PHYFLOW: PHYSICS-AWARE RECTIFIED FLOW WITH FINE-GRAINED SEMANTIC ALIGNMENT FOR TEXT-TO-MOTION GENERATION

ABSTRACT

Text-to-motion generation aims to synthesize human motion sequences that are semantically faithful to natural-language descriptions while remaining realistic and physically plausible. Recent latent diffusion methods such as ARDT achieve competitive performance but rely on pooled sentence-level representations and sequence-level alignment, limiting fine-grained semantic control and temporal grounding. Iterative sampling is also computationally costly, while adversarial training does not explicitly enforce physical consistency. We propose **PhyFlow**, a semantic- and physics-aware rectified-flow framework for efficient and controllable text-to-motion generation. PhyFlow retains a two-stage latent architecture, in which a VAE compresses motion into a continuous latent space and a Rectified Flow Transformer learns a direct velocity field from noise to motion latents. A frozen Qwen3 text encoder provides token-level representations, while an Action Parsing Module extracts action phrases, body-part cues, directional modifiers, and temporal relations. Fine-Grained Semantic Alignment grounds action-level tokens to corresponding latent motion frames, while physics guidance regularizes foot contact, bone length, joint limits, and temporal smoothness. Non-adversarial moment and maximum mean discrepancy objectives reduce distributional gaps. Experiments on HumanML3D show improved semantic retrieval over ARDT, achieving Top-1/Top-2/Top-3 R-precision of 0.573/0.756/0.843 and an MM-Dist of 2.788. On KIT-ML, PhyFlow reduces FID from 0.334 to 0.128 while maintaining diversity close to real motion. These results demonstrate the potential of combining structured semantic alignment, efficient rectified-flow generation, and physics-aware guidance for controllable text-to-motion synthesis.

1 INTRODUCTION

Text-to-motion generation aims to synthesize three-dimensional sequences of human movement from natural-language descriptions. Formally, given a text description, the goal is to generate motion that is semantically faithful to the input, consistent over time, physically plausible, and sufficiently diverse across multiple generations for the same input (Guo et al., 2022).

Motion descriptions are often compositional. A sentence may specify multiple actions occurring simultaneously or sequentially (Athanasiou et al., 2023), the body parts involved, modifiers such as speed or direction, and temporal relationships between sub-actions. These descriptions require models to identify actions, involved body parts, temporal order, duration, and relations between sub-actions. The generated sequence must also satisfy fundamental physical constraints, including consistent bone lengths, bounded joint angles, and stable foot–ground contact. Foot sliding and unnatural jitter remain perceptually apparent even when the motion is semantically correct (Yuan et al., 2023; Li et al., 2025b).

Text-to-motion methods use either continuous latent representations or discrete motion tokens. Continuous latent diffusion models (Chen et al., 2023) can use adversarial supervision to improve motion detail and realism (Uchida et al., 2025). Continuous latent methods such as ARDT (Li et al., 2026) achieve competitive performance but rely on sentence-level semantic conditioning and iterative generation. Discrete models generate motion tokens autoregressively or through masked mod-

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.

eling (Zhang et al., 2023; Guo et al., 2024), avoiding iterative denoising but potentially introducing quantization artifacts (Cho et al., 2025).

However, continuous latent diffusion models still face limitations in fine-grained semantic control, sampling efficiency, and physical consistency. Global CLIP embeddings and whole-sequence alignment provide limited action-to-motion correspondence, while iterative diffusion and adversarial training introduce computational and optimization costs (Radford et al., 2021; Tevet et al., 2022; Petrovich et al., 2023; Li et al., 2026; Ho et al., 2020; Goodfellow et al., 2014).

We propose **PhyFlow**, a continuous latent text-to-motion model developed from ARDT. PhyFlow retains the two-stage latent framework of ARDT, a motion VAE followed by a text-conditioned latent generator, and redesigns the second stage to address the issues above. A semantic conditioning module combines a frozen token-level language encoder (Yang et al., 2025) with an action-parsing stage to represent actions, body parts, and temporal information. A fine-grained semantic alignment mechanism associates action-level representations with corresponding temporal segments of the generated motion. A rectified-flow transformer is trained under a flow-matching objective to reduce the number of sampling steps. Explicit physics-based constraints supervise physical plausibility, while non-adversarial distribution-matching objectives match latent- and feature-level representations without generator–discriminator minimax optimization. Thus, the adversarial and sentence-level contrastive objectives of ARDT are replaced rather than extended. We evaluate PhyFlow on HumanML3D (Guo et al., 2022) and KIT-ML (Plappert et al., 2016).

PhyFlow makes four main contributions: **(1)** structured semantic conditioning for action, body-part, and temporal information; **(2)** fine-grained alignment between action-level semantics and temporal motion segments; **(3)** efficient continuous latent generation with rectified flow; and **(4)** physics-based and non-adversarial distribution-matching objectives for physical plausibility and distributional realism.

2 Related Work

Continuous and discrete text-to-motion generation. Diffusion models generate motion by denoising Gaussian noise in raw pose space (Tevet et al., 2023; Zhang et al., 2024) or a learned latent space (Chen et al., 2023). MoLA (Uchida et al., 2025) and ARDT (Li et al., 2026) improve motion detail in latent diffusion through adversarial supervision, with ARDT additionally using temporal semantic contrastive learning. Vector-quantized methods represent motion as discrete tokens generated autoregressively or through masked modeling (Zhang et al., 2023; Guo et al., 2024; Pinyoanuntapong et al., 2024; 2025), avoiding iterative denoising but potentially introducing quantization artifacts (Cho et al., 2025). Cho et al. (2025) further reduce these effects by decoding discrete tokens through continuous rectified flow.

Text conditioning and fine-grained alignment. Continuous latent models often use a single global embedding from a frozen vision–language model such as CLIP (Radford et al., 2021), which does not explicitly disentangle actions, body parts, or modifiers. Li et al. (2025a) pretrain language representations on paired text–motion data to reduce the domain gap of vision-oriented encoders. Existing alignment objectives operate between whole sentences and whole motion sequences (Tevet et al., 2022; Petrovich et al., 2023). Finer-grained approaches parse body-part and relational information (Wang et al., 2025), map actions to body parts (Athanasiou et al., 2023), or align text with kinematic body-part groups (Zhang et al., 2025). However, these methods do not directly supervise the correspondence between individual actions and their temporal motion segments.

Sampling efficiency and flow-based generation. Diffusion models require sequential denoising steps, and deterministic samplers reduce but do not eliminate this cost (Ho et al., 2020; Song et al., 2021). Flow matching and rectified flow learn continuous transport between noise and data distributions and support fewer-step generation (Lipman et al., 2023; Liu et al., 2023). PhyFlow combines flow matching with action-to-temporal-segment grounding and explicit physics guidance.

Non-adversarial distribution matching. Adversarial discriminators can improve motion detail (Uchida et al., 2025; Li et al., 2026) but introduce a potentially unstable minimax optimization problem (Goodfellow et al., 2014). PhyFlow instead matches latent and evaluator-feature statistics using moment and maximum mean discrepancy objectives (Gretton et al., 2012), without adversarial training.

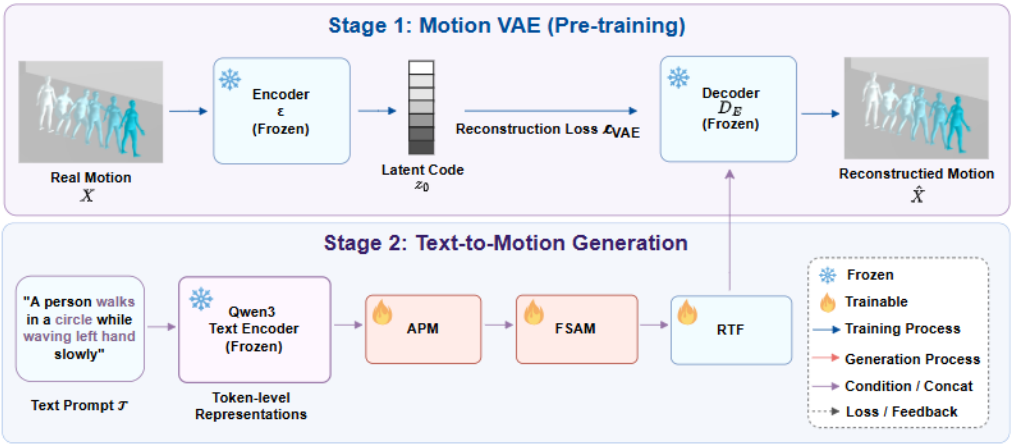

Figure 1: Overview of the proposed framework. Stage 1 pre-trains a motion VAE that maps real motions into a compact latent space and reconstructs them back to the motion feature space. Stage 2 freezes the VAE and performs text-to-motion generation in the latent space. A frozen Qwen3 text encoder extracts token-level semantic representations, which are structured by the Action Parsing Module (APM), aligned by the Fine-Grained Semantic Alignment Module (FSAM), and used to condition the Rectified Flow Transformer (RFT). The generated latent is decoded by the frozen VAE decoder to obtain the final motion.

Physics-aware motion generation. Yuan et al. (2023) incorporate a physics simulator into the diffusion sampling process, while Li et al. (2025b) use a simulator-trained physics module to refine motion before fine-tuning the generator. PhyFlow instead applies physical constraints directly during training, jointly with semantic alignment, without a physics simulator.

3 METHOD

Given a natural-language prompt $\mathcal{T}$, our goal is to generate a human motion sequence $X \in \mathbb{R}^{T\times D}$ that is semantically consistent, temporally coherent, and physically plausible. As shown in Fig. 1, PhyFlow follows a two-stage latent generation pipeline: a motion VAE first learns a compact latent space, followed by a semantically structured Rectified Flow Transformer for text-to-motion generation.

3.1 STAGE 1: MOTION VAE PRE-TRAINING

Let $X = \{x_t\}_{t=1}^{T}$ denote a motion sequence in the HumanML3D or KIT-ML feature space. A motion encoder E_ϕ maps X to a latent sequence:

$$z_0 = E_\phi(X), \qquad z_0 \in \mathbb{R}^{L\times d_z}, \tag{1}$$

where L is the latent temporal length and d_z is the latent dimension. The decoder D_ψ reconstructs the motion sequence:

$$\hat{X} = D_\psi(z_0). \tag{2}$$

The VAE is trained with a reconstruction objective and a latent regularization term:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{rec}}(X, \hat{X}) + \beta_{\text{KL}} D_{\text{KL}}\big(q_\phi(z_0|X) \,\|\, p(z)\big). \tag{3}$$

After pre-training, both E_ϕ and D_ψ are frozen.

3.2 STRUCTURED TEXT CONDITIONING

A limitation of global text conditioning is that the full sentence is compressed into a single vector. This is insufficient for compositional prompts such as “a person walks in a circle while waving the

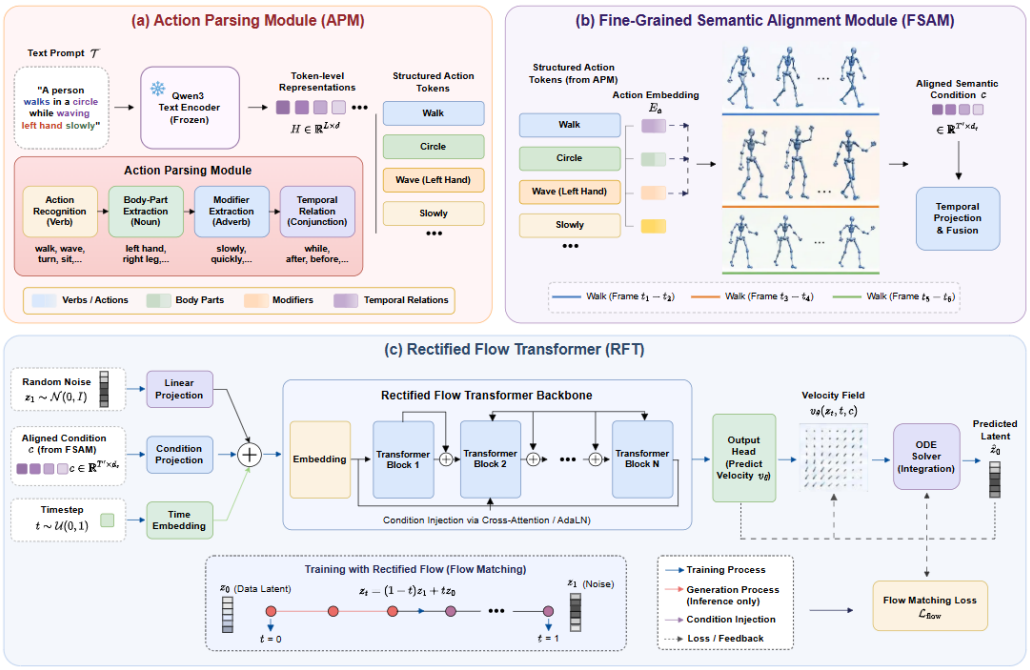

Figure 2: Detailed architecture of the proposed method. (a) APM parses token-level Qwen3 representations into structured action tokens. (b) FSAM aligns action-level semantic tokens with latent motion frames through action-frame alignment and temporal projection. (c) RFT learns a conditional velocity field between Gaussian noise and real motion latents using rectified flow matching. During inference, the learned velocity field is integrated with a small number of ODE steps to generate the predicted motion latent.

left hand", where different semantic units correspond to different body parts and temporal segments. To address this, we construct a structured semantic condition from the input prompt.

We first encode the prompt using a frozen Qwen3 text encoder (Yang et al., 2025):

$$h_{\mathrm{g}}, H = f_{\mathrm{text}}(\mathcal{T}), \tag{4}$$

where $h_{\mathrm{g}} \in \mathbb{R}^{d_h}$ is the pooled sentence representation and $H = \{h_i\}_{i=1}^{N}$ are token-level representations. The text encoder is frozen, while lightweight projection layers map these features into the motion generator dimension:

$$c_{\mathrm{g}} = W_{\mathrm{g}} h_{\mathrm{g}}, \qquad C_{\mathrm{text}} = W_{\mathrm{t}} H. \tag{5}$$

Action Parsing Module. As shown in Fig. 2(a), the Action Parsing Module (APM) extracts weakly structured action phrases from the prompt. In our implementation, APM is a lightweight deterministic parser that identifies verbs/actions, body parts, directions, modifiers, and temporal connectives. For a prompt $\mathcal{T}$, APM produces up to K action-level phrases:

$$\mathcal{A} = \{a_k\}_{k=1}^{K}, \qquad K \le K_{\max}. \tag{6}$$

Each action phrase is encoded by the same frozen text encoder and combined with a learnable action-slot embedding:

$$e_k = W_{\mathrm{g}} f_{\mathrm{text}}(a_k) + s_k, \tag{7}$$

where s_k is the learnable embedding of the k-th action slot. The final conditioning tokens are formed by concatenating action tokens and original text tokens:

$$C = [E_{\mathrm{act}}; C_{\mathrm{text}}], \qquad E_{\mathrm{act}} = \{e_k\}_{k=1}^{K_{\max}}. \tag{8}$$

This representation preserves global sentence semantics while exposing explicit action-level control signals to the generator.

3.3 FINE-GRAINED SEMANTIC ALIGNMENT

The original sentence-level contrastive alignment only encourages the whole motion to match the whole text. It does not explicitly indicate which action should control which temporal segment. We therefore introduce a Fine-Grained Semantic Alignment Module (FSAM), shown in Fig. 2(b).

Given a motion latent sequence $z \in \mathbb{R}^{L \times d_z}$, we project each latent frame to the semantic space:

$$m_t = W_{\mathrm{m}} z_t, \qquad M = \{m_t\}_{t=1}^{L}. \tag{9}$$

We first use a global sentence-motion contrastive loss based on InfoNCE (van den Oord et al., 2018):

$$\mathcal{L}_{\mathrm{global}} = \mathrm{InfoNCE}\left(c_{\mathrm{g}}, \frac{1}{L}\sum_{t=1}^{L} m_t\right). \tag{10}$$

For action-frame alignment, APM provides weak temporal priors. For each parsed action a_k, we construct a soft target distribution $q_{k,t}$ over latent frames, using temporally ordered Gaussian windows for sequential actions and overlapping windows for concurrent actions indicated by while" or as". The action-frame similarity is:

$$s_{k,t} = \frac{\langle \bar{e}_k, \bar{m}_t \rangle}{\tau}, \tag{11}$$

where $\bar{e}_k$ and $\bar{m}_t$ are ℓ_2-normalized embeddings and τ is the temperature. The frame-level alignment loss is:

$$\mathcal{L}_{\mathrm{frame}} = -\frac{1}{|\mathcal{V}|}\sum_{k\in\mathcal{V}}\sum_{t=1}^{L} q_{k,t} \log \frac{\exp(s_{k,t})}{\sum_{t'=1}^{L}\exp(s_{k,t'})}, \tag{12}$$

where $\mathcal{V}$ denotes valid parsed action slots.

To encourage temporally ordered actions, we compute the expected frame center of each action:

$$\mu_k = \sum_{t=1}^{L} t \cdot \mathrm{softmax}(s_{k,t}), \tag{13}$$

and penalize inversions:

$$\mathcal{L}_{\mathrm{order}} = \frac{1}{K-1}\sum_{k=1}^{K-1} \max(0, \mu_k - \mu_{k+1} + \delta). \tag{14}$$

The full semantic alignment objective is:

$$\mathcal{L}_{\mathrm{align}} = \mathcal{L}_{\mathrm{global}} + \mathcal{L}_{\mathrm{frame}} + \lambda_{\mathrm{ord}}\mathcal{L}_{\mathrm{order}}. \tag{15}$$

3.4 RECTIFIED FLOW TRANSFORMER

As shown in Fig. 2(c), we formulate latent motion generation as conditional rectified flow matching. Let $z_0 \sim p_{\mathrm{data}}(z)$ be a real motion latent from the frozen VAE encoder and $z_1 \sim \mathcal{N}(0, I)$ be Gaussian noise. For $t \sim \mathcal{U}(0, 1)$, we define the linear interpolation path:

$$z_t = (1-t)z_1 + tz_0. \tag{16}$$

The target velocity along this path is:

$$v^\star = z_0 - z_1. \tag{17}$$

A Rectified Flow Transformer v_θ predicts the velocity field conditioned on the text:

$$\hat{v}_\theta = v_\theta(z_t, t, c_{\mathrm{g}}, C, \ell), \tag{18}$$

where ℓ denotes the normalized motion-length condition. The flow matching objective is:

$$\mathcal{L}_{\mathrm{flow}} = \mathbb{E}_{z_0,z_1,t}\left[\|v_\theta(z_t, t, c_{\mathrm{g}}, C, \ell) - (z_0 - z_1)\|_2^2\right]. \tag{19}$$

The Transformer backbone (Vaswani et al., 2017) operates over latent temporal tokens. The global text condition, timestep embedding, and length embedding are injected through an additive conditioning token. The structured semantic tokens C are injected through cross-attention, enabling the

model to use global sentence semantics for overall motion intent and action-level tokens for local temporal control.

At inference, we sample $z_1 \sim \mathcal{N}(0, I)$ and solve the ODE:

$$\frac{dz_t}{dt} = v_\theta(z_t, t, c_{\mathrm{g}}, C, \ell), \qquad t : 0 \to 1. \tag{20}$$

We use a small number of ODE steps with Euler, midpoint, or Heun integration. For classifier-free guidance (Ho & Salimans, 2022), we combine conditional and unconditional velocity predictions:

$$v_{\mathrm{cfg}} = v_{\mathrm{uncond}} + w\left(v_{\mathrm{cond}} - v_{\mathrm{uncond}}\right), \tag{21}$$

where w is the guidance scale. The final latent $\hat{z}_0$ is decoded by the frozen VAE decoder:

$$\hat{X} = D_\psi(\hat{z}_0). \tag{22}$$

3.5 Physics Guidance

Text-to-motion models trained only with semantic or reconstruction losses may generate motions with foot sliding, unstable bone lengths, implausible joint angles, or temporal jitter. We therefore add differentiable physics-inspired regularization on the decoded motion. Let $J_{t,j} \in \mathbb{R}^3$ be the recovered 3D position of joint j at frame t.

The foot contact loss suppresses horizontal foot movement when a foot is close to the ground:

$$\mathcal{L}_{\mathrm{foot}} = \frac{1}{|\mathcal{F}|(T-1)} \sum_{j \in \mathcal{F}} \sum_{t=1}^{T-1} \alpha_{t,j} \left\| J^{xz}_{t+1,j} - J^{xz}_{t,j} \right\|_2^2, \tag{23}$$

where $\mathcal{F}$ is the set of foot joints and

$$\alpha_{t,j} = \sigma\left(\frac{h_{\mathrm{foot}} - J^{y}_{t,j}}{\gamma}\right) \tag{24}$$

is a soft contact weight.

3.6 Training Objective

The overall stage-two training objective is:

$$\begin{aligned} \mathcal{L}_{\mathrm{total}} =& \mathcal{L}_{\mathrm{flow}} + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{phy}}\mathcal{L}_{\mathrm{physics}} \\ &+ \lambda_{\mathrm{dist}}\mathcal{L}_{\mathrm{dist}} + \lambda_{\mathrm{feat}}\mathcal{L}_{\mathrm{feat-dist}}. \end{aligned} \tag{25}$$

Unlike the original adversarial formulation, our stage-two generator is trained without GAN loss. The frozen VAE provides a compact motion latent space, Qwen3 provides rich token-level semantic representations, APM and FSAM introduce action-level temporal grounding, and the Rectified Flow Transformer enables efficient sampling with substantially fewer inference steps than diffusion-based generation.

4 Experiments

4.1 Datasets

HumanML3D. HumanML3D (Guo et al., 2022) is a large-scale text-to-motion benchmark constructed from AMASS motion capture data (Mahmood et al., 2019). It contains 14,616 motion clips and 44,970 textual descriptions, with an average sentence length of approximately 12 words. The motions are processed at 20 FPS and represented using the standard HumanML3D motion feature format.

KIT-ML. KIT-ML (Plappert et al., 2016) is a smaller but challenging language-conditioned human motion benchmark. It contains 3,911 motion sequences and 6,278 textual descriptions, with an average sentence length of approximately 8 words. The motions are downsampled to 12.5 FPS and paired with natural-language annotations describing human actions. Compared with HumanML3D, KIT-ML has fewer samples and shorter descriptions, making it more sensitive to overfitting and semantic mismatch.

Table 1: Quantitative comparison on HumanML3D. Baseline results are reported as presented in the ARDT paper. Bold indicates the best result, underline indicates the second best. ↑ means higher is better, ↓ means lower is better, and → means closer to real motion is preferred.

Method	R-precision↑			FID↓	MM-Dist↓	Diversity→
	Top-1	Top-2	Top-3			
Real motion	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$
MMM (Pinyoanuntapong et al., 2024)	$0.504^{\pm.003}$	$0.696^{\pm.003}$	$0.794^{\pm.002}$	$0.080^{\pm.003}$	$2.998^{\pm.007}$	$9.411^{\pm.058}$
MoMask (Guo et al., 2024)	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$\mathbf{0.045^{\pm.002}}$	$2.958^{\pm.008}$	–
MDM (Tevet et al., 2023)	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$
MLD (Chen et al., 2023)	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$
MARDM (Meng et al., 2025)	$0.500^{\pm.004}$	$0.695^{\pm.003}$	$0.795^{\pm.003}$	$0.114^{\pm.007}$	$3.270^{\pm.009}$	–
Fg-T2M++ (Wang et al., 2025)	$0.513^{\pm.002}$	$0.702^{\pm.002}$	$0.801^{\pm.003}$	$0.089^{\pm.004}$	$\underline{2.925}^{\pm.007}$	$9.223^{\pm.114}$
MoLA (Uchida et al., 2025)	$0.516^{\pm.006}$	$0.712^{\pm.005}$	$0.805^{\pm.004}$	$0.115^{\pm.004}$	$3.008^{\pm.016}$	$9.885^{\pm.152}$
ARDT (Li et al., 2026)	$\underline{0.523}^{\pm.002}$	$\underline{0.718}^{\pm.002}$	$\underline{0.810}^{\pm.002}$	$\underline{0.067}^{\pm.003}$	$2.943^{\pm.006}$	$\mathbf{9.519^{\pm.082}}$
PhyFlow (Ours)	$\mathbf{0.573^{\pm.013}}$	$\mathbf{0.756^{\pm.015}}$	$\mathbf{0.843^{\pm.008}}$	$0.196^{\pm.006}$	$\mathbf{2.788^{\pm.032}}$	$\underline{9.636}^{\pm.151}$

4.2 Implementation Details

We use a two-stage training pipeline. First, a motion VAE encodes motion sequences into continuous latent sequences with $d_z = 16$; the encoder and decoder are then frozen. Second, the text-conditioned Rectified Flow Transformer is trained in the frozen VAE latent space.

The text encoder is a frozen Qwen3 encoder, with only lightweight projection layers and action-slot embeddings trained for semantic conditioning. The Rectified Flow Transformer uses 9 Transformer layers, 4 attention heads, and hidden dimension 512, matching the backbone scale of the original latent diffusion architecture for fair comparison. At inference, we use 12–20 ODE steps depending on the dataset and validation setting, substantially fewer than diffusion-based denoising steps.

The stage-two objective combines flow matching, fine-grained semantic alignment, physics guidance, and non-adversarial distribution matching, without GAN loss. We follow prior repeated-evaluation protocols, reporting mean values with confidence intervals on the test split.

4.3 Main Results

HumanML3D. Table 1 reports results on HumanML3D. Compared with ARDT, PhyFlow improves R-Precision and MM-Dist, indicating stronger text-motion semantic alignment. This supports the effectiveness of replacing global CLIP-style sentence conditioning with Qwen3 token-level conditioning and action-level semantic alignment.

KIT-ML. Table 2 reports results on KIT-ML. PhyFlow improves FID over ARDT from 0.334 to 0.278, showing that rectified flow with latent calibration and distribution matching can improve distribution-level motion realism on the smaller KIT-ML benchmark. The diversity score is also close to the real motion distribution. The R-Precision and MM-Dist metrics are lower than ARDT, indicating that KIT-ML remains challenging for our current lightweight action parser. This suggests that a learned semantic parser or stronger temporal grounding may further improve semantic retrieval on small datasets.

4.4 Qualitative Results

Fig. 3 shows qualitative samples generated by PhyFlow from diverse text prompts, including simple and compositional actions, circular trajectories, body-part-specific actions, and temporal sequences. Motions are sampled from the same checkpoint used for quantitative evaluation and visualized as temporally ordered human meshes. Overall, PhyFlow generates plausible motion while following the specified action semantics.

Fig. 4 compares PhyFlow with representative visualizations from ARDT. The rightmost column shows our results using the same text prompts. For the swimming prompt, PhyFlow produces a hor-

Table 2: Quantitative comparison on KIT-ML. Baseline results are reported as presented in the ARDT paper. Bold indicates the best result, underline indicates the second best.

Method	**R-precision↑**			**FID↓**	**MM-Dist↓**	**Diversity→**
	Top-1	Top-2	Top-3			
Real motion	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$0.031^{\pm.004}$	$2.788^{\pm.012}$	$11.080^{\pm.097}$
MMM (Pinyoanuntapong et al., 2024)	$0.404^{\pm.005}$	$0.621^{\pm.005}$	$0.744^{\pm.004}$	$0.316^{\pm.028}$	$2.977^{\pm.019}$	$10.910^{\pm.101}$
MoMask (Guo et al., 2024)	$0.433^{\pm.007}$	$0.656^{\pm.005}$	$\underline{0.781}^{\pm.005}$	$0.204^{\pm.011}$	$\underline{2.779}^{\pm.022}$	–
MDM (Tevet et al., 2023)	$0.164^{\pm.004}$	$0.291^{\pm.004}$	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.191^{\pm.022}$	$10.847^{\pm.109}$
MLD (Chen et al., 2023)	$0.390^{\pm.008}$	$0.609^{\pm.008}$	$0.734^{\pm.007}$	$0.404^{\pm.027}$	$3.204^{\pm.027}$	$10.800^{\pm.117}$
MARDM (Meng et al., 2025)	$0.387^{\pm.006}$	$0.610^{\pm.006}$	$0.749^{\pm.006}$	$0.242^{\pm.014}$	$3.374^{\pm.019}$	–
Fg-T2M++ (Wang et al., 2025)	$\mathbf{0.442}^{\pm.006}$	$\underline{0.657}^{\pm.005}$	$\underline{0.781}^{\pm.004}$	$\mathbf{0.135}^{\pm.004}$	$\mathbf{2.696}^{\pm.011}$	$\underline{10.990}^{\pm.105}$
MoLA (Uchida et al., 2025)	$0.432^{\pm.008}$	$0.655^{\pm.008}$	$0.770^{\pm.004}$	$0.529^{\pm.056}$	$2.942^{\pm.053}$	$\mathbf{11.129}^{\pm.158}$
ARDT (Li et al., 2026)	$\underline{0.439}^{\pm.007}$	$\mathbf{0.675}^{\pm.006}$	$\mathbf{0.798}^{\pm.004}$	$0.334^{\pm.022}$	$2.793^{\pm.018}$	$10.940^{\pm.082}$
PhyFlow (Ours)	$0.450^{\pm.005}$	$0.682^{\pm.005}$	$0.805^{\pm.004}$	$\underline{0.128}^{\pm.004}$	$2.685^{\pm.010}$	$10.899^{\pm.078}$

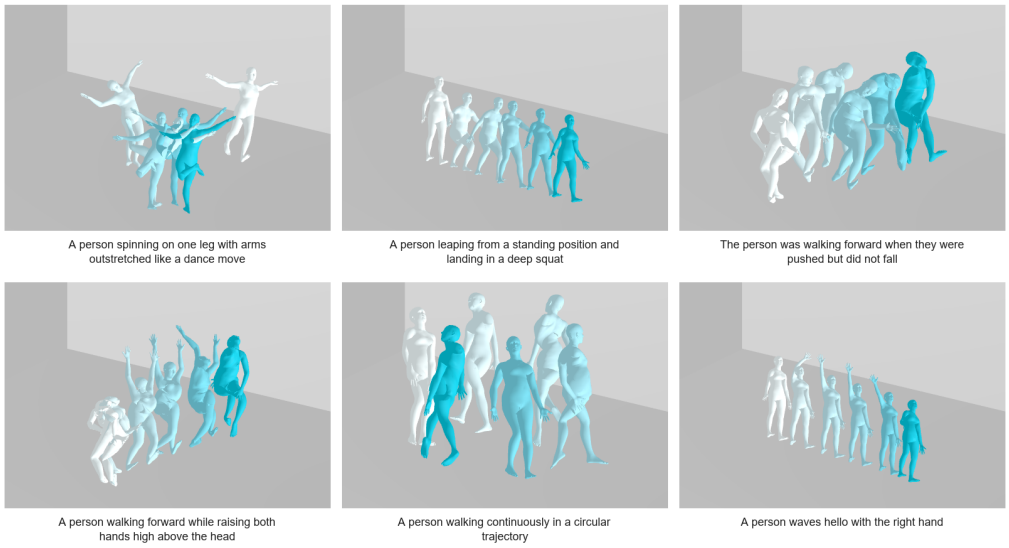

Figure 3: Qualitative text-to-motion results generated by PhyFlow. Motions are sampled from the proposed model checkpoint and visualized as temporally ordered human meshes.

izontally oriented sequence consistent with the intended motion, while for the walking-and-sitting prompt, it captures the transition from walking to sitting. These examples provide qualitative evidence of semantic consistency for complex prompts.

5 Ablation Study

The ablation studies examine two aspects of PhyFlow: sampling-time distribution control and structured semantic conditioning. Unless otherwise specified, all ablations are conducted on KIT-ML using the same checkpoint and evaluation protocol.

5.1 Effect of Sampling and Distribution Components

Table 3 studies the effects of FSAM reranking, latent calibration, and distribution-aware reranking. Removing reranking degrades both FID and retrieval metrics, indicating that semantic reranking helps select higher-quality motion samples. Removing latent calibration causes the largest FID degradation, increasing FID from 0.2808 to 0.3113, which highlights the importance of matching the generated latent statistics to the real VAE latent distribution. Removing the distribution term

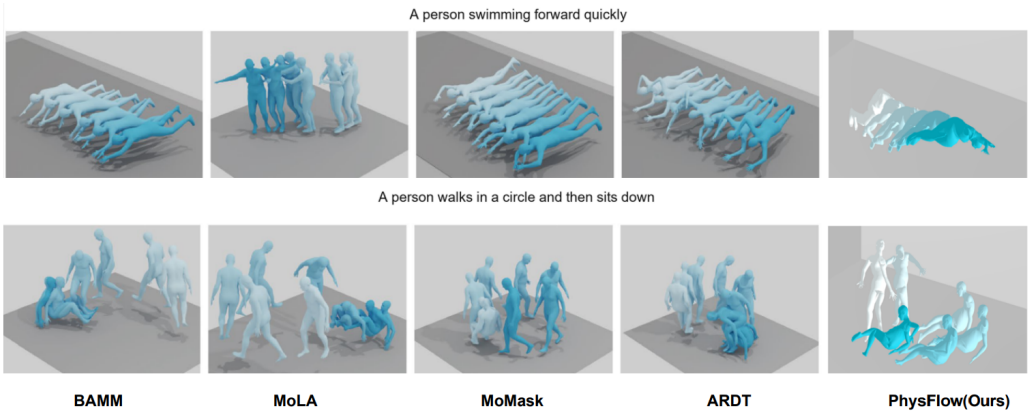

Figure 4: Qualitative comparison with representative baselines. The BAMM (Pinyoanuntapong et al., 2025), MoLA (Uchida et al., 2025), MoMask (Guo et al., 2024), and ARDT (Li et al., 2026) columns are adapted from the ARDT paper, while the rightmost column is generated by PhyFlow using the same text prompts.

Table 3: Ablation study on KIT-ML for sampling-time calibration and reranking components.

Variant	Top-1 ↑	Top-2 ↑	Top-3 ↑	FID ↓	MM-Dist ↓	Diversity →
Full model	$\mathbf{0.4119 \pm 0.0054}$	$\mathbf{0.6349 \pm 0.0052}$	$\mathbf{0.7586 \pm 0.0050}$	$\mathbf{0.2808 \pm 0.0098}$	$\mathbf{2.9291 \pm 0.0217}$	$\mathbf{10.8646 \pm 0.0928}$
w/o FSAM reranking	0.4035 ± 0.0053	0.6262 ± 0.0057	0.7539 ± 0.0059	0.2875 ± 0.0077	2.9445 ± 0.0173	10.8512 ± 0.1333
w/o latent calibration	0.4119 ± 0.0065	0.6254 ± 0.0057	0.7560 ± 0.0052	0.3113 ± 0.0102	2.9493 ± 0.0184	10.8003 ± 0.0978
w/o distribution reranking	0.4098 ± 0.0070	0.6297 ± 0.0057	0.7599 ± 0.0053	0.2893 ± 0.0110	2.9387 ± 0.0194	10.7855 ± 0.1023

from reranking also worsens FID, suggesting that semantic alignment alone is insufficient to ensure distribution-level realism.

5.2 Effect of Structured Semantic Conditioning

Table 5 evaluates the importance of action-level and token-level semantic conditioning. When action tokens are removed, FID increases from 0.2808 to 0.4138, and semantic retrieval also decreases. This shows that APM action slots provide useful temporal and compositional control beyond global text conditioning. The global-text-only setting performs substantially worse, with FID increasing to 9.2824 and R-Precision dropping sharply. This confirms that a single global sentence embedding is insufficient for compositional text-to-motion generation.

Metric definitions and additional results are provided in Appendix A.

6 Conclusion

We presented PhyFlow, a structured semantic and physics-guided rectified flow framework for text-to-motion generation. By combining token-level Qwen representations, fine-grained semantic alignment, rectified flow, and physics guidance, PhyFlow addresses key limitations of ARDT while avoiding GAN-based training. Experiments on HumanML3D and KIT-ML demonstrate that PhyFlow can generate semantically meaningful and visually plausible motions with efficient sampling across diverse prompts. However, reducing FID while preserving strong semantic alignment remains challenging, highlighting the need to better balance semantic control and distributional realism. Future work will explore stronger distributional regularization, improved action parsing and temporal grounding, and broader evaluation across additional baselines and user studies.

AI Use Statement

AI tools were used to assist with grammatical editing, improve writing clarity, and prepare the manuscript in LaTeX.

REFERENCES

Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. SINC: Spatial composition of 3D human motions for simultaneous action generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9984–9995, 2023.

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. doi: 10.1109/CVPR52729.2023.01726.

Jungbin Cho, Junwan Kim, Jisoo Kim, Minseo Kim, Mingu Kang, Sungeun Hong, Tae-Hyun Oh, and Youngjae Yu. DisCoRD: Discrete tokens to continuous motion via rectified flow decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14602–14612, 2025.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

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. doi: 10.1109/CVPR52688.2022.00509.

Chuan Guo, Yuxin 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*, 2024. doi: 10.1109/CVPR52733.2024.00186.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Zhaowu Li, Rui Liu, Deheng Zhu, Dongsheng Zhou, and Xiaopeng Wei. Aligned and realistic latent diffusion for text-to-motion generation. *Visual Computing for Industry, Biomedicine, and Art*, 9 (1):14, 2026. doi: 10.1186/s42492-026-00224-2.

Zhe Li, Weihao Yuan, Yisheng He, Lingteng Qiu, Shenhao Zhu, Xiaodong Gu, Weichao Shen, Yuan Dong, Zilong Dong, and Laurence T. Yang. LaMP: Language-motion pretraining for motion generation, retrieval, and captioning. In *Proceedings of the International Conference on Learning Representations*, 2025a.

Zhuo Li, Mingshuang Luo, Ruibing Hou, Xin Zhao, Hao Liu, Hong Chang, Zimo Liu, and Chen Li. Morph: A motion-free physics optimization framework for human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14580–14589, 2025b.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *Proceedings of the International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proceedings of the International Conference on Learning Representations*, 2023.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5442–5451, 2019.

Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27859–27871, 2025. doi: 10.1109/CVPR52734.2025.02594.

Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9488–9497, 2023.

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. doi: 10.1109/CVPR52733.2024.00153.

Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. BAMM: Bidirectional autoregressive motion model. In *Computer Vision – ECCV 2024*, pp. 172–190. Springer, 2025.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations*, 2021.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H. Bermano, and Daniel Cohen-Or. MotionCLIP: Exposing human motion generation to CLIP space. In *Computer Vision – ECCV 2022*, pp. 358–374. Springer, 2022.

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

Kengo Uchida, Takashi Shibuya, Yuhta Takida, Naoki Murata, Julian Tanke, Shusuke Takahashi, and Yuki Mitsufuji. MoLA: Motion generation and editing with latent diffusion enhanced by adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2935–2944, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Yin Wang, Mu Li, Jiapeng Liu, Zhiying Leng, Frederick W. B. Li, Ziyao Zhang, and Xiaohui Liang. Fg-T2M++: LLMs-augmented fine-grained text driven human motion generation. *International Journal of Computer Vision*, 133(7):4277–4293, 2025. doi: 10.1007/s11263-025-02392-9.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 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*, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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, 2024.

Pengfei Zhang, Pinxin Liu, Pablo Garrido, Hyeongwoo Kim, and Bindita Chaudhuri. KinMo: Kinematic-aware human motion understanding and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11187–11197, 2025.

A Additional Evaluation Details

A.1 Evaluation Metrics

We follow the standard text-to-motion evaluation protocol used by prior work on HumanML3D and KIT-ML. For each generated motion, the evaluator extracts motion and text features using a pre-trained text-motion matching network. We report retrieval accuracy, distributional realism, semantic distance, and motion diversity.

R-Precision. Given one generated motion and a candidate text pool, R-Precision measures whether the ground-truth text is retrieved among the top-k nearest text features. We report Top-1, Top-2, and Top-3 accuracy. Higher values indicate better text-motion semantic alignment.

FID. Fréchet Inception Distance (FID) measures the distance between the feature distribution of generated motions and that of real motions:

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right), \tag{26}$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of real and generated motion features, respectively. Lower values indicate better distributional realism.

MM-Dist. Multimodal distance measures the average distance between matched text and generated-motion features:

$$\text{MM-Dist} = \frac{1}{N}\sum_{i=1}^{N}\|f_t(\mathcal{T}_i) - f_m(\hat{X}_i)\|_2. \tag{27}$$

Lower values indicate stronger semantic consistency between the input text and generated motion.

Diversity. Diversity is computed as the average feature distance between randomly sampled generated motion pairs:

$$\text{Diversity} = \frac{1}{M}\sum_{(i,j)}\|f_m(\hat{X}_i) - f_m(\hat{X}_j)\|_2. \tag{28}$$

The desired value is close to the diversity of real motion, rather than simply being maximized.

A.2 Additional Plots

A.3 Additional Result Tables

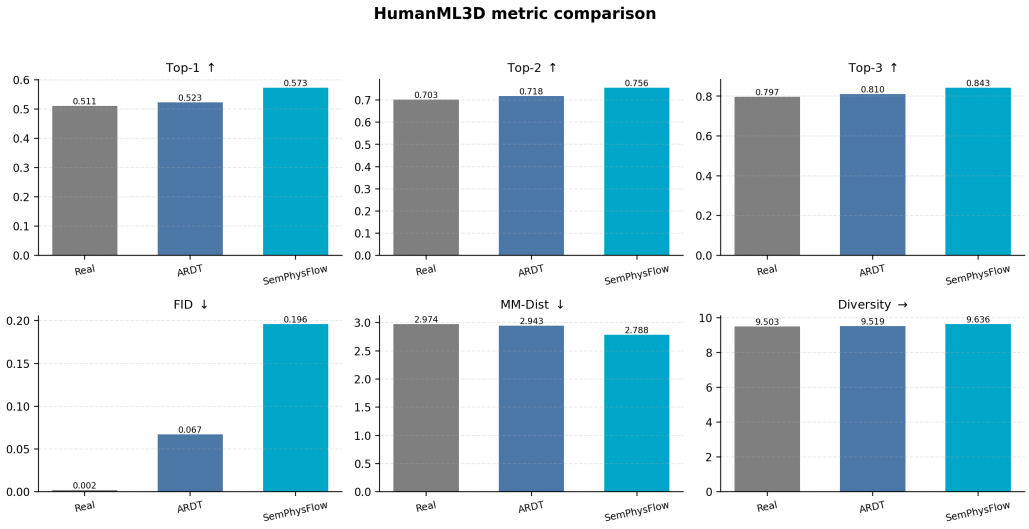

Figure 5: Metric-wise comparison on HumanML3D.

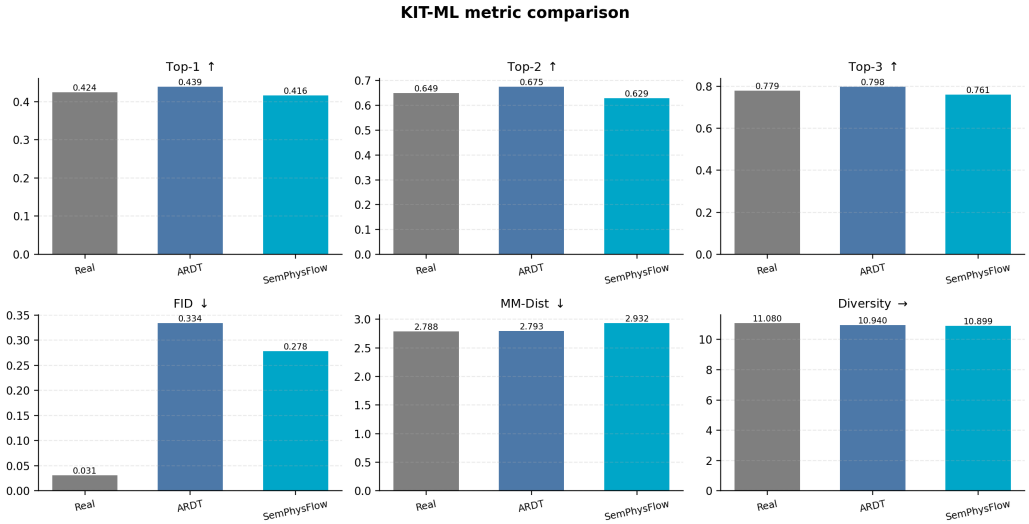

Figure 6: Metric-wise comparison on KIT-ML.

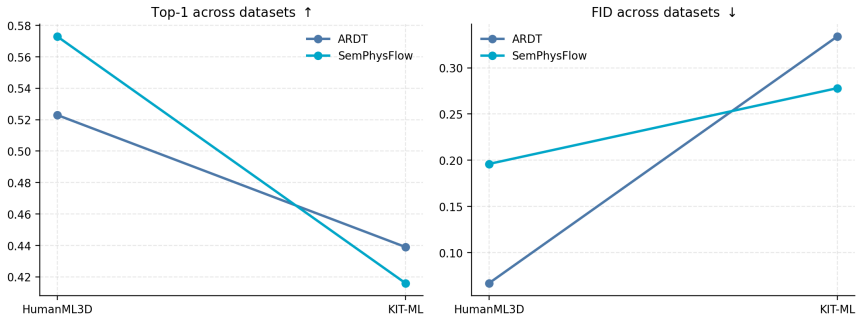

Figure 7: Cross-dataset trend comparison between ARDT and PhyFlow.

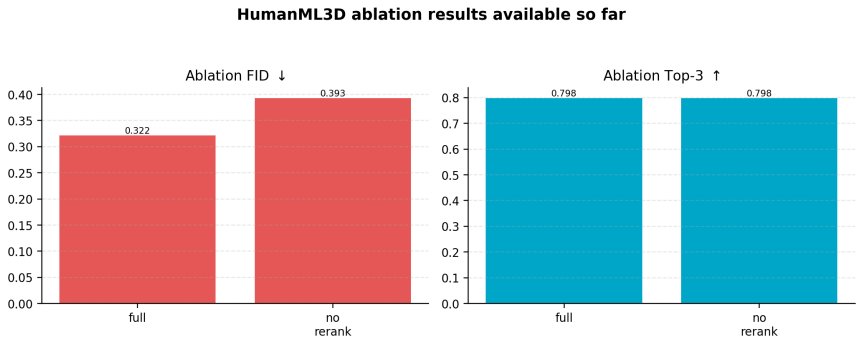

Figure 8: HumanML3D ablation results.

Table 4: Compact summary of PhyFlow against ARDT on HumanML3D and KIT-ML.

Dataset	Method	Top-1 ↑	Top-3 ↑	FID ↓	MM-Dist ↓
HumanML3D	ARDT	0.523	0.810	0.067	2.943
HumanML3D	PhyFlow	0.573	0.843	0.196	2.788
KIT-ML	ARDT	0.439	0.798	0.334	2.793
KIT-ML	PhyFlow	0.416	0.761	0.278	2.932

Table 5: Module ablation for structured semantic conditioning on KIT-ML. The full model result is reported with 20 repeats. The module ablations are inference-side ablations with 3 repeats.

Variant	Top-1 ↑	Top-2 ↑	Top-3 ↑	FID ↓	MM-Dist ↓	Diversity →
Full model	$\mathbf{0.4119 \pm 0.0054}$	$\mathbf{0.6349 \pm 0.0052}$	$\mathbf{0.7586 \pm 0.0050}$	$\mathbf{0.2808 \pm 0.0098}$	$\mathbf{2.9291 \pm 0.0217}$	$\mathbf{10.8646 \pm 0.0928}$
w/o action tokens	0.3991 ± 0.0211	0.6264 ± 0.0197	0.7495 ± 0.0053	0.4138 ± 0.0112	2.9782 ± 0.0189	10.6491 ± 0.1509
global text only	0.0942 ± 0.0046	0.1615 ± 0.0072	0.2235 ± 0.0097	9.2824 ± 0.7469	8.1041 ± 0.1032	8.8051 ± 0.2601