CAN TEXT-TO-3D HUMAN MOTION MODELS REALLY UNDERSTAND PHYSICAL ACTIONS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-3D human motion generation aims to translate natural-language intentions into 3D human motion, serving as an important bridge between language understanding and human behavior generation, with broad applications in digital humans, computer animation, and human–computer interaction. Existing methods primarily rely on large-scale motion data to learn motion representations and generative distributions, while improving generation quality and motion naturalness through generative modeling, motion priors, or post-training optimization. However, these approaches provide limited explicit supervision over the physical regularities and latent physical distributions underlying human motion during training. Consequently, generated motion trajectories can still deviate substantially from the distribution of real human motion. To address this problem, we propose MoPhys (Motion Physics Modeling), a time–frequency modeling and physics-expert collaborative framework for physically consistent text-to-3D human motion generation. Specifically, MoPhys employs multi-scale time–frequency interactions to capture motion variations across different temporal scales, and introduces physics-inspired experts constructed from body observables to adaptively refine generated motions. MoPhys further transforms structural and temporal information from real motions into direct supervision for the generative model, guiding it to learn physical constraints embedded in the training data and to generate motions that better conform to real human motion distributions. Extensive experiments show that MoPhys consistently improves motion kinematics and spectral statistics across diverse data distributions and multiple human motion generation paradigms, demonstrating the effectiveness and generalizability of the proposed framework.

1 INTRODUCTION

Text-driven 3D human motion generation aims to translate natural-language descriptions into semantically aligned and temporally coherent 3D motion sequences, and has become an important research direction connecting language understanding with human behavior generation. As a key technology bridging high-level semantic understanding and low-level motion control, 3D motion generation has substantial application value in digital content creation, virtual human interaction, embodied intelligence, and related areas. However, in addition to correctly expressing action semantics, generated motion sequences should preserve plausible body structure, joint coordination, and temporal dynamics. Prior work has shown that motions with high visual quality can still exhibit physically implausible artifacts such as foot sliding, floating, and penetration (Yuan et al., 2023). This motivates a more fundamental question: when a model can generate semantically correct human motion, has it also sufficiently learned the latent kinematic regularities of real human motion?

Existing text-driven human motion generation research mainly improves motion representation and generation quality through generative modeling, quality alignment, and physics-guided optimization. As shown in Figure 1, the motions generated by the model better conform to physical laws after physics-based correction. For generative modeling, representative methods learn complex text-conditioned motion distributions through diffusion in motion or latent spaces, or through sequence modeling over discrete motion representations (Tevet et al., 2023; Chen et al., 2023; Zhang et al., 2023; Jiang et al., 2023; Guo et al., 2024; Barquero et al., 2024). For quality alignment, recent studies further employ preference learning, reward guidance, post-training optimization, and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

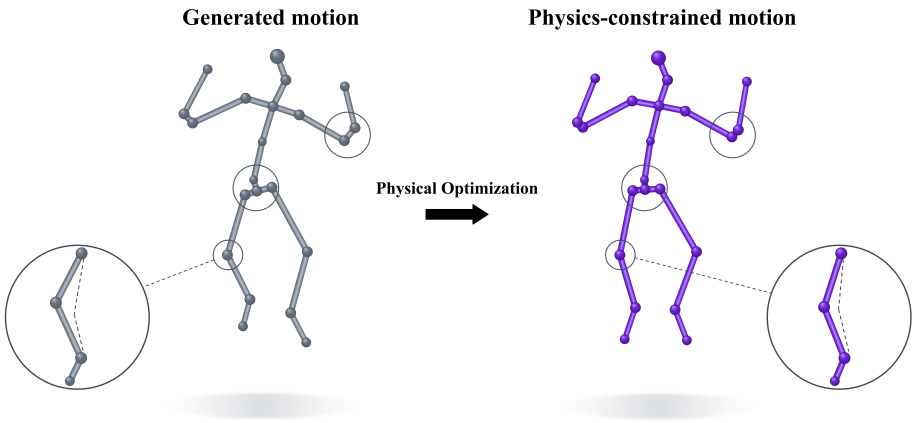


Figure 1: Conceptual illustration of our motivation. Left: a generated jumping motion with irregular local joint configurations. Right: a corresponding motion with more plausible joint configurations under physical constraints, while retaining similar semantics and overall pose. Circles and magnified views highlight local differences.

cross-modal semantic alignment to improve overall motion quality and text–motion consistency (Tan et al., 2025; Weng et al., 2026; Han et al., 2026). For physics guidance, several works incorporate physical constraints into generation or optimization using physics simulation, motion tracking, or control feedback (Yuan et al., 2023; Juravsky et al., 2024; Tevet et al., 2025; Yue et al., 2026). Nevertheless, existing methods still largely optimize motion representations through data fitting and generation-quality alignment, while observable physical signals contained in real human motion are not sufficiently transformed into effective training supervision. As a result, a noticeable gap can remain between generated motions and the motion distribution of real humans.

To address this issue, we propose MoPhys (Motion Physics Modeling), a post-training framework for physically consistent text-driven 3D human motion generation that improves motion representations through two complementary directions: time–frequency modeling and physics-inspired expert collaboration. Specifically, we introduce a Multi-scale Time-Frequency Modeling (MTFM) module to jointly model motion dynamics at different temporal scales, improving the representation of complex motion patterns and their temporal regularities. We further propose a Physics-Inspired Expert Collaboration (PEC) module, which constructs complementary experts from observable body information and transforms structural and motion information from real data into direct supervision for adaptive refinement. By combining time–frequency representation learning with physics-inspired supervision, MoPhys encourages generated motions to better follow real human motion regularities. Extensive experiments on three human motion datasets and multiple mainstream generation models show that MoPhys consistently improves kinematic and spectral statistics, demonstrating its effectiveness and generalization across motion distributions and generation paradigms.

Our main contributions are summarized as follows:

1. We introduce a generation-optimization perspective centered on explicit supervision from observable physical signals in real human motion, providing a new way to reduce the kinematic distribution gap between generated and real motions.

2. We propose MoPhys, which unifies multi-scale time–frequency modeling and physics-inspired expert collaboration to combine dynamic motion representation learning with physics-inspired supervision.

3. Extensive experiments across three human motion datasets and multiple generation models show consistent improvements in kinematic and spectral motion statistics, validating the effectiveness of MoPhys across data distributions and generation paradigms.

2 Related Work

2.1 Text-Driven 3D Human Motion Generation

Text-driven 3D human motion generation has gradually expanded from continuous motion modeling to more structured discrete generation paradigms. T2M-GPT quantizes continuous human motion into discrete motion tokens and models text-conditioned motion sequences with an autoregressive Transformer (Zhang et al., 2023). MMM further adopts masked motion modeling to predict motion tokens in parallel using bidirectional context, improving generation efficiency while supporting motion completion and editing (Pinyoanuntapong et al., 2024). These studies suggest that effective motion representations and explicit modeling of dependencies among motion units are important for improving text-driven motion generation.

Beyond motion representations, recent work has increasingly focused on fine-grained alignment between language semantics and human motion. ATM characterizes semantic misalignment both across motions and within individual motions, and improves the response of generated motion to language descriptions through fine-grained alignment mechanisms (Han et al., 2026). Such methods mainly strengthen supervision from the language-semantic side, whereas our work further investigates supervision from the kinematic information inherent in real human motion.

2.2 Temporal Modeling and Physical Consistency in Human Motion

Human motion exhibits pronounced short- and long-term temporal dependencies, making the modeling of dynamics at different temporal scales an important problem in motion generation. FlowMDM combines absolute and relative positional encodings to model global coherence and local transitions in long motion sequences, enabling seamless composition under multiple text conditions (Barquero et al., 2024). Motion Mamba employs hierarchical state-space modeling to enhance temporal dependencies and spatial relationships for long-sequence motion generation (Zhang et al., 2024b). These methods primarily improve temporal dependency modeling through sequence structures and network design, while the interaction among motion variations at different temporal scales remains underexplored.

Meanwhile, visual quality and temporal coherence alone do not guarantee physically plausible human motion. PhysDiff introduces physics-simulation-based motion projection into the diffusion process to reduce artifacts such as sliding, floating, and penetration (Yuan et al., 2023). SuperPADL learns language-directed physics-based character control through reinforcement learning and supervised distillation (Juravsky et al., 2024). CLoSD further combines a motion generator with a physics-based controller to enable closed-loop execution in physical environments (Tevet et al., 2025). In contrast to approaches that rely on additional physics simulators or controllers, MoPhys directly constructs training supervision from observable kinematic information in real motion and combines it with multi-scale time–frequency modeling, jointly optimizing dynamic motion representation and body-motion regularities.

3 Method

This section presents MoPhys, which unifies dynamic motion representation learning and physics-inspired supervision through multi-scale time–frequency modeling and physics-inspired expert collaboration, thereby improving the ability of generated motion to capture regularities of real human movement. The overall framework is illustrated in Figure 2.

3.1 Multi-scale Time-Frequency Modeling

Human motion contains both persistent global movement and short-term local variations. To jointly model motion patterns at different scales, we introduce Multi-scale Time-Frequency Modeling (MTFM), which combines time–frequency decomposition with cross-band interaction to integrate information from multiple temporal scales into the motion representation.

Given a text condition c, the model generates a motion sequence $X \in \mathbb{R}^{T \times d_x}$ of length T. Let $H \in \mathbb{R}^{T \times d}$ denote the frame-level hidden representation of the backbone after dimensional alignment,

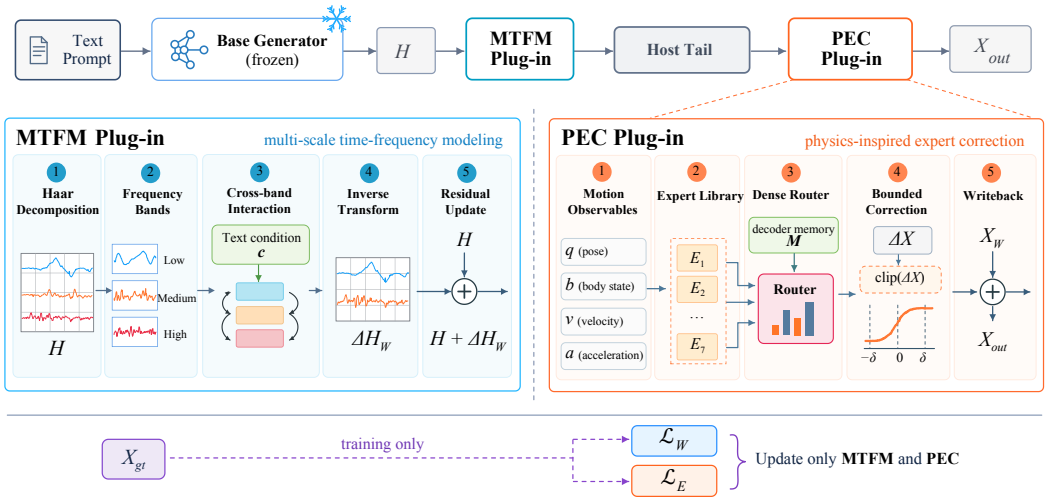


Figure 2: Overview of MoPhys. The top row shows the motion generation and refinement pipeline, where Host Tail denotes the remaining generation or decoding layers of the backbone. The bottom row expands the Multi-scale Time-Frequency Modeling (MTFM) module and the Physics-Inspired Expert Collaboration (PEC) module. The dashed branch at the bottom denotes supervision used during training.

and let $m \in \{0,1\}^T$ be the valid-frame mask. We apply a two-level Haar decomposition along the temporal axis to separate slowly varying components from local variations at different scales. Let $L^{(0)} = H$. The decomposition at level ℓ is

$$L_k^{(\ell)} = \frac{L_{2k}^{(\ell-1)} + L_{2k+1}^{(\ell-1)}}{\sqrt{2}}, \qquad D_k^{(\ell)} = \frac{L_{2k}^{(\ell-1)} - L_{2k+1}^{(\ell-1)}}{\sqrt{2}}, \qquad \ell \in \{1,2\}. \tag{1}$$

This yields three complementary bands, $\{L^{(2)}, D^{(2)}, D^{(1)}\}$, representing long-term variation trends and local details at two temporal scales. For variable-length sequences, the decomposition is computed only over valid frames, with boundary extension using the last valid frame to prevent padding from contaminating the time–frequency representation.

After obtaining the multi-scale representation, we explicitly model interactions across bands so that local variations can be interpreted jointly with longer-range temporal context. Specifically, the normalized band coefficients are concatenated into U, while band encodings and temporal-support encodings are combined into E to preserve each coefficient's scale and temporal support. A text condition vector g is injected through learnable mappings, and the cross-band update and temporal reconstruction are written as

$$\Delta C = \mathrm{MHA}(U + E + G_Q(g),\, U + E + G_K(g),\, U;\, m_{\mathcal{B}}), \qquad H^+ = H + m \odot \mathcal{W}^{-1}(\mathrm{Split}(\Delta C)), \tag{2}$$

where G_Q and G_K are text-conditioning projections, $m_{\mathcal{B}}$ is the valid-band mask, and $\mathcal{W}^{-1}$ denotes the inverse Haar transform. Attention aggregates information across valid temporal positions and frequency bands, after which the inverse transform restores the update to the original frame grid, allowing information from multiple scales to jointly refine frame-level representations.

MTFM uses residual connections and a zero-initialized output projection so that the original generation behavior is preserved at the beginning of training and time–frequency corrections are learned gradually. For discrete-motion and latent generative models, MTFM operates on frame-level features in the continuous decoding process; for direct motion diffusion models, it operates on motion-frame features within the denoising network. Different backbones retain their original generation

objectives and sampling procedures while sharing the same time–frequency modeling mechanism through dimensional adapters.

3.2 Physics-Inspired Expert Collaboration

While MTFM improves motion representations across temporal scales, the Physics-Inspired Expert Collaboration (PEC) module further refines motion outputs using observable body quantities. Let X_m denote the motion after time–frequency modeling. We first recover its original coordinate representation and extract root-relative body positions $q_{t,j}$, parent–child bone vectors $b_{t,j}$, and temporal differences. Here, t and j index time and joints, respectively; the bone vector is $b_{t,j} = q_{t,j} - q_{t,\pi(j)}$, where $\pi(j)$ denotes the parent joint. The first- and second-order differences are defined as $v_{t,j} = f\Delta q_{t,j}$ and $a_{t,j} = f^2\Delta^2 q_{t,j}$, where f is the actual frame rate. These quantities describe structural and motion changes in the local root-relative body coordinate system and provide kinematically meaningful inputs for different experts.

Based on these observables, we construct seven experts for body configuration, body center, rotational coordination, end-effector chains, bilateral coordination, relative kinematics, and temporal variation. Each expert receives a dedicated set of descriptors, which are scale-normalized and encoded by an independent multilayer perceptron to obtain $h_{t,j}^{(e)}$, where $e \in \{1, \ldots, 7\}$. This attribute-specific decomposition allows the refinement process to jointly account for skeletal structure, local dynamics, and inter-joint coordination rather than predicting all corrections from a single undifferentiated motion feature.

To keep body refinement consistent with the intended motion semantics, we form a query from expert features and body context, retrieve condition information from the semantic memory M associated with the sample, and inject it into each expert through a learnable gate to obtain $\widetilde{h}_{t,j}^{(e)}$. The semantic memory is provided by the corresponding backbone: sentence-level representations provide global conditioning, while sequence representations retain finer contextual information. This interaction enables the experts to determine refinement directions jointly from body state and semantic context.

Each expert then predicts a displacement candidate $d_{t,j}^{(e)}$, and a router assigns fusion weights according to all expert features and body context $r_{t,j}$. To avoid excessive modification of the original motion, the fused correction is explicitly bounded:

$$\begin{aligned} \alpha_{t,j} &= \mathrm{softmax}\left(\frac{R([\widetilde{h}_{t,j}^{(1:7)}; r_{t,j}]) + \eta}{\tau}\right), \\ \delta_{t,j} &= \sigma(\beta)\,\Pi_{\mathcal{B}_\rho}\left[\rho \tanh\left(\frac{\sum_{e=1}^{7} \alpha_{t,j}^{(e)} d_{t,j}^{(e)}}{\rho}\right)\right]. \end{aligned} \tag{3}$$

Here, R denotes the routing network, η is a routing prior, τ is the temperature, $\sigma(\beta)$ is a learnable gain, and $\Pi_{\mathcal{B}_\rho}$ projects onto an Euclidean ball of radius ρ, ensuring $\|\delta_{t,j}\|_2 \le \rho$. Corrections are written only to non-root joint positions, while other motion channels remain unchanged; temporal differences used after refinement are recomputed from the corrected positions.

To transform body regularities from real motion into direct supervision, we construct training objectives on joint positions, bone vectors, and first- and second-order temporal differences. Let ℓ_i^{in} and ℓ_i^{out} denote the normalized SmoothL1 errors of descriptor type i before and after refinement, respectively. The body supervision is defined as

$$\mathcal{L}_{\mathrm{body}} = \frac{\sum_i w_i \left(\ell_i^{\mathrm{out}} + \lambda_{\mathrm{nd}}[\ell_i^{\mathrm{out}} - \mathrm{sg}(\ell_i^{\mathrm{in}})]_+\right)}{\sum_i w_i}. \tag{4}$$

Here, w_i weights the different descriptors, sg denotes stop-gradient, and $[x]_+ = \max(x, 0)$. The first term supervises the refined body state, while the second penalizes increases in error caused by refinement. We additionally use routing-balance, expert-diversity, and residual-magnitude regularization, with warm-up and magnitude control so that optimization remains primarily driven by body supervision.

The overall objective combines the native backbone objective with the PEC objective, i.e., $\mathcal{L} = \mathcal{L}_{\mathrm{native}} + \mathcal{L}_{\mathrm{body}} + \mathcal{R}$, where $\mathcal{R}$ denotes the warmed-up and bounded auxiliary regularization.

Discrete-motion and latent models use reconstructions corresponding to ground-truth motions to form supervised pairs, while direct motion diffusion models use clean-motion estimates from the same denoising training samples. The motion and external semantic memory entering PEC are stop-gradient inputs, so body supervision updates only the expert module, while the backbone and MTFM remain optimized by their respective native objectives. During inference, PEC is applied once to the final generated motion.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We conduct the main experiments on HumanML3D and KIT-ML. HumanML3D contains 14,616 human motion sequences paired with 44,970 text descriptions and covers diverse daily activities (Guo et al., 2022). KIT-ML contains 3,911 motion sequences and 6,278 language annotations collected from heterogeneous motion-capture sources (Plappert et al., 2016).

Backbone Models. We evaluate eight backbone configurations. MDM generates continuous human motion by conditional denoising directly in motion space (Tevet et al., 2023). MLD performs diffusion in a compressed motion latent space and decodes the resulting latents into motion sequences (Chen et al., 2023). MotionDiffuse models text-conditioned motion distributions with diffusion and supports fine-grained motion control (Zhang et al., 2024a). MotionLCM uses latent consistency modeling for one- or few-step motion generation (Dai et al., 2024). MotionGPT discretizes human motion into motion tokens and models language and motion in a unified framework (Jiang et al., 2023). MoMask combines residual quantization with masked prediction to generate hierarchical discrete motion representations (Guo et al., 2024). Motion-Agent uses language models and parameter-efficient adaptation for motion generation and interaction (Wu et al., 2025). Our Motion-LLM configuration uses the motion-language-model backbone associated with IRG-MotionLLM, which performs text-to-motion generation through motion-token prediction and continuous decoding (Li et al., 2026).

Post-Training and Generation Alignment Methods. PPO performs constrained policy updates using reward feedback to optimize generated outputs (Schulman et al., 2017). DPO directly optimizes a generative policy from preference pairs without separately training an explicit reward model (Rafailov et al., 2023). AToM uses event-level feedback from a vision-language model to improve alignment between generated motion and text descriptions (Han et al., 2025). SoPo combines offline preferred motions with online non-preferred motions for semi-online preference optimization (Tan et al., 2025). ReAlign guides diffusion sampling with step-aware rewards to improve text–motion consistency (Weng et al., 2026). EasyTune performs reward-driven optimization at the denoising-step level to reduce the cost of backpropagating through complete sampling trajectories (Tan et al., 2026). IRG alternates motion generation, assessment, and refinement to improve motion generation through cross-stage feedback (Li et al., 2026).

Evaluation Metrics. We use FID, Matching, and R-Precision to measure motion-feature distribution distance, paired text–motion embedding distance, and retrieval consistency, respectively. To further evaluate body-motion statistics, we introduce Kinematic Fréchet Distance (KFD) and Power-Spectrum Wasserstein Distance (PSWD), which compare motion-statistics distributions and temporal spectral distributions, respectively.

KFD follows the Fréchet feature-distribution comparison (Heusel et al., 2017). We construct normalized kinematic features from the temporal means of squared horizontal velocity, squared vertical velocity, and acceleration magnitude for each joint in root-removed world trajectories, and compute

$$\mathrm{KFD} = \|\mu_g - \mu_r\|_2^2 + \mathrm{Tr}\left[\Sigma_g + \Sigma_r - 2\left(\Sigma_g^{1/2}\Sigma_r\Sigma_g^{1/2}\right)^{1/2}\right], \tag{5}$$

where (μ_g, Σ_g) and (μ_r, Σ_r) are the mean and covariance of generated and reference kinematic features, respectively.

PSWD follows the spectral comparison principle of the Wasserstein–Fourier distance (Cazelles et al., 2020). We estimate windowed power spectra from root-removed joint velocities, aggregate

motions with equal weight, normalize each joint spectrum into a frequency distribution, and compute

$$\mathrm{PSWD} = \frac{1}{J-1} \sum_{j=1}^{J-1} \frac{W_1(\nu_j^g, \nu_j^r)}{f/2}, \tag{6}$$

where ν_j^g and ν_j^r denote the generated and reference frequency distributions of joint j, J is the number of joints, f is the frame rate, and W_1 is the first-order Wasserstein distance. Lower KFD and PSWD are better.

Training Settings. Each backbone retains its native generation objective and sampling procedure, together with the corresponding optimizer, learning rate, and scheduling strategy. The newly introduced modules are optimized through the training interface of each backbone, with residual branches initialized to zero. Each dataset uses its own skeleton and normalization statistics. Comparisons under the same backbone share data splits and random seeds, and checkpoints are selected according to validation FID.

4.2 Main Results

Table 1 summarizes ten model–dataset combinations over two datasets and eight backbones, with the best result within each comparison group highlighted in bold. Overall, MoPhys reduces KFD, PSWD, and Matching in all combinations, showing consistent improvements in body-kinematic distributions, temporal spectra, and text–motion matching. It also improves FID and retrieval metrics in most settings. These results span direct motion diffusion, latent-space generation, and discrete motion modeling, indicating that MoPhys provides effective kinematic refinement across diverse generation paradigms.

Table 1: Main results on HumanML3D and KIT-ML. Results using the same backbone on the same dataset form one comparison group; the best result in each group is shown in bold.

Dataset / Backbone	Method	FID↓	Matching↓	R@1↑	R@2↑	R@3↑	KFD↓	PSWD↓
HumanML3D / MDM	Base	0.5366	3.6539	0.4133	0.5996	0.7086	20.5596	0.0131
	MoPhys (ours)	**0.4910**	**2.5235**	**0.5576**	**0.8333**	**0.8758**	**4.1220**	**0.0120**
HumanML3D / MotionGPT	PPO	0.1838	3.9175	0.4111	0.5777	0.6695	3.7227	0.0157
	MoPhys (ours)	**0.1778**	**3.5428**	**0.4639**	**0.6514**	**0.6799**	**3.4864**	**0.0140**
HumanML3D / MLD	ReAlign	**0.2262**	2.6543	0.5684	**0.7649**	0.8566	13.9905	0.0483
	MoPhys (ours)	0.2295	**2.4841**	**0.5804**	0.7442	**0.9064**	**12.6699**	**0.0442**
HumanML3D / MotionLCM	Base	0.0902	2.7641	**0.5484**	0.7431	**0.8364**	2.4766	0.0219
	MoPhys (ours)	**0.0892**	**2.6395**	0.5384	**0.7599**	0.8122	**2.3371**	**0.0206**
HumanML3D / Motion-Agent	Base	0.1700	2.9000	**0.5200**	**0.7150**	0.8090	3.1940	0.0164
	MoPhys (ours)	**0.1623**	**2.8638**	0.5104	0.7077	**0.8743**	**2.9857**	**0.0154**
HumanML3D / MotionLLM	IRG Base	**0.1991**	3.0290	0.4998	0.6950	0.7847	4.2576	0.0176
	MoPhys (ours)	0.2015	**2.5945**	**0.5593**	**0.7779**	**0.8660**	**3.9214**	**0.0158**
KIT-ML / MLD	SoPo	6.1538	6.9624	0.1510	0.2671	0.3534	186.1612	0.2052
	MoPhys (ours)	**5.9635**	**6.5143**	**0.1707**	**0.2803**	**0.4052**	**174.5012**	**0.1873**
KIT-ML / MDM	ReAlign	**0.4554**	3.0035	0.4052	**0.6257**	**0.7422**	2.6566	0.0294
	MoPhys (ours)	0.4683	**2.9714**	**0.4323**	0.6188	0.7297	**2.4564**	**0.0279**
KIT-ML / MoMask	SoPo	**0.2197**	2.7656	0.4496	0.6655	0.7915	19.3186	0.0901
	MoPhys (ours)	0.2250	**2.5880**	**0.4618**	**0.7099**	**0.8297**	**17.8751**	**0.0837**
KIT-ML / MotionDiffuse	EasyTune Base	1.7681	2.9690	**0.4160**	**0.6189**	0.7389	6.1362	0.0371
	MoPhys (ours)	**1.6860**	**2.7228**	0.4132	0.6179	**0.8102**	**5.6650**	**0.0339**

On HumanML3D, all six backbones improve in body statistics and text–motion matching; for example, Matching on MotionGPT decreases by 9.57% relative to PPO. The same trend holds across the four KIT-ML backbones, where KFD for MoMask decreases by 7.47% relative to SoPo. Overall, the results show that MoPhys captures motion regularities from two complementary perspectives—multi-scale temporal variation and body kinematics—and effectively improves generated motion.

4.3 Ablation Studies

To analyze the independent contributions of each module and its key design choices, we conduct ablations on the representative HumanML3D/MotionGPT and KIT-ML/MoMask settings. We evaluate four variant types, yielding eight ablation configurations in total: MTFM only, PEC only, and two structural alternatives for the core designs, together with the corresponding baselines and the full model. As shown in Table 2, MTFM and PEC exhibit clear complementarity: MTFM contributes more strongly to FID and text–motion matching, whereas PEC produces larger gains in body-kinematic statistics; combining them yields the strongest joint performance on semantic retrieval and body-motion statistics. Replacing the structured PEC experts with parameter-matched generic MLP branches (*MTFM + 7 Generic MLPs*) degrades Matching, KFD, and PSWD relative to the full model, indicating that the gains do not arise merely from extra branch capacity but from expert specialization over distinct body attributes. Likewise, replacing multi-scale time–frequency modeling with frame-level temporal attention (*Flat Temporal Attention + PEC*) degrades text matching and body statistics, showing that the benefit of MTFM cannot be attributed solely to additional temporal attention. These results demonstrate that MTFM and PEC contribute through complementary mechanisms—multi-scale motion representation and kinematic body refinement—and jointly improve semantic consistency and kinematic quality.

Table 2: Ablation results for the proposed modules and their key structural designs.

Setting	Method	FID↓	Matching↓	R@1↑	R@2↑	R@3↑	KFD↓	PSWD↓
HumanML3D / MotionGPT	PPO	0.1838	3.9175	0.4111	0.5777	0.6695	3.7227	0.0157
	w/o PEC	0.1805	3.7059	0.4418	0.6163	0.6748	3.6502	0.0152
	w/o MTFM	0.1826	3.8208	0.4217	0.5908	0.6719	3.5486	0.0145
	MTFM + 7 Generic MLPs	0.1795	3.6189	0.4542	0.6412	0.6778	3.5228	0.0143
	Flat Temporal Attention + PEC	0.1802	3.6286	0.4516	0.6384	0.6769	3.5185	0.0143
	MoPhys	**0.1778**	**3.5428**	**0.4639**	**0.6514**	**0.6799**	**3.4864**	**0.0140**
KIT-ML / MoMask	SoPo	0.2197	2.7656	0.4496	0.6655	0.7915	19.3186	0.0901
	w/o PEC	**0.2192**	2.6578	0.4562	0.6884	0.8096	19.0054	0.0886
	w/o MTFM	0.2260	2.7104	0.4521	0.6746	0.7989	18.2687	0.0856
	MTFM + 7 Generic MLPs	0.2258	2.6308	0.4580	0.6987	0.8181	18.1849	0.0851
	Flat Temporal Attention + PEC	0.2263	2.6346	0.4568	0.6994	0.8204	18.0369	0.0848
	MoPhys	0.2250	**2.5880**	**0.4618**	**0.7099**	**0.8297**	**17.8751**	**0.0837**

4.4 Generalization to Game Motion

We further evaluate MoPhys on CMP (CombatMotionProcessed), a game-motion dataset released by the AnimationGPT project that contains 8,700 combat animations from game assets and 26,100 text descriptions (Liao et al., 2024). We use MDM, MLD, and MotionGPT as backbones and follow the training and evaluation settings in Section 4.1 and the comparison protocol in Section 4.2. As shown in Table 3, MoPhys reduces Matching, KFD, and PSWD on all three backbones, and further improves FID on MLD and MotionGPT, indicating better text matching, body-kinematic statistics, and temporal spectral distributions. These results show that MoPhys generalizes beyond daily motion to game combat animation and can capture physical distributional regularities characterized by body kinematics and temporal spectra in game-motion samples.

Table 3: Results on the CMP game-motion dataset. Results using the same backbone form one comparison group; the best result in each group is shown in bold.

Dataset / Backbone	Method	FID↓	Matching↓	R@1↑	R@2↑	R@3↑	KFD↓	PSWD↓
CMP / MDM	Base	**9.0348**	8.1904	0.0507	0.1024	**0.1542**	62.7852	0.4152
	MoPhys (ours)	9.1331	**7.4920**	**0.0529**	**0.1097**	0.1501	**25.6887**	**0.3939**
CMP / MLD	Base	7.8662	10.1524	**0.0420**	0.0781	0.1149	35.7907	0.0521
	MoPhys (ours)	**6.5738**	**9.9991**	0.0410	**0.0796**	**0.1237**	**20.9149**	**0.0493**
CMP / MotionGPT	NativeCMP	0.4754	4.3508	0.2868	0.4580	0.5807	69.1444	0.0855
	MoPhys (ours)	**0.4395**	**3.9496**	**0.3108**	**0.4856**	**0.5832**	**43.1323**	**0.0798**

4.5 Computational Cost

Figure 3 reports training and evaluation time on HumanML3D and KIT-ML, together with the total runtime for each method. Overall, among the nine model–dataset combinations compared, MoPhys achieves lower total runtime than all compared methods in six settings. On HumanML3D, for example, the total runtime with MotionLCM decreases from 303.4 minutes for ReAlign to 56.8 minutes, a reduction of 81.3%. On KIT-ML, the total runtime with MLD decreases from 148.9 minutes for SoPo to 14.5 minutes, a reduction of 90.3%. Together with the improvements in generation quality and kinematic statistics, these results indicate that MoPhys also offers favorable computational cost across multiple backbones.

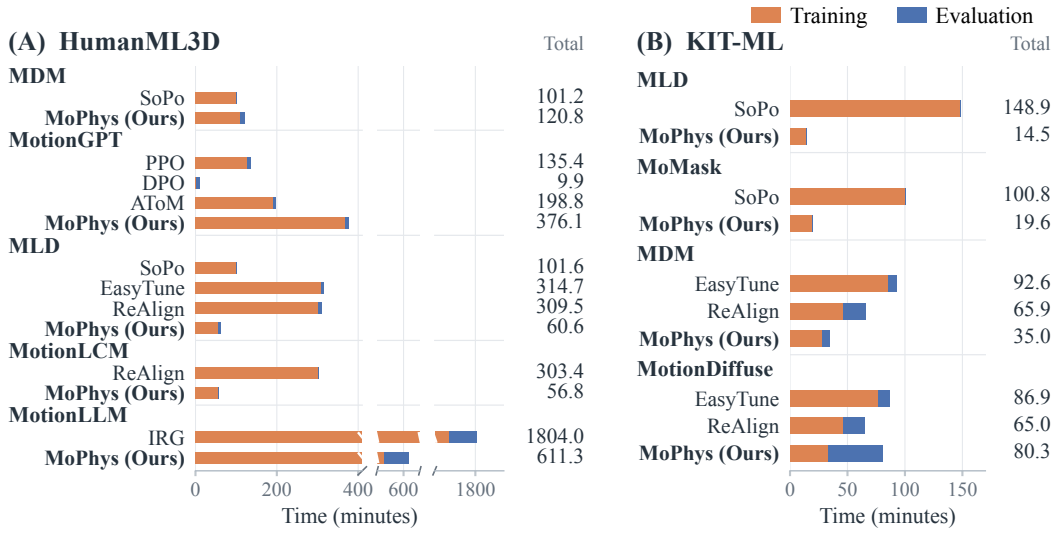


Figure 3: Training and evaluation runtime on (A) HumanML3D and (B) KIT-ML. Orange and blue bars denote training and evaluation time, respectively, and the total runtime (minutes) is listed on the right. Panel (A) uses a broken x-axis.

5 Conclusion

This paper addresses the lack of effective supervision for real human motion regularities in text-driven 3D human motion generation. We propose MoPhys, which unifies dynamic motion representation learning with observable physical signals from real motion through multi-scale time–frequency modeling and physics-inspired expert collaboration. MTFM models motion dynamics across temporal scales, while PEC adaptively refines generated motion using body structure and kinematic information. Experiments across multiple generation paradigms and backbones on HumanML3D and KIT-ML show that MoPhys consistently improves Matching, KFD, and PSWD, while also improving generation quality and text–motion consistency in most settings. Ablation studies validate the effectiveness of multi-scale time–frequency modeling, structured physics-inspired experts, and their complementary interaction. On the CMP game-combat dataset, whose motion distribution differs substantially from daily human motion, MoPhys also improves body-kinematic and temporal spectral distributions, demonstrating cross-distribution generalization. Computational-cost experiments further show that the framework can achieve these gains with favorable training and evaluation overhead on multiple backbones. Overall, transforming observable physical information from real motion into direct supervision effectively reduces the gap between generated and real human motion distributions, providing a practical direction for building text-driven 3D motion generation models that better respect human motion regularities while retaining broad applicability and efficiency.

REFERENCES

German Barquero, Sergio Escalera, and Cristina Palmero. Seamless human motion composition with blended positional encodings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 457–469, 2024.

Elsa Cazelles, Arnaud Robert, and Felipe Tobar. The wasserstein-fourier distance for stationary time series. *arXiv preprint arXiv:1912.05509*, 2020.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18000–18010, 2023.

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 *European Conference on Computer Vision*, pp. 390–408, 2024. doi: 10.1007/978-3-031-72640-8_22.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5152–5161, 2022.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024.

Haonan Han, Xiangzuo Wu, Huan Liao, Zunnan Xu, Zhongyuan Hu, Ronghui Li, Yachao Zhang, and Xiu Li. Atom: Aligning text-to-motion model at event-level with gpt-4vision reward. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22746–22755, 2025.

Ke Han, Yueming Lyu, Weichen Yu, and Nicu Sebe. Atm: Enhanced alignment for text-to-motion generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6862–6872, 2026. doi: 10.1109/WACV61042.2026.00663.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. Superpadl: Scaling language-directed physics-based control with progressive supervised distillation. In *SIGGRAPH 2024 Conference Papers*, 2024.

Yuan-Ming Li, Qize Yang, Nan Lei, Shenghao Fu, Ling-An Zeng, Jian-Fang Hu, Xihan Wei, and Wei-Shi Zheng. Irg-motionllm: Interleaving motion generation, assessment and refinement for text-to-motion generation. In *European Conference on Computer Vision*, 2026. arXiv:2512.10730.

Yihao Liao, Yiyu Fu, Ziming Cheng, and Jiangfeiyang Wang. AnimationGPT: An aigc tool for generating game combat motion assets. `https://github.com/fyyakaxyy/AnimationGPT`, 2024.

Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1546–1555, 2024.

Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big Data*, 4(4):236–252, 2016. doi: 10.1089/big.2016.0028.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Xiaofeng Tan, Hongsong Wang, Xin Geng, and Pan Zhou. Sopo: Text-to-motion generation using semi-online preference optimization. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2029.

Xiaofeng Tan, Wanjiang Weng, Haodong Lei, and Hongsong Wang. Easytune: Efficient step-aware fine-tuning for diffusion-based motion generation. In *International Conference on Learning Representations*, 2026.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023.

Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit H. Bermano, and Michiel van de Panne. CLoSD: Closing the loop between simulation and diffusion for multi-task character control. In *International Conference on Learning Representations*, 2025.

Wanjiang Weng, Xiaofeng Tan, Junbo Wang, Guo-Sen Xie, Pan Zhou, and Hongsong Wang. Realign: Text-to-motion generation via step-aware reward-guided alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10621–10629, 2026. doi: 10.1609/aaai.v40i13.38035.

Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Motion-agent: A conversational framework for human motion generation with llms. In *International Conference on Learning Representations*, 2025.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16010–16021, 2023.

Junpeng Yue, Zepeng Wang, Jiangxing Wang, Yuxuan Wang, Yu Zhang, Xinrun Xu, Bin Cao, Sipeng Zheng, Ziluo Ding, and Zongqing Lu. Rl from physical feedback: Aligning large motion models with humanoid control. In *European Conference on Computer Vision*, 2026. arXiv:2506.12769.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14730–14740, 2023.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4115–4128, 2024a. doi: 10.1109/TPAMI.2024.3355414.

Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation. In *European Conference on Computer Vision*, 2024b.