Causal Motion Diffusion Models for Autoregressive Motion Generation

Supplementary Material

A. Implementation Details

A.1. MAC-VAE

The proposed MAC-VAE consists of seven causal convolutional layers and two causal ResNet blocks with left padding in both the encoder and decoder to ensure strict temporal causality. Each convolutional layer uses a kernel size of 3 and a stride of 1, followed by ReLU activation. The latent feature dimension is set to 64, and motion sequences are downsampled/upsampled by a factor of 4 along the temporal axis using stride-2 convolutional layers within the ResNet blocks.

To achieve semantic alignment between motion and text, we modify Part-TMR [20] to extract frame-level motion–language embeddings. Part-TMR uses a `[class]` token to aggregate frames into a global feature, whereas we directly extract features from each frame and align them with the corresponding text features via contrastive learning, which serves as the supervision signal for MAC-VAE. The loss weighting coefficient is set to β=1.0, and the margin parameters are set to m_1=0.5 and m_2=0.25.

We train MAC-VAE using the AdamW optimizer with a learning rate of 1×10^{-4} and a batch size of 128 for 50 epochs on a single NVIDIA A100 GPU. The learning rate follows a cosine decay schedule, and gradient clipping with a maximum norm of 1.0 is applied for training stability.

A.2. Causal-DiT

The Causal-DiT is implemented as a lightweight transformer-based denoiser with 8 layers, 4 attention heads, and a hidden dimension of 512. Causal self-attention is applied using a lower-triangular mask to enforce temporal order, while cross-attention conditions motion latents on text embeddings extracted from DistilBERT [15]. We incorporate Adaptive Layer Normalization (AdaLN) [9] and Rotary Positional Encoding (ROPE) [16] to embed timestep information and stabilize long-horizon attention. During training, the text condition is randomly dropped with a probability of 0.1 to enable classifier-free guidance. The model is optimized using AdamW with the same hyperparameter settings as MAC-VAE. The scale of classifier-free guidance is set to 3.0 during inference.

A.3. Causal Diffusion Forcing

In CMDM, causal diffusion forcing is employed to enable temporally ordered denoising while maintaining frame-level stochasticity. During training, each frame t is perturbed with an independent noise level $k_t \in [0, K]$, where K=1000 denotes the total number of diffusion steps. The Causal-DiT serves as the denoiser, learning to predict noise residuals $\boldsymbol{\epsilon}_\theta(\tilde{\mathbf{z}}_{\leq t}, k_t, \mathbf{c})$ conditioned on all preceding latent frames and the text embedding $\mathbf{c}$. This formulation ensures that each frame is denoised based solely on its causal history, thereby enforcing strict temporal dependencies. The overall training process is summarized in Algorithm 1.

During inference, we adopt the Frame-Wise Sampling Schedule (FSS) with diffusion steps K=50 and uncertainty scale L=2. In this setting, the denoising of frame t+1 begins at step $K-L$ of frame t, allowing partially denoised frames to guide subsequent generations. This causal scheduling mechanism significantly accelerates inference by reducing redundant diffusion steps while maintaining temporal consistency across frames. The overall inference process with FSS is summarized in Algorithm 2.

B. Additional Quantitative Results

B.1. Experiments on BABEL

We further evaluate CMDM on the BABEL dataset [13] to assess its generalization ability to diverse motion compositions. BABEL contains densely annotated sequences with multiple actions and transitions, making it suitable for long-horizon motion synthesis and evaluation. We train CMDM by constructing training samples from adjacent subsequences in BABEL, where each pair of consecutive segments is used to learn motion continuation across long sequences. As shown in Table 5, our method achieves the best overall performance across both subsequence and transition metrics, demonstrating the advantage of CMDM in maintaining consistency across action boundaries and generating smooth, continuous motions.

B.2. Evaluation on Other Motion Features

To further examine the generalization ability of CMDM, we conduct experiments using motion features with redundant dimensions removed, following the analysis in [7]. As discussed in prior work, the standard HumanML3D motion representation contains redundant components such as local joint rotations and contact features that do not directly influence the final human pose. Removing these redundant features yields a more compact and physically meaningful representation better suited for continuous diffusion modeling.

Table 6 reports the results on HumanML3D using only essential motion features. Compared to the baseline methods, CMDM consistently improves generation quality and semantic alignment under both autoregressive (AR) and diffusion (FSS) configurations. Specifically, **CMDM w/**

Algorithm 1 CMDM Training with Causal Diffusion Forcing

Require: Pretrained MAC-VAE encoder E_ϕ, text embedding $\mathbf{c}$, Causal-DiT ϵ_θ, diffusion schedule $\{\alpha_k, \bar{\alpha}_k\}_{k=0}^{K}$

for `each minibatch` **do**
  Encode motion sequence: $\mathbf{z}_{1:T} \leftarrow E_\phi(\mathbf{x}_{1:T})$
  for $t = 1$ to T **do**
    Sample independent noise level $k_t \sim \mathcal{U}\{0, 1, \ldots, K\}$
    Diffuse latent: $\tilde{\mathbf{z}}_t^{k_t} \leftarrow \sqrt{\bar{\alpha}_{k_t}}\mathbf{z}_t + \sqrt{1-\bar{\alpha}_{k_t}}\boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$
    Predict noise with causal conditioning: $\hat{\boldsymbol{\epsilon}}_t \leftarrow \epsilon_\theta(\tilde{\mathbf{z}}_{\leq t}, k_t, \mathbf{c})$
  end for
  Compute loss: $\mathcal{L}_{\text{DF}} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \|\boldsymbol{\epsilon}_t - \hat{\boldsymbol{\epsilon}}_t\|_2^2$
  Update $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}_{\text{DF}}$
end for

Algorithm 2 CMDM Streaming Generation with Frame-wise Sampling Schedule (FSS)

Require: Causal-DiT ϵ_θ, MAC-VAE decoder D_ψ, text embedding $\mathbf{c}$, schedule matrix $\mathbf{K} \in \mathbb{N}^{M\times T}$

Initialize $\tilde{\mathbf{z}}_t^K \sim \mathcal{N}(0, \mathbf{I})$ for t=$1, \ldots, T$
for $m = 1, \ldots, M$ **do**
  for $t = 1, \ldots, T$ **do**
    Obtain noise level $k \leftarrow K_{m,t}$
    Predict noise with previous frames: $\hat{\boldsymbol{\epsilon}}_t \leftarrow \epsilon_\theta(\tilde{\mathbf{z}}_{<t}^{k}, k, \mathbf{c})$
    Denoise the current frame: $\tilde{\mathbf{z}}_t^{k-1} \leftarrow \frac{1}{\sqrt{\alpha_k}}\left(\tilde{\mathbf{z}}_t^k - \frac{1-\alpha_k}{\sqrt{1-\bar{\alpha}_k}}\hat{\boldsymbol{\epsilon}}_t\right) + \sigma_k\,\mathbf{w},\ \mathbf{w}\sim\mathcal{N}(0, \mathbf{I})$
    if $k = 0$ **then**
      Decode final clean latent: $\hat{\mathbf{x}}_t \leftarrow D_\psi(\tilde{\mathbf{z}}_{\leq t}^{0})$
    end if
  end for
end for
return Decoded motion $\hat{\mathbf{x}}_{1:T}$ (or latents $\tilde{\mathbf{z}}_{1:T}^{0}$)

FSS achieves the best overall performance, reaching an R-Precision of 0.563/0.759/0.849 for Top-1/Top-2/Top-3 accuracy and the lowest FID of 0.078, confirming that our causal diffusion formulation effectively models temporally coherent motion even in compact feature spaces. These results demonstrate that CMDM remains robust across different motion representations, further validating its adaptability to feature compression and reparameterized motion distributions.

B.3. Compositional Motion Generation

We evaluate CMDM on the compositional motion generation task following the protocol of Multi-Track Timeline (MTT) [11], which requires generating coherent motions conditioned on multiple temporally structured text descriptions. This task evaluates both semantic composition, *i.e.*, correctly realizing multiple concepts within a single sequence, and temporal composition, *i.e.*, ensuring smooth and consistent transitions across segments.

Specifically, following prior work [11, 22], we report per-crop semantic correctness metrics (R@1, R@3, and TMR-Score for M2T and M2M), as well as realism metrics including FID and transition distance. As shown in Table 7, CMDM, under the single-track multi-crop setting, consistently outperforms EnergyMoGen and other compositional baselines across all metrics. Notably, CMDM achieves substantial improvements in semantic alignment while simultaneously reducing FID and transition distance, demonstrating stronger long-horizon consistency and smoother transitions between composed motion segments.

B.4. Latency Analysis

To evaluate the practical efficiency of different causal motion generation methods, we measure the latency for generating each token (4 frames) on a single NVIDIA A100 GPU. MARDM [7], MotionStreamer [18], and CMDM w/ AR require approximately 210 ms, 360 ms, and 150 ms, respectively, to generate the first token, with similar latency for each subsequent token. This is because these autoregressive diffusion methods perform full diffusion denoising for each token independently, requiring multiple denoising steps per frame regardless of its temporal position. In contrast, CMDM w/ FSS takes about 220 ms for the first token but only 30 ms per subsequent token, achieving a 5×–12× speedup for streaming generation. This dramatic reduction in per-token latency stems from our frame-wise sampling

Methods	Subsequence				Transition			
	R-prec↑	FID↓	Div→	MM-Dist↓	FID↓	Div→	PJ→	AUJ↓
GT	$0.715^{\pm0.003}$	$0.00^{\pm0.00}$	$8.42^{\pm0.15}$	$3.36^{\pm0.06}$	$0.00^{\pm0.00}$	$6.20^{\pm0.06}$	$0.02^{\pm0.00}$	$0.00^{\pm0.00}$
FlowMDM	$0.702^{\pm0.004}$	$0.99^{\pm0.04}$	$8.36^{\pm0.13}$	$3.45^{\pm0.02}$	$2.61^{\pm0.06}$	$\mathbf{6.47^{\pm0.05}}$	$0.06^{\pm0.00}$	$0.13^{\pm0.00}$
Ours	$\mathbf{0.711^{\pm0.005}}$	$\mathbf{0.90^{\pm0.06}}$	$\mathbf{8.47^{\pm0.20}}$	$\mathbf{3.39^{\pm0.05}}$	$\mathbf{2.45^{\pm0.05}}$	$6.73^{\pm0.05}$	$\mathbf{0.05^{\pm0.00}}$	$\mathbf{0.11^{\pm0.01}}$

Table 5. Comparison of long-horizon motion generation on BABEL. Subsequence metrics evaluate motion quality and diversity within segments, while transition metrics assess temporal continuity and smoothness between segments.

Methods	Framework	R-Precision↑			FID↓	MM-Dist↓	MModality↑	CLIP-score↑
		Top 1	Top 2	Top 3				
T2M-GPT [21]	VQ	$0.470^{\pm.003}$	$0.659^{\pm.002}$	$0.758^{\pm.002}$	$0.335^{\pm.003}$	$3.505^{\pm.017}$	$2.018^{\pm.053}$	$0.607^{\pm.005}$
MMM [12]		$0.487^{\pm.003}$	$0.683^{\pm.002}$	$0.782^{\pm.002}$	$0.132^{\pm.004}$	$3.359^{\pm.019}$	$2.241^{\pm.073}$	$0.635^{\pm.003}$
MoMask [4]		$0.490^{\pm.004}$	$0.687^{\pm.003}$	$0.786^{\pm.003}$	$0.116^{\pm.006}$	$3.353^{\pm.010}$	$1.263^{\pm.079}$	$0.637^{\pm.003}$
MDM-50Step [17]	Diffusion	$0.440^{\pm.007}$	$0.636^{\pm.006}$	$0.742^{\pm.004}$	$0.518^{\pm.032}$	$3.640^{\pm.028}$	$\mathbf{3.604^{\pm.031}}$	$0.578^{\pm.003}$
MotionDiffuse [17]		$0.450^{\pm.006}$	$0.641^{\pm.005}$	$0.753^{\pm.005}$	$0.778^{\pm.035}$	$3.490^{\pm.023}$	$3.179^{\pm.046}$	$0.606^{\pm.004}$
MLD [2]		$0.461^{\pm.004}$	$0.651^{\pm.004}$	$0.750^{\pm.003}$	$0.431^{\pm.014}$	$3.445^{\pm.019}$	$\underline{3.506}^{\pm.031}$	$0.615^{\pm.003}$
ReMoDiffuse [23]		$0.468^{\pm.003}$	$0.653^{\pm.003}$	$0.754^{\pm.005}$	$0.883^{\pm.021}$	$3.414^{\pm.020}$	$2.703^{\pm.154}$	$0.621^{\pm.003}$
SALAD [5]		$\underline{0.552}^{\pm.003}$	$\underline{0.748}^{\pm.003}$	$\underline{0.839}^{\pm.002}$	$0.124^{\pm.005}$	$2.990^{\pm.010}$	$1.833^{\pm.081}$	$0.671^{\pm.001}$
MARDM-DDPM [7]	Autoregressive	$0.492^{\pm.006}$	$0.690^{\pm.005}$	$0.790^{\pm.005}$	$0.116^{\pm.004}$	$3.349^{\pm.010}$	$2.470^{\pm.053}$	$0.637^{\pm.005}$
MARDM-SiT [7]	Transformer	$0.500^{\pm.004}$	$0.695^{\pm.003}$	$0.795^{\pm.003}$	$0.114^{\pm.007}$	$3.270^{\pm.009}$	$2.231^{\pm.071}$	$0.642^{\pm.002}$
Ours w/ AR	Autoregressive	$0.550^{\pm.004}$	$0.747^{\pm.004}$	$0.838^{\pm.003}$	$\underline{0.085}^{\pm.004}$	$\underline{2.987}^{\pm.011}$	$1.810^{\pm.068}$	$\underline{0.675}^{\pm.001}$
Ours w/ FSS	Diffusion	$\mathbf{0.563^{\pm.004}}$	$\mathbf{0.759^{\pm.003}}$	$\mathbf{0.849^{\pm.002}}$	$\mathbf{0.078^{\pm.003}}$	$\mathbf{2.920^{\pm.007}}$	$1.827^{\pm.094}$	$\mathbf{0.685^{\pm.001}}$

Table 6. Results of text-to-motion generation on HumanML3D without redundant features. The average is reported over 10 runs with 95% confidence intervals. **Bold** indicates the best result, and underline denotes the second-best result.

Method	Input type		Per-crop semantic correctness				Realism	
	#tracks	#crops	R@1↑	R@3↑	M2T↑	M2M↑	FID↓	Transition↓
GT	-	-	55.0	73.3	0.748	1.000	0.000	1.5
EnergyMoGen [22]	Single	Single	15.9	28.0	0.591	0.567	0.604	1.6
MDM-SMPL [11]	Single	Single	12.1	23.5	0.573	0.578	0.484	1.8
w/ DiffCollage [11]	Single	Multi	29.1	49.7	0.675	0.656	0.446	1.2
w/ STMC [11]	Multi	Multi	30.5	50.9	0.675	0.665	0.459	**0.9**
CMDM (Ours)	Single	Multi	**41.7**	**57.9**	**0.690**	**0.672**	**0.438**	1.2

Table 7. Comparison with prior compositional motion generation methods on the Multi-track timeline (MTT) dataset [11].

schedule, which allows each frame to be predicted from partially denoised preceding frames rather than requiring full iterative refinement.

B.5. Ablation Studies

Architecture of MAC-VAE. We evaluate several configurations of MAC-VAE to analyze the effects of latent dimension and temporal downsampling rate on both reconstruction and generation performance. The notation (d, r) denotes the latent dimension d and the temporal downsampling rate r. As shown in Table 8, increasing the latent dimension improves reconstruction accuracy but also introduces redundancy that slightly affects generation quality in terms of FID. Conversely, larger temporal downsampling rates (*e.g.*, $r = 1/8$) reduce temporal resolution and lead to minor degradation in R-Precision and MM-Dist due to information loss. Among all configurations, MAC-VAE with $(64, 1/4)$ achieves the best balance between reconstruction fidelity (FID$=0.000$, MPJPE$=0.012$) and generation quality (R-Top1$=0.588$, FID$=0.068$, MM-Dist$=2.620$),

Model	Config.	Reconstruction		Generation		
		FID↓	MPJPE↓	R-Top1↑	FID↓	MM-Dist↓
VAE	64, 1/4	0.001	0.016	0.561	0.107	2.706
C-VAE	64, 1/4	0.000	0.012	0.575	0.070	2.650
MAC-VAE	64, 1/4	**0.000**	**0.012**	**0.588**	0.068	**2.620**
MAC-VAE	32, 1/4	0.002	0.033	0.583	0.065	2.628
MAC-VAE	16, 1/4	0.011	0.077	0.573	0.071	2.647
MAC-VAE	64, 1/8	0.002	0.035	0.570	0.069	2.664
MAC-VAE	32, 1/8	0.006	0.060	0.566	0.057	2.704
MAC-VAE	16, 1/8	0.024	0.101	0.561	**0.054**	2.709

Table 8. Comparison of reconstruction and generation performance on HumanML3D. MPJPE is measured in millimeters. The notation (d, r) denotes the latent dimension d and the temporal downsampling rate r.

which we adopt as the default setting in all subsequent experiments. These results confirm that a compact latent space with moderate temporal compression effectively captures semantic and temporal dependencies for downstream motion generation.

Motion-Language Models To evaluate the effectiveness of different motion–language alignment strategies, we compare several pretrained motion–language models integrated into the MAC-VAE framework, including TMR [10], MotionPatches [19], and Part-TMR [20]. As shown in Table 9, all motion–language models improve generation quality while maintaining reconstruction performance compared to the baseline VAE and C-VAE, demonstrating the effectiveness of semantic alignment between motion and text.

Model	Reconstruction		Generation		
	FID↓	MPJPE↓	R-Top1↑	FID↓	MM-Dist↓
VAE	0.001	0.016	0.561	0.107	2.706
C-VAE	0.001	0.012	0.575	0.070	2.650
Part-TMR [20]	**0.000**	**0.012**	**0.588**	**0.068**	**2.620**
MotionPatches [19]	0.000	0.013	0.586	0.070	2.622
TMR [10]	0.001	0.013	0.580	0.070	2.638

Table 9. Comparison of motion-language models in MAC-VAE on HumanML3D. MPJPE is measured in millimeters.

Model Size	Config.	R-Precision↑			FID↓	MM-Dist↓
		Top1	Top2	Top3		
S (19M)	H2, L4, D512	0.543	0.738	0.834	0.247	2.845
M (38M)	H4, L8, D512	0.588	0.778	0.860	0.068	2.620
L (129M)	H6, L12, D768	0.585	0.779	0.859	0.044	2.621
XL (304M)	H8, L16, D1024	**0.590**	**0.779**	**0.861**	**0.042**	**2.610**

Table 10. Comparison of model sizes on HumanML3D. The notation (H, L, D) denotes the number of attention heads H, layers L, and hidden dimension D.

Text Encoder	Embedding	R-Precision↑			FID↓	MM-Dist↓
		Top1	Top2	Top3		
DistilBERT [15]	Word	**0.588**	**0.778**	**0.860**	**0.068**	**2.620**
CLIP [14]	Word	0.556	0.751	0.843	0.086	2.717
CLIP [14]	Sentence	0.527	0.717	0.809	0.145	2.941
Sentence-T5 [8]	Sentence	0.564	0.754	0.841	0.126	2.737

Table 11. Comparison of text encoders on HumanML3D.

Among them, Part-TMR achieves the best overall performance with the lowest reconstruction error (FID$=0.000$, MPJPE$=0.012$) and the highest R-Precision (0.588), confirming its strong ability to capture fine-grained part-level correspondences between text and motion. These results validate the choice of Part-TMR as the alignment backbone in MAC-VAE, enabling more semantically coherent and temporally consistent motion generation.

Model Size of Causal-DiT. We investigate the impact of model size on generation quality by varying the number of attention heads (H), layers (L), and hidden dimensions (D) in Causal-DiT. As shown in Table 10, larger models generally achieve better performance due to increased representational capacity. The medium-sized model (38M parameters) already provides strong results with an R-Precision of 0.588 and FID of 0.068, balancing quality and efficiency. Further scaling to 304M parameters yields marginal improvements (R-Precision$=0.590$, FID$=0.042$), demonstrating that Causal-DiT scales effectively while maintaining computational practicality. Unless otherwise specified, we use the medium (38M) configuration in all main experiments.

Text Encoder We compare several pretrained language models as text encoders to evaluate their impact on semantic alignment and motion quality. As shown in Table 11, the choice of text encoder influences both text–motion correspondence (R-Precision) and visual realism (FID). DistilBERT [15], which provides word-level embeddings, achieves the best overall performance with the highest R-Precision (0.588) and lowest FID (0.068), demonstrating its ability to capture fine-grained semantic cues that align well with motion features. Using the CLIP-based encoder, the word-level variant, which is identical to that employed in StableMoFusion [6], also outperforms StableMoFusion, further confirming the benefits of word-level representations. Such token-level embeddings are crucial for maintaining causal dependencies between linguistic tokens and motion frames, which is necessary for stable autoregressive generation in CMDM. In contrast, sentence-level embeddings from CLIP [14] exhibit reduced precision and higher FID due to the loss of temporal granularity. Meanwhile, Sentence-T5 [8] performs better than the CLIP-based models and also outperforms MotionLCM V2 [6], despite MotionLCM V2 also using Sentence-T5. These findings validate our choice of DistilBERT as the text encoder for CMDM, as it effectively preserves local semantics and enables causally consistent motion–language modeling.

C. Additional Qualitative Results

To further demonstrate the effectiveness of CMDM, we provide additional qualitative comparisons on long-horizon and text-to-motion generation. Fig. 4 and Fig. 5 compare CMDM with FlowMDM [1] and MARDM [7] on long-horizon motion generation for HumanML3D and SnapMoGen, respectively. CMDM produces temporally coherent and semantically accurate motions without content drift or skeleton flipping, whereas previous methods often suffer from static poses, incorrect transitions, or inconsistent actions across segments. These examples highlight the ability of CMDM to maintain smooth temporal dynamics and causal consistency throughout extended sequences.

Fig. 6 presents qualitative results on HumanML3D. Compared with MoMask [4], MotionLCM [3], and StableMoFusion [6], CMDM generates motions that more faithfully reflect fine-grained textual semantics (*e.g.*, arm rotations, leg movements, or walking direction) while preserving natural body articulation. Fig. 7 shows additional results on SnapMoGen, where CMDM directly uses the raw text prompts without LLM-based augmentation and still produces more realistic motions than prior methods. ***Please refer to the videos on the project homepage https://yu1ut.com/CMDM-HP/ for full-length visualizations.***

D. Sample Code

The code will be released at https://github.com/YU1ut/CMDM. We provide the training code for building and evaluating the proposed CMDM with the HumanML3D dataset. ***Please refer to the README file in the code directory for details.***

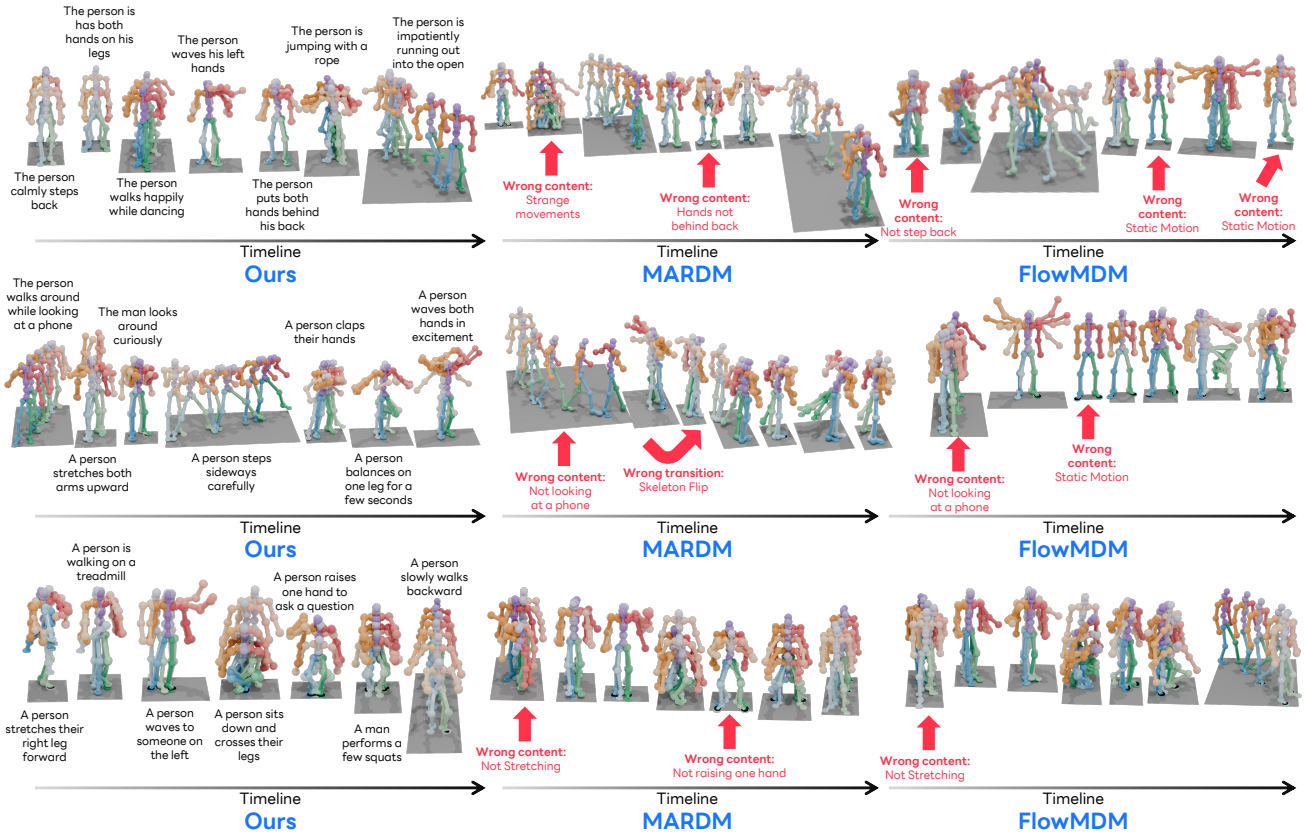

Figure 4. Qualitative results of long-horizon motion generation on HumanML3D. Comparison between our CMDM and previous methods. The generated motion is continuous and seamless; for visualization purposes, we split each long sequence into shorter segments corresponding to their captions. Please refer to the videos in the supplementary materials for the complete motion sequences.

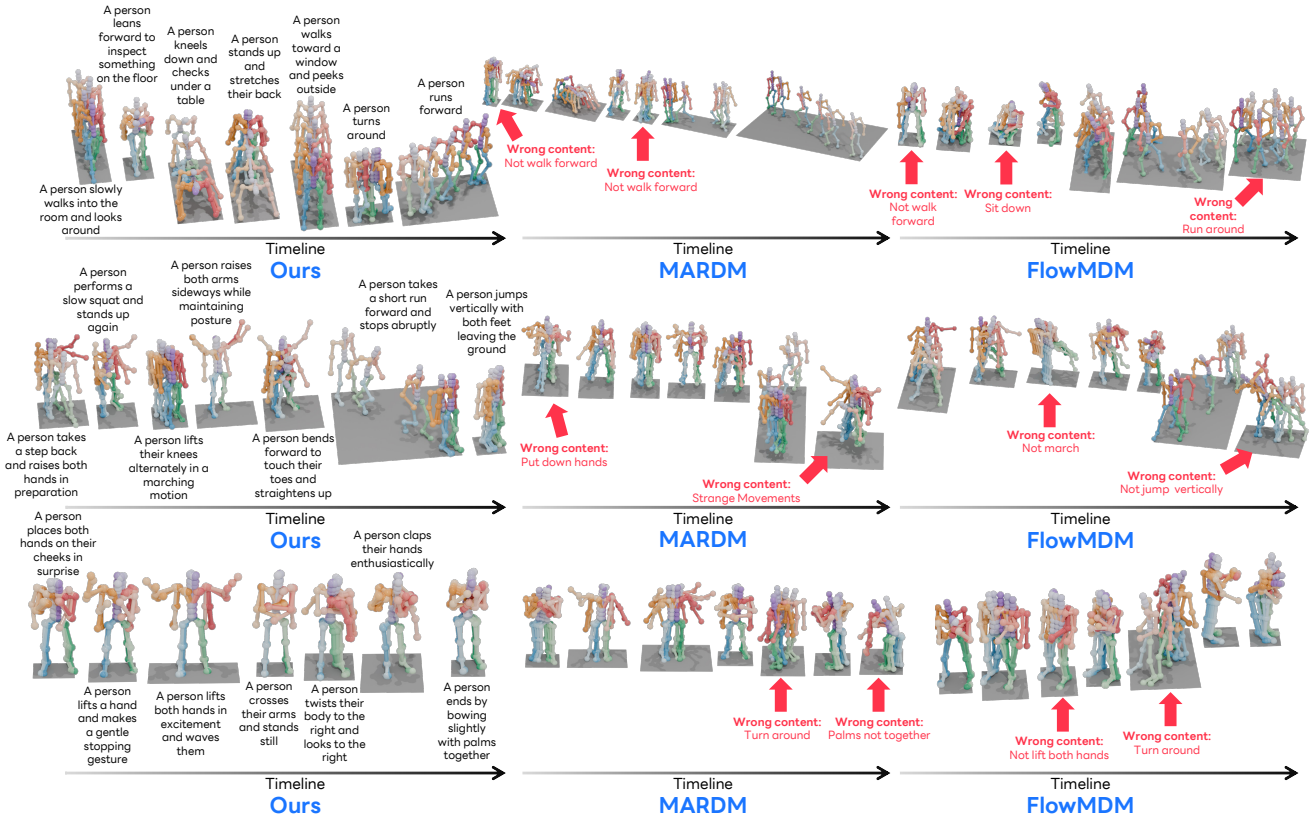

Figure 5. Qualitative results of long-horizon motion generation on SnapMoGen. Comparison between our CMDM and previous methods. The generated motion is continuous and seamless; for visualization purposes, we split each long sequence into shorter segments corresponding to their captions. Please refer to the videos in the supplementary materials for the complete motion sequences.

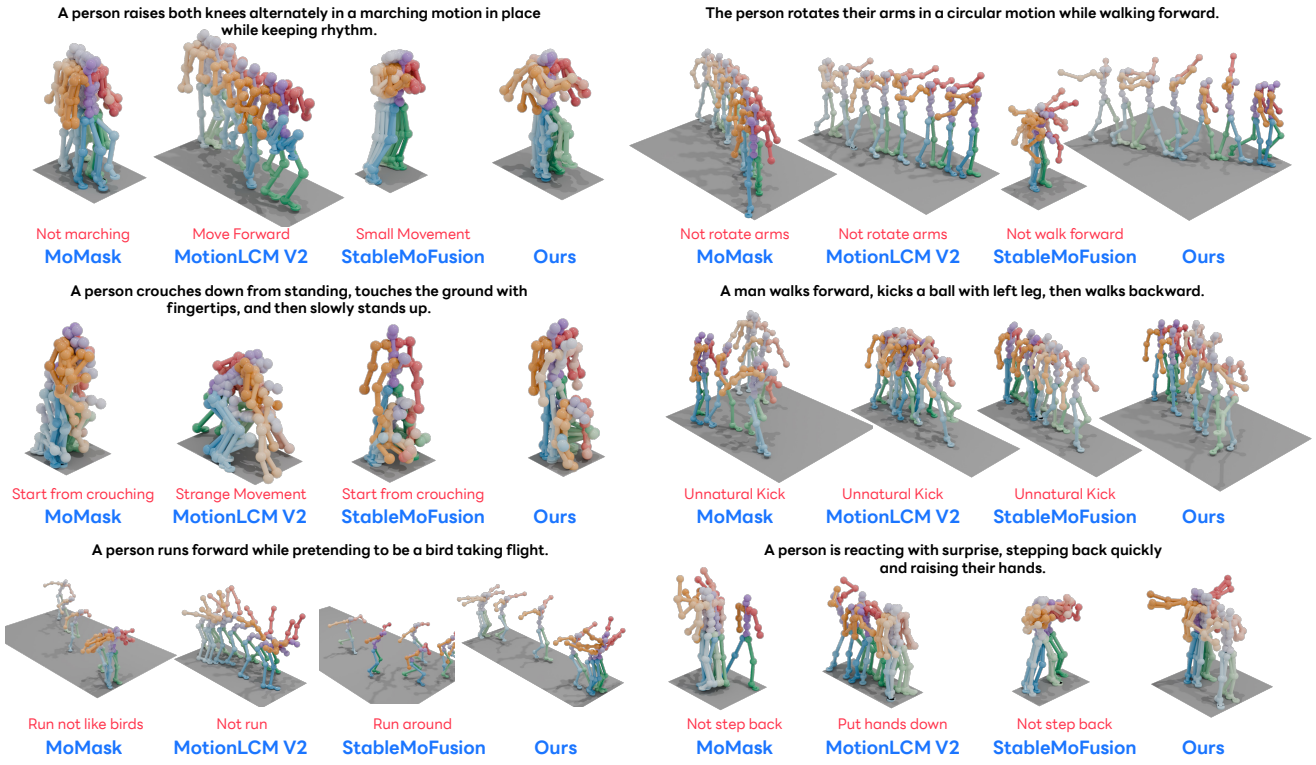

Figure 6. Qualitative results of text-to-motion generation on HumanML3D. CMDM produces motions that better capture fine-grained textual semantics and maintain natural body articulation compared to previous methods. Please refer to the supplementary videos for clearer visualization.

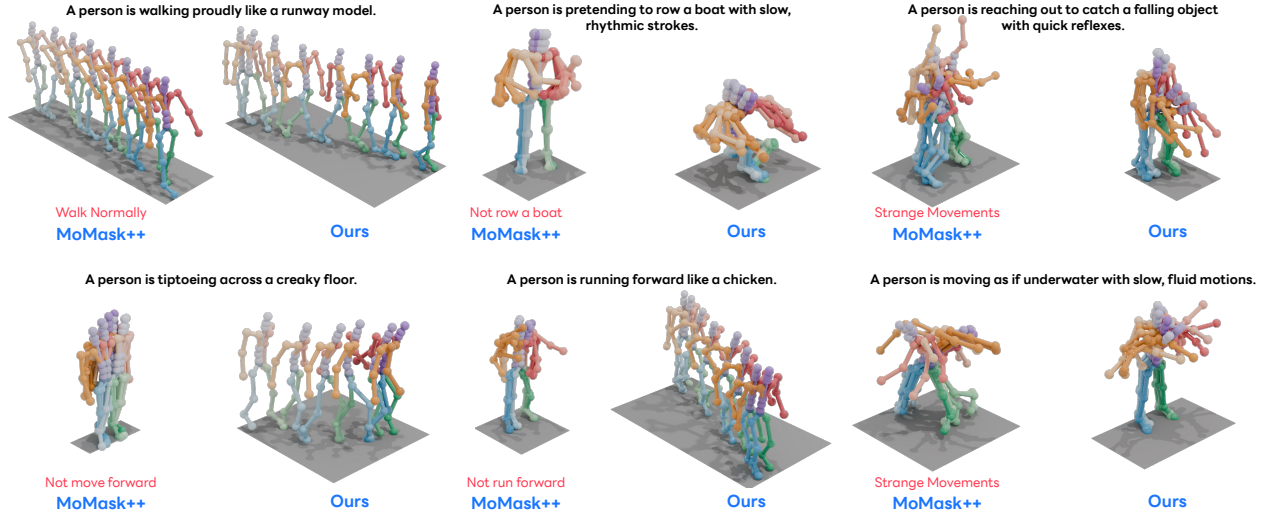

Figure 7. Qualitative results of text-to-motion generation on SnapMoGen. Comparison between our CMDM and previous methods. We directly use the raw text prompts without any LLM-based augmentation and CMDM still achieves strong generation quality. Please refer to the supplementary videos for clearer visualization.

References

[1] German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *CVPR*, 2024. 4

[2] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, 2023. 3

[3] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *ECCV*, 2024. 4

[4] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *CVPR*, 2024. 3, 4

[5] Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. In *CVPR*, 2025. 3

[6] Yiheng Huang, Hui Yang, Chuanchen Luo, Yuxi Wang, Shibiao Xu, Zhaoxiang Zhang, Man Zhang, and Junran Peng. Stablemofusion: Towards robust and efficient diffusion-based motion generation framework. In *ACMMM*, 2024. 4

[7] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation. In *CVPR*, 2024. 1, 2, 3, 4

[8] Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *ACL*, 2022. 4

[9] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *CVPR*, 2023. 1

[10] Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *ICCV*, 2023. 3, 4

[11] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J Black, Gul Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *CVPR-W*, 2024. 2, 3

[12] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *CVPR*, 2024. 3

[13] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *CVPR*, 2021. 1

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[15] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS-W*, 2019. 1, 4

[16] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 1

[17] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. In *ICLR*, 2023. 3

[18] Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. In *ICCV*, 2025. 2

[19] Qing Yu, Mikihiro Tanaka, and Kent Fujiwara. Exploring vision transformers for 3d human motion-language models with motion patches. In *CVPR*, 2024. 3, 4

[20] Qing Yu, Mikihiro Tanaka, and Kent Fujiwara. Remogpt: Part-level retrieval-augmented motion-language models. In *AAAI*, 2025. 1, 3, 4

[21] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023. 3

[22] Jianrong Zhang, Hehe Fan, and Yi Yang. Energymogen: Compositional human motion generation with energy-based diffusion model in latent space. In *CVPR*, 2025. 2, 3

[23] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *ICCV*, 2023. 3