أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Physics-Informed Continuous Autoencoding for Compact 3D Human Motion

Osama Alshammari, **Omar Hamad**, **Dr. Kerven Durdymyradov**

zacx.college.ai@gmail.com, omarhamadevl@gmail.com, kerventkm@gmail.com

Abstract

Human motion is temporally redundant, but compact representations must preserve articulated pose, global trajectory, and ground contact. Frame-wise representations retain detail at the cost of long sequences, while discrete tokenizers can introduce quantization error. We present a continuous autoencoder that maps variable-length 3D human motion to one 256-dimensional token per eight frames. Each token reserves 224 learned dimensions for articulated motion and 32 deterministic dimensions for planar-root velocity, yaw velocity, root-height statistics, and soft foot-support cues. The carrier is part of the reported token and is decoded jointly with learned features; it is not an external trajectory stream. The model is trained for 300 epochs on HumanML3D with reconstruction, kinematic, geometric, contact, and trajectory objectives. On 4,210 eligible clips from the official HumanML3D test split, using no prediction-side alignment, Ours obtains 8.940 mm global MPJPE, 8.887 mm root-relative MPJPE, 62.763 mm/s velocity error, 24.667 mm/s contact-foot speed, and effectively zero terminal root drift. Under the same reconstruction protocol, MotionGPT obtains 172.085 mm MPJPE and MLD-1 obtains 104.408 mm. These results establish accurate 8× temporal compression for reconstruction; they do not yet establish equal-bitrate superiority or downstream text-to-motion performance.

Date: 18 August 2026

1 Introduction

Compact motion representations reduce the sequence length processed by motion generators, retrieval systems, and language-conditioned models. The difficulty is that human motion is not merely a sequence of independent poses. Small errors in planar-root velocity, heading, body configuration, or foot contact accumulate over time and appear as trajectory drift, foot sliding, and implausible ground interaction.

Two common compression strategies expose different trade-offs. Frame-wise continuous features preserve local information but produce long sequences. Vector-quantized models shorten the sequence and provide discrete symbols, but every segment must be represented through a finite codebook. A very small global continuous latent can be even shorter, yet may discard local temporal detail. This work instead targets a fixed 8× reduction in temporal steps while retaining a continuous, locally structured latent sequence.

Our model is a variable-length continuous motion autoencoder built around an explicit in-token physical carrier. Every group of eight frames maps to one 256-dimensional token. The majority of the token is learned, while 32 dimensions preserve the root and contact quantities whose errors accumulate most strongly. The decoder consumes the entire token jointly and constrains learned root-height residuals so that they cannot overwrite carrier endpoints and block means.

The main contributions are:

- a mask-consistent hierarchical autoencoder that supports 40–192 frames at a fixed 8× temporal rate;
- a 256-dimensional continuous token containing 224 learned values and a 32-dimensional physical carrier, with no external side channel;
- physics-informed training that supervises reconstruction, velocity, acceleration, bone length, contact, foot sliding, penetration, and terminal planar motion at 20 FPS; and
- a full official HumanML3D test evaluation on 4,210 eligible clips against released MotionGPT and MLD-1 autoencoders under a shared, alignment-free reconstruction protocol.

The reported comparison concerns reconstruction quality and temporal token rate. It is not an equal-bitrate comparison: Ours stores continuous values, MotionGPT stores codebook indices, and MLD-1 uses one global continuous slot.

2 Related Work

Motion data and continuous representations.. HumanML3D pairs a large collection of 3D motions with natural-language descriptions and standardizes a 22-joint representation at 20 FPS [1]. It is derived from motion sources including AMASS [2]. Continuous variational autoencoders have been used for action-conditioned motion and latent diffusion, including ACTOR [3] and MLD [4]. Such models avoid codebook quantization, but aggressively global latents can remove local detail.

Discrete motion tokenization.. Vector-quantized representation learning [5] has enabled motion to be modeled as discrete token sequences. T2M-GPT [6], MotionGPT [7], MMM [9], and MoMask [8] use discrete motion representations for text-conditioned generation. MoMask improves reconstruction through residual quantization, while ParTY adds temporal-aware aggregation and body-part guidance for expressive synthesis [10]. Published HumanML3D studies of these tokenizers cover both 4× and 8× temporal compression, providing useful context for our 8× continuous representation. Our method asks whether alignment-free continuous reconstruction

can retain more local and physical detail at that longer temporal rate.

Structured and physics-aware latents.. PRISM factorizes motion latents by joint and time, showing that latent organization can materially affect both reconstruction and generation quality [11]. It reports 9.72 mm HumanML3D reconstruction MPJPE with 4× temporal downsampling and a 1.5× scalar information-compression ratio. Our method instead partitions each temporal token by function: a learned articulated-motion subspace and a deterministic physical carrier. The term *physics-informed* in this paper refers to kinematic, geometric, and contact constraints. Unlike simulator-based reward systems such as PhyMotion [12], our model does not estimate forces or torques and does not use MuJoCo during training or evaluation.

3 Method

3.1 Input representation and temporal hierarchy

For each frame, the native input is a 70-dimensional continuous vector

$$x_t = [p_t^{\text{local}},\ v_t^{xy},\ \omega_t,\ h_t] \in \mathbb{R}^{66+2+1+1}, \tag{1}$$

where p^{local} contains 22 local-joint coordinates, v^{xy} is planar-root velocity, ω is yaw velocity, and h is root height. Train-only means and standard deviations normalize these channels.

Three hierarchical stages, each containing two masked Transformer blocks followed by stride-2 temporal downsampling, map a valid sequence of T frames to

$$L = \left\lceil \frac{T}{8} \right\rceil \tag{2}$$

latent positions. Explicit frame and latent masks are propagated through attention, downsampling, upsampling, decoding, and every loss term. Training batches contain only fully valid or fully padded eight-frame blocks, preventing partially observed carrier tokens.

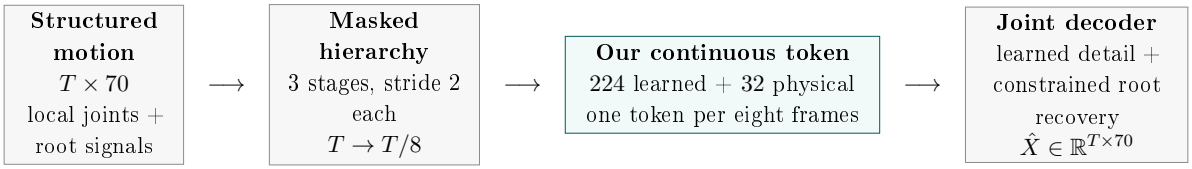


Figure 1 Overview of our autoencoder. The 32-dimensional physical carrier remains inside each reported 256-dimensional token; there is no external root or contact side channel.

3.2 In-token physical carrier

Each token is partitioned as

$$z_k = [z_k^{\text{learned}}; c_k^{\text{physical}}] \in \mathbb{R}^{224+32}. \tag{3}$$

For the corresponding eight-frame block, the carrier contains all eight planar-root velocity pairs (16 values), all eight yaw velocities (8 values), root height at the block start and end plus its mean (3 values), mean soft support confidence for four foot joints (4 values), and root-height standard deviation (1 value). Soft support is derived differentiably from input foot height and speed; ground-truth contact labels are not encoder inputs.

The carrier is concatenated to the learned code before the decoder projection, so both components jointly condition articulated reconstruction. Planar and yaw velocities are recovered

directly from the carrier. Root height is reconstructed from the carrier endpoints and mean plus a learned within-block residual whose endpoints and block mean are projected to zero. This constraint preserves long-range planar trajectory exactly within floating-point tolerance while allowing learned vertical detail.

3.3 Physics-informed objective

The training objective is

$$\begin{aligned}\mathcal{L} = {} & 2\mathcal{L}_{\text{local}} + 10\mathcal{L}_{\text{global}} + 0.5\mathcal{L}_{\text{SE(2)}-\text{vel}} + 0.1\mathcal{L}_{\text{heading}} + 2\mathcal{L}_{\text{height}} \\ & + 0.05\mathcal{L}_{\text{contact}} + 0.05\mathcal{L}_{\text{vel}} + 5 \times 10^{-4}\mathcal{L}_{\text{acc}} + 2\mathcal{L}_{\text{bone}} + 0.1\mathcal{L}_{\text{slide}} \\ & + 10\mathcal{L}_{\text{anchor}} + 5\mathcal{L}_{\text{penetration}} + \mathcal{L}_{\text{terminal}-\text{SE(2)}}. \end{aligned} \tag{4}$$

Velocity and acceleration are evaluated in SI units using the HumanML3D rate $f = 20$ FPS. Representative terms are

$$\mathcal{L}_{\text{global}} = \frac{1}{|M|J} \sum_{t,j} M_t \|\hat{p}_{t,j} - p_{t,j}\|_1, \tag{5}$$

$$\mathcal{L}_{\text{vel}} = \frac{1}{|M'|J} \sum_{t,j} M'_t \|f\Delta\hat{p}_{t,j} - f\Delta p_{t,j}\|_1, \tag{6}$$

$$\mathcal{L}_{\text{slide}} = \frac{1}{|M'|} \sum_{t,j\in\mathcal{F}} M'_t q_{t,j} \|f\Delta\hat{p}^{xy}_{t,j}\|_2, \tag{7}$$

where M is the valid-frame mask, $\mathcal{F}$ contains four foot joints, and q denotes target contact supervision used by the loss. The complete model has 22.59 million parameters and retains a 256-dimensional latent width.

4 Experimental Protocol

4.1 Dataset and split discipline

Experiments use the official HumanML3D split and 22-joint representation at 20 FPS [1]. Only the training split updates model parameters, the validation split selects checkpoints, and the test split is used once for final evaluation. Training samples variable sequence lengths from 40 to 192 frames. The selected checkpoint is the epoch-300 exponential-moving-average model with the best validation MPJPE.

The official test split contains 4,384 identifiers. A shared comparison requires at least 40 frames, leaving 4,210 eligible clips and excluding 174 short clips; no short clip is padded. Each model receives the same centered crop, chosen as the longest length divisible by eight and capped at 192 frames. All 4,210 requested clips are evaluated with zero skips.

4.2 Baselines and coordinate protocol

MotionGPT uses its released HumanML3D VQ-VAE weights at 4× temporal compression [7]. MLD-1 uses its released HumanML3D VAE weights with one global latent slot [4]. Ours uses 24 tokens for a 192-frame crop, MotionGPT uses 48 discrete tokens, and MLD-1 uses one global continuous slot.

Evaluation recovers canonical HumanML3D joints in metres. A target-derived canonical transform is shared by target and prediction; there is no prediction-specific root, heading, rigid, or Procrustes alignment. Independent target recovery is required to agree within 5×10^{-6} m mean error and 2×10^{-5} m maximum error. This check validates coordinate conversion and does not align predictions.

4.3 Metrics and optimization

Global MPJPE measures absolute joint reconstruction, while root-relative MPJPE removes global root translation. Velocity and acceleration errors evaluate temporal derivatives. Contact-foot speed measures sliding during target contact, penetration measures the fraction of joints more than 5 mm below the ground plane, and terminal root drift measures final trajectory error. Lower is better for every reported metric. Micro results weight motion elements across the full dataset; macro results weight clips equally.

Our model is trained for 300 epochs with batch size 128, AdamW learning rate 10^{-4}, weight decay 10^{-4}, five warm-up epochs, gradient clipping at 1.0, EMA decay 0.999, bfloat16 precision, and random seed 42. Train-only normalization uses 10,000 samples.

5 Results

5.1 Full HumanML3D test reconstruction

Table 1 reports the recommended element-weighted aggregate over 4,210 official test clips. Ours obtains 8.940 mm global MPJPE and 8.887 mm root-relative MPJPE. Relative to MotionGPT, global MPJPE is 94.8% lower despite using half as many temporal positions; relative to MLD-1 it is 91.4% lower. The carrier makes terminal planar-root drift numerically negligible.

Table 1 HumanML3D full-test micro aggregate over 4,210 eligible clips. Lower is better.

Model	Global MPJPE (mm)	Root-relative (mm)	Velocity (mm/s)	Acceleration (m/s^2)	Contact speed (mm/s)	Penetration (%)	Terminal drift (mm)
Ours 8×	**8.940**	**8.887**	**62.763**	**1.369**	**24.667**	**0.019**	**0.000**
MotionGPT VQ 4×	172.085	55.987	212.885	2.934	105.854	0.345	251.731
MLD-1	104.408	43.924	223.479	2.271	66.276	0.038	126.876

To place the result in the broader HumanML3D reconstruction literature, Table 2 collects reported MPJPE values for compact autoencoders evaluated on the same benchmark. It includes the available 4× and 8× settings. These values are taken from the respective papers rather than recomputed under our alignment-free protocol, so they provide benchmark context rather than a controlled replacement for Tables 1 and 3. MMM reports reconstruction FID, but not reconstruction MPJPE, and is therefore not included in this MPJPE-only table [9].

Table 2 Published HumanML3D reconstruction MPJPE at 4× and 8× temporal compression. Lower is better. Protocols and motion representations follow the respective papers.

Model	Temporal compression	MPJPE (mm)
Ours	8×	**8.940**
PRISM [11]	4×	9.72
MoMask + ParTY temporal modules [10]	4×	11.0
MoMask + ParTY temporal modules [10]	8×	14.0
ParTY holistic temporal-aware VQ-VAE [10]	8×	19.0
MoMask RVQ-VAE [8]	4×	29.5
MoMask VQ-VAE, larger window [10]	8×	55.0
T2M-GPT VQ-VAE [6, 8]	4×	58.0

Among the reported HumanML3D results in Table 2, Ours has the lowest MPJPE while operating at 8× temporal compression. PRISM is the closest result at 9.72 mm and 4× temporal compression; its joint-factorized latent has a reported scalar information-compression ratio of

only 1.5×, whereas our 70-to-256 design has a 2.19× scalar ratio over each eight-frame block. The ParTY results also show that temporal-aware aggregation substantially narrows the reconstruction penalty when the window is enlarged from four to eight frames.

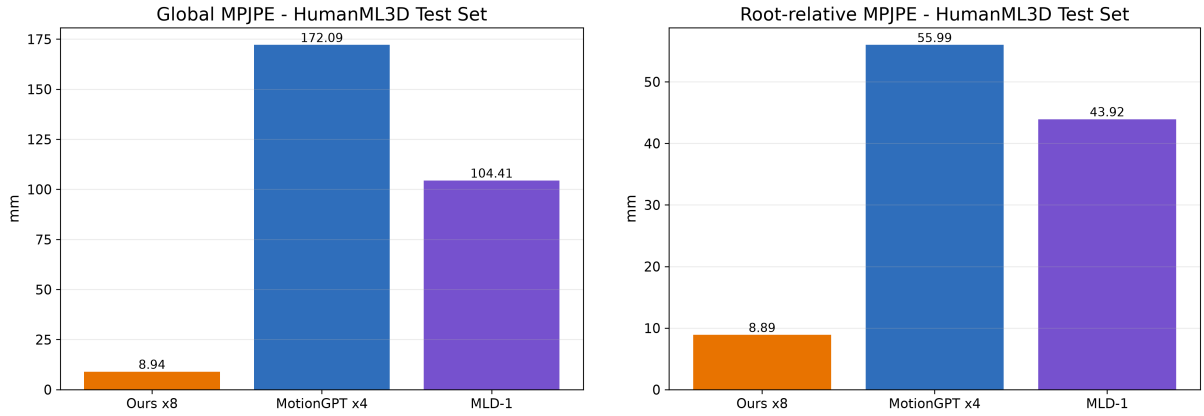


Figure 2 Global and root-relative MPJPE on the official HumanML3D test split.

The improvement is not limited to static position error. Ours reduces velocity error by 70.5% and contact-foot speed by 76.7% relative to MotionGPT. Against MLD-1, the corresponding reductions are 71.9% and 62.8%. Penetration remains below 0.02% of evaluated joints.



Figure 3 Temporal and contact fidelity on the full HumanML3D test set.

The equal-clip macro aggregate follows the same ranking (Table 3), indicating that the result is not produced only by longer clips.

Table 3 HumanML3D full-test macro aggregate, weighting every clip equally.

Model	Global MPJPE (mm)	Root-relative (mm)	Velocity (mm/s)	Acceleration (m/s^2)	Contact speed (mm/s)	Penetration (%)	Terminal drift (mm)
Ours 8×	**8.996**	**8.932**	**66.621**	**1.502**	**31.700**	**0.024**	**0.000**
MotionGPT VQ 4×	153.652	52.816	211.865	3.044	117.709	0.280	251.731
MLD-1	90.506	39.362	212.416	2.369	84.242	0.036	126.876

5.2 Per-clip distribution and qualitative reconstruction

The per-clip distribution in Figure 4 shows that Ours has a broad advantage rather than one explained by a small number of outliers.

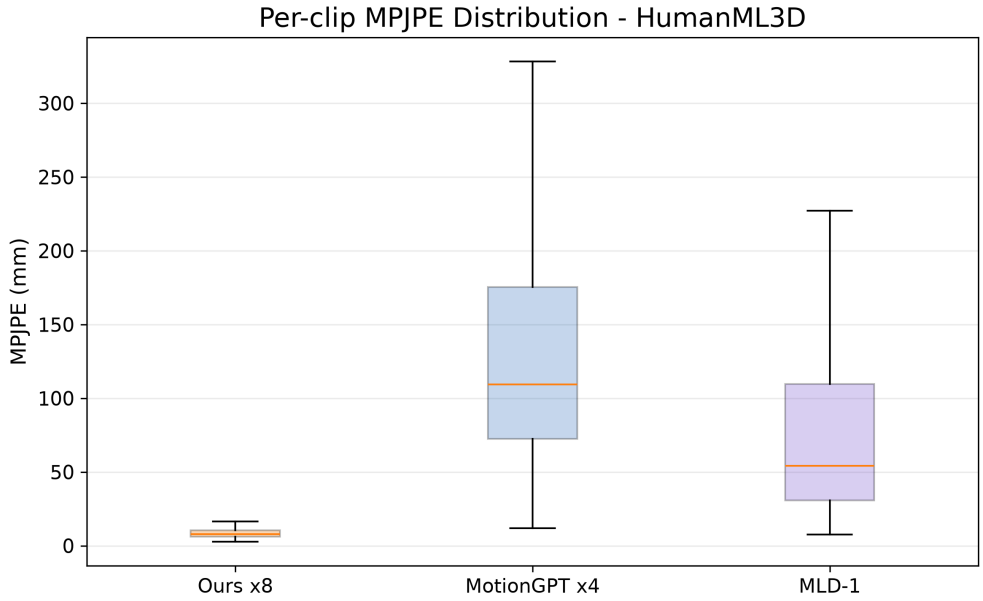


Figure 4 Distribution of per-clip global MPJPE across the eligible HumanML3D test set.

Figure 5 provides eight evenly spaced, synchronized poses from a single 192-frame held-out sequence. It is included as a qualitative illustration of reconstruction across time; the dataset-level conclusions remain those reported in Table 1.

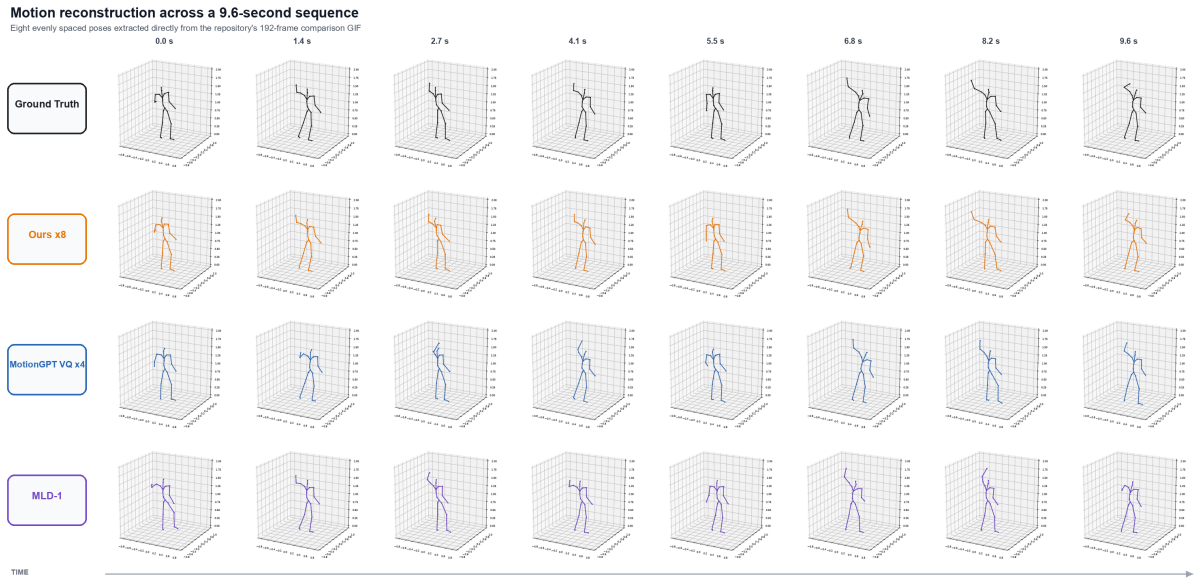


Figure 5 Synchronized qualitative reconstruction sequence for ground truth, Ours, MotionGPT, and MLD-1, with no post-hoc alignment.

6 Discussion and Limitations

The full-test results support the central hypothesis: explicitly preserving signals whose errors accumulate can improve an otherwise learned continuous bottleneck. Ours has root-relative MPJPE close to its global MPJPE, and terminal root drift is effectively zero, consistent with exact in-token planar and yaw velocity recovery. Its improvement in velocity and contact-foot

speed indicates that the result is not merely a coordinate-system advantage. The broader HumanML3D comparison in Table 2 places the 8.940 mm result below reported 4× and 8× reconstruction MPJPE values from PRISM, T2M-GPT, MoMask, and ParTY-based tokenizers. This is encouraging evidence that the gain persists beyond the two released baselines evaluated directly in this work.

Several boundaries remain important. First, 8× describes temporal-token reduction, not total bitrate. Our token contains 256 continuous values, whereas a MotionGPT token is a codebook index and MLD-1 uses one global continuous slot. PRISM, for example, uses 4× temporal downsampling but retains a grid of 23 joint-associated latent tokens, yielding a reported 1.5× scalar information-compression ratio. Our 8× temporal design yields a 2.19× scalar ratio under the corresponding input-to-latent scalar calculation. The current comparison therefore measures reconstruction at each model's representation, not rate-distortion at equal stored bits. Second, the published values in Table 2 share the HumanML3D benchmark but follow their papers' coordinate representations and evaluation details; Tables 1 and 3 remain the controlled alignment-free comparison. Third, the task is autoencoding: text-to-motion FID, retrieval, diversity, and semantic alignment are outside the present evidence. Fourth, the physical carrier is deterministic and derived from the motion input; downstream generative models must learn to generate valid carrier values. Fifth, zero planar terminal drift follows from the representation design and does not imply that every local pose or contact is exact. Finally, the losses are kinematic and geometric rather than force- or torque-level simulation.

Future evaluation should include a matched continuous 8× control without the carrier, equal-bitrate discrete and continuous controls, downstream text-to-motion generation, latency and FLOP profiling under identical hardware, and robustness to generated or noisy carrier values. Per-action and per-length analyses would also identify the motion regimes in which the carrier is most beneficial.

7 Conclusion

Our continuous 3D human-motion autoencoder compresses eight frames into one 256-dimensional token while preserving high-value physical signals inside the token itself. A mask-consistent hierarchy captures articulated motion, a 32-dimensional carrier retains planar-root, yaw, root-height, and support information, and physics-informed losses supervise temporal and contact quality. On 4,210 eligible clips from the official HumanML3D test split, Ours achieves 8.940 mm global MPJPE with effectively zero terminal root drift and substantially lower temporal and contact errors than the evaluated released MotionGPT and MLD-1 autoencoders. The same result is also below published HumanML3D reconstruction MPJPE values from compact 4× and 8× PRISM, T2M-GPT, MoMask, and ParTY-based representations. This demonstrates strong reconstruction under 8× temporal compression, while matched-protocol equal-bitrate and downstream generative evaluation remain necessary before broader representation claims.

References

[1] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating Diverse and Natural 3D Human Motions From Text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[2] N. Mahmood, N. Ghorbani, N. Troje, G. Pons-Moll, and M. Black, "AMASS: Archive of Motion Capture as Surface Shapes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[3] M. Petrovich, M. Black, and G. Varol, "Action-Conditioned 3D Human Motion Synthesis with Transformer VAE," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[4] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, and G. Yu, "Executing Your Commands via Motion Diffusion in Latent Space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[5] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural Discrete Representation Learning," in *Advances in Neural Information Processing Systems*, 2017.

[6] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, and X. Shen, "T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[7] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen, "MotionGPT: Human Motion as a Foreign Language," in *Advances in Neural Information Processing Systems*, 2023.

[8] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng, "MoMask: Generative Masked Modeling of 3D Human Motions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[9] E. Pinyoanuntapong, P. Wang, M. Lee, and C. Chen, "MMM: Generative Masked Motion Model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[10] K. Heo, S. Kim, Y. Gwon, Y. Kim, and M. Cho, "ParTY: Part-Guidance for Expressive Text-to-Motion Synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[11] Z. Ling, Q. Shuai, T. Zhang, S. Li, B. Han, and C. Zou, "PRISM: Per-unit Kinematic Motion Generation in a Structured Latent Manifold," arXiv:2603.08590, 2026.

[12] Y. Huang, Z. Wang, H. Lin, D.-K. Kim, S. Omidshafiei, J. Yoon, J. Cho, Y. Zhang, and M. Bansal, "PhyMotion: Structured 3D Motion Reward for Physics-Grounded Human Video Generation," arXiv:2605.14269, 2026.