This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

ATM: Enhanced Alignment for Text-to-Motion Generation

Ke Han[1] Yueming Lyu[2*] Weichen Yu[3] Nicu Sebe[1]
[1] University of Trento [2] Nanjing University [3] Carnegie Mellon University
ke.han.aca@gmail.com, ymlv@nju.edu.cn, wyu3@andrew.cmu.edu, niculae.sebe@unitn.it

Abstract

Existing text-to-motion (T2M) generation methods primarily rely on regression-based objectives, such as minimizing positional errors. However, they lack effective semantic supervision and correction mechanisms, often leading to substantial misalignment between text and motion. To address this, we propose ***Aligned Text-to-Motion (ATM)****, a semantics-aware generation framework that automatically identifies and corrects text-motion misalignment. ATM incorporates two key components: (1)* ***Inter-motion alignment****, which detects semantic contradictions across motions and applies adaptive corrections based on the degree of semantic discrepancy, flexibly handing diverse misalignments and ensuring global text-motion consistency; (2)* ***Intra-motion alignment****, which refines locally missing or inaccurate motion semantics in an unsupervised manner by inferring semantic proxies, effectively addressing the absence of localized textual annotations. ATM is model-agnostic and can be seamlessly integrated into various T2M methods as a plug-and-play module. Extensive experiments on HumanML3D and KIT demonstrate that ATM consistently improves both generation quality and text-motion alignment. Code is available at https://github.com/ke-han-aca/ATM.git.*

1. Introduction

Text-to-motion (T2M) generation aims to create a sequence of human movements that aligns semantically with a provided text description. It has significant potential for applications such as animation, filmmaking and robotics, due to the user-friendly nature and semantic richness of natural language descriptions [3, 18, 26]. Recent advances in deep generative models [12, 36, 39] have significantly improved motion quality, enabling the generation of increasingly natural and coherent human motions.

However, the motion sequences generated by existing methods still often exhibit significant semantic misalignment with the given texts, as shown in Fig. 1. We categorize this misalignment into two types: 1) *inter-motion* misalignment, where semantically distinct text descriptions produce overly similar motions; and 2) *intra-motion* misalignment, where desired actions in the text are missing, inaccurately generated, or replaced by unintended actions.

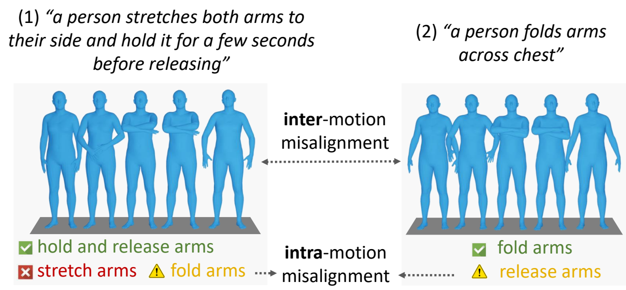

Figure 1. Semantic misalignment generated by MoMask [9]. *Inter-motion* misalignment highlights instances where semantically different text descriptions lead to overly similar motions, while *intra-motion* misalignment refers to the presence of inaccurate semantics. Green, red, and yellow words represent correct, incorrect, and undesired motion semantics, respectively.

These issues primarily stem from the reliance on regression-based objectives, such as the Mean Squared Error (MSE) [28, 42, 56], or Negative Log-Likelihood (NLL) losses [9, 30, 31]. While effective for reducing positional errors, these objectives lack explicit semantic supervision: 1) *how to identify semantic misalignment*; 2) *which semantic components are missing or incorrect*; and 3) *how to correct them*. As a result, significant inconsistencies between text and motion persist.

Achieving effective semantic guidance in T2M remains challenging. For inter-motion misalignment, although prior works incorporate contrastive learning to align text-motion pairs [23, 41], they do not explicitly handle diverse and complex semantic relationships. For instance, both "fold arm" and "jump" may serve as negative samples for "fold and then release arms," yet their semantic relevance differs significantly. Existing methods typically impose uniform penalties across all negative pairs, making it difficult to balance semantic separability and similarity. For intra-motion

*Corresponding author

misalignment, unlike full sequences paired with textual annotations, motion clips lack localized textual supervision, making it challenging to address locally missing or inaccurate motion semantics.

To tackle these challenges, we introduce **Aligned Text-to-Motion (ATM)**, a semantics-aware framework designed to identify and correct misalignment in T2M generation.

Specifically, ATM comprises two key components: 1) **Inter-motion alignment**, which identifies inter-motion semantic contradictions—cases where generated motions are overly similar despite distinct textual semantics. To handle diverse misalignment relationships, we propose semantics-aware adaptive supervision that dynamically adjusts optimization strength according to the degree of semantic discrepancy. This instance-specific correction enables the model to learn structured semantic boundaries while preserving essential semantic correlations. 2) **Intra-motion alignment**, which corrects local semantic inaccuracies in an unsupervised manner. In the absence of localized text annotations, we introduce an inferred semantic proxy by retrieving semantically aligned motion clips and refining the misaligned clip towards this proxy. Adaptive supervision is further applied to regularize clip-level semantics, allowing flexible and precise refinement.

ATM is simple yet effective, easily integrating into existing text-to-motion generation models as a plug-and-play solution. It consistently enhances generation quality and alignment performance, achieving state-of-the-art results on HumanML3D [8] and KIT [32] benchmarks.

The contributions of this work are summarized below.

- We propose ATM, a framework for text-to-motion generation that explicitly identifies and corrects semantic misalignment.
- We introduce semantics-aware adaptive supervision for inter-motion alignment, addressing diverse semantic relationships with instance-specific correction.
- We propose unsupervised intra-motion alignment using inferred semantic proxies, effectively refining local semantics without requiring localized annotations.
- Our method can be seamlessly integrated with various existing models to boost their alignment capacity, achieving state-of-the-art generation results.

2. Related Work

Text-to-Motion (T2M) Generation. Inspired by the successful applications of diffusion models [12], text-to-motion works such as MDM [42] and MotionDiffuse [56] adopt diffusion techniques to improve the quality of motion generation. Subsequent research focuses on refining diffusion architectures [13, 16, 20, 35, 44, 48], accelerating inference speed [4, 54, 61], and enhancing controllability [38, 50, 53], etc. Additionally, drawing inspiration from large language models [5, 25, 27, 43] and quantized variational autoencoders [45], methods like T2M-GPT [55], MotionGPT [15], AvatarGPT [62], and LMM [58] use discrete motion token representations, and unify multiple text-motion tasks into a single, cohesive system. Further explorations extend human motion generation to open-vocabulary scenarios [19, 21, 22], incorporate scene embeddings [47], and enable multi-motion generation [2, 6], etc. Despite these advancements, the challenge of achieving precise text-motion alignment still persists in existing text-to-motion methods.

Text-Motion Alignment. To improve alignment quality in T2M generation, works such as TEMOS [28], MotionCLIP [41] and T2M [8] align the distributions of motion and language latent spaces via utilizing Kullback-Leibler (KL) divergence. Methods like GraphMotion [17] and Fg-T2M [46] refine text codes by leveraging linguistic semantic graphs to achieve fine-grained generation. However, these approaches lack explicit semantic supervision on the generated motions to ensure alignment with the input texts. To address this, the works [23, 51] incorporate contrastive learning [29, 33, 52] by encouraging positive text-motion pairs while separating negatives. They generally apply uniform penalties to all negative pairs, failing to adapt to varying semantic relationships. Moreover, such supervision operates only at the sequence level and does not resolve local semantic misalignment due to the absence of localized text annotations. In contrast, our approach introduces dynamic, semantics-aware supervision for both global and local alignment, adaptively handling diverse semantic relationships across sequences and refining local misalignments in an unsupervised manner.

Metric Learning. Metric learning is widely used in computer vision tasks such as face recognition [40, 49] and person re-identification [10, 60]. Conventional metric learning techniques (*e.g.* contrastive learning [33], triplet loss [37]) are typically *semantics-agnostic*, applying uniform penalties to all negative pairs without explicitly considering their semantic relevance. This limits their ability to handle the heterogeneous nature of semantic misalignment in T2M generation. In contrast, our method retains a metric learning framework while introducing semantics-aware adaptive supervision. By scaling the optimization objective according to the degree of semantic discrepancy, our approach enables instance-specific alignment tailored to varying semantic relationships.

3. Method

In this work, we propose an Aligned Text-to-Motion (ATM) generation framework to address semantic misalignment, by automatically identifying and correcting inconsistencies between text and motion.

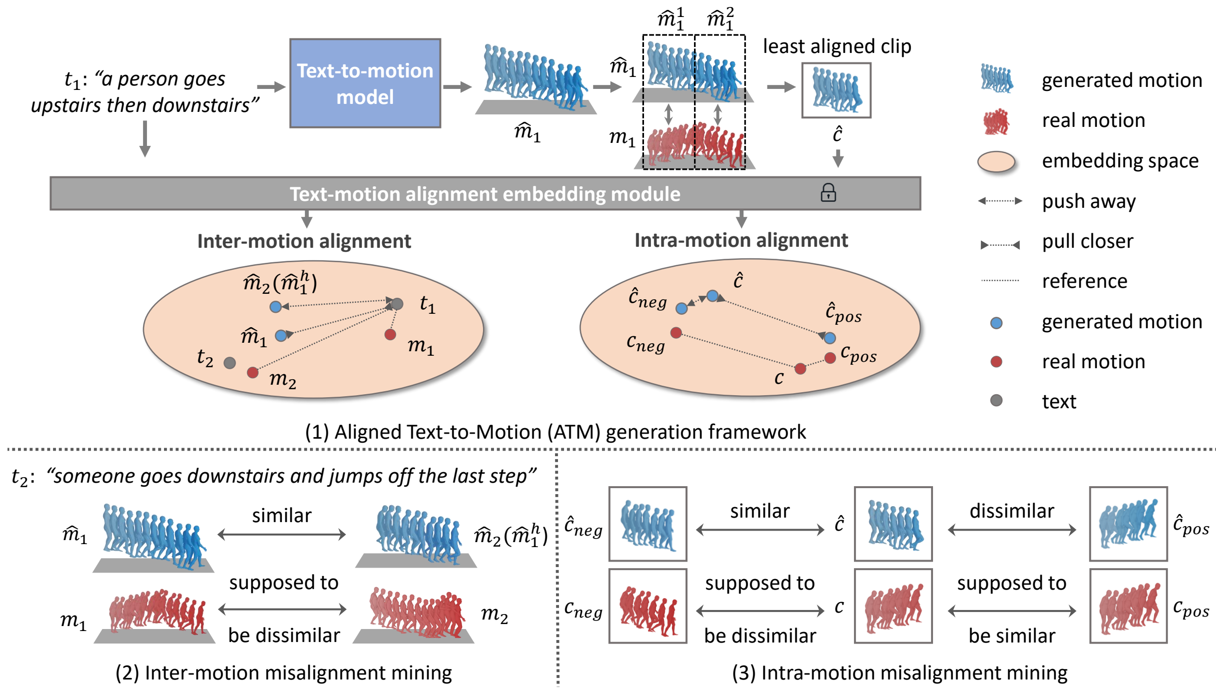

Figure 2. (1) The Aligned Text-to-Motion (ATM) generation framework. In a pre-trained text-motion alignment space, ATM performs inter-motion alignment by correcting motion sequences that are structurally similar but semantically distinct, as illustrated in (2), and intra-motion alignment by refining semantically inaccurate motion clips using positive and negative proxy clips, as shown in (3).

3.1. Overview

As shown in Fig. 2, ATM integrates a text-to-motion generation model with a pre-trained text-motion alignment embedding module. It comprises two key components: 1) *Inter-motion alignment*, which identifies and adaptively corrects semantic misalignment across motion sequences to address global, sequence-level inconsistencies; and 2) *Intra-motion alignment*, which leverages clip-level semantic proxies to refine and correct local semantic inaccuracies within motion sequences.

Given a text description $\boldsymbol{t}$ of a human motion, the model aims to generate a sequence of human poses $\boldsymbol{m} \in \mathbb{R}^{F \times J \times D}$ that aligns with the text description, where F, J and D denote the number of frames, joints, and the dimension of joint representations, respectively. The model is trained on a dataset $\mathcal{D} = \{(\boldsymbol{t}_i, \boldsymbol{m}_i)\}_{i=1}^{N}$, where $(\boldsymbol{t}_i, \boldsymbol{m}_i)$ is a text-motion pair, and N represents the total number of such pairs.

ATM is a plug-and-play training framework, compatible with various existing methods [28, 56, 57] as its text-to-motion model. In this work, we use the diffusion-based model MDM [42] as our baseline due to its concise and efficient structure. The model first regularizes its output using mean squared error (MSE) loss to learn motion generation, defined as:

$$\mathcal{L}_{mse} = \|\boldsymbol{m}_i - \hat{\boldsymbol{m}}_i\|_2^2, \tag{1}$$

where $\hat{\boldsymbol{m}}_i$ represents the generated motion from the text $\boldsymbol{t}_i$.

However, models trained solely with MSE loss often generate overly similar motions for distinct textual semantics, leading to severe inter-motion semantic misalignment, as illustrated in Fig. 1. This occurs because the MSE loss focuses on frame-level skeletal position regression, without modeling relationships across different instances, thereby failing to effectively capture semantic similarities between motions. To mitigate this, we introduce an inter-motion alignment strategy that explicitly identifies and corrects such misalignment.

3.2. Inter-Motion Alignment

Inter-motion misalignment mining. To identify inter-motion misalignment, we first pre-train a fundamental text-motion alignment embedding module [23], to construct an embedding space where representations of paired text and motion $(\boldsymbol{t}_i, \boldsymbol{m}_i)$ are aligned. Then, for a generated motion sequence $\hat{\boldsymbol{m}}_i$, we define its inter-motion misalignment example $\hat{\boldsymbol{m}}_i^h$ in a mini batch as

$$\hat{\boldsymbol{m}}_i^h = \hat{\boldsymbol{m}}_{j^*}, \text{where } j^* = \arg\min_{j \neq i} \frac{\mathcal{D}(\hat{\boldsymbol{m}}_i, \hat{\boldsymbol{m}}_j)}{\mathcal{D}(\boldsymbol{m}_i, \boldsymbol{m}_j)}, \tag{2}$$

where $\mathcal{D}$ represents the Euclidean distance between two motion representations in the embedding space. As shown in Fig. 2 (2), a selected pair of misaligned examples is characterized by the minimum ratio between $\mathcal{D}(\hat{\boldsymbol{m}}_i, \hat{\boldsymbol{m}}_j)$ and $\mathcal{D}(\boldsymbol{m}_i, \boldsymbol{m}_j)$, satisfying the following criteria: 1) the distance between two ground-truth motions $\mathcal{D}(\boldsymbol{m}_i, \boldsymbol{m}_j)$ is relatively large. This implies that these motions likely have distinct semantics, making them valid negative examples for differentiation; and 2) the distance between the corresponding generated motions $\mathcal{D}(\hat{\boldsymbol{m}}_i, \hat{\boldsymbol{m}}_j)$ is relatively small, suggesting that the model struggles to capture the semantic differences between the two motions, resulting in overly similar outputs instead.

In criterion 1), exceptions may arise when similar texts produce diverse yet semantically consistent motions. To address this, we exclude instances with highly similar texts by setting $\mathcal{D}(\hat{\boldsymbol{m}}_i, \hat{\boldsymbol{m}}_j)$ to infinity when $\cos(\boldsymbol{t}_i, \boldsymbol{t}_j) > \epsilon$, where $\cos(\boldsymbol{t}_i, \boldsymbol{t}_j)$ represents the cosine similarity between two texts. This similarity is computed by encoding the texts into embeddings using a pre-trained language model [34]. The threshold ϵ defines the maximum similarity score for a text pair to be considered as valid negative examples.

Inter-motion adaptive alignment. To address the diverse and complex semantic relationships among misaligned pairs, we introduce an adaptive inter-motion alignment loss, defined as:

$$\begin{aligned}\mathcal{L}_{inter} = \max\{0,\ \mathcal{D}(\boldsymbol{t}_i, \hat{\boldsymbol{m}}_i) - \mathcal{D}(\boldsymbol{t}_i, \hat{\boldsymbol{m}}_i^h) + \phi_{ter}\},\\ \text{where } \phi_{ter} = \mathcal{D}(\boldsymbol{t}_i, \boldsymbol{m}_i^h) - \mathcal{D}(\boldsymbol{t}_i, \boldsymbol{m}_i),\end{aligned} \tag{3}$$

$\mathcal{D}$ denotes the Euclidean distance between the embeddings of text and motion. $\mathcal{L}_{inter}$ is structured as a triplet loss but offers two key advantages over conventional triplet loss or metric learning formulations [11, 37, 51].

First, conventional approaches focus on finding the closest negatives in feature space. However, in T2M, structurally similar motions (*e.g.*, "run" and "jog") may be semantically compatible and should not be penalized. $\mathcal{L}_{inter}$ redefines negatives based on *semantic contradiction*, *i.e.*, motions that are structurally similar yet semantically divergent, enabling semantics-aware mining tailored to T2M. Second, existing approaches usually apply uniform penalties across all negatives, overlooking the varying degrees of semantic misalignment. In contrast, $\mathcal{L}_{inter}$ introduces adaptive supervision by scaling the optimization target according to the severity of semantic inconsistency. This instance-specific correction enables the model to learn structured semantic boundaries, promoting effective semantic differentiation while preserving inherent semantic correlations.

3.3. Intra-Motion Alignment

The inter-motion alignment focuses on sequence-level motion-semantic optimization, but it may fall short when the overall generated sequence satisfies the distance relationship formulated in Eq. (3), whereas local misalignment persists in certain clips within the sequence. For example, one of desired actions is missing or semantically incorrect, which may not significantly affect sequence-level alignment but can limit generation precision. To address this, we also introduce an intra-motion alignment approach to refine clip-level local misalignment.

Intra-motion misalignment mining. To identify local misalignment, we first divide both the generated and real motion sequences into a series of successive clips, denoted as $\hat{\boldsymbol{m}}_i = \{\hat{\boldsymbol{m}}_i^1, \hat{\boldsymbol{m}}_i^2, ...\}$ and $\boldsymbol{m}_i = \{\boldsymbol{m}_i^1, \boldsymbol{m}_i^2, ...\}$, respectively, where each clip has a fixed length L. The least aligned clip $\hat{\boldsymbol{c}}$ within the generated motion $\hat{\boldsymbol{m}}_i$ is identified as the clip with the lowest similarity to its corresponding real clip, formulated as:

$$\hat{\boldsymbol{c}} = \hat{\boldsymbol{m}}_i^{k^*},\ \boldsymbol{c} = \boldsymbol{m}_i^{k^*}, \text{where } k^* = \arg\max_k \mathcal{D}(\hat{\boldsymbol{m}}_i^k, \boldsymbol{m}_i^k). \tag{4}$$

This criterion assumes that the generated motion is generally temporally aligned with the real motion, under the supervision of the MSE loss $\mathcal{L}_{mse}$. However, unlike full motion sequences paired with textual annotations, motion clips lack direct localized textual correspondence, making it challenging to resolve local misalignment. To address this, the intra-motion alignment approach models semantic relationships across a large pool of clips and identifies suitable positive and negative proxies to guide the semantic correction.

Intra-motion adaptive alignment. For a misaligned clip $\hat{\boldsymbol{c}}$, its positive proxy $\hat{\boldsymbol{c}}_{pos}$ and negative proxy $\hat{\boldsymbol{c}}_{neg}$ are defined as:

$$\hat{\boldsymbol{c}}_{pos} = \hat{\boldsymbol{c}}_{k^*}, \text{where } k^* = \arg\max_k \frac{\mathcal{D}(\hat{\boldsymbol{c}}, \hat{\boldsymbol{c}}_k)}{\mathcal{D}(\boldsymbol{c}, \boldsymbol{c}_k)}, \tag{5}$$

$$\hat{\boldsymbol{c}}_{neg} = \hat{\boldsymbol{c}}_{k^*}, \text{where } k^* = \arg\min_k \frac{\mathcal{D}(\hat{\boldsymbol{c}}, \hat{\boldsymbol{c}}_k)}{\mathcal{D}(\boldsymbol{c}, \boldsymbol{c}_k)}, \tag{6}$$

where $\hat{\boldsymbol{c}}_{k^*} \in \{\hat{\boldsymbol{c}}_k\}$, and the set $\{\hat{\boldsymbol{c}}_k\}$ represents all generated motion clips within the mini-batch, excluding $\hat{\boldsymbol{c}}$ itself. $\boldsymbol{c}_k$ denotes their corresponding ground truth clip. As depicted in Fig. 2 (3), the criterion for selecting $\hat{\boldsymbol{c}}_{pos}$ identifies a motion clip that should be semantically similar to $\hat{\boldsymbol{c}}$ but is currently dissimilar, highlighting the motion semantics $\hat{\boldsymbol{c}}$ should move closer to. Conversely, $\hat{\boldsymbol{c}}_{neg}$ identifies a clip that should differ from $\hat{\boldsymbol{c}}$ but is currently similar, indicating the motion semantics that $\hat{\boldsymbol{c}}$ should distance itself from. An intra-motion alignment loss is formulated as

$$\begin{aligned}\mathcal{L}_{intra} = \max\{0,\ \mathcal{D}(\hat{\boldsymbol{c}}, \hat{\boldsymbol{c}}_{pos}) - \mathcal{D}(\hat{\boldsymbol{c}}, \hat{\boldsymbol{c}}_{neg}) + \phi_{tra}\},\\ \text{where } \phi_{tra} = \mathcal{D}(\boldsymbol{c}, \boldsymbol{c}_{neg}) - \mathcal{D}(\boldsymbol{c}, \boldsymbol{c}_{pos})),\end{aligned} \tag{7}$$

where $\hat{\boldsymbol{c}}$ and $\boldsymbol{c}$, $\hat{\boldsymbol{c}}_{pos}$ and $\boldsymbol{c}_{pos}$, $\hat{\boldsymbol{c}}_{neg}$ and $\boldsymbol{c}_{neg}$ denote pairs of generated and real motion clips. Similar to Eq. (3), $\mathcal{L}_{intra}$ also functions as a triplet loss with an adaptive margin for each pair of examples, enabling flexible refinement for misaligned motion semantics based on their specific semantic relationships. To enhance the effectiveness of intra-motion alignment, $\mathcal{L}_{intra}$ is applied exclusively when the entire generated and real motion sequences exhibit substantial semantic differences. This condition is defined as $\cos(\hat{\boldsymbol{m}}_i, \boldsymbol{m}_i) < \gamma$, where cos denotes the cosine similarity between the two motions in the text-motion alignment embedding space, and γ represents the maximum similarity threshold.

Methods	R-Precision ↑			FID ↓	MM-Dist ↓	Diversity →	MModality ↑
	Top-1	Top-2	Top-3				
Real motions	$0.511^{\pm0.003}$	$0.703^{\pm0.003}$	$0.797^{\pm0.003}$	$0.002^{\pm0.000}$	$2.974^{\pm0.008}$	$9.503^{\pm0.065}$	-
MDM [42]	$0.320^{\pm0.005}$	$0.498^{\pm0.004}$	$0.611^{\pm0.007}$	$0.544^{\pm0.044}$	$5.566^{\pm0.027}$	$9.559^{\pm0.086}$	$2.799^{\pm0.072}$
GraphMotion [17]	$0.504^{\pm0.003}$	$0.699^{\pm0.002}$	$0.785^{\pm0.002}$	$0.116^{\pm0.004}$	$3.070^{\pm0.008}$	$9.692^{\pm0.067}$	$2.766^{\pm0.096}$
MMM [31]	$0.515^{\pm0.002}$	$0.708^{\pm0.002}$	$0.804^{\pm0.002}$	$0.089^{\pm0.005}$	$2.926^{\pm0.007}$	$9.577^{\pm0.050}$	$1.226^{\pm0.035}$
Motion Mamba [59]	$0.502^{\pm0.003}$	$0.693^{\pm0.002}$	$0.792^{\pm0.002}$	$0.281^{\pm0.009}$	$3.060^{\pm0.058}$	$9.871^{\pm0.084}$	$2.294^{\pm0.058}$
ParCo [63]	$0.515^{\pm0.003}$	$0.706^{\pm0.003}$	$0.801^{\pm0.002}$	$0.109^{\pm0.005}$	$2.927^{\pm0.008}$	$9.576^{\pm0.088}$	$1.382^{\pm0.060}$
CoMo [14]	$0.502^{\pm0.002}$	$0.692^{\pm0.007}$	$0.790^{\pm0.002}$	$0.262^{\pm0.004}$	$3.032^{\pm0.015}$	$9.936^{\pm0.066}$	$1.013^{\pm0.046}$
BAMM [30]	$0.522^{\pm0.003}$	$0.715^{\pm0.003}$	$0.808^{\pm0.003}$	$0.055^{\pm0.002}$	$2.936^{\pm0.077}$	$9.636^{\pm0.009}$	$1.732^{\pm0.055}$
Baseline	$0.425^{\pm0.013}$	$0.615^{\pm0.011}$	$0.712^{\pm0.009}$	$0.570^{\pm0.090}$	$3.635^{\pm0.045}$	$9.761^{\pm0.127}$	$2.532^{\pm0.079}$
ATM	$0.519^{\pm0.004}$	$0.709^{\pm0.007}$	$0.809^{\pm0.044}$	$0.320^{\pm0.037}$	$3.032^{\pm0.027}$	$9.463^{\pm0.122}$	$2.765^{\pm0.072}$
MLD [1]	$0.481^{\pm0.003}$	$0.673^{\pm0.003}$	$0.772^{\pm0.002}$	$0.473^{\pm0.013}$	$3.196^{\pm0.010}$	$9.724^{\pm0.082}$	$2.413^{\pm0.079}$
+ATM	$0.522^{\pm0.003}$	$0.703^{\pm0.004}$	$0.796^{\pm0.004}$	$0.266^{\pm0.011}$	$2.989^{\pm0.012}$	$9.633^{\pm0.122}$	$2.698^{\pm0.063}$
MotionDiffuse [56]	$0.491^{\pm0.001}$	$0.681^{\pm0.001}$	$0.782^{\pm0.001}$	$0.630^{\pm0.001}$	$3.113^{\pm0.001}$	$9.410^{\pm0.049}$	$1.553^{\pm0.042}$
+ATM [56]	$0.531^{\pm0.005}$	$0.719^{\pm0.004}$	$0.810^{\pm0.005}$	$0.351^{\pm0.002}$	$3.082^{\pm0.012}$	$9.453^{\pm0.098}$	$2.824^{\pm0.035}$
ReMoDiffuse [57]	$0.510^{\pm0.005}$	$0.698^{\pm0.006}$	$0.795^{\pm0.004}$	$0.103^{\pm0.004}$	$2.974^{\pm0.016}$	$9.018^{\pm0.075}$	$1.795^{\pm0.043}$
+ATM	$0.533^{\pm0.006}$	$0.716^{\pm0.003}$	$0.806^{\pm0.007}$	$0.062^{\pm0.003}$	$2.930^{\pm0.009}$	$9.412^{\pm0.084}$	$2.809^{\pm0.041}$
MoMask [9]	$0.521^{\pm0.002}$	$0.713^{\pm0.002}$	$0.807^{\pm0.002}$	$0.045^{\pm0.002}$	$2.958^{\pm0.008}$	-	$1.241^{\pm0.040}$
+ATM	$0.528^{\pm0.009}$	$0.715^{\pm0.004}$	$0.807^{\pm0.006}$	$0.043^{\pm0.007}$	$2.944^{\pm0.012}$	$9.426^{\pm0.077}$	$2.638^{\pm0.074}$

Table 1. Results on HumanML3D [8]. "↑", "↓" and "→" indicate that higher or lower values, or values closer to real motion are better, respectively. "+ATM" indicates incorporating ATM with corresponding models. The baseline is trained only with $\mathcal{L}_{mse}$. Red and blue highlight the top two results.

Methods	R-Precision ↑			FID ↓	MM-Dist ↓	Diversity →	MModality ↑
	Top-1	Top-2	Top-3				
Real motions	$0.424^{\pm0.005}$	$0.649^{\pm0.006}$	$0.779^{\pm0.006}$	$0.031^{\pm0.004}$	$2.788^{\pm0.012}$	$11.08^{\pm0.097}$	-
MDM [42]	$0.164^{\pm0.004}$	$0.291^{\pm0.004}$	$0.396^{\pm0.004}$	$0.497^{\pm0.021}$	$9.191^{\pm0.022}$	$10.85^{\pm0.109}$	$1.907^{\pm0.214}$
GraphMotion [17]	$0.429^{\pm0.007}$	$0.648^{\pm0.006}$	$0.769^{\pm0.006}$	$0.313^{\pm0.013}$	$3.076^{\pm0.022}$	$11.12^{\pm0.135}$	$3.627^{\pm0.113}$
MMM [31]	$0.404^{\pm0.005}$	$0.621^{\pm0.005}$	$0.744^{\pm0.004}$	$0.316^{\pm0.028}$	$2.977^{\pm0.019}$	$10.91^{\pm0.101}$	$1.232^{\pm0.039}$
Motion Mamba [59]	$0.419^{\pm0.006}$	$0.645^{\pm0.005}$	$0.765^{\pm0.006}$	$0.307^{\pm0.041}$	$3.021^{\pm0.025}$	$11.02^{\pm0.098}$	$1.678^{\pm0.064}$
ParCo [63]	$0.430^{\pm0.004}$	$0.649^{\pm0.007}$	$0.772^{\pm0.006}$	$0.453^{\pm0.027}$	$2.820^{\pm0.028}$	$10.95^{\pm0.094}$	$1.245^{\pm0.022}$
CoMo [14]	$0.422^{\pm0.009}$	$0.638^{\pm0.007}$	$0.765^{\pm0.011}$	$0.332^{\pm0.045}$	$2.873^{\pm0.021}$	$10.95^{\pm0.196}$	$1.249^{\pm0.008}$
BAMM [30]	$0.436^{\pm0.007}$	$0.660^{\pm0.006}$	$0.791^{\pm0.005}$	$0.200^{\pm0.011}$	$2.714^{\pm0.016}$	$10.91^{\pm0.097}$	$1.517^{\pm0.058}$
Baseline	$0.392^{\pm0.011}$	$0.605^{\pm0.007}$	$0.731^{\pm0.008}$	$0.566^{\pm0.052}$	$3.134^{\pm0.026}$	$10.86^{\pm0.151}$	$1.921^{\pm0.211}$
ATM	$0.434^{\pm0.010}$	$0.649^{\pm0.009}$	$0.776^{\pm0.006}$	$0.354^{\pm0.041}$	$2.803^{\pm0.028}$	$10.95^{\pm0.083}$	$2.783^{\pm0.011}$
MLD [1]	$0.390^{\pm0.008}$	$0.609^{\pm0.008}$	$0.734^{\pm0.007}$	$0.404^{\pm0.027}$	$3.204^{\pm0.027}$	$10.80^{\pm0.117}$	$2.192^{\pm0.071}$
+ATM	$0.429^{\pm0.020}$	$0.642^{\pm0.043}$	$0.770^{\pm0.005}$	$0.372^{\pm0.016}$	$2.986^{\pm0.007}$	$10.81^{\pm0.068}$	$2.633^{\pm0.110}$
MotionDiffuse [56]	$0.417^{\pm0.004}$	$0.621^{\pm0.004}$	$0.739^{\pm0.004}$	$1.954^{\pm0.062}$	$2.958^{\pm0.005}$	$11.10^{\pm0.143}$	$0.730^{\pm0.013}$
+ATM	$0.437^{\pm0.004}$	$0.651^{\pm0.004}$	$0.779^{\pm0.011}$	$0.235^{\pm0.021}$	$2.799^{\pm0.011}$	$11.12^{\pm0.125}$	$3.265^{\pm0.076}$
ReMoDiffuse [57]	$0.427^{\pm0.014}$	$0.641^{\pm0.004}$	$0.765^{\pm0.055}$	$0.155^{\pm0.006}$	$2.814^{\pm0.012}$	$10.80^{\pm0.105}$	$1.239^{\pm0.028}$
+ATM	$0.434^{\pm0.006}$	$0.657^{\pm0.004}$	$0.785^{\pm0.014}$	$0.193^{\pm0.028}$	$2.747^{\pm0.015}$	$11.23^{\pm0.198}$	$2.224^{\pm0.095}$
MoMask [9]	$0.433^{\pm0.007}$	$0.656^{\pm0.005}$	$0.781^{\pm0.005}$	$0.204^{\pm0.011}$	$2.779^{\pm0.022}$	-	$1.131^{\pm0.043}$
+ATM	$0.436^{\pm0.011}$	$0.655^{\pm0.006}$	$0.785^{\pm0.006}$	$0.198^{\pm0.026}$	$2.763^{\pm0.032}$	$10.97^{\pm0.148}$	$1.252^{\pm0.052}$

Table 2. Results on the KIT [32] dataset, using the same notations as in Table 1.

Insights. The intra-motion alignment approach can also be understood as inter-action alignment. The insight behind exploring inter-action alignment is that a motion sequence is often composed of multiple fundamental action elements, such as walking, turning, jumping, etc. Compared to entire, more complex motion sequences, these individual actions are generally better learned by current text-to-motion methods. Misalignment often occurs when models fail to capture crucial semantic details from the input texts, such as key actions or their temporal order, leading to missing or inaccurate actions in the generated motion sequences.

The intra-motion alignment approach formulates semantic relationships between clip-level actions, aiming to identify actions that are respectively close to the current inaccurate state and the desired target state (*i.e.*, negative and positive proxies). The model is guided to refine misaligned actions by optimizing from the current inaccurate semantic state towards the desired target semantic state via $\mathcal{L}_{intra}$. As shown in Fig. 2 (3), the selected positive and negative proxies are ideally aligned with their respective ground

truth. Even when these clips are misaligned, our method can handle such cases by computing an adaptive margin based on ground-truth clips, helping stabilize the training process.
Overall loss. The text-to-motion model is trained with the overall loss $\mathcal{L}_{all}$, defined as

$$\mathcal{L}_{all} = \mathcal{L}_{mse} + \lambda \cdot (\mathcal{L}_{inter} + \mathcal{L}_{intra}), \tag{8}$$

where λ is a weight factor. $\mathcal{L}_{mse}$ ensures precise skeleton position regression, while $\mathcal{L}_{inter}$ and $\mathcal{L}_{intra}$ guarantee text-motion semantic alignment at the global sequence and local clip levels, respectively. They are only applied to the final denoised motion sequence, where a complete motion can be decoded and backpropagation remains differentiable.

4. Experiments

4.1. Experiment Settings

Dataset. We conduct experiments on two standard human motion datasets **HumanML3D** [8] and **KIT** [32] for model training and testing. HumanML3D comprises 14,616 motion sequences sourced from AMASS [24] and HumanAct12 [7], annotated with 44,970 textual descriptions. Each sequence spans 2 to 10 seconds at a frame rate of 20 frames per second (FPS). The KIT dataset contains 3,911 motion sequences paired with 6,278 text descriptions, with each sequence recorded at 12.5 FPS.
Evaluation Protocols. Following the established evaluation protocol [8], we assess performance using five evaluation metrics. **R-Precision** and **Multimodal Distance (MMDist)** quantify how well the generated motions align with the input prompts, with R-Precision reported at Top-1, Top-2, and Top-3 accuracy levels. The **Frechet Inception Distance (FID)** measures the distributional difference between generated and ground truth motion features. **Diversity** is computed by averaging the Euclidean distances of 300 randomly sampled motion pairs. **MultiModality (MModality)** reflects the average variance for a single text prompt, calculated by measuring the Euclidean distances across 10 generated motion pairs.
Implementation Details. We initiate model training by pretraining a text-motion alignment embedding module, adopting the module structure and training strategy from the work [23], and subsequently freeze the module. Texts and motions are encoded into 256-dim embeddings. To reduce training time, the embeddings of texts, ground truth motion sequences and clips are precomputed and stored, rather than being processed during each training step. The text-to-motion model is trained for 400,000 steps with a batch size of 64, following the settings of training MDM [42]. The hyper-parameters are set as follows: the weight factor $\lambda = 0.01$, similarity threshold $\epsilon = 0.85$, and $\gamma = 0.8$. The length L of motion clips depends on the dataset, experimentally set to match a duration of 2 seconds, corresponding to 40 frames on HumanML3D and 25 frames on KIT. Motion clips shorter than 1 second are discarded. Experiments are conducted on one NVIDIA GeForce RTX 4090 GPU.

4.2. Comparison with State-of-the-Art Methods

Quantitative comparison. We present the performance comparison of state-of-the-art (SOTA) T2M models on the HumanML3D [8] and KIT [32] datasets in Table 1 and Table 2, respectively. Compared to the most competitive methods MoMask [9] and BAMM [30], ATM achieves comparable results in terms of R-Precision, MM-Dist, and diversity, while demonstrating superior MModality scores and inferior FID scores. ATM serves as a plug-and-play solution and can be easily integrated with various approaches. Incorporating ATM into MLD [1], MotionDiffuse [56] and ReMoDiffuse [57] greatly enhances their generation performance. In contrast, integrating ATM with VQ-VAE-based T2M models, such as MoMask, yields only marginal improvements. This limitation arises because MoMask generates discrete motion tokens, resulting in a non-differentiable mapping from the discrete motion space to the continuous text-motion embedding space (as further discussed in Sec. 5). Consequently, the losses can only be applied to the motion reconstruction phase to refine the codebook, rather than directly supervising the generated motion representations, which restricts the strength of semantic alignment.
Performance improvement analysis. From Tables 1 and 2, we observe that ATM, both independently and in combination with other methods, improves performance across all evaluation metrics. Specifically, text-motion alignment metrics, such as R-Precision and MModality, show the most significant gains, achieving new SOTA results on the HumanML3D dataset. This aligns with our primary goal of enhancing text-motion alignment. Additionally, the diversity score is notably improved, as our approach emphasizes semantic supervision rather than focusing solely on skeleton position regression, thereby promoting greater diversity in the generated motions. In contrast, FID shows only marginal improvements and does not demonstrate an advantage over existing SOTA methods.
Visual comparison. We visualize motion sequences generated by SOTA methods in Fig. 3. In (1), MoMask incorrectly generates “sit down” instead of “squat”, as these two actions are highly similar in terms of human skeleton positions. This highlights the limitation of focusing solely on skeleton position regression without semantic supervision, leading to semantic misalignment. BAMM correctly generate “squat” and “stand up” but in the wrong order in (1), while MoMask generates “stumble to the right” but also produces an unintended “stumble to the left” in (2). These examples support our argument in Subsection *Insights* (Section 3.3) that current methods struggle to achieve sequence-level alignment but can generate individual actions well. In contrast, ATM and ATM+MotionDiffuse produce motions that align better with the given texts, demonstrating the ef-

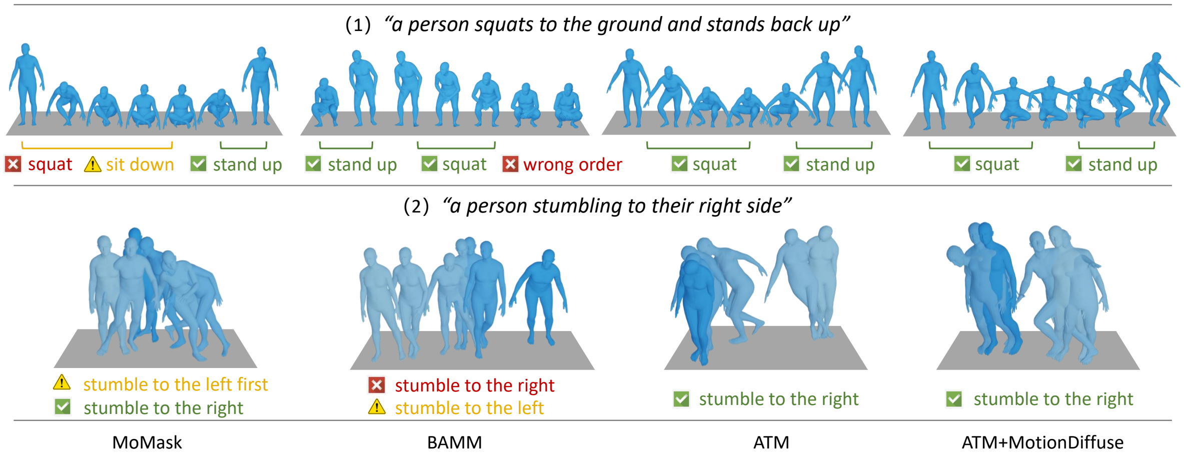

Figure 3. Examples generated by SOTA methods. In (1), the unchanged color indicates remaining in the same position, while in (2), the darkening color highlights position change. Green, red, and yellow words denote correct, incorrect, and undesired semantics, respectively.

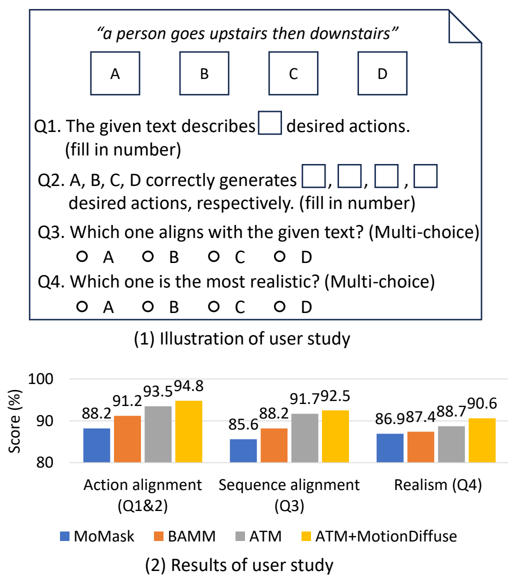

Figure 4. User study. In (1), A, B, C, D denote motion videos generated by different methods. In (2), the action alignment score is calculated as the ratio of users' responses to Q2 relative to Q1. The sequence alignment score and realistic quality are derived from the statistical analysis of users' answers to Q3 and Q4, respectively.

fectiveness of our approach.

User study. We conducted a user study to evaluate the subject quality of motion generation. Two questionnaires were designed, each containing 25 groups of randomly selected text descriptions paired with four motion sequences generated by MoMask, BAMM, ATM and MotionDiffuse. Fig. 4 (1) shows an example, where Q1 and Q2 assess action-level alignment, while Q3 evaluates sequence-level alignment. Notably, even if all desired actions are correctly generated, the overall sequence may still misalign with the text due to incorrect ordering or the inclusion of undesired actions. We

$\mathcal{L}_{mse}$	$\mathcal{L}_{inter}$	$\mathcal{L}_{intra}$	R-Precision (Top 1) ↑	FID ↓	MM-Dist ↓	Diversity →
✓	✗	✗	0.425	0.570	3.635	9.761
✓	✓	✗	0.502	0.343	3.071	9.592
✓	✗	✓	0.476	0.376	3.125	9.652
✓	✓	✓	**0.519**	**0.320**	**3.032**	**9.463**

Table 3. Ablation study of loss functions on HumanML3D.

distributed the questionnaires through public social channels. A total of 20 users participated in the study, and each participant was randomly assigned to one of the two questionnaires. Most participants are postgraduate students or researchers with expertise in computer vision, which helps ensure the reliability of the evaluation results. The results presented in Fig. 4 (2) show that, compared to MoMask and BAMM, ATM and ATM+MotionDiffuse achieve obviously higher scores in both action and sequence alignment, along with slightly higher realism scores. This demonstrates that our method can substantially improve alignment quality while preserving high fidelity in generated motions.

4.3. Ablation Study

Inter-motion and intra-motion alignment. Table 3 presents the results of the ablation study on $\mathcal{L}_{inter}$ and $\mathcal{L}_{intra}$. The removal of either $\mathcal{L}_{inter}$ or $\mathcal{L}_{intra}$ leads to a noticeable decline in performance across all four evaluation metrics. Additionally, Fig. 5 showcases motion sequences generated by models trained without these respective losses. In the first row, the model trained solely with $\mathcal{L}_{mse}$ generates the action "bring arms down" for both input texts, even though it is not specified in text (2), indicating an inter-motion misalignment issue. When $\mathcal{L}_{inter}$ is introduced, as shown in the second row, this unintended action is corrected in (2). However, key actions such as "clap", "stand", "turn" and "walk away" are still missing, revealing an intra-motion misalignment issue. By further

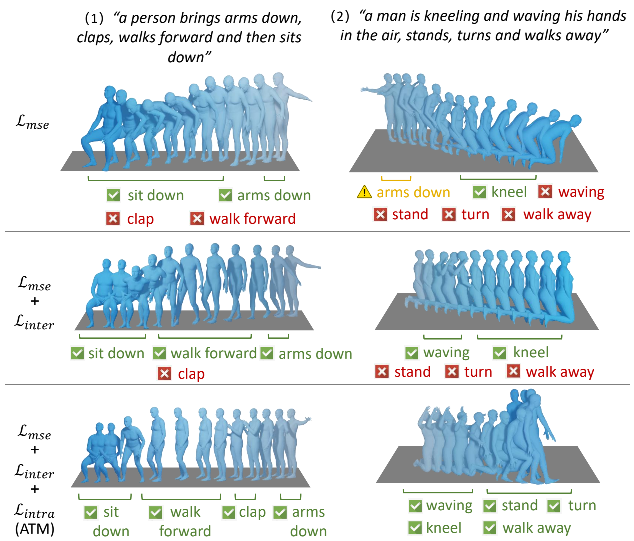

Figure 5. Motion sequences generated by different components of our method. Darker colors indicate motion progression.

incorporating $\mathcal{L}_{intra}$, the model successfully captures these missing semantics, achieving accurate alignment in the last row. These results not only validate the effectiveness of inter-motion and intra-motion alignment, but also demonstrate their complementary roles in enhancing alignment.

Adaptive semantic margin. To evaluate the effectiveness of the adaptive semantic margins, we compare them with fixed-margin baselines, as shown in Table 4. Adaptive margins yield substantial improvements by dynamically adjusting to each negative pair based on their semantic relationship and degree of misalignment. This enables more flexible separation of semantically contradictory examples while preserving meaningful semantic correlations.

Comparison with metric learning techniques. We also compare $\mathcal{L}_{inter}$ with other metric learning losses in Table 5. Contrastive loss [23] maximizes the ratio of positive-to-negative similarity in the text-motion embedding space. Triplet loss [37] ensures the distance between negative pairs exceeds that of positive pairs by a set margin across all example pairs, while hard triplet loss [11] leverages only the closest negative example, and DropTriplet loss [51] excludes negative examples with high similarity to the anchor. To facilitate a fair comparison, we integrate each of these losses into our baseline to replace $\mathcal{L}_{inter}$, remove $\mathcal{L}_{intra}$, and apply the proposed *adaptive margin* across all triplet-based losses. In this setup, the primary difference between $\mathcal{L}_{inter}$ and the other triplet losses lies in the selection of negative examples. The results show that $\mathcal{L}_{inter}$ significantly outperforms alternative losses, underscoring its effectiveness in identifying semantically meaningful misalignments tailored to T2M.

Margin	Inter-motion			Intra-motion		
	Top 1 ↑	FID ↓	Diversity →	Top 1 ↑	FID ↓	Diversity →
ϕ=1	0.323	1.572	9.659	0.349	1.795	9.875
ϕ=5	0.342	1.563	9.752	0.389	2.754	9.735
ϕ=10	0.420	0.982	9.843	0.342	1.853	9.844
ϕ=20	0.346	1.825	9.855	0.286	0.955	9.746
Adaptive (ours)	**0.519**	**0.320**	**9.463**	**0.519**	**0.320**	**9.463**

Table 4. Comparison between the adaptive margin and predefined margin values on HumanML3D.

Loss	R-Precision (Top 1) ↑	FID ↓	MM-Dist ↓	Diversity →
Contrastive [23]	0.473	0.352	3.348	9.635
Triplet [37]	0.420	0.682	4.342	9.843
Hard triplet [11]	0.382	0.536	5.675	9.675
DropTriplet [51]	0.467	0.433	3.215	9.732
$\mathcal{L}_{inter}$	**0.502**	**0.343**	**3.071**	**9.592**

Table 5. Comparison between $\mathcal{L}_{inter}$ and other metric learning losses on HumanML3D.

5. Limitations

The limitations of our approach are discussed below.

1) Our alignment losses are applied in a continuous text-motion embedding space obtained via a motion encoder. Therefore, the generated motion representation must be differentiable to allow gradient back-propagation from the embedding space to the motion generator. However, VQ-VAE-based T2M models (*e.g.*, MoMask) operate on discrete motion tokens, where the mapping from token prediction (via argmax) to the learned codebook is non-differentiable. As a result, directly applying our alignment losses to the token-level prediction is infeasible, since gradients cannot flow through the discrete quantization process.

2) The intra-motion alignment relies on the assumption of temporal alignment between generated and real motions. Temporal misalignment could impact the accuracy of similarity computations between generated and real clips.

Future work will address these limitations by 1) exploring reparameterization strategies to enable optimization and enhance compatibility with discrete motion representations, and 2) developing mechanisms to ensure temporal alignment across motion clips.

6. Conclusion

In this paper, we propose a simple yet effective text-to-motion framework, Aligned Text-to-Motion (ATM), to significantly improve semantic alignment in human motion generation. ATM introduces two complementary components: an adaptive inter-motion alignment mechanism that corrects global semantic inconsistencies across motion sequences, and an intra-motion alignment strategy that refines local misalignment at the clip level using unsupervised semantic proxies. Extensive experiments demonstrate that our method can be integrated with various T2M models to consistently enhance their generation and alignment quality.

Acknowledgments

This work was supported by the EU Horizon Project "ELIAS - European Lighthouse of AI for Sustainability" (No. 101120237), the FIS Project GUIDANCE (Debugging Computer Vision Models via Controlled Cross-modal Generation) (No. FIS2023-03251), the National Natural Science Foundation of China (No. 62502200), the Jiangsu Provincial Science and Technology Major Project (No. BG2024042), and the Natural Science Foundation of Jiangsu Province (No. BK20251203).

References

[1] 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*, pages 18000–18010, 2023. 5, 6

[2] Seunggeun Chi, Hyung-gun Chi, Hengbo Ma, Nakul Agarwal, Faizan Siddiqui, Karthik Ramani, and Kwonjoon Lee. M2d2m: Multi-motion generation from text with discrete diffusion models. *European Conference on Computer Vision*, 2024. 2

[3] Peishan Cong, Ziyi Wang, Yuexin Ma, and Xiangyu Yue. Semgeomo: Dynamic contextual human motion generation with semantic and geometric guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17561–17570, 2025. 1

[4] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. *European Conference on Computer Vision*, 2024. 2

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019. 2

[6] Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Freemotion: A unified framework for number-free text-to-motion synthesis. *European Conference on Computer Vision*, 2024. 2

[7] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the ACM International Conference on Multimedia*, pages 2021–2029, 2020. 6

[8] 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*, pages 5152–5161, 2022. 2, 5, 6

[9] 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*, pages 1900–1910, 2024. 1, 5, 6

[10] Ke Han, Shaogang Gong, Yan Huang, Liang Wang, and Tieniu Tan. Clothing-change feature augmentation for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22066–22075, 2023. 2

[11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 4, 8

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2

[13] Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. In *Conference on Computer Vision and Pattern Recognition*, pages 7158–7168. IEEE, 2025. 2

[14] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing. In *European Conference on Computer Vision*, pages 180–196, 2024. 5

[15] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2023. 2

[16] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Runyi Yu, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Local action-guided motion diffusion model for text-to-motion generation. *European Conference on Computer Vision*, 2024. 2

[17] Peng Jin, Yang Wu, Yanbo Fan, Zhongqian Sun, Wei Yang, and Li Yuan. Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic graphs. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 5

[18] Boeun Kim, Hea In Jeong, JungHoon Sung, Yihua Cheng, Jeongmin Lee, Ju Yong Chang, Sang-Il Choi, Younggeun Choi, Saim Shin, Jungho Kim, and Hyung Jin Chang. Personabooth: Personalized text-to-motion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22756–22765, 2025. 1

[19] Han Liang, Jiacheng Bao, Ruichi Zhang, Sihan Ren, Yuecheng Xu, Sibei Yang, Xin Chen, Jingyi Yu, and Lan Xu. Omg: Towards open-vocabulary motion generation via mixture of controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–493, 2024. 2

[20] Chang Liu, Mengyi Zhao, Bin Ren, Mengyuan Liu, Nicu Sebe, et al. Spatio-temporal graph diffusion for text-driven human motion generation. In *British Machine Vision Conference*, pages 722–729, 2023. 2

[21] Hanchao Liu, Xiaohang Zhan, Shaoli Huang, Tai-Jiang Mu, and Ying Shan. Programmable motion generation for open-set motion control tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1399–1408, 2024. 2

[22] Jinpeng Liu, Wenxun Dai, Chunyu Wang, Yiji Cheng, Yansong Tang, and Xin Tong. Plan, posture and go: Towards open-world text-to-motion generation. *arXiv preprint arXiv:2312.14828*, 2023. 2

[23] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. In *International Conference on Machine Learning*, 2024. 1, 2, 3, 6, 8

[24] 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*, pages 5442–5451, 2019. 6

[25] Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020. 2

[26] 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 Computer Vision and Pattern Recognition Conference*, pages 27859–27871, 2025. 1

[27] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 2

[28] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497, 2022. 1, 2, 3

[29] 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*, pages 9488–9497, 2023. 2

[30] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. Bamm: bidirectional autoregressive motion model. In *European Conference on Computer Vision*, pages 172–190, 2024. 1, 5, 6

[31] 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*, pages 1546–1555, 2024. 1, 5

[32] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2, 5, 6

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021. 2

[34] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. 4

[35] Zeping Ren, Shaoli Huang, and Xiu Li. Realistic human motion generation with cross-diffusion models. *European Conference on Computer Vision*, 2023. 2

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015. 2, 4, 8

[38] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *International Conference on Learning Representations*, 2024. 2

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1

[40] Zichang Tan, Ajian Liu, Jun Wan, Hao Liu, Zhen Lei, Guodong Guo, and Stan Z Li. Cross-batch hard example mining with pseudo large batch for id vs. spot face recognition. *IEEE Transactions on Image Processing*, 31:3224–3235, 2022. 2

[41] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374, 2022. 1, 2

[42] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023. 1, 2, 3, 5, 6

[43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[44] 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 *Conference on Computer Vision and Pattern Recognition Workshops*, pages 2910–2919, 2025. 2

[45] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[46] Yin Wang, Zhiying Leng, Frederick WB Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22035–22044, 2023. 2

[47] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say, interact as you can: Language-guided human motion generation with scene af-

fordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[48] Dong Wei, Huaijiang Sun, Bin Li, Xiaoning Sun, Shengxiang Hu, Weiqing Li, and Jianfeng Lu. Nerm: Learning neural representations for high-framerate human motion synthesis. In *International Conference on Learning Representations*, 2024. 2

[49] Degui Xiao, Jiazhi Li, Jianfang Li, Shiping Dong, and Tao Lu. Ihem loss: Intra-class hard example mining loss for robust face recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7821–7831, 2022. 2

[50] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *International Conference on Learning Representations*, 2024. 2

[51] Sheng Yan, Yang Liu, Haoqiang Wang, Xin Du, Mengyuan Liu, and Hong Liu. Cross-modal retrieval for motion and text via droptriple loss. In *Proceedings of the ACM International Conference on Multimedia in Asia*, pages 1–7, 2023. 2, 4, 8

[52] Kangning Yin, Shihao Zou, Yuxuan Ge, and Zheng Tian. Tri-modal motion retrieval by learning a joint embedding space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1596–1605, 2024. 2

[53] Ling-An Zeng, Gaojie Wu, Ancong Wu, Jian-Fang Hu, and Wei-Shi Zheng. Progressive human motion generation based on text and few motion frames. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 9205–9217, 2025. 2

[54] Ling-An Zeng, Guohong Huang, Gaojie Wu, and Wei-Shi Zheng. Light-t2m: A lightweight and fast model for text-to-motion generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 2

[55] 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. 2

[56] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 1, 2, 3, 5, 6

[57] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 364–373, 2023. 3, 5, 6

[58] Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, et al. Large motion model for unified multi-modal motion generation. *European Conference on Computer Vision*, 2024. 2

[59] 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*, pages 265–282, 2024. 5

[60] Cairong Zhao, Xinbi Lv, Zhang Zhang, Wangmeng Zuo, Jun Wu, and Duoqian Miao. Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification. *IEEE Transactions on Multimedia*, 22(12): 3180–3195, 2020. 2

[61] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast, high-quality motion generation. *European Conference on Computer Vision*, 2024. 2

[62] Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1357–1366, 2024. 2

[63] Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji. Parco: Part-coordinating text-to-motion synthesis. In *European Conference on Computer Vision*, pages 126–143, 2024. 5