Supplementary Material of M2D2M: Multi-Motion Generation from Text with Discrete Diffusion Models

Seunggeun Chi[1,2*], Hyung-gun Chi[2*], Hengbo Ma[‡], Nakul Agarwal[1], Faizan Siddiqui[1], Karthik Ramani[2†], and Kwonjoon Lee[1†]

[1]Honda Research Institute USA [2]Purdue University
{chi65, chi45}@purdue.edu,
ramani@purdue.edu, kwonjoon_lee@honda-ri.com

In the supplementary material, we offer additional details and experiments that are not included in the main paper due to the page limit. This includes implementation specifics and architectural design, along with baseline implementation methodologies (Appendix A). Additionally, we describe the generation of test sets for the multi-motion generation task in Appendix B, present further ablation studies in Appendix C, report comprehensive comparison in Appendix D, and provide in-depth analysis of our work in Appendix E. Lastly, we include extra qualitative results in Appendix F.

A Additional Details

A.1 M2D2M

Motion VQ-VAE. In developing the Motion VQ-VAE, we adopt the architecture proposed by Zhang *et al*. [63]. We construct both the encoder and decoder of the Motion VQ-VAE using a CNN-based architecture, specifically employing 1D convolutions. Additionally, we adhere to the same hyperparameters and training procedures as outlined in their study.

Denoising Transformer. The denoising transformer configuration is specified as follows: 12 layers, 16 attention heads, 512 embedding dimensions, 2048 hidden dimensions, and a dropout rate of 0. Also, we designed action sentence conditioning for the denoising transformer to enable the multi-motion generation task with the HumanML3D dataset and KIT-ML dataset. We focus on the action verbs within a sentence (i.e., 'walk', 'turn around') of datasets, because they offer clear information about the type of motion involved. Therefore, we further break down the sentence using action verbs and then enrich them to form a complete action description, like 'a person walking,' which serves as the basis for conditioning the motion generation as illustrated in Fig. 1. For a joint sampling of Two-Phase Sampling (TPS), which aims to create a seamless motion sequence, we concatenate action tokens from successive actions for conditioning. This forms a compound condition that infuses the motion generation with

[*] Co-first authors. [†] Senior authors. [‡] Work done at Honda Research Institute.

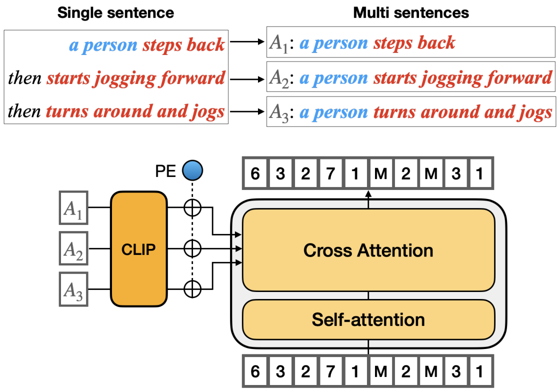

Fig. 1: Overview of action sentence conditioning of M2D2M. We initially decompose sentences to extract action verbs and subsequently utilize these verbs to construct new sentences. These newly formed sentences then serve as conditions for generating human motion sequences.

contextual information, ensuring the resulting sequence is both cohesive and reflective of the intended actions.

Implementation Details. Our model adheres to the hyper-parameter settings of VQ-Diffusion [17] unless otherwise stated, encompassing the configurations for the transition matrix parameters, namely $\bar{\alpha}_t$ and $\bar{\gamma}_t$. We linearly increase the $\bar{\gamma}_t$ and decrease the $\bar{\alpha}_t$. The loss coefficient is set at $\lambda = 5.0{\times}10^{-4}$ as per Eq. (3), and the diffusion process is defined over $T = 100$ timesteps. Optimization is carried out using the AdamW optimizer with a learning rate of 2.0×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.99$, and weight decay 4.5×10^{-2}. We trained the model for 110 epochs, and the learning rate decayed to 2.0×10^{-5} at the 100th epoch. We use the guidance scale of $s = 4$ for single motion generation, and $s = 2$ for multi-motion generation. When generating multi-motion sequence, we use $T_s = 90$ for TPS. For generating single motions, we apply a Dynamic Transition Probability scale factor of $\eta = 0.5$, and for multi-action generation, we adjust the scale factor to $\eta = 0.25$.

A.2 Baselines for Multi-Motion Generation

We evaluated the baseline methods of T2M-GPT[‡] [63] and PriorMDM[§] [49] for the task of multi-motion generation based on the code provided from the original papers. For a fair comparison with T2M-GPT, we modified the model to produce codebooks matching the specified ground truth length by disabling the end-token output. These codebooks were then concatenated for each motion and fed into the decoder. In the case of PriorMDM and Handshake [49], we set the hyper-parameter to match the illustration of Fig. 3 in the main paper for the fair comparison, employing a handshake size of 40 and transition margins

[‡] https://github.com/Mael-zys/T2M-GPT [§] https://githubwcom/priorMDM/priorMDM

Table 1: Comparison table for Multi-motion generation performance with different classifier-free scales on HumanML3D dataset.

Classifier-free	Individual Motion				Transition (40 frames)		
Guidance Scale (s)	R-Top3↑	FID↓	MMdist↓	Div→	FID↓	Div→	Jerk→
Ground Truth (Single)	$0.791^{\pm.002}$	$0.002^{\pm.000}$	$2.707^{\pm.008}$	$9.820^{\pm.065}$	$0.003^{\pm.002}$	$9.574^{\pm.054}$	$1.192^{\pm.005}$
Ground Truth (Concat)	-	-	-	-	-	-	$1.371^{\pm.004}$
1.0	$0.628^{\pm.005}$	$0.350^{\pm.021}$	$3.836^{\pm.019}$	$9.573^{\pm.156}$	$3.299^{\pm.152}$	$8.395^{\pm.142}$	$1.246^{\pm.006}$
1.5	$0.705^{\pm.004}$	$\mathbf{0.254}^{\pm.017}$	$3.063^{\pm.017}$	$9.777^{\pm.170}$	$3.293^{\pm.177}$	$8.545^{\pm.115}$	$1.242^{\pm.009}$
2.0	$0.733^{\pm.003}$	$\mathbf{0.254}^{\pm.016}$	$3.165^{\pm.019}$	$\mathbf{9.806}^{\pm.158}$	$\mathbf{3.276}^{\pm.173}$	$8.599^{\pm.154}$	$\mathbf{1.238}^{\pm.008}$
2.5	$0.746^{\pm.006}$	$0.262^{\pm.025}$	$3.063^{\pm.017}$	$9.844^{\pm.148}$	$3.321^{\pm.178}$	$8.622^{\pm.124}$	$1.252^{\pm.009}$
3.0	$\mathbf{0.751}^{\pm.006}$	$0.270^{\pm.020}$	$\mathbf{3.042}^{\pm.023}$	$9.795^{\pm.147}$	$3.400^{\pm.194}$	$\mathbf{8.648}^{\pm.130}$	$1.263^{\pm.007}$

of 20. For the other hyper-parameters, we follow the setup of PriorMDM [49]. For the SLERP algorithm, unlike the TEACH [6] setup, we first independently generate individual motions with half-transition length shorter than the given ground truth length, then apply SLERP as illustrated in Fig. 3 of the main paper. We computed the FID score based on their prescribed method, for both individual motions and transitions.

B Multi-Motion Generation Test Set

Due to the absence of distinct motion boundaries in multi-action verb annotations within the HumanML3D and KIT-ML datasets used in our experiments, we opted for test sets that exclusively consist of single action verbs. In the curated test sets, each sentence includes only one action verb, such as ‘walk’ or ‘run’. We then randomly selected N action descriptions from this pool of single-action verb sentences, ensuring no overlap, to create our test set for the multi-motion generation task. Specifically, for $N = 4$, the test set from the HumanML3D dataset comprises 1448 motions, each associated with a single-verb annotation. Similarly, the test set from the KIT-ML dataset includes 532 motions, all characterized by single action verb annotations.

C Additional Ablation Studies

In this section, we present a series of additional ablation studies that were not included in Sec. 5.4 of the main paper due to the page limit. It includes 1) exploring different classifier-free guidance scales (Appendix C.1), 2) assessing our model's performance with varying numbers of actions in multi-motion generation tasks (Appendix C.2), 3) examining the smoothness-fidelity trade-off at different independent sampling steps in TPS (Appendix C.4), and finally, 4) evaluating the Dynamic Transition Probability scale η (Appendix C.5).

C.1 Classifier-free Guidance Scale

We first focus on the effect of different classifier guidance scales s, which is described in Eq. (9). To evaluate the performance of our model in multi-motion generation and single-motion generation, we utilize the HumanML3D dataset,

Table 2: Single-motion generation performance on the different classifier-free guidance scale on HumanML3D.

Classifier-free Guidance Scale (s)	R-Top 3↑	FID↓	MM-Dist↓	Diversity→
Ground Truth	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$
0.0	$0.686^{\pm.003}$	$0.107^{\pm.005}$	$3.690^{\pm.008}$	$9.580^{\pm.088}$
1.0	$0.786^{\pm.003}$	$0.146^{\pm.002}$	$3.084^{\pm.008}$	$9.897^{\pm.088}$
2.0	$\mathbf{0.804}^{\pm.003}$	$0.139^{\pm.004}$	$2.995^{\pm.008}$	$9.886^{\pm.082}$
3.0	$0.803^{\pm.002}$	$0.107^{\pm.003}$	$\mathbf{2.980}^{\pm.006}$	$9.815^{\pm.089}$
4.0	$0.799^{\pm.002}$	$\mathbf{0.087}^{\pm.003}$	$3.018^{\pm.008}$	$9.672^{\pm.086}$
5.0	$0.787^{\pm.002}$	$0.127^{\pm.007}$	$3.089^{\pm.007}$	$\mathbf{9.439}^{\pm.086}$

Table 3: Single motion generation performance on different distance functions for $d(\cdot, \cdot)$ on Human3D dataset.

Methods	R-Top3↑	FID↓	MM-Dist↓	Diversity→	MModality↑
L2	$0.798^{\pm.002}$	$0.098^{\pm.005}$	$3.018^{\pm.008}$	$\mathbf{9.623}^{\pm.085}$	$2.115^{\pm.079}$
L2 Rank	$0.799^{\pm.002}$	$\mathbf{0.087}^{\pm.004}$	$3.018^{\pm.008}$	$9.672^{\pm.086}$	$2.132^{\pm.073}$
Cosine	$\mathbf{0.801}^{\pm.002}$	$0.092^{\pm.004}$	$\mathbf{3.011}^{\pm.008}$	$9.670^{\pm.084}$	$\mathbf{2.137}^{\pm.084}$
Cosine Rank	$0.797^{\pm.002}$	$0.099^{\pm.005}$	$3.026^{\pm.008}$	$9.669^{\pm.085}$	$2.125^{\pm.069}$

and provide results presented in Table 2. This experiment reveals that the optimal balance between accuracy and fidelity for these metrics is achieved at a classifier guidance scale of $s = 4$ for single-motion generation, and best smoothness at $s = 2$ for multi-motion generation.

C.2 Number of Action in Multi-Motion Generation

In order to explore our model's effectiveness in generating long-term motion, we evaluate the performance of our model by progressively increasing the number of actions (N) using the HumanML3D dataset. The results of these evaluations are detailed in Table 4. We found that as N increases, R-Top3 and FID scores of individual motion demonstrate a decline, indicating a reduction in fidelity with more actions. Despite this, it's noteworthy that our model's performance on the transition part remains comparably effective to that of real single motions, even at $N = 32$, a considerably long motion sequence. This highlights our model's proficiency in generating long-term motion with smooth and coherent transitions.

C.3 Different Distance metrics for Dynamic Transition Probability

In Table 3, we conduct a comparative analysis of different distance functions for $d(\cdot, \cdot)$, utilized in defining the codebook distance for Eq. (8). Specifically, we evaluate the performance of L2 and Cosine Distance, focusing on their effectiveness as distance functions. Our findings indicate that the L2 Rank distance function yields the best FID score, highlighting its superiority in this context.

C.4 Effect of Two-Phase Sampling

In Table 5, we explore the impact of Two-Phase Sampling. Our analysis also includes adjustments in the ratio of independent denoising steps (T_s) to the total

Table 4: Multi-motion generation performance on the different number of actions (N) on HumanML3D.

The number of actions (N)	Individual Motion				Transition (40 frames)		
	R-Top3↑	FID↓	MMdist↓	Div→	FID↓	Div→	Jerk→
Ground Truth (Single)	$0.791^{\pm.002}$	$0.002^{\pm.000}$	$2.707^{\pm.008}$	$9.820^{\pm.065}$	$0.003^{\pm.002}$	$9.574^{\pm.054}$	$1.192^{\pm.005}$
Ground Truth (Concat)	-	-	-	-	-	-	$1.371^{\pm.004}$
$N = 1$	$0.751^{\pm.008}$	$0.196^{\pm.003}$	$3.012^{\pm.018}$	$9.894^{\pm.057}$	$3.340^{\pm.219}$	$8.751^{\pm.005}$	$1.248^{\pm.005}$
$N = 2$	$0.737^{\pm.007}$	$0.198^{\pm.025}$	$3.127^{\pm.031}$	$9.870^{\pm.064}$	$3.430^{\pm.431}$	$8.497^{\pm.121}$	$1.244^{\pm.013}$
$N = 4$	$0.733^{\pm.003}$	$0.254^{\pm.016}$	$3.165^{\pm.019}$	$9.806^{\pm.158}$	$3.276^{\pm.173}$	$8.599^{\pm.154}$	$1.238^{\pm.008}$
$N = 8$	$0.733^{\pm.005}$	$0.307^{\pm.027}$	$3.153^{\pm.028}$	$9.624^{\pm.137}$	$3.343^{\pm.092}$	$8.675^{\pm.121}$	$1.255^{\pm.010}$
$N = 16$	$0.725^{\pm.004}$	$0.312^{\pm.031}$	$3.193^{\pm.018}$	$9.557^{\pm.066}$	$3.380^{\pm.109}$	$8.455^{\pm.165}$	$1.245^{\pm.011}$
$N = 32$	$0.731^{\pm.005}$	$0.350^{\pm.040}$	$3.192^{\pm.023}$	$9.555^{\pm.069}$	$3.336^{\pm.145}$	$8.537^{\pm.182}$	$1.248^{\pm.013}$

Table 5: Multi-motion generation performance across a different number of independent denoising steps (T_s) of Two-Phase Sampling on HumanML3D.

Methods	Individual Motion				Transition (40 frames)		
	R-Top3↑	FID↓	MMdist↓	Div→	FID↓	Div→	Jerk→
Ground Truth (Single)	$0.791^{\pm.002}$	$0.002^{\pm.000}$	$2.707^{\pm.008}$	$9.820^{\pm.065}$	$0.003^{\pm.002}$	$9.574^{\pm.054}$	$1.192^{\pm.005}$
Ground Truth (Concat)	-	-	-	-	-	-	$1.371^{\pm.004}$
w/o TPS	$\mathbf{0.755}^{\pm.007}$	$\mathbf{0.173}^{\pm.010}$	$3.015^{\pm.024}$	$9.950^{\pm.076}$	$3.455^{\pm.142}$	$8.554^{\pm.081}$	$1.402^{\pm.005}$
$T_s = 100$	$0.751^{\pm.008}$	$0.196^{\pm.003}$	$\mathbf{3.012}^{\pm.018}$	$9.894^{\pm.057}$	$3.340^{\pm.219}$	$8.751^{\pm.005}$	$1.248^{\pm.005}$
$T_s = 95$	$0.737^{\pm.004}$	$0.232^{\pm.028}$	$3.105^{\pm.017}$	$9.772^{\pm.167}$	$3.289^{\pm.243}$	$8.643^{\pm.132}$	$1.253^{\pm.007}$
$T_s = 90$	$0.733^{\pm.003}$	$0.254^{\pm.016}$	$3.165^{\pm.019}$	$\mathbf{9.806}^{\pm.158}$	$\mathbf{3.276}^{\pm.173}$	$8.599^{\pm.154}$	$\mathbf{1.238}^{\pm.008}$
$T_s = 80$	$0.725^{\pm.006}$	$0.284^{\pm.024}$	$3.194^{\pm.029}$	$9.767^{\pm.129}$	$3.338^{\pm.129}$	$\mathbf{8.691}^{\pm.114}$	$1.247^{\pm.007}$
$T_s = 50$	$0.709^{\pm.006}$	$0.371^{\pm.034}$	$3.315^{\pm.018}$	$9.665^{\pm.125}$	$3.282^{\pm.263}$	$8.595^{\pm.144}$	$1.254^{\pm.010}$

Table 6: Multi-motion generation on different smoothing methods with MDM on HumanML3D.

Methods	Individual Motion				Transition (40 frames)		
	R-Top3↑	FID↓	MMdist↓	Div→	FID↓	Div→	Jerk→
Ground Truth (Single)	0.791	0.002	2.707	9.820	0.003	9.574	1.192
Ground Truth (Concat)	-	-	-	-	-	-	1.371
MDM [54] + Handshake [49]	0.586	0.832	5.901	**9.543**	**3.351**	**8.801**	0.476
MDM [54] + **TPS (Ours)**	**0.640**	**0.582**	**5.287**	9.321	3.376	8.070	**0.634**

number of denoising steps (T). This examination reveals a clear trade-off in motion generation between smoothness and fidelity. As discussed in Sec. 4.3, phases of independent sampling enhance the fidelity of individual motions, while phases of joint sampling improve the fidelity and smoothness of transitions between motions. Implementing the Two-Phase Sampling algorithm and reducing the number of independent sampling steps (T_s) tends to improve smoothness metrics (e.g., Jerk), but simultaneously, fidelity metrics such as R-Top3 and FID begin to deteriorate. This observation emphasizes the intrinsic trade-off between smoothness and fidelity in motion generation, identifying an optimal $T_s = 90$ for the smoothness metric being identified.

In Table 6, we evaluate multi-motion generation algorithms on non-latent diffusion models. We applied Handshake [49] and TPS to MDM [54], a diffusion model operating in Cartesian space with 3D skeletal coordinates. We observe that the effectiveness of TPS is not confined to its designed latent space; it also functions effectively in the Cartesian domain. The results show that TPS achieves better FID and R-Precision for individual motions, albeit with reduced

Table 7: Multi-motion generation performance across a different number of independent denoising steps (T_s) of Two-Phase Sampling on HumanML3D.

Transition Probability Methods		Individual Motion R-Top3↑	FID↓	MMdist↓	Div→	Transition (40 frames) FID↓	Div→	Jerk→
Ground Truth (Single)		$0.791^{\pm.002}$	$0.002^{\pm.000}$	$2.707^{\pm.008}$	$9.820^{\pm.065}$	$0.003^{\pm.002}$	$9.574^{\pm.054}$	$1.192^{\pm.005}$
Ground Truth (Concat)		-	-	-	-	-	-	$1.371^{\pm.004}$
$\beta(t)$	-	$0.738^{\pm.009}$	$0.253^{\pm.002}$	$3.164^{\pm.021}$	$9.822^{\pm.051}$	$3.483^{\pm.029}$	$8.625^{\pm.044}$	$1.265^{\pm.005}$
$\beta(d,t)$	$\eta = 1.00$	$0.730^{\pm.005}$	$0.264^{\pm.026}$	$3.152^{\pm.028}$	$9.808^{\pm.162}$	$3.315^{\pm.225}$	$8.654^{\pm0.064}$	$1.252^{\pm.007}$
$\beta(d,t)$	$\eta = 0.50$	$0.733^{\pm.003}$	$\mathbf{0.244}^{\pm.016}$	$3.156^{\pm.029}$	$9.830^{\pm.160}$	$3.278^{\pm.138}$	$8.586^{\pm.127}$	$1.250^{\pm.008}$
$\beta(d,t)$	$\eta = 0.33$	$0.732^{\pm004}$	$0.245^{\pm.010}$	$\mathbf{3.150}^{\pm.173}$	$\mathbf{9.815}^{\pm.152}$	$3.312^{\pm.171}$	$\mathbf{8.675}^{\pm.134}$	$1.246^{\pm.009}$
$\beta(d,t)$	$\eta = 0.25$	$\mathbf{0.734}^{\pm.003}$	$0.253^{\pm.016}$	$3.165^{\pm.019}$	$9.806^{\pm.158}$	$\mathbf{3.276}^{\pm.017}$	$8.599^{\pm.154}$	$\mathbf{1.238}^{\pm.008}$
$\beta(d,t)$	$\eta = 0.20$	$0.724^{\pm005}$	$0.254^{\pm.010}$	$3.194^{\pm.026}$	$9.803^{\pm.152}$	$3.330^{\pm.205}$	$8.519^{\pm.162}$	$1.247^{\pm.008}$

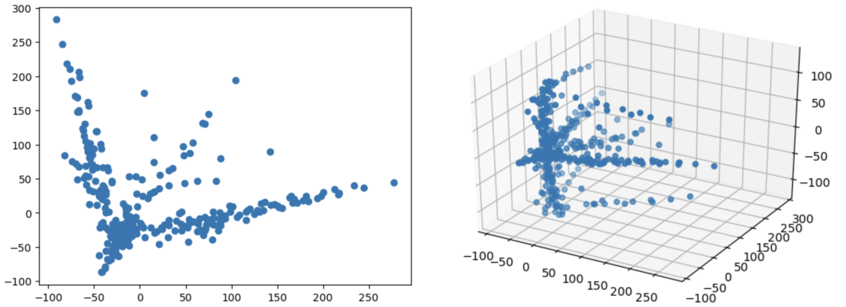

Fig. 2: PCA plot representing motion tokens from the codebook of Motion VQ-VAE, visualized in 2D (**Left**) and 3D (**Right**) space.

diversity. For the transition part, TPS demonstrates comparable FID results while exhibiting improved smoothness as measured by Jerk.

C.5 The scale of Transition Probability Matrix

We investigated the impact of dynamic transition probability on the generation of multiple motions by conducting an ablation study that varied the transition probability scale, η. In Table 7, we noted that the dynamic transition probability, $\beta(d,t)$, outperforms the traditional method of $\beta(t)$. Additionally, the results indicate a trend where the smoothness metric (Jerk) becomes closer to ground truth single motion as η is reduced.

D Comprehensive Comparison of Single-Motion Generation

In Tables 8 and 9, we present a comprehensive comparison of single-motion generation results to demonstrate the effectiveness of our method.

Table 8: Single-motion generation performance on HumanML3D. The figures highlighted in **bold** and **blue** denote the best and second-best results, respectively.

Methods	R-Top 3↑	FID↓	MM-Dist↓	Diversity→	MModality↑
Ground Truth	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
VQ-VAE (reconstruction)	$0.785^{\pm.002}$	$0.070^{\pm.001}$	$3.072^{\pm.009}$	$9.593^{\pm.079}$	-
Seq2Seq [37]	$0.396^{\pm.002}$	$11.75^{\pm.035}$	$5.529^{\pm.007}$	$6.223^{\pm.061}$	-
J2LP [2]	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$6.223^{\pm.058}$	-
Text2Gesture [10]	$0.345^{\pm.002}$	$5.012^{\pm.030}$	$6.030^{\pm.008}$	$7.676^{\pm.071}$	-
Hier [16]	$0.552^{\pm.004}$	$6.532^{\pm.024}$	$5.012^{\pm.018}$	$6.409^{\pm.042}$	-
MoCoGAN [55]	$0.106^{\pm.001}$	$94.41^{\pm.021}$	$9.643^{\pm.006}$	$8.332^{\pm.008}$	$0.019^{\pm.000}$
Dance2Music [34]	$0.097^{\pm.001}$	$66.98^{\pm.016}$	$8.116^{\pm.006}$	$0.462^{\pm.011}$	$0.043^{\pm.001}$
TEMOS [42]	$0.722^{\pm.002}$	$3.734^{\pm.028}$	$3.703^{\pm.008}$	$0.725^{\pm.071}$	$0.368^{\pm.018}$
TM2T [21]	$0.729^{\pm.002}$	$1.501^{\pm.017}$	$3.467^{\pm.011}$	$8.973^{\pm.076}$	$2.424^{\pm.093}$
MLD [11]	$0.736^{\pm.002}$	$1.087^{\pm.021}$	$3.347^{\pm.008}$	$8.589^{\pm.083}$	$2.219^{\pm.074}$
Guo *et al.* [19]	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$9.175^{\pm.082}$	$2.413^{\pm.079}$
MDM [54]	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.724^{\pm.086}$	$2.799^{\pm.072}$
MotionDiffuse [64]	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$\mathbf{9.410}^{\pm.049}$	$1.553^{\pm.042}$
T2M-GPT [63]	$0.775^{\pm.002}$	$0.116^{\pm.004}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$
AttT2M [67]	$0.786^{\pm.006}$	$0.112^{\pm.006}$	$3.038^{\pm.007}$	$9.700^{\pm.090}$	$\mathbf{2.452}^{\pm.051}$
MAA [8]	$0.675^{\pm.002}$	$0.774^{\pm.007}$	-	$8.230^{\pm.064}$	-
M2DM [32]	$0.763^{\pm.003}$	$0.352^{\pm.005}$	$3.134^{\pm.010}$	$9.926^{\pm.073}$	$\mathbf{3.587}^{\pm.072}$
M2D2M (w/ β_t)	$\mathbf{0.796}^{\pm.002}$	$\mathbf{0.115}^{\pm.006}$	$\mathbf{3.036}^{\pm.008}$	$9.680^{\pm.074}$	$2.193^{\pm.077}$
M2D2M (w/ $\beta(t,d)$)	$\mathbf{0.799}^{\pm.002}$	$\mathbf{0.087}^{\pm.004}$	$\mathbf{3.018}^{\pm.008}$	$\mathbf{9.672}^{\pm.086}$	$2.115^{\pm.079}$

Table 9: Single-motion generation performance on KIT-ML. The figures highlighted in **bold** and **blue** denote the best and second-best results, respectively.

Methods	R-Top3↑	FID↓	MM-Dist↓	Diversity→	MModality↑
Ground Truth	$0.779^{\pm.006}$	$0.031^{\pm.004}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
VQ-VAE (reconstruction)	$0.740^{\pm.006}$	$0.472^{\pm.011}$	$2.986^{\pm.027}$	$10.994^{\pm.120}$	-
Seq2Seq [37]	$0.241^{\pm.006}$	$24.86^{\pm.348}$	$7.960^{\pm.031}$	$6.744^{\pm.106}$	-
J2LP [2]	$0.483^{\pm.005}$	$6.545^{\pm.072}$	$5.147^{\pm.030}$	$9.073^{\pm.100}$	-
Text2Gesture [10]	$0.338^{\pm.005}$	$12.12^{\pm.183}$	$6.964^{\pm.029}$	$9.334^{\pm.079}$	-
Hier [16]	$0.531^{\pm.007}$	$5.203^{\pm.107}$	$4.986^{\pm.027}$	$9.563^{\pm.072}$	-
MoCoGAN [55]	$0.063^{\pm.003}$	$82.69^{\pm.242}$	$10.47^{\pm.012}$	$3.091^{\pm.043}$	$0.250^{\pm.009}$
Dance2Music [34]	$0.086^{\pm.003}$	$115.4^{\pm.240}$	$10.40^{\pm.016}$	$0.241^{\pm.004}$	$0.062^{\pm.002}$
TEMOS [42]	$0.687^{\pm.002}$	$3.717^{\pm.028}$	$3.417^{\pm.008}$	$10.84^{\pm.004}$	$0.532^{\pm.018}$
TM2T [21]	$0.587^{\pm.005}$	$3.599^{\pm.051}$	$4.591^{\pm.019}$	$9.473^{\pm.100}$	$\mathbf{3.292}^{\pm.034}$
Guo *et al.* [19]	$0.681^{\pm.007}$	$3.022^{\pm.107}$	$3.488^{\pm.028}$	$10.72^{\pm.145}$	$2.052^{\pm.107}$
MLD [11]	$0.734^{\pm.007}$	$\mathbf{0.404}^{\pm.027}$	$3.204^{\pm.027}$	$10.80^{\pm.117}$	$2.192^{\pm.071}$
MDM [54]	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.191^{\pm.022}$	$10.847^{\pm.109}$	$1.907^{\pm.214}$
MotionDiffuse [64]	$0.739^{\pm.004}$	$1.954^{\pm.062}$	$\mathbf{2.958}^{\pm.005}$	$\mathbf{11.10}^{\pm.143}$	$0.730^{\pm.013}$
T2M-GPT [63]	$0.737^{\pm.006}$	$0.717^{\pm.041}$	$3.053^{\pm.026}$	$\mathbf{10.862}^{\pm.094}$	$1.912^{\pm.036}$
AttT2M [67]	$0.751^{\pm.006}$	$0.870^{\pm.039}$	$3.309^{\pm.021}$	$10.96^{\pm.123}$	$2.281^{\pm.047}$
M2DM [32]	$\mathbf{0.743}^{\pm.004}$	$0.515^{\pm.029}$	$3.015^{\pm.017}$	$11.417^{\pm.097}$	$\mathbf{3.325}^{\pm037}$
M2D2M (w/ β_t)	$\mathbf{0.743}^{\pm.006}$	$\mathbf{0.404}^{\pm.022}$	$3.018^{\pm.019}$	$10.749^{\pm.102}$	$2.063^{\pm.066}$
M2D2M (w/ $\beta(t,d)$)	$\mathbf{0.753}^{\pm.006}$	$\mathbf{0.378}^{\pm.023}$	$\mathbf{3.012}^{\pm.021}$	$10.709^{\pm.121}$	$2.061^{\pm.067}$

E Analysis

E.1 Codebook visualization

To examine relationships within the codebook, which inspired our design of dynamic transition probabilities as detailed in Sec. 4.2, we have visualized the tokens from the Motion VQ-VAE's codebook in Fig. 2. This visualization reveals that certain tokens are more closely correlated, as evidenced by their clustering or alignment along implicit lines. Unlike the uniform transition strategy used in

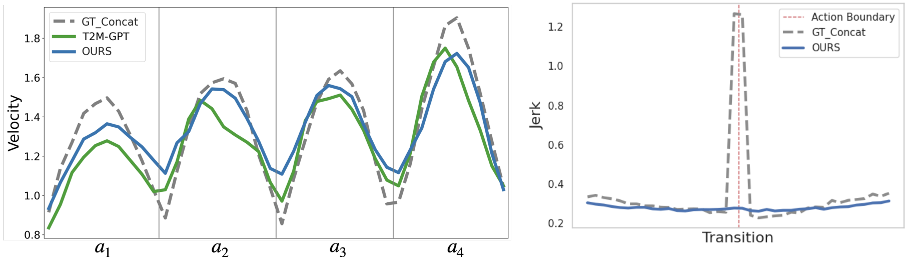

Fig. 3: **(Left)** Plot of Mean Velocity and **(Right)** plot of Mean Jerk of all transitions (40 frames) across all test sets in Multi-Motion Generation with $N = 4$. 'GT' represents concatenated real single motions.

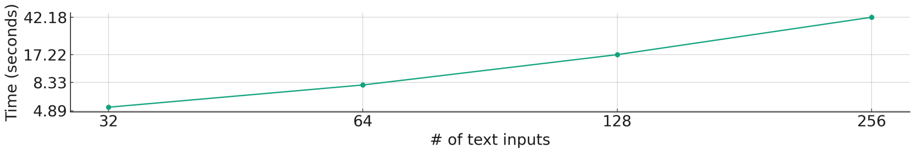

Fig. 4: Inference time scaling with action sequence length. Measured with a single NVIDIA RTX A6000 GPU.

the VQ-Diffusion model, our method starts with a broad, exploratory range of transitions to encourage diversity by considering token proximity. These results justify our design of transition probabilities for the discrete diffusion.

E.2 Mean Velocity & Jerk Plot of Generated Multi-Motion

To assess the smoothness of our M2D2M model, we plotted the mean velocity of the generated multi-motion sequences across all test sets for multi-motion generation, as shown in Fig. 3). In this figure, concatenated real single motions serve as the ground truth (GT). It is evident that the GT demonstrates discrete transitions between motions, while our M2D2M model (OURS) achieves smoother transitions with reduced jerk in the transitional phases.

E.3 Running time

We calculate inference time based on the number of actions and visualize the results in Fig. 4. This illustration demonstrates that our method is practical for generating multi-motion sequences with reasonable computational cost. We set each action to have 196 frames; thus, 256 text prompts generate 50,176 frames. The gradient of the plotted line is nearly linear, as the joint sampling step is limited to T_s, allowing most other steps to be executed in parallel within a batch.

F Additional Qualitative Results of Generated Multi-Motion from M2D2M

Further qualitative results showcasing the capabilities of M2D2M in multi-motion generation, akin to the examples in Fig. 1 in the main paper, are provided as animations (GIFs) in the supplementary materials.