Ablation Study of Diffusion Transformers for Human Motion Generation

Daniel Elgarici
Computer Science
Technion
elgarici-dan@campus.technion.ac.il

Gil Kizner
Computer Science
Technion
kizner.gil@campus.technion.ac.il

Eden Dembinsky
Computer Science
Technion
edendem@campus.technion.ac.il

Abstract

We present a systematic ablation study on the *Motion Diffusion Model* (MDM), a diffusion-based framework for human motion synthesis. We investigate how replacing the baseline transformer in MDM with *Diffusion Transformer* (DiT) blocks affects motion quality, and then conduct controlled ablations over (i) the self-attention mechanism and (ii) the temporal tokenization scheme. Our results show that incorporating DiT blocks with *AdaLN-Zero* conditioning improves convergence and fidelity in motion generation. Among the attention mechanisms, temporal *Sliding-Window* attention with a moderate window size provides the best trade-off between coherence and smoothness. Finally, a comparative study of temporal tokenizations, including per-frame, patch-based, and convolutional, reveals that learnable temporal aggregation can enhance motion diversity without sacrificing temporal consistency. These findings indicate that architectural priors from vision Diffusion Transformers can be effectively transferred to the domain of human motion modeling.

1 Introduction

Diffusion models have achieved state-of-the-art performance across modalities including images, audio, and motion. The *Motion Diffusion Model* (MDM) [1] demonstrated that text-conditioned diffusion can generate realistic human motion. However, the original MDM transformer stack predates several stabilization and conditioning techniques that later became standard in diffusion backbones.

We revisit MDM through a targeted ablation program. First, we integrate architectural components from the *Diffusion Transformer* (DiT) [2], originally proposed for image generation, and select the best-performing DiT block variant. Next, we ablate the self-attention mechanism (temporal Sliding-Window vs. spatial Neighborhood). Finally, fixing the attention, we study how tokenization granularity (Frame / Joint / Part) impacts learning dynamics and diversity.

Contributions. (1) A clean integration of DiT blocks into MDM. (2) An attention ablation isolating the effect of local temporal context vs. local spatial neighborhoods. (3) A tokenization ablation revealing the trade-offs between frame-level, joint-level, and part-level representations.

2 Background and Related Work

MDM. Tevet et al. [1] generate motion via a denoising diffusion process over pose trajectories with a transformer backbone and text/action conditioning.

Evaluation Metrics. The MDM framework evaluates generated motions using a combination of quantitative and perceptual metrics that reflect both realism and semantic consistency. The three primary measures are: *(1) R-Precision* (Top-1/Top-2/Top-3), which quantifies how well generated motions match their textual or action descriptions based on retrieval accuracy in a joint embedding space. *(2) Fréchet Inception Distance (FID)*, which measures the distributional distance between real and generated motion features, capturing perceptual realism and smoothness, *(3) Multimodality* which measures how varied the generated outputs are for the same condition. (Higher = the model better captures multiple valid possibilities instead of repeating the same pattern). and *(4) Diversity*, which computes the average feature variance among generated samples to assess motion variety and prevent model collapse. Together, these metrics provide complementary insights: R-Precision measures semantic alignment, FID captures visual quality, Multimodality captures variance and diversification and Diversity quantifies expressive variability, allowing a comprehensive evaluation of motion generation performance [1].

Diffusion Transformers. DiT [2] systematically explored how transformer architectures can replace U-Nets as diffusion model backbones. The authors proposed three core variants of DiT blocks, differing in how conditioning and normalization are applied: *(1) Standard DiT*, which directly inserts timestep and class embeddings through cross-attention layers, similar in spirit to class-conditional U-Nets. *(2) DiT-AdaLN*, which replaces traditional normalization with *Adaptive Layer Normalization (AdaLN)*, where each block's normalization parameters are modulated by a learned function of the diffusion timestep and class embedding, and *(3) DiT-AdaLN-Zero*, an initialization scheme where the modulation parameters are initialized to zero, ensuring the network initially performs identity denoising and gradually learns conditional structure during training. This design improves stability, convergence, and generalization across multiple image diffusion benchmarks. While DiT was originally developed for visual data, its architectural principles, explicit conditioning through AdaLN, transformer-based token processing, and scalable depth, are model-agnostic and transfer naturally to sequential domains such as 3D motion.

Efficient Attention. The computational bottleneck of standard self-attention lies in its $\mathcal{O}(N^2)$ complexity with respect to sequence length N, which becomes prohibitive for long motion trajectories containing hundreds of frames or spatial tokens. To address this, recent transformer variants introduce locality priors that restrict the receptive field of each token. Two complementary strategies are particularly relevant for motion modeling:

(1) Temporal locality via Sliding-Window Attention (SWA): instead of computing global attention across all tokens, SWA attends only to a fixed temporal window of w neighboring frames on each side, reducing complexity to $\mathcal{O}(Nw)$. This retains short-term temporal dependencies while enabling scalable long-horizon modeling. Such windowed attention has proven effective in long-sequence transformers such as the Longformer [3] and the Swin Transformer, where overlapping windows allow contextual flow across local regions.

(2) Spatial locality via Neighborhood Attention (NA): in motion or graph-structured data, tokens can represent spatial elements (e.g., body joints). NA restricts attention to a subset of spatially adjacent nodes within a k-hop neighborhood on the skeleton graph [4]. This formulation preserves local spatial coherence, joints within the same limb or torso naturally share context, while reducing redundant interactions between distant parts.

Each approach encodes a distinct inductive bias: SWA emphasizes temporal smoothness and continuity, which is crucial for realistic motion transitions, whereas NA enforces structured spatial reasoning aligned with human body topology. When integrated into diffusion-based motion models, these locality-aware attentions provide a principled way to balance global context with computational efficiency, allowing richer modeling of long sequences without the quadratic cost of full self-attention.

Tokenization for Motion. In transformer-based diffusion models, the way a motion sequence is divided into tokens strongly influences the model's temporal reasoning and computational efficiency. Unlike natural language, where tokens are discrete symbols, motion data are continuous spatiotemporal signals that must be segmented along the time axis before entering the transformer. Several

tokenization strategies have been proposed in prior work: *frame-based* tokenization, where each frame is represented as an individual token, *temporal patch* tokenization, which aggregates consecutive frames into a single representation to capture short-term dynamics and reduce sequence length, and *convolutional temporal* tokenization, which applies a 1D convolution across time to form tokens with learnable temporal pooling. These approaches define how local or global temporal dependencies are exposed to the attention layers, and thus form an essential design dimension in adapting diffusion transformers to motion generation.

3 Methodology

3.1 Base Model: MDM

We follow the public MDM setup [1]: motion uses continuous 6D joint rotations. conditioning uses CLIP embeddings per dataset, the denoiser is a transformer encoder predicting x_0 (signal) with classifier-free guidance. When predicting rotations, we add FK-based geometric losses (positions/velocity/foot-contact), and enable inference-time inpainting (temporal gaps or specific joints) by clamping known segments during sampling.

3.2 DiT Block Integration

We replace the baseline transformer block with DiT-style blocks. Each block contains: (i) AdaLN-Zero conditioning that modulates normalization via time/condition embeddings, (ii) multi-head attention. and (iii) a 2-layer MLP (GELU). We evaluate three DiT variants: (a) AdaLN-Zero, (b) Cross-Attention conditioning, and (c) In-Context conditioning.

3.3 Attention Mechanisms

After selecting the best DiT variant, in terms of Diversity, we ablate the self-attention: (1) **Sliding-Window Attention** (SWA) with temporal windows of size 4, 8, 12, 16. and (2) **Neighborhood Attention** (NA) with k-hop spatial neighborhoods (k = 16, 32) over a skeleton adjacency graph. For computational stability with Joint/Part tokenizations, SWA is implemented per-joint along the temporal axis to avoid quadratic scaling over $T \times J$.

3.4 Tokenization Strategies

We compare three temporal tokenizations that differ only in how frames are aggregated into tokens, leaving the DiT blocks and attention unchanged.

Frame tokens. Each frame is projected to a latent token (one token per frame). This keeps the longest sequence but preserves per-frame resolution.

Temporal patch tokens. We aggregate P consecutive frames into a single token via a learned linear projection of the concatenated frames. We evaluate two settings: $P{=}4$ with no overlap and $P{=}8$ with an overlap of 4 frames. Overlapping patches enable contextual flow while reducing the overall token count.

Conv-temporal tokens. We apply a 1D convolution over time on flattened per-frame features to produce a downsampled token sequence. We evaluate kernel-stride pairs $(5, 2)$ and $(7, 3)$, which control the temporal downsampling factor and the size of the learnable temporal receptive field.

4 Experiments

Dataset and Metrics. We evaluate using the HumanML3D [5] protocol, following the four standard metrics described earlier: R-Precision (Top-1/2/3) for text–motion alignment, Fréchet Inception Distance (FID) for perceptual realism, Diversity for inter-sample variation, and Multimodality for intra-condition variability.

Training. Unless noted otherwise, models are trained until convergence and with matched optimization hyperparameters (AdamW, cosine learning rate schedule with linear warm-up). We report mean $\pm$ std. over repeated runs where applicable.

4.1 Ablation 1: DiT Variants

Table 1: Quantitative comparison between MDM variants and real motion data. R-Precision measures retrieval accuracy (higher is better), while FID measures generation realism (lower is better). Diversity and Multimodality capture output variety.
DiT Block with adaLN-Zero is MDM-1
DiT Block with Cross-Attention is MDM-2
DiT Block with In-Context Conditioning is MDM-3

Model	R-Prec (top 1) ↑	R-Prec (top 2) ↑	R-Prec (top 3) ↑	FID ↓	Diversity →	Multimodality ↑
Real	0.4541 ± 0.0060	0.6612 ± 0.0065	0.7704 ± 0.0054	0.0010 ± 0.0001	9.1322 ± 0.0943	–
MDM-Original	0.3883 ± 0.0158	0.5828 ± 0.0091	0.6945 ± 0.0026	$\mathbf{0.5243 \pm 0.1885}$	9.3325 ± 0.0594	$\mathbf{2.6270 \pm 0.0539}$
MDM-1	0.4495 ± 0.0035	0.6576 ± 0.0060	$\mathbf{0.7681 \pm 0.0053}$	0.6466 ± 0.0774	$\mathbf{9.7104 \pm 0.1696}$	2.5973 ± 0.1703
MDM-2	$\mathbf{0.4546 \pm 0.0029}$	$\mathbf{0.6606 \pm 0.0033}$	$\mathbf{0.7681 \pm 0.0040}$	0.6042 ± 0.0868	9.0701 ± 0.1454	2.4932 ± 0.0947
MDM-3	0.3979 ± 0.0127	0.5877 ± 0.0123	0.6992 ± 0.0052	0.5602 ± 0.0817	9.3052 ± 0.1416	2.5170 ± 0.0678

4.2 Ablation 2: Attention Mechanisms

Table 2: Comparison between Sliding Window Attention (SWA) and Neighborhood Attention (NA) configurations. R-Precision measures retrieval accuracy (higher is better), FID measures realism (lower is better), and Diversity and Multimodality quantify output variety.

Attention Variant	R-Prec (top 1) ↑	R-Prec (top 2) ↑	R-Prec (top 3) ↑	FID ↓	Diversity →	Multimodality ↑
SWA-(with window size 4)	0.4102 ± 0.0081	0.6059 ± 0.0083	0.7205 ± 0.0087	1.1091 ± 0.1846	9.3799 ± 0.1220	2.4070 ± 0.0826
SWA-(with window size 8)	0.4100 ± 0.0135	0.6146 ± 0.0077	0.7254 ± 0.0055	0.7480 ± 0.1373	9.5032 ± 0.1696	2.2632 ± 0.0623
SWA-(with window size 12)	$\mathbf{0.4869 \pm 0.0070}$	$\mathbf{0.6805 \pm 0.0080}$	$\mathbf{0.7947 \pm 0.0076}$	$\mathbf{0.6248 \pm 0.0621}$	$\mathbf{9.6316 \pm 0.1798}$	2.6038 ± 0.1162
SWA-(with window size 16)	0.4057 ± 0.0063	0.6029 ± 0.0070	0.7211 ± 0.0037	0.9196 ± 0.1833	9.3885 ± 0.1943	2.4265 ± 0.1204
NA-16	0.3732 ± 0.0121	0.5723 ± 0.0093	0.6869 ± 0.0075	1.1105 ± 0.1239	9.1334 ± 0.1387	$\mathbf{2.6973 \pm 0.0513}$
NA-32	0.3857 ± 0.0121	0.5814 ± 0.0072	0.6973 ± 0.0071	0.6952 ± 0.0600	9.2068 ± 0.1127	2.5957 ± 0.0464

4.3 Ablation 3: Tokenization Strategies

For this ablation, we fix the attention to SWA with a window size of 16 (selected for its stability during development) and vary the tokenization granularity. We experimented with smaller window (12) and got slightly better results, but since the performance difference was minor, we opted for 16 to provide broader temporal context and more stable training. Results are presented in Table 3.

Table 3: Tokenization ablations under a fixed SWA configuration. We compare different temporal and spatial tokenization schemes. R-Precision measures retrieval accuracy (higher is better), FID evaluates realism (lower is better), and Diversity and Multimodality quantify output variety. The table presents comparisons between the chosen base model, the temporal patch tokenizations (with $P{=}4, O{=}0$ and $P{=}8, O{=}4$), and the conv-temporal tokenizations with kernel/stride pairs $(5, 2)$ and $(7, 3)$.

Tokenization Variant	R-Prec (top 1) ↑	R-Prec (top 2) ↑	R-Prec (top 3) ↑	FID ↓	Diversity →	Multimodality ↑
SWA-16 (base)	0.4057 ± 0.0063	0.6029 ± 0.0070	0.7211 ± 0.0037	0.9196 ± 0.1833	9.3885 ± 0.1943	2.4265 ± 0.1204
tubelet_p4o0	$\mathbf{0.4121 \pm 0.0158}$	$\mathbf{0.6109 \pm 0.0205}$	$\mathbf{0.7240 \pm 0.0161}$	$\mathbf{0.6275 \pm 0.0840}$	$\mathbf{9.5540 \pm 0.1772}$	2.2371 ± 0.0495
tubelet_p8o4	0.4064 ± 0.0112	0.6092 ± 0.0089	0.7209 ± 0.0054	0.8067 ± 0.0866	9.2741 ± 0.2226	2.4800 ± 0.0705
convk5s2	0.4111 ± 0.0147	0.6086 ± 0.0133	0.7232 ± 0.0079	0.9910 ± 0.1098	9.2075 ± 0.1155	2.4582 ± 0.1058
convk7s3	0.4062 ± 0.0064	0.6041 ± 0.0181	0.7166 ± 0.0137	0.9615 ± 0.1339	9.3084 ± 0.1009	$\mathbf{2.4867 \pm 0.1116}$

5 Discussion

AdaLN-Zero conditioning. We observed improvements in the final motion quality. The zero-initialized adaptive modulation allows the network to initially behave as a simple denoiser, then progressively learn complex conditional dependencies, stabilizing diffusion training.

Temporal attention locality. Temporal Sliding-Window attention consistently outperforms purely spatial neighborhood mechanisms in our experiments, indicating that maintaining short-term temporal coherence is more critical for realistic motion synthesis than enforcing local spatial correlations across joints. This supports the view that temporal context dominates motion fidelity in diffusion-based sequence generation.

Tokenization effects. The results in Table 3 reveal that temporal tokenization has a clear impact on motion quality and diversity. Temporal patch tokens ($P{=}4, O{=}0$ and $P{=}8, O{=}4$) improve both retrieval precision and FID compared to the baseline, indicating that aggregating short frame windows helps the model capture local temporal coherence while reducing noise in frame-wise representations. Conv-temporal tokens, which perform learnable temporal aggregation via 1D convolutions, further enhance multimodality and maintain competitive FID, suggesting that the convolutional kernels adaptively emphasize motion patterns over flexible timescales. Overall, learned temporal aggregation provides smoother and more diverse motion generation by balancing fine-grained detail with broader contextual awareness.

6 Conclusion and Future Work

We conducted a systematic ablation of Diffusion Transformer (DiT) components within the Motion Diffusion Model (MDM) framework. The best-performing configuration combines DiT with AdaLN-Zero conditioning and a temporal Sliding-Window attention using a moderate window size ($w{=}12$). Under this setup, temporal patch and conv-temporal tokenizations achieved the most balanced trade-off between fidelity, realism, and diversity, highlighting the importance of learnable temporal aggregation for high-quality motion synthesis.

Future Work. Future directions include extending the Neighborhood Attention experiments to larger window sizes beyond the tested $w{\in}\{16, 32\}$, as broader receptive fields may capture longer-range spatial dependencies. Additionally, exploring a wider range of parameters for both temporal patch and conv-temporal tokenizations (e.g., larger patch sizes or alternative convolutional strides) could further clarify their effect on motion smoothness and diversity. Beyond these, hybrid temporal, spatial attention mechanisms and hierarchical multi-scale DiT backbones and experimenting with varying number of attention heads, as each head can capture different aspects of the motion representation. All of these offer promising methods for modeling longer and more complex motion sequences.

Reproducibility. All implementation details and reproduction instructions are available in our public GitHub repository: https://github.com/DanDanielElgarici/transformers-project/tree/main.

References

[1] G. Tevet, S. Hertz, B. Shafir, et al. Human Motion Diffusion Model. In *International Conference on Learning Representations (ICLR)*, 2023.

[2] W. Peebles and S. Xie. DiT: Diffusion Models with Transformer Backbones. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[3] I. Beltagy, M. Peters, and A. Cohan. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[4] A. Hassani and S. Walton. Neighborhood Attention Transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[5] C. Guo, C. Jiang, J. Lan, et al. HumanML3D: A Large-Scale 3D Humanoid Motion Dataset for Motion Understanding and Generation. In *European Conference on Computer Vision (ECCV)*, 2022.