Supplementary Material
Controllable Long-term Motion Generation with Extended Joint Targets

A. Dataset Details

COMET is trained on two complementary motion-capture datasets: AMASS and CIRCLE. AMASS provides diverse, long-horizon behaviors essential for modeling natural navigation, while CIRCLE offers densely annotated, contact-rich reaching motions. Both datasets are converted into a unified SMPL-X representation.

A.1. AMASS

The AMASS dataset [8] serves as the primary source for learning general human motion patterns. It is a large-scale repository containing over 17,000 sequences spanning a broad range of movements, with extensive coverage of locomotion behaviors. This diversity is crucial for training COMET to synthesize natural navigation dynamics, particularly when the character must traverse large distances to reach a target. Since AMASS lacks explicit goal annotations, we introduce a *pseudo-goal* strategy by randomly selecting a future frame and using its ground-truth joint position as the target. This enables COMET to acquire goal-directed behaviors from unannotated sequences. To ensure quality and consistency, we remove clips with excessive foot lifting and uniformly resample all motions to 30 frames per second (fps).

A.2. CIRCLE

The CIRCLE dataset [1] provides task-specific data for learning fine-grained goal-reaching behaviors. It contains approximately 10 hours of motion capture data, comprising more than 7,000 sequences from five subjects performing whole-body reaching tasks with both the right and left hands. The motions include a wide variety of reaching-related actions, such as bending, crawling, crouching, and kneeling, performed within a static environment. Unlike AMASS, CIRCLE offers explicit annotations of initial and target conditions, which are critical for teaching COMET precise upper-body and arm movements necessary for accurate interaction with target locations.

A.3. Preprocessing

All motion clips undergo a shared preprocessing pipeline. First, we compute joint positions for translation-dependent features and to re-ground each sequence so that the lowest joint touches z=0. Second, sequences are trimmed into 8 second windows that match the temporal receptive field of the transformer modules. Third, we compute dataset-wide statistics for every feature channel (root motion, joint rotations, goal descriptors, and joint coordinates), which are later used to standardize the inputs at training and inference time. Finally, the combined motion dataset is randomly partitioned into disjoint splits with an 80/10/10 ratio for training, validation, and testing. This pipeline guarantees that heterogeneous sources provide consistent supervision for COMET's goal-aware motion synthesis.

Table 1. Key architectural specifications of COMET.

Component	Specification
Latent space dimensionality (z)	64
MLP layers (pose + delta features)	16
Regularization	LayerNorm + Dropout
Conditional intention embedding	Linear layer (1 layer)
Transformer layers	4
Model dimension	64
Attention heads	8
Feed-forward network dimension	64
Positional encoding	Sinusoidal

B. Implementation Details

B.1. Model Architecture

The COMET architecture is built upon a , where both the encoder and decoder are realized using Transformer Encoder modules. The latent space is configured with a dimensionality of $z = 64$. Input motion features, including pose and delta (δ) representations, are first processed through a 16-layer Multi-Layer Perceptron (MLP). Simultaneously, the conditional intention vector, which provides task-specific guidance, is projected via a single linear layer before integration into the main Transformer pipeline. The Transformer modules employ standard sinusoidal positional encoding and are configured as detailed in Table 1.

This design enables COMET to effectively model complex, temporally coherent motion sequences while maintaining computational efficiency suitable for real-time generation.

B.2. Training Details

The COMET model was trained using the PyTorch on a workstation equipped with four NVIDIA A6000 GPUs. The full training process required approximately 36 hours to complete a total of 1,800 epochs. We used a batch size of 512 and maintained a fixed learning rate of 1×10^{-4} throughout training. To mitigate exposure bias, which commonly arises in autoregressive sequence generation, we employed a **scheduled sampling** strategy. This strategy was activated at epoch 10, after which the number of autoregressive (AR) steps was progressively increased over subsequent epochs until epoch 50, where it was capped at a maximum of 10 AR steps. This gradual scheduling allowed the model to smoothly transition from relying on ground-truth inputs to generating longer sequences based on its own predictions, thereby improving stability and long-horizon motion synthesis.

C. Reference-guided Feedback Details

To model the distribution of plausible human poses, we train a Gaussian Mixture Model (GMM) using walking motion data from the training set. The GMM is configured with 50 mixture components and optimized using the Expectation-Maximization (EM) algorithm, with a maximum of 1,000 iterations to ensure convergence. For our benchmark model, training typically completes in under one minute.

Algorithm 1 outlines the full Reference-Guided Feedback (RGF) process. At each timestep, COMET predicts a delta pose that updates the current state. When feedback is active, the predicted joint rotations and pelvis height are softly corrected toward the closest GMM component mean, as determined by the Mahalanobis distance. This correction step (lines 10–12) nudges the motion toward the learned manifold of natural human poses, mitigating drift and preserving long-term stability. The feedback loop is deactivated once the character enters the vicinity of the goal, allowing for precise final adjustments without interference from the reference signal.

C.1. Choice of GMM Components

To determine the appropriate number of Gaussian components K for the Reference-Guided Feedback (RGF) module, we conducted an experiment by varying K over a wide range. As shown in Table 2, performance saturates beyond $K = 50$ and slightly declines at $K = 100$. Therefore, we select $K = 50$ as the setting for our benchmark model.

C.2. Feedback Scale α

The feedback scale parameter α is a critical hyperparameter in COMET, as it controls the strength of the reference-guided feedback applied to the generated motion. The value of α directly influences performance. When α is too small, the corrective signal becomes weak, leading to insufficient error correction and reduced motion stability. In this case, the generated poses may drift away from the learned distribution of plausible human motions. Conversely, an excessively large α can cause the model to overfit to specific reference poses, resulting in mode collapse and degraded motion quality, such as exaggerated foot skating artifacts. Through empirical evaluation, we found that $\alpha = 0.01$ pro-

Algorithm 1 Reference-Guided Feedback (RGF)

```
Require: COMET decoder $p_\theta$, initial pose $\mathbf{p}_1$, control joints $J_c$,
      joint goals $\{\mathbf{G}_j\}_{j \in J_c}$, GMM params $\{(\mu_k, \Sigma_k)\}_{k=1}^{K}$,
      feedback scalar $\alpha$, sequence length $T$, stop distance $d_{\text{goal}}$
 1: Initialize sequence $\hat{\mathcal{M}} \leftarrow \{\mathbf{p}_1\}$,    feedback_active ← True
 2: for $i = 1, \ldots, T-1$ do
 3:     Update intentions $\mathbf{I}_i$
 4:     Predict delta:
        $\hat{\delta}_{i+1} \leftarrow p_\theta(\delta_{i+1} | \mathbf{p}_i, \mathbf{I}_i, \mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}))$
 5:     Pose update: $\hat{\mathbf{p}}_{i+1} \leftarrow \mathbf{p}_i + \hat{\delta}_{i+1}$
 6:     if feedback_active then
 7:         Extract features: $\hat{\mathbf{f}}_{i+1} \subset \hat{\mathbf{p}}_{i+1}$
 8:         Select closest GMM component:
                $k^* = \arg\min_k \sqrt{(\hat{\mathbf{f}}_{i+1} - \mu_k)^{\mathrm{T}} \Sigma_k^{-1} (\hat{\mathbf{f}}_{i+1} - \mu_k)}$
 9:         Correct features:
                $\hat{\mathbf{f}}_{i+1}^{\text{corrected}} = \hat{\mathbf{f}}_{i+1} + \alpha(\mu_{k^*} - \hat{\mathbf{f}}_{i+1})$
10:         Update pose with corrected features:
                $\hat{\mathbf{p}}_{i+1} = \text{Update}\big(\hat{\mathbf{p}}_{i+1}, \hat{\mathbf{f}}_{i+1}^{\text{corrected}}\big)$
11:         if $\|\mathbf{P}_{i+1}^{xy} - \mathbf{G}_{\text{avg}}^{xy}\|_2 < d_{\text{goal}}$ then
12:             feedback_active ← False
13:         end if
14:     end if
15:     $\hat{\mathcal{M}} \leftarrow \hat{\mathcal{M}} \cup \{\hat{\mathbf{p}}_{i+1}\}$
16: end for
17: return generated sequence $\hat{\mathcal{M}}$
```

Table 2. Ablation on the number of Gaussian components K for the Reference-Guided Feedback (RGF).

# Components K	SR (%) (↑)	FS (%) (↓)	DTG (cm) (↓)
0	31.47	18.97	59.25
3	49.20	13.62	21.00
5	51.36	13.63	21.23
10	51.25	13.93	21.44
30	49.44	**13.59**	23.16
50	**52.67**	13.95	**20.30**
100	51.73	14.03	21.68

vides the best balance between stability and naturalness, delivering consistent improvements across all benchmarks. For the motion in-betweening task, we use a higher value of $\alpha = 0.05$, where the reference pose is fixed to the given target pose to ensure precise convergence to the final keyframe.

C.3. Feedback Stopping Distance

In COMET, the reference-guided feedback is applied during the main transit phases of motion but is intentionally deactivated as the model approaches its target. This design prevents interference with the final fine-grained adjustments needed to accurately align the controlled joints with their target positions. Persisting with feedback near the goal can be counterproductive, as it may push the motion toward the closest reference pose rather than the precise target configuration. We define the feedback stopping distance as the threshold at which feedback is disabled. Empirically, we determined that a threshold of 1 meter yields the optimal trade-off, ensuring both accurate goal attainment and natural, stable final postures.

D. Evaluation Details

D.1. Single-Joint Control (Right Wrist)

The single-joint control experiment assesses the model's ability to generate motion from unseen initial poses towards specific target goals. This setup follows the evaluation methodology employed in WANDR [3], utilizing goals uniformly distributed within a 3D cylindrical space. This approach allows for an effective validation of the model's generalization capabilities to inputs not encountered during training. The specific experimental parameters are as follows: (1) Angle, 360 degrees divided into 5 uniform increments (i.e., 72-degree intervals). (2) Height, the range from 0.5m to 1.8m, divided into 5 uniform increments. (3) Distance, the range from 0.5m to 5m, divided into 5 uniform increments. (4) Initial Pose: 6 distinct unseen initial poses are used. (5) For each combination of the above parameters, 5 trials are sampled. The total number of cases is $5 \times 5 \times 5 \times 6 \times 5 = 3750$. The duration for all generated motions is set to 8 seconds. In this setup, the orientation intention is defined as the direction from the current pelvis position towards the target goal.

D.2. Multi-Joint Control

To evaluate COMET's ability to control multiple joints simultaneously, we sampled random target poses from the held-out test set. Target positions for the pelvis in the XY plane were selected from five discrete directions and five distance settings, following the same protocol as in the single-joint control evaluation. The vertical (Z) position was implicitly determined by the selected final pose. We used six unseen final target poses, with each paired with six different initial poses. For each unique target–initial pairing, five trials were conducted, providing a comprehensive evaluation across diverse joint configurations. Metrics were first computed individually for each controlled joint and then averaged to report the final performance. Each motion sequence was fixed to 8 seconds in duration.

D.3. Long-term Sequential Target Control

To evaluate the capacity for sustained long-distance locomotion and its adaptability to dynamically shifting objectives, we introduce a method for evaluation, navigating towards a sequence of three consecutive target goals, each positioned at a fixed right wrist height of 1.0 meter. Each segment of the sequence requires covering a distance of 5.0 meters. The evaluation is conducted using 6 distinct initial poses, randomly selected from the test set. For each initial pose, the motion generation involves selecting one of 5 uniform directions for each of the three consecutive target segments. This methodology yields $5^3 = 125$ unique locomotion paths per initial pose, resulting in a total of $6 \times 125 = 750$ distinct test cases. For each test case, 5 trials are sampled, bringing the total number of evaluated motions to $750 \times 5 = 3750$. The duration for each of the three motion segments is set to 8 seconds.

D.4. Evaluation for Motion In-betweening

For the motion in-betweening task, we follow the evaluation protocol introduced by Harvey et al. [4], with several key modifications. While the original study evaluated models only on fixed-length intervals, our evaluation considers in-betweening over arbitrary intervals up to the maximum sequence length supported by the models. In each trial, the first and last keyframes of the interval are provided to the models as boundary conditions. For CondMDI [2], we ensured fairness by training it exclusively on scenarios where the first and last frames are given, rather than allowing it to learn arbitrary frame predictions. For DNO [6], we adopted the ODE step size of 100 as reported in the original paper and generated outputs using both 100-step and 300-step optimization settings for comparison. Both CondMDI [2] and DNO [6] were reimplemented and retrained under identical conditions using the same training dataset as COMET. Furthermore, their generation window was set to 240 frames (8 seconds) to match the temporal receptive field used during COMET's training.

D.5. Evaluation for Plug-and-Play Motion Stylization

As stylization is difficult to quantify using conventional numeric metrics, classifier-based measures such as Style Recognition Accuracy (SRA) first introduced in Motion Puzzle [5] have been also adopted in MoST [7] and

(a) Success Rate (SR, %) (↑)

# Joints	Pelvis	L. Ankle	R. Ankle	Head	L. Wrist	R. Wrist	Mean
1	94.80	66.93	64.13	86.53	66.13	51.20	71.62
2	94.06	65.39	54.42	74.70	72.39	61.53	70.41
3	96.38	68.98	52.64	69.78	72.45	66.73	71.16
4	94.80	66.49	51.29	64.03	72.82	67.07	69.42
5	95.11	62.80	49.36	66.86	73.42	64.94	68.75
6	95.33	59.87	49.47	69.60	73.07	65.87	68.87

(b) Foot Skate (FS) (↓)

# Joints	Pelvis	L. Ankle	R. Ankle	Head	L. Wrist	R. Wrist	Mean
1	15.94	13.74	14.32	13.80	12.07	12.33	13.70
2	15.27	15.05	15.24	14.82	14.73	14.52	14.94
3	15.26	15.29	15.23	14.89	14.75	14.84	15.04
4	15.29	15.49	15.36	15.41	15.27	15.32	15.36
5	15.46	15.54	15.47	15.50	15.44	15.49	15.48
6	15.32	15.32	15.32	15.32	15.32	15.32	15.32

(c) Distance-to-Goal (DTG, cm) (↓)

# Joints	Pelvis	L. Ankle	R. Ankle	Head	L. Wrist	R. Wrist	Mean
1	3.30	12.56	12.84	8.81	26.41	34.02	16.32
2	4.16	11.94	16.89	9.60	11.30	16.15	11.67
3	3.02	10.08	15.42	9.40	9.11	11.83	9.81
4	3.45	9.98	14.14	10.34	9.19	11.61	9.79
5	3.16	10.17	13.60	9.47	8.63	11.37	9.40
6	2.67	10.57	13.36	8.17	8.33	11.37	9.08

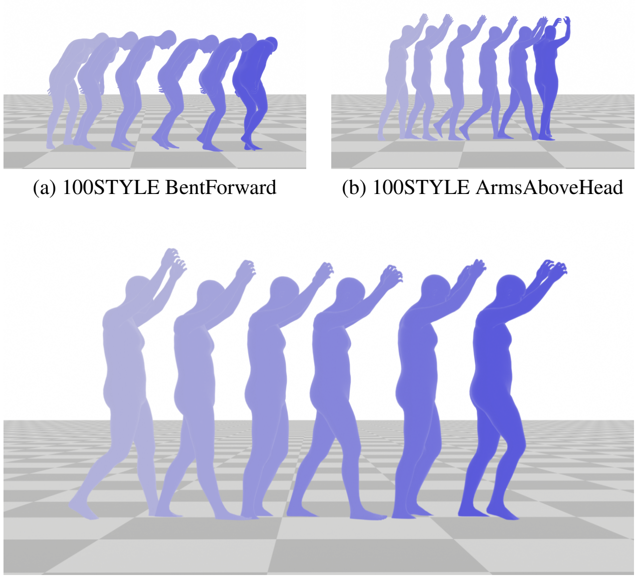
Figure 1. Evaluation results for multi-joint control with varying numbers of simultaneously controlled joints.

SMooDi [13]. However, these classifier-based metrics are fundamentally misaligned with the design of COMET, limiting their reliability as evaluators. Instead, stylization was assessed qualitatively through a user-preference study conducted on domain experts, including motion-capture specialists and professional animators.

The reliability of classifier-based metrics in motion stylization, namely SRA [5], acquires reliability based on two assumptions: (1) training and evaluation use labeled data sets that share a common classification scheme, ensuring results are compared under the same classification criterion; (2) because both the model and the metric depend on the same scheme, even if one trains on a labeled data set different from the one used in the evaluation, the model tends to achieve similar scores, providing reproducibility. Motion Puzzle [5], MoST [7] and SMooDi [13] follow this setting and, in that context, appropriately report quantitative results with SRA. In contrast, COMET's Reference-Guided Feedback does not rely on a pre-defined label taxonomy. It rapidly learns style motions composed of the required actions and injects them in a plug-and-play manner, reducing dependence on large labeled datasets and improving practical usability. Moreover, in motion data the conceptual separation between content and style is often inherently ambiguous or mutually entangled, so the chosen classification scheme can strongly influence classifier-based metrics. Consequently, such metrics are valid for comparing models that share the same scheme but are not suitable for evaluating COMET, which operates independently of dataset-specific style taxonomies. Additionally, since COMET operates without a content-labeling scheme, it is incompatible with metrics that presuppose content classification, such as Style Consistency (SC) [10]. Therefore, stylization evaluation relies on a user study that qualitatively compares the results obtained with the same style input motion to SMooDi.

E. User Preference Study

While we make extensive efforts to quantitatively evaluate motion quality using established metrics, these measures may not fully capture perceptual differences as judged by humans. To address this gap, we conducted a user preference study to directly assess the perceptual quality of our generated motions. A total of 30 participants with professional experience in motion-related fields, including motion capture engineers, 3D artists, and character animators, were recruited to ensure the reliability of the subjective evalua-



(a) 100STYLE BentForward

(b) 100STYLE ArmsAboveHead

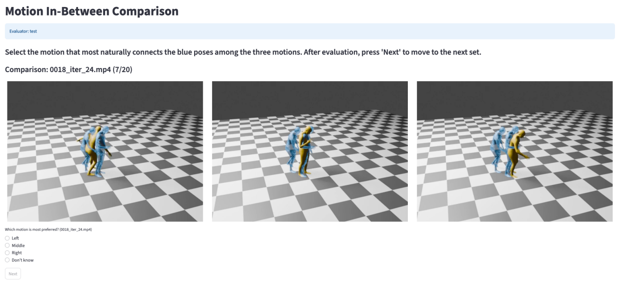
(c) COMET generated motion with ArmsAoveHead as style input

Figure 2. Comparison between Original 100STYLE dataset and stylized motion generated with COMET. Despite COMET faithfully capturing the style, a classifier with a one-layer Transformer architecture similar with used in SMooDi [13], classified COMET's output illustrated in (c) as BentForward. This suggests that the classifier can easily overfit to non-salient cues rather than the primary characteristics.

tion. The study consisted of two tasks:

Motion In-betweening. Participants were presented with side-by-side comparisons of motion sequences generated by three methods: COMET, CondMDI [2], and DNO [6]. All methods were evaluated under identical boundary conditions to ensure fairness. Each participant viewed 20 test cases and was asked to select the motion they perceived as having the highest overall quality. To reduce noise from uncertain judgments, a *"Not sure"* option was provided for cases where the differences between sequences were ambiguous.

Motion Stylization. For stylization, COMET was compared against the baseline method SMoodi [13]. Participants evaluated 24 generated sequences, corresponding to 12 distinct target styles with two sequences generated per style. Each pair consisted of one motion produced by COMET and one by the baseline, and participants were asked to choose which output exhibited higher stylistic quality and naturalness. Similar to the in-betweening task, a *"Not sure"* option was included to handle ambiguous cases where no clear preference could be determined.



Motion In-Between Comparison

Evaluator: test

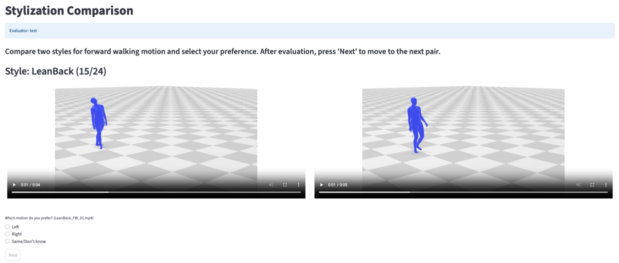
Select the motion that most naturally connects the blue poses among the three motions. After evaluation, press 'Next' to move to the next set.

Comparison: 0018_iter_24.mp4 (7/20)

Which motion is most preferred? (0018_iter_24.mp4)

○ Left
○ Middle
○ Right
○ Don't know

Figure 3. Motion In-betweening User Study



Stylization Comparison

Evaluator: test

Compare two styles for forward walking motion and select your preference. After evaluation, press 'Next' to move to the next pair.

Style: LeanBack (15/24)

Which motion do you prefer? (LeanBack_FW_01.mp4)

○ Left
○ Right
○ Same/Don't know

Figure 4. Stylization User Study

Table 3. Motion generation speed comparison (seconds). Lower values indicate faster generation. For OmniControl, the time remains nearly constant as it generates the entire sequence in a single pass, regardless of sequence length within its processing capacity. DartControl results are measured under its optimization-based setting.

Method	30 Frames	60 Frames	150 Frames
WANDR	0.51 ± 0.07	1.00 ± 0.14	2.42 ± 0.18
OmniControl	57.13 ± 2.82	62.49 ± 4.56	56.76 ± 1.77
DartControl	82.04 ± 0.23	160.44 ± 1.22	397.81 ± 1.22
Ours (COMET)	0.64 ± 0.18	1.30 ± 0.19	3.20 ± 0.23

F. Comparison with Diffusion-based Joint Control Model

We adopt a Variational Autoencoder (VAE) as the core architecture of COMET due to its superior efficiency, which enables real-time controllable motion generation. In contrast, recent diffusion-based approaches, while achieving remarkable accuracy and generating fine-grained motion details, incur substantially higher sampling costs. For example, OmniControl [11] requires 1,000 denoising steps to generate a sequence, making it roughly 90 times slower than COMET for producing 30 frames. DartControl [12] reduces the denoising process to 10 DDIM [9] steps but introduces an additional 100-step iterative optimization procedure to refine motion quality. This repetitive optimization causes the sampling time to grow linearly with sequence length.

Consequently, diffusion-based methods are powerful but are better suited for offline generation scenarios, where their significant computational demands are less restrictive. A detailed speed comparison is provided in Table 3.

References

[1] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21211–21221, 2023. 1

[2] Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. Flexible motion in-betweening with diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024. 3, 5

[3] Markos Diomataris, Nikos Athanasiou, Omid Taheri, Xi Wang, Otmar Hilliges, and Michael J Black. Wandr: Intention-guided human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 927–936, 2024. 3

[4] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020. 3

[5] Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics (TOG)*, 41(3):1–16, 2022. 3, 4

[6] Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1334–1345, 2024. 3, 5

[7] Boeun Kim, Jungho Kim, Hyung Jin Chang, and Jin Young Choi. Most: Motion style transformer between diverse action contents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1705–1714, 2024. 3, 4

[8] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 1

[9] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[10] Yu-Hui Wen, Zhipeng Yang, Hongbo Fu, Lin Gao, Yanan Sun, and Yong-Jin Liu. Autoregressive stylized motion synthesis with generative flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13612–13621, 2021. 4

[11] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *The Twelfth International Conference on Learning Representations*, 2024. 5

[12] Kaifeng Zhao, Gen Li, and Siyu Tang. Dartcontrol: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *The Thirteenth International Conference on Learning Representations*, 2024. 5

[13] Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. Smoodi: Stylized motion diffusion model. In *European Conference on Computer Vision*, pages 405–421. Springer, 2024. 4, 5