detailed evaluation protocol for our six quantitative metrics, including FID, R-Top3, SRA, Diversity, PJ, and AUJ. For the FID and R-Top3 metrics, we first extract high-level motion embeddings using a pretrained motion encoder from HumanML3D, enabling us to measure both realism and semantic consistency. During evaluation, we encode the generated motions and the real motions into the same latent feature space, and compute the Fréchet distance between their Gaussian distributions to obtain the FID score, which reflects global motion realism. In parallel, we perform a cross-modal retrieval experiment for R-Top3 by encoding both the input condition (e.g., text) and the generated motion, and computing their similarity; we then check whether the correct motion is retrieved within the top-3 candidates, which measures semantic alignment.

For SRA and Diversity, we evaluate the rhythmic and stylistic fidelity as well as the variation across generated samples. Specifically, SRA is computed by measuring the correlation between the temporal energy curves of generated motions and their style references, thereby assessing how well rhythmic and pose-level periodic patterns are preserved. During evaluation, we randomly select a content text from the HumanML3D dataset and a motion style sequence from the 100Style dataset to generate the stylized motion. A pre-trained style classifier is then used to compute the SRA for the generated motion, providing an objective measure of style reflection. This multi-faceted evaluation approach ensures a robust assessment of both the fidelity and creativity of the motion generation process. Diversity is evaluated by computing the average pairwise distance among the feature embeddings of multiple generated samples, reflecting the richness and spread of the generative distribution.

Finally, for PJ and AUJ, we quantify both short-term and long-term temporal stability. PJ measures local jitter by computing second-order frame differences of joint positions, capturing high-frequency instability. AUJ fur-

ther evaluates accumulated unnatural motion by computing the third-order temporal derivative (jerk) over the entire sequence, providing a sensitive measure of long-range smoothness and autoregressive error accumulation. Together, these six metrics offer a comprehensive quantitative assessment of realism, semantic consistency, stylistic accuracy, diversity, and temporal stability for long-horizon motion generation.

### 4. User Study Details

This section details our user study. We designed and collected questionnaires using the Wenjuanxing website (Wenjuanxing [1]). Our user study included six unique long-sequence style transfer combinations. Each combination corresponded to different text and style actions, all of which were subsequently converted into video format. Participants are asked to rate results generated by these methods on a scale of 1 (significantly inaccurate) to 5 (significantly accurate), based on three metrics: (1) Content Preservation: the level of the stylized motion to preserve the content information from content motion, (2) Style Performance: the level of the stylized motion to perform the style features from style motion, and (3) Motion Smoothness: the level to smoothly transition between different motions when generating long-sequence motions. Figure 2 shows the user interface we designed, where users rated each stylized action in a specific style transfer group.

As for user background, the study involves 40 participants of various backgrounds, including 25 students, 2 sales staff, 4 production workers, 3 teachers, and 6 individuals of other professions. Among them, there are 28 male users and 12 female users, including 2 under 18 years old, 28 between 18 and 25 years old, 6 between 26 and 30 years old, and 4 over 30 years old.

As shown in Table 6, our method consistently achieves the highest scores across all three evaluation metrics: Content Preservation, Style Performance, and Motion Smoothness. To further verify the statistical significance of these differences, we conduct a one-way ANOVA test on the user study ratings. The results reveal substantial between-group variations for Content Preservation ( $F = 4.48$ ,  $p < 0.01$ ), Style Performance ( $F = 59.76$ ,  $p < 0.01$ ), and Motion Smoothness ( $F = 17.89$ ,  $p < 0.01$ ). Post-hoc pairwise comparisons indicate that our method is significantly preferred over all baselines across the three metrics (all  $p < 0.01$ ). These statistical results strongly confirm the superior perceptual quality of our approach in long-sequence stylized motion generation.



## 5. Implementation Details of Baselines

In this section, we detail how we modified the baseline method to adapt it to our long-temporal stylized motion generation task.

### 5.1. AutoMLD+SMooDi

To enable long-horizon stylized motion generation using SMooDi, we extend the original model—originally designed for single text–style pairs and fixed-length outputs—by introducing an autoregressive segmented generation framework together with enhanced conditioning strategies. First, we incorporate a historical-frame conditioning module that stores the last  $N_h$  ( $N_h = 5$ ) frames of the previously generated segment and embeds them through a dedicated linear layer before concatenating them with the current latent variable. This modification ensures temporal continuity between adjacent segments and prevents abrupt motion transitions. Second, we implement an iterative sequence synthesis procedure in which each segment is generated sequentially by diffusing and decoding its latent variable while continuously updating the historical-frame buffer; this allows the model to synthesize arbitrarily long sequences  $\mathbf{m}_1, \dots, \mathbf{m}_K$  by recursively setting  $\mathbf{h}_i = \mathbf{m}_i[-N_h:]$  and concatenating all decoded outputs. Third, we expand the original dual-conditional guidance into a three-way guidance mechanism during the diffusion reverse process. The predicted noise is computed by combining unconditional, style-conditioned, and text-conditioned predictions, weighted by  $w_s = 1.5$  and  $w_t = 7.5$ , respectively, and further supplemented by a style-gradient term derived from a pretrained style classifier. This additional gradient—applied only in the early denoising steps—encourages the generated motion to match the reference style in the classifier’s intermediate feature space using an L1-based loss across multiple feature layers. Together, these modifications allow SMooDi to support long-sequence synthesis with segment-level style control, improved temporal smoothness, and more reliable style preservation across extended motion durations.

### 5.2. ControlNet+FlowMDM

To enable long-sequence stylized motion generation, we systematically extend the FlowMDM model and propose ControlNet+FlowMDM. Our approach retains the original FlowMDM backbone with its Blended Positional Encodings (BPE) while introducing a ControlNet-inspired style control branch. Specifically, a Transformer-based style encoder extracts style features from reference motions, producing 256-dimensional embeddings, which are further processed by a zero-initialized convolutional control branch to generate

multi-layer, fine-grained control signals. During style encoding, segment-wise shuffling (16 frames per segment) and length-aware processing are applied to enhance robustness and prevent overfitting to temporal order. Style conditions are globally injected into noisy motion features during training, whereas during inference, multiple style references can be combined flexibly to achieve long-sequence stylization. The BPE mechanism works in conjunction with style control: the APE stage (absolute positional encoding) restores global coherence and restricts attention within subsequences, while the RPE stage (relative positional encoding) employs a local attention window to produce smooth transitions, ensuring both local fluidity and global consistency. Multi-layer control signals are independently processed through zero-convolution layers and passed to the backbone network to realize hierarchical style modulation. To preserve FlowMDM’s pretrained capabilities, only the style control branch is trained while the backbone parameters are frozen, enabling long-sequence stylized motion generation without compromising the original generation quality.

## 6. Network Architecture

In order to enable our method to be successfully reproduced, we elaborate our MoCoDiff network structure in Table 8.

## References

- [1] Wenjuanxing. <https://www.wjx.cn/>. 4
- [2] German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 457–469, 2024. 3
- [3] CMU. Carnegie-mellon mocap database. <http://mocap.cs.cmu.edu/>. 3
- [4] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 2
- [5] Yunhui Guo, Chaofeng Wang, Stella X Yu, Frank McKenna, and Kincho H Law. Adaln: a vision transformer for multidomain learning and predisaster building information extraction from images. *Journal of Computing in Civil Engineering*, 36(5):04022024, 2022. 1
- [6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2
- [7] Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics*, 41(3):1–16, 2022. 3, 4
- [8] Boeun Kim, Hea In Jeong, JungHoon Sung, Yihua Cheng, Jeongmin Lee, Ju Yong Chang, Sang-Il Choi,

Component	Network Details
Text Encoder	Frozen CLIP ViT-B/32 TransformerEncoderLayer(d_model=256, num_heads=4, dim_feedforward=2048) $\times$ 4 LayerNorm(normalized_shape=256)
Timestep Embedder	PositionalEncoding(d_model=512, max_len=5000) Linear(in_features=512, out_features=2048) Mish() Linear(in_features=2048, out_features=512)
History Encoder	Linear(in_features=263, out_features=256) ReLU(inplace=True)
Style Encoder	Frozen MotionCLIP Encoder Linear(in_features=512, out_features=256)
UNet Encoder	<b>Down Block 1:</b> Conv1D(in_channels=263, out_channels=512, cond_dim=256, time_dim=512) + CrossAttention(num_heads=8, dropout=0.1) Conv1DBlock(512 $\rightarrow$ 512, cond_dim=256, time_dim=512) ResidualTemporalBlock Downsample1d(dim=512) <b>Down Block 2:</b> Conv1DBlock(512 $\rightarrow$ 1024, cond_dim=256, time_dim=512) ResidualTemporalBlock + CrossAttention Downsample1d(dim=1024) <b>Down Block 3:</b> Conv1DBlock(1024 $\rightarrow$ 2048, cond_dim=256, time_dim=512) ResidualTemporalBlock + CrossAttention Downsample1d(dim=2048) <b>Down Block 4:</b> Conv1DBlock(2048 $\rightarrow$ 4096, cond_dim=256, time_dim=512) ResidualTemporalBlock + CrossAttention Downsample1d(dim=4096)
UNet Middle	Conv1DBlock(4096 $\rightarrow$ 4096, cond_dim=256, time_dim=512) ResidualTemporalBlock + CrossAttention(num_heads=8)
UNet Decoder	<b>Up Block 1:</b> Upsample1d(4096 $\rightarrow$ 2048) ConvTranspose1d(4096, 2048, 4, 2, 1) Conv1DBlock(2048 $\rightarrow$ 2048, cond_dim=256) <b>Up Block 2:</b> Upsample1d(2048 $\rightarrow$ 1024) Conv1DBlock(1024 $\rightarrow$ 1024, cond_dim=256) <b>Up Block 3:</b> Upsample1d(1024 $\rightarrow$ 512) Conv1DBlock(512 $\rightarrow$ 512, cond_dim=256) <b>Up Block 4:</b> Upsample1d(512 $\rightarrow$ 263) Conv1DBlock(263 $\rightarrow$ 263, cond_dim=256)
Output Layer	Conv1d(in_channels=263, out_channels=263, kernel_size=1, bias=False) (zero-initialized)
CrossAttention	LayerNorm(latent_dim=263/512/1024/2048/4096) Linear(query/key/value projection) Multi-Head Attention (num_heads=8) Dropout=0.1

Table 8. **Architecture of our method.** Detailed network architecture including encoders, UNet backbone, output layers, and conditioning modules.

- 363 Younggeun Choi, Saim Shin, Jungho Kim, et al. Per-  
364 sonabooth: Personalized text-to-motion generation. In  
365 *Proceedings of the Computer Vision and Pattern Recog-  
366 nition Conference*, pages 22756–22765, 2025. 3
- 367 [9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya  
368 Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sas-  
369 try, Amanda Askell, Pamela Mishkin, Jack Clark, et al.  
370 Learning transferable visual models from natural lan-  
371 guage supervision. In *International Conference on Ma-  
372 chine Learning*, pages 8748–8763, 2021. 2
- 373 [10] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano,  
374 and Daniel Cohen-Or. Motionclip: Exposing human mo-  
375 tion generation to clip space. In *Proceedings of the Eu-  
376 ropean Conference on Computer Vision*, pages 358–374,  
377 2022. 2
- 378 [11] Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica  
379 Hodgins. Realtime style transfer for unlabeled heteroge-  
380 neous human motion. *ACM Transactions on Graphics*,  
381 34(4):1–10, 2015. 3, 4
- 382 [12] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala.  
383 Adding conditional control to text-to-image diffusion  
384 models. In *Proceedings of the IEEE/CVF international  
385 conference on computer vision*, pages 3836–3847, 2023.  
386 2
- 387 [13] Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun,  
388 and Huaizu Jiang. Smoodi: Stylized motion diffusion  
389 model. In *European Conference on Computer Vision*,  
390 pages 405–421. Springer, 2024. 3