electronics

MDPI

Article

A Latent-Guided Framework for Text-Based Full-Body Human Motion Generation

Jannatul Nayeem [1], Hak-Bum Lee [1] and Young-Ho Seo [1,2,*]

[1] Department of Electronic Materials Engineering, Kwangwooon University, Seoul 01897, Republic of Korea; jannatulnayeem@kw.ac.kr (J.N.); hblee@kw.ac.kr (H.-B.L.)
[2] Omotion Inc., Seoul 999007, Republic of Korea
* Correspondence: yhseo@kw.ac.kr

Abstract

Text-to-motion generation aims to synthesize realistic human motion sequences that accurately reflect natural language descriptions. While recent approaches have improved motion quality, achieving strong semantic alignment between text and motion, especially for fine-grained articulations, remains a significant challenge. In this work, we propose a latent-guided text-to-motion generation framework that strengthens the interaction between textual representations and motion latent sequences. The proposed method integrates a structured motion latent space with a text-conditioned variational generation module, enhanced by a cross-modal attention mechanism. This design enables the model to effectively capture both global motion dynamics and detailed semantic information from text. Extensive experiments on the Motion-X dataset demonstrate that the proposed approach achieves strong semantic alignment, as reflected by improved R-precision and competitive matching performance. In addition, the model improves multi-modality, indicating its ability to generate diverse motion patterns under the same textual condition. Qualitative results further show that the generated motions preserve core action semantics and exhibit coherent temporal dynamics across different motion categories. Overall, the proposed framework provides an effective solution for improving text–motion alignment in high-dimensional motion spaces, highlighting the importance of latent-guided modeling for realistic and semantically consistent motion generation.

Keywords: text-to-motion generation; motion-x; latent representation; cross-modal attention; semantic alignment; multi-modal motion synthesis

Check for updates

Academic Editors: Sotiris Diplaris, Nefeli Georgakopoulou, Stefanos Vrochidis, Giuseppe Amato, Maurice Benayoun and Beatrice De Gelder

Received: 11 May 2026
Revised: 16 June 2026
Accepted: 18 June 2026
Published: 22 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Human motion synthesis from natural language, commonly referred to as text-to-motion generation, has gained significant attention due to its applications in animation, virtual reality, human–computer interaction, and robotics. The goal is to generate realistic, temporally coherent motion sequences that accurately reflect the semantics of the input text. Despite recent progress, achieving strong text–motion alignment while maintaining realistic motion dynamics remains a challenging problem. Recent methods have shown promising results using sequence-to-sequence models, variational frameworks, and attention mechanisms. For instance, TEMOS [1] learns a shared latent space between text and motion, while attention-based methods such as AttT2M [2] improve semantic consistency through cross-modal interactions. Diffusion-based approaches, including Motion-GPT [3], further enhance motion realism and diversity, and unified frameworks like UniTMGE [4]

improve multi-modal alignment and controllability. Robustness and cross-modal representation learning have also been explored to address inconsistencies in text–motion mapping. The SATO framework [5] introduces stability-aware mechanisms to handle variations in textual input, while retrieval-based approaches such as TMR [6] learn structured joint embeddings to improve semantic alignment. However, integrating fine-grained semantics, robustness, and high-quality motion generation within a single framework remains an ongoing challenge. This challenge becomes more significant with large-scale datasets such as Motion-X [7,8], which provide high-dimensional motion representations with detailed full-body and hand articulations. While these datasets enable modeling of complex motions, they also increase the difficulty of maintaining alignment, realism, and diversity simultaneously. To address these challenges, we propose a latent-guided text-to-motion generation framework that enhances interaction between textual representations and motion latent sequences. The model first learns a structured motion latent space capturing global dynamics and local joint behavior. A text-conditioned generation module with cross-modal attention and a variational prior and posterior formulation is then employed to align text and motion effectively. The key idea is to guide motion generation through structured latent representations, enabling stable learning and improved semantic consistency. Operating in latent space allows for better modeling of temporal dependencies and reduces noise, while attention mechanisms ensure fine-grained alignment between text and motion. We evaluated our method on the Motion-X dataset. Experimental results demonstrate strong semantic alignment, reflected in retrieval-based metrics such as R-precision, along with competitive motion-generation quality and diversity across different action categories. The main contributions of this work are summarized as follows:

- We propose a latent-guided text-to-motion generation framework that enhances semantic alignment by explicitly modeling structured motion latent representations.
- We introduce a text-conditioned variational generation module with cross-modal attention, enabling effective interaction between textual features and motion dynamics.
- We design the framework to capture both global motion patterns and fine-grained joint-level semantics, particularly for complex full-body and hand articulations in Motion-X.
- We conducted extensive experiments on the Motion-X dataset, demonstrating improved R-precision, competitive matching scores, and strong multi-modality compared to baseline methods.

The remainder of this paper is organized as follows. Section 2 reviews related work on text-to-motion generation and cross-modal learning. Section 3 presents the proposed latent-guided framework, including the structured motion latent space and text-conditioned generation module. Section 4 describes the experimental setup and evaluation protocols and reports quantitative and qualitative results on the Motion-X data set. Finally, Section 5 concludes the article and discusses future directions.

2. Related Works

Early research in text-to-motion generation relied primarily on encoder–decoder architectures and joint embedding learning between text and motion. Methods such as TEMOS [1] learn a shared latent space using variational frameworks to enable diverse motion synthesis, while approaches such as TEACH [9] extend this paradigm by incorporating temporal context for long-sequence generation. Transformer-based models such as T2M-GPT [10] further reformulate motion generation as a token prediction problem, enabling improved sequence modeling and scalability. More recent works have explored diffusion-based generative models to enhance motion quality and diversity. Methods such as Motion-Diffuse [11] and MDM [12] generate temporally coherent motion sequences through iterative denoising processes, achieving strong realism and stability. Latent dif-

fusion approaches further reduce computational cost by operating in compressed motion spaces. However, despite these advances, diffusion-based methods often struggle with complex textual descriptions and maintaining fine-grained semantic alignment across long sequences.

To address alignment issues, several studies focus on improving cross-modal interaction between text and motion. Attention-based models introduce global and local attention mechanisms to better capture relationships between textual semantics and motion dynamics. Additionally, robustness to linguistic variations has been explored in frameworks such as SATO [5], which enforce stability under paraphrased or diverse textual inputs, highlighting the importance of consistent behaviour across different language expressions. In parallel with generation approaches, text-to-motion retrieval methods aim to learn structured joint embedding spaces to improve semantic alignment. TMR [6] introduces contrastive learning to align text and motion representations, enabling effective cross-modal matching. Extensions such as TMR++ [13] further enhance robustness using language-model-based augmentation. Moreover, multi-modal frameworks like LAVIMO [14] incorporate additional modalities such as video to strengthen alignment and retrieval performance. Despite these improvements, retrieval-based approaches often suffer from limited generalization when handling complex or unseen textual descriptions.

Recent research has also explored whole-body motion generation, emphasizing the importance of fine-grained control over body parts such as hands and facial expressions. For example, Human-TOMATO [15] proposes a hierarchical framework combining discrete motion representations and GPT-based generation to produce text-aligned whole-body motions, demonstrating improved realism and alignment. However, such approaches still rely heavily on hierarchical token prediction and may struggle with flexible interaction between text and motion dynamics.

Another emerging direction focuses on controllable motion generation. For instance, timeline-based methods introduce structured inputs to specify multiple actions and their temporal relationships. The multi-track timeline control framework [16] enables fine-grained temporal and spatial composition of motions by decomposing complex prompts into multiple sub-actions, improving controllability and realism. Nevertheless, these methods often depend on pre-trained diffusion models and lack unified latent modeling for structured motion generation. In addition, earlier foundational work such as HumanML3D [17] introduces large-scale datasets and motion representation learning strategies, including motion snippet encoding and temporal VAEs, which significantly improve diversity and realism in generated motions. These contributions have laid the groundwork for modern text-to-motion systems but still face challenges in achieving strong alignment and controllability simultaneously. Recent advances in text-to-motion generation have been driven not only by improved modeling techniques but also by the availability of richer datasets and more controllable generation frameworks. While earlier datasets, such as HumanML3D, enabled learning semantic alignment between text and motion, they are limited in representing fine-grained body dynamics. In contrast, the Motion-X [7] dataset introduces a large-scale, high-dimensional SMPL-X [18] representation that captures full-body motion including hands and fingers, enabling more realistic synthesis but also increasing modeling complexity. To address the limitations of single-text inputs, recent works have explored controllable generation paradigms such as multi-track timeline control, which organizes multiple text prompts along temporal intervals to support both sequential and simultaneous actions. In parallel, compositional motion generation methods have investigated temporal and spatial decomposition strategies to better handle complex actions by modeling sequences of motions or independent body-part interactions.

To overcome the limitations of both generation and retrieval paradigms, hybrid approaches have been proposed. Methods such as MRAG [19] integrate retrieval-augmented generation to leverage prior motion knowledge, while ReMo-Diffuse [20] incorporates retrieval mechanisms into diffusion models to refine generation quality. These approaches demonstrate that incorporating external knowledge can improve semantic alignment and robustness, but they still struggle to jointly achieve fine-grained alignment, robustness to language variations, and high-quality motion generation.

Existing latent-based methods mainly focus on learning compact motion representations, while variational frameworks primarily model motion diversity through latent distribution learning. Attention-based approaches improve text-motion interaction through cross-modal feature alignment, and recent whole-body generation methods extend motion synthesis to detailed body and hand articulation. In contrast, the proposed framework combines structured latent representation learning, cross-modal attention, and latent prior refinement within a unified Motion-X-based generation framework. This design enables simultaneous modeling of semantic alignment, latent consistency, and high-dimensional full-body motion generation, including detailed hand articulation. Despite substantial progress, existing methods still struggle to jointly achieve fine-grained semantic alignment, robustness to linguistic variations, and high-quality motion generation, particularly for high-dimensional datasets such as Motion-X. This motivates the need for a unified framework that effectively integrates structured latent modeling with strong cross-modal interaction.

3. Proposed Motion Generation

Our proposed framework is a text-to-motion generation system designed to produce realistic full-body human motion, including fine-grained finger articulation, from natural language descriptions. As illustrated in Figure 1, the overall architecture follows a latent-guided generation paradigm consisting of three key stages: text encoding, latent motion representation learning, and motion decoding.

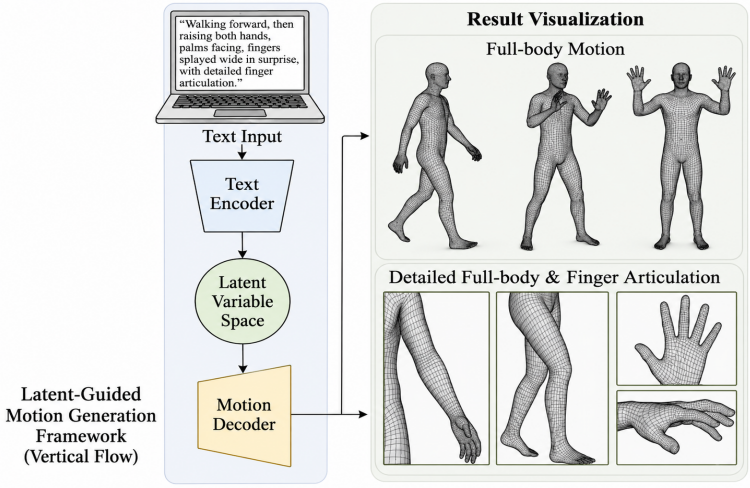


Figure 1. Overview of the proposed latent-guided text-to-motion generation framework. The input text is encoded into a semantic representation and mapped to a latent motion space, from which the motion decoder generates full-body motion sequences. The results illustrate both overall motion generation and detailed articulation, including fine-grained finger movements.

The proposed latent-guided text-to-motion generation framework consists of two main components: a motion autoencoder and a text-conditioned generation module designed to produce realistic full-body human motion, as illustrated in Figure 2. To further clarify the technical distinction of the proposed framework from existing text-to-motion generation approaches, summarizes the key architectural characteristics of representative latent-based,

variational, attention-based, and whole-body motion generation methods. As shown in Table 1, existing approaches primarily focus on latent representation learning, variational motion modeling, attention-based text-motion alignment, or whole-body motion synthesis. In contrast, the proposed framework uniquely introduces a latent-guided prior refinement mechanism that operates directly on the learned motion latent representation before motion decoding. This additional refinement stage aims to improve semantic consistency between textual descriptions and generated motions while preserving motion realism and diversity.

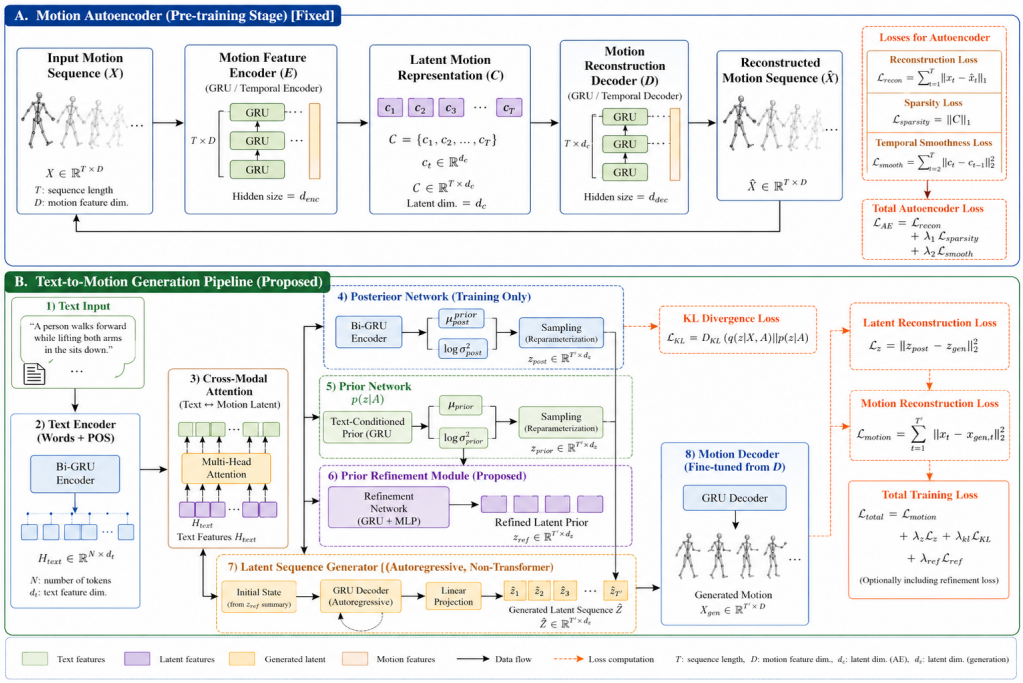


Figure 2. Overview of the proposed framework. The motion autoencoder encodes an input motion sequence into a compact latent representation and reconstructs it to preserve both global dynamics and joint-level details, guided by reconstruction, smoothness, and sparsity constraints. The text-conditioned generation module encodes textual input and aligns it with the motion latent space through cross-modal interaction, enabling the generation of coherent and semantically consistent motion sequences from text.

Table 1. Technical comparison between representative text-to-motion frameworks and the proposed method.

Method	Latent Rep.	Variational	Attention	Whole-Body	Prior Ref.
TEMOS	Yes	Yes	No	No	No
TMR	Yes	Yes	No	No	No
MotionDiffuse	Yes	No	Yes	No	No
MDM	Yes	No	Yes	No	No
HumanTOMATO	Yes	No	Yes	Yes	No
Proposed Method	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

The motion autoencoder learns a structured latent representation of human motion by encoding input motion sequences into a compact latent space and reconstructing them through a decoder. This design enables the model to capture both global motion dynamics and local joint-level variations. To ensure high-quality representations, the model is optimized using a motion reconstruction loss. In addition, temporal smoothness regularization is applied to enforce continuity across motion frames, while a sparsity constraint is introduced to promote compact and interpretable latent representations. These constraints collectively contribute to stable and coherent motion encoding.

Built upon the learned latent space, the text-conditioned generation module aligns textual semantics with motion dynamics through a cross-modal attention mechanism. Given a text input, a text representation encoder extracts semantic features, which are then fused with motion representations to guide the generation process. During training, a posterior latent inference network learns motion-aware latent distributions conditioned on both text and ground-truth motion, enabling accurate modeling of motion variability. At inference time, a text-conditioned prior network predicts latent distributions directly from textual input, allowing motion generation without access to ground-truth motion. To further enhance the quality of the generated motions, a prior refinement module is introduced to improve the consistency and expressiveness of the latent representations.

The refined latent sequence is then passed to a latent sequence generator and a motion decoder to produce the final motion sequence. The entire framework is optimized using a combination of motion reconstruction losses (for both latent and final motion outputs) and a KL divergence loss to align the prior and posterior distributions. This design enables the model to generate semantically consistent, temporally smooth, and physically realistic full-body motions, including fine-grained finger articulation.

3.1. Motion Representation Learning

Before text-conditioned motion generation, we first learn a compact motion representation space from raw motion sequences. This stage serves as a motion decomposition module that transforms high-dimensional motion features into a structured latent representation while preserving the original motion dynamics through reconstruction. Unlike conventional motion autoencoders used in existing text-to-motion frameworks, the proposed motion representation module is specifically designed for high-dimensional Motion-X full-body motion generation with detailed hand articulation. The novelty of this module lies in integrating structured latent representation learning with joint reconstruction, sparsity regularization, and temporal smoothness constraints into a unified latent-guided optimization framework. This design enables the latent space to preserve both global motion dynamics and fine-grained articulated motion details while improving temporal stability and semantic consistency for complex Motion-X motion representations.

Given an input motion sequence $\mathbf{x} \in \mathbb{R}^{T \times D}$, we first normalize the input using the dataset mean and standard deviation. The normalized sequence is then fed into a motion feature encoder, which maps the input motion into a lower-dimensional latent representation, as shown in Equation (1).

$$\mathbf{z} = E(\mathbf{x}) \tag{1}$$

where $E(\cdot)$ denotes the motion feature encoder and $\mathbf{z}$ represents the latent motion representation. The latent sequence is subsequently passed through a motion reconstruction decoder to recover the original motion, as defined in Equation (2), where $D(\cdot)$ denotes the decoder and $\hat{\mathbf{x}}$ is the reconstructed motion sequence.

$$\hat{\mathbf{x}} = D(\mathbf{z}) \tag{2}$$

The objective of this phase is not only to reconstruct the input motion accurately, but also to encourage the latent representation to be compact and temporally stable. To achieve this, the model is trained with three complementary losses, each enforcing a specific property of the latent space.

First, we use an ℓ_1 reconstruction loss between the reconstructed motion and the original motion, as shown in Equation (3). This loss computes the element-wise absolute difference between the reconstructed motion $\hat{\mathbf{x}}$ and the ground-truth motion $\mathbf{x}$, ensuring

that the model preserves both global motion dynamics and fine-grained joint-level details during reconstruction.

$$\mathcal{L}_{\text{rec}} = \|\hat{\mathbf{x}} - \mathbf{x}\|_1 \tag{3}$$

Second, we impose a sparsity constraint on the latent sequence to encourage compact and efficient motion encoding, as defined in Equation (4). This loss penalizes large activations in the latent representation $\mathbf{z}$, encouraging the model to retain only the most informative features and reduce redundancy in the latent space.

$$\mathcal{L}_{\text{sp}} = \frac{1}{N}\sum|\mathbf{z}| \tag{4}$$

Third, we apply a temporal smoothness constraint on adjacent latent steps to reduce abrupt changes in the learned representation, as shown in Equation (5). This term enforces consistency between consecutive latent vectors $\mathbf{z}_t$ and $\mathbf{z}_{t-1}$, promoting smooth temporal transitions that are essential for realistic human motion.

$$\mathcal{L}_{\text{sm}} = \|\mathbf{z}_t - \mathbf{z}_{t-1}\|_1 \tag{5}$$

The final training objective for motion representation learning is defined as shown in Equation (6), where λ_{sp} and λ_{sm} control the strengths of the sparsity and smoothness regularization terms, respectively. These coefficients balance the trade-off between accurate reconstruction and structured latent representation learning. The regularization coefficients were empirically determined through preliminary validation experiments. Specifically, the sparsity regularization coefficient and smoothness regularization coefficient were set to $\lambda_{sp} = 0.001$ and $\lambda_{sm} = 0.001$, respectively. These values provided a favorable trade-off between reconstruction accuracy, latent compactness, and temporal smoothness, resulting in stable motion representation learning.

$$\mathcal{L}_{\text{motion}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{sp}}\mathcal{L}_{\text{sp}} + \lambda_{\text{sm}}\mathcal{L}_{\text{sm}} \tag{6}$$

During training, the encoder and decoder are optimized jointly using the Adam optimizer. Gradient clipping is applied to stabilize optimization. After training, this module provides a structured latent motion space that is later used by the text-conditioned generation stage.

3.2. Text-Conditioned Motion Generation

After learning a structured motion latent space, we introduce a text-conditioned generation module to synthesize motion sequences from natural language descriptions. This module operates in the latent motion space and models the interaction between text features and motion dynamics through a sequence-to-sequence variational framework. Given a textual description, the input text is first encoded into contextual representations using a bidirectional recurrent encoder. Specifically, the text encoder produces a sequence of hidden states and a global sentence embedding shown in Equation (7), where $\mathbf{w}$ denotes the input word embeddings and parts-of-speech (POS) features.

$$\mathbf{h} = \text{TextEncoder}(\mathbf{w}) \tag{7}$$

To align text and motion, we employ a cross-modal attention mechanism that dynamically extracts relevant textual information at each motion step. Given the current decoding state, an attention vector is computed as Equation (8).

$$\mathbf{a}_t = \text{Attention}(\mathbf{s}_t, \mathbf{h}) \tag{8}$$

This attention vector allows the model to focus on different parts of the text during motion generation, improving semantic alignment between motion and language. The model adopts a variational formulation with both a posterior and a prior network defined in the latent space. During training, the posterior network incorporates ground-truth motion information and text features to produce latent variables as shown in Equation (9).

$$\mathbf{z}_t^{post} \sim q(\mathbf{z}_t \mid \mathbf{x}_t, \mathbf{a}_t) \tag{9}$$

The prior network, on the other hand, predicts latent variables conditioned only on text features as shown in Equation (10).

$$\mathbf{z}_t^{pri} \sim p(\mathbf{z}_t \mid \mathbf{a}_t) \tag{10}$$

Both networks are implemented using sequential decoders that model temporal dependencies across motion steps. At each time step, the model generates motion latent conditioned on previous states and textual information. During training, we use the posterior latent variables, while at inference, the model relies on the prior, which is shown as Equation (11).

$$\mathbf{z}_t = \begin{cases} \mathbf{z}_t^{post}, & \text{training} \\ \mathbf{z}_t^{pri}, & \text{inference} \end{cases} \tag{11}$$

The latent variables are then used to generate motion representations sequentially as in Equation (12).

$$\mathbf{m}_t = \text{Decoder}(\mathbf{m}_{t-1}, \mathbf{a}_t, \mathbf{z}_t) \tag{12}$$

The generated latent motion sequence is decoded into full motion features using the motion representation module learned in the first stage, where the encoder is fixed and the decoder is fine-tuned during training. We calculate it as in Equation (13).

$$\hat{\mathbf{x}} = D(\mathbf{m}) \tag{13}$$

3.3. Training

The model is trained with a combination of motion reconstruction losses and a Kullback–Leibler (KL) divergence term, as shown in Euations (14)–(16). Each loss component is designed to enforce consistency in different representation spaces while ensuring stable latent distribution learning.

First, we enforce reconstruction in the motion space using an ℓ_1 loss, as defined in Equation (14). This term measures the absolute difference between the generated motion $\hat{\mathbf{x}}$ and the ground-truth motion $\mathbf{x}$ across all frames and feature dimensions. Minimizing this loss ensures that the final generated motion sequence preserves both global motion dynamics and fine-grained joint-level details, leading to realistic and semantically accurate motion synthesis.

$$\mathcal{L}_{mot} = \|\hat{\mathbf{x}} - \mathbf{x}\|_1 \tag{14}$$

Second, we enforce reconstruction in the latent motion space, as defined in Equation (15). This loss minimizes the discrepancy between the predicted latent motion sequence $\hat{\mathbf{m}}$ and the target latent representation $\mathbf{m}$. By constraining the latent space directly, the model is encouraged to learn structured and temporally consistent latent trajectories, which improves the stability of the generation process and reduces error accumulation during sequential decoding.

$$\mathcal{L}_{mov} = \|\hat{\mathbf{m}} - \mathbf{m}\|_1 \tag{15}$$

Finally, we minimize the KL divergence between the posterior and prior distributions in the latent space, as shown in Equation (16). The posterior distribution $q(\mathbf{z} \mid \mathbf{x})$ is conditioned on both motion and text during training, while the prior distribution $p(\mathbf{z})$ is conditioned only on textual features. Minimizing this divergence aligns the two distributions, ensuring that the prior network can generate meaningful latent representations at inference time without access to ground-truth motion. This regularization is essential for stable training and enables the model to produce diverse yet coherent motion sequences.

$$\mathcal{L}_{kld} = D_{KL}\left(q(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})\right) \tag{16}$$

The overall training objective is defined as in Equation (17), where λ_1, λ_2, and λ_3 are weighting coefficients that balance the contributions of motion reconstruction, latent reconstruction, and distribution alignment, respectively. These coefficients allow the model to trade off between reconstruction accuracy and latent regularization. The loss weighting coefficients were empirically determined through preliminary validation experiments and training-stability analysis. Specifically, the coefficients were set to $\lambda_1 = 1$, $\lambda_2 = 1$, and $\lambda_3 = 0.005$ in the final model. This configuration achieved a favorable balance among motion reconstruction accuracy, latent reconstruction consistency, and latent distribution regularization, while ensuring stable optimization throughout training. Consequently, it provided the best overall trade-off between reconstruction quality and motion generation stability.

$$\mathcal{L}_{gen} = \lambda_1 \mathcal{L}_{mot} + \lambda_2 \mathcal{L}_{mov} + \lambda_3 \mathcal{L}_{kld} \tag{17}$$

By combining cross-modal attention with a latent prior–posterior formulation, the proposed model effectively aligns textual descriptions with motion generation. The attention mechanism dynamically extracts relevant semantic features from the input text at each time step, enabling fine-grained grounding of motion generation. Meanwhile, the variational structure regularizes the latent space, improves generalization, and allows the model to generate diverse motion sequences conditioned solely on textual input during inference. The latent representation dimension was set to 256, and both the prior and posterior networks were implemented using multi-layer GRU architectures with hidden dimensions of 512. The text encoder employed a bidirectional GRU using word embedding and parts-of-speech (POS) representations as input features. The cross-modal attention module was implemented using hidden-state-based attention between textual features and latent motion sequences. During training, the framework was optimized using the Adam optimizer with a learning rate of 2×10^{-4} and a batch size of 32. The model was trained for 300 epochs with gradient clipping and learning rate scheduling to improve optimization stability. The weighting coefficients λ_1, λ_2, and λ_3 were empirically selected to balance the motion reconstruction, latent reconstruction, and latent distribution regularization losses.

4. Experimental Results

4.1. Experimental Setup

The proposed framework was implemented using PyTorch 1.10 version and trained in a GPU-based environment on the Motion-X dataset. All motion samples were preprocessed following the procedure described in Section 3, including sequence normalization, temporal filtering, random cropping, and length alignment. To improve generalization, each motion sequence was paired with a selected caption during training. Optimization was performed using the Adam optimizer, and the model was trained for a fixed number of epochs until convergence. Gradient clipping was applied during training to improve stability. The training process was conducted in two stages. In the first stage, the motion representation module was trained to learn a compact and temporally stable latent space

from raw motion sequences. In the second stage, the text-conditioned generation module was trained in the learned latent space, where the motion encoder was kept fixed, and the decoder was fine-tuned jointly with the text-guided generation components.

For text processing, each caption was tokenized and represented using word embeddings together with parts-of-speech features. These textual features were encoded into contextual representations and used to guide motion latent generation through the proposed cross-modal attention and prior-posterior learning framework. All experiments were conducted under the same training and evaluation protocol to ensure fair and consistent analysis. To provide a clear summary of the experimental configuration, the training and evaluation settings are presented in Table 2.

Table 2. Training and Evaluation Environment.

Component	Setting/Configuration
Framework	PyTorch
Hardware	GPU-based environment (NVIDIA GPU)
Dataset	Motion-X
Motion Representation	623-dimensional SMPL-X feature
Sequence Length	40–200 frames
Preprocessing	Normalization, filtering, random cropping, alignment
Text Representation	Word embeddings + POS encoding
Text Encoder	Bi-GRU
Optimizer	Adam
Training Strategy	Two-stage training (Autoencoder → Generation)
Gradient Clipping	Applied
Training Protocol	Unified for all experiments
Evaluation Metrics	Matching, R-precision, FID, Diversity, Multimodality

Computational Complexity: All Motion-X experiments were conducted on a single NVIDIA RTX 4090 GPU (24 GB). The proposed method and the re-implemented baseline were trained and evaluated under the same Motion-X pre-processing pipeline, 52-joint full-body representation, and experimental protocol. To further analyze the computational characteristics of the proposed framework, we evaluated the storage cost, training cost, and evaluation-time cost under the Motion-X setting. Table 3 summarizes the computational requirements of the proposed framework. The decomposition checkpoint size is approximately 61 MB, while the baseline composition model and the proposed prior-refinement composition model require approximately 421 MB and 434 MB of storage, respectively. These results indicate that the proposed prior-refinement mechanism introduces only a small additional storage overhead compared with the overall model size.

Table 3. Computational complexity analysis of the proposed framework.

Component	Value
Decomposition checkpoint size	61 MB
Baseline composition model size	421 MB
Proposed composition model size	434 MB
GPU	NVIDIA RTX 4090 (24 GB)
Batch size	32
Decomposition training time	5.2 h
Composition training time	2.0 h
Text-motion matching training time	4 min
Total training time	7.3 h
Evaluation runtime (20 replications)	22 min
Runtime per replication	66 s

As shown in Table 3, the proposed method introduces only a modest computational overhead compared with the baseline model while providing improved text-motion alignment and motion quality. The complete evaluation process required approximately 22 min for 20 replications, corresponding to roughly 66 s per replication. The complete training pipeline required approximately 7.3 h on a single RTX 4090 GPU. Peak GPU memory usage was not explicitly logged during the original experiments; however, all training and evaluation runs were completed within the 24 GB memory budget using a batch size of 32. These results demonstrate that the proposed framework can be trained and evaluated on a single consumer-grade GPU while maintaining practical computational requirements for Motion-X-based full-body motion generation.

4.2. Dataset and Pre-Processing

We conduct our experiments on the Motion-X: A Large-Scale Dataset for Human Motion Generation dataset, which provides large-scale human motion sequences paired with natural language descriptions. Each motion sample is represented as a sequence of pose features and is associated with one or more textual captions describing the action. Each motion sequence is represented as a continuous time series $\mathbf{x} \in \mathbb{R}^{T \times D}$, where T denotes the number of frames and D represents the motion feature dimension based on joint representations. The motion features include root rotation and translation, joint positions and rotations, local joint velocities, and foot contact signals. In the Motion-X dataset, each motion sequence is represented using a structured skeletal model consisting of 52 joints, covering full-body motion and detailed hand articulation. Specifically, the skeleton includes a 22-joint body structure and 30 additional joints for both hands (15 joints per hand), enabling fine-grained modeling of finger movements. The hand joints are kinematically attached to the wrist joints, forming a unified full-body kinematic tree.

Each frame is represented by a 623-dimensional motion feature vector, which encodes both global motion and local joint dynamics. This representation includes root motion (4D), rotation-invariant local joint positions (153D), continuous 6D joint rotations (306D), local joint velocities (156D), and foot contact signals (4D). This structured representation allows the model to effectively capture both global motion patterns and fine-grained local movements. To ensure data quality, we apply several pre-processing steps. Motion sequences with fewer than 40 frames or longer than 200 frames are removed, and sequences containing invalid (non-finite) values are discarded. For samples with temporal annotations, we extract sub-sequences using the provided start and end time tags, resulting in a cleaned set of valid motion-text pairs. To stabilize training, we apply Z-normalization shown in Equation (18), where μ and σ denote the dataset mean and standard deviation, respectively. Different feature groups are scaled with separate factors to balance their influence, including root motion components, joint rotations, local velocities, and foot contact signals.

$$\hat{\mathbf{x}} = \frac{\mathbf{x} - \mu}{\sigma} \tag{18}$$

Since motion sequences have variable lengths, we apply random temporal cropping during training and normalize sequence lengths to a fixed maximum. Additionally, motion lengths are aligned to multiples of a predefined unit length to ensure consistency across batches while preserving motion diversity. Each motion is associated with one or more textual descriptions. We process text by tokenizing sentences into word-level tokens, adding special tokens (start-of-sequence and end-of-sequence), and padding or truncating to a fixed maximum length. Each token is then converted into a word embedding and a parts-of-speech (POS) one-hot representation. During training, each motion sequence is

paired with a randomly selected caption, and motion segments are randomly cropped to improve generalization. All inputs are normalized before being fed into the model.

4.3. Training Results

The training dynamics of the proposed Motion-X framework are illustrated in Figure 3, covering both the motion autoencoder and the full generation pipeline. Overall, the model exhibits stable and efficient optimization throughout training. All loss curves decrease rapidly during the initial phase, indicating fast convergence, and stabilize after approximately 150–200 epochs. The close alignment between training and validation curves, with only a minimal gap, demonstrates strong generalization and no evident overfitting. For the motion autoencoder, the reconstruction, smoothness, and sparsity losses converge smoothly without oscillation. These components ensure accurate motion reconstruction, temporal consistency, and compact latent representation, enabling effective modeling of both global body dynamics and fine-grained finger movements. In the generation pipeline, the total generation loss steadily decreases, while movement and motion reconstruction losses confirm efficient learning of both coarse and detailed motion patterns. The Kullback–Leibler Divergence (KLD) loss decreases gradually without collapse, indicating proper latent space regularization. Overall, the smooth and stable loss trajectories confirm the robustness and training efficiency of the proposed approach, even under the high-dimensional Motion-X representation.

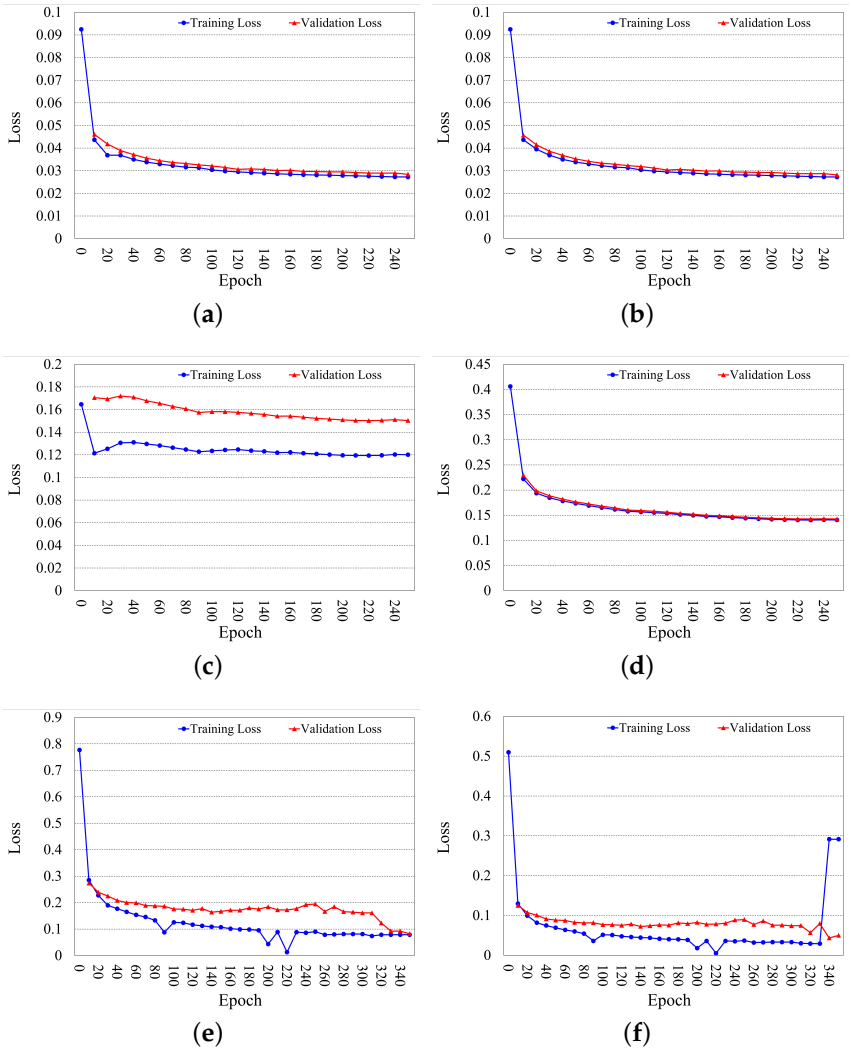


Figure 3. *Cont.*

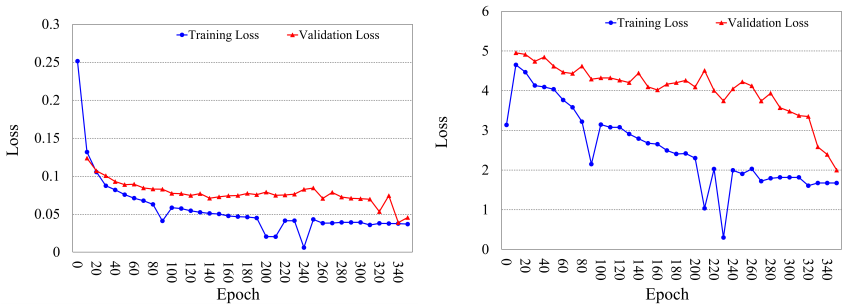


Figure 3. Training dynamics of the proposed model showing different loss components across epochs: (**a**) total loss, (**b**) reconstruction loss, (**c**) smooth loss, (**d**) sparsity loss of motion auto encoder, (**e**) total generation loss, (**f**) motion reconstruction loss, (**g**) movement reconstruction loss, and (**h**) KLD loss of training and validation performance.

4.4. Evaluation Metrics

The performance of the proposed method is evaluated using standard metrics commonly adopted in text-to-motion generation tasks, including Matching Score, R-precision, Fréchet Inception Distance (FID), Diversity, and Multimodality. The Matching Score measures the distance between text and motion embeddings obtained from a pretrained text–motion matching network. A lower Matching Score indicates better semantic alignment between the generated motion and the input text. The R-precision evaluates retrieval accuracy by checking whether the correct motion appears within the top-k retrieved results for a given text query. We reports, Top 1, Top 2, and Top 3 accuracy. Higher values indicate better text–motion correspondence. The Fréchet Inception Distance (FID) measures the similarity between the distribution of generated motions and real motions. A lower FID score indicates that the generated motions are more realistic and closer to the ground-truth distribution. The Diversity metric measures the variability among generated motion samples by computing the average distance between randomly sampled motion pairs. Higher diversity indicates that the model produces a wider range of motion patterns. The Multi-modality metric evaluates the variation of generated motions under the same text condition by measuring the average distance between multiple motions generated from the same input text. A higher value indicates that the model can generate multiple plausible motions for a single textual description. All metrics are computed by following the standard evaluation protocol to ensure consistency and comparability with prior work.

4.5. Quantitative Results

Table 4 presents the quantitative evaluation results of the proposed method on the Motion-X dataset. Compared with HumanML3D, Motion-X provides a significantly more challenging benchmark due to its high-dimensional full-body motion representations, including detailed hand and finger articulation. Since most existing text-to-motion generation methods are primarily evaluated on HumanML3D using different skeletal structures, motion representations, and evaluation protocols, direct numerical comparison is not fully aligned with our experimental setting. Therefore, we primarily focus on a fair in-dataset evaluation on Motion-X. In addition, direct experimental comparison with many recent text-to-motion generation frameworks remains challenging because most existing approaches are developed and evaluated under HumanML3D-based settings using different motion representations, skeletal structures, pre-processing strategies, and evaluation protocols. In contrast, the proposed framework is specifically designed for the Motion-X dataset, which introduces high-dimensional full-body motion together with detailed hand and finger articulation. Due to these substantial differences in dataset characteristics and

evaluation settings, direct numerical comparison may not provide a fully fair benchmark. Therefore, the proposed work primarily emphasizes controlled in-dataset evaluation under a unified Motion-X experimental setting while providing broader reference comparisons only for contextual analysis.

For fair comparison, we additionally implemented a baseline model based on the original text-to-motion generation framework [17]. The baseline follows the same overall encoder–decoder architecture and training protocol without the proposed latent-guided refinement and cross-modal enhancement modules. Both the baseline and the proposed model were trained under identical Motion-X pre-processing, training, and evaluation settings to ensure reproducibility and consistent assessment.

Table 4. Overall quantitative evaluation on the Motion-X dataset.

Method	Matching ↓	R@1 ↑	R@2 ↑	R@3 ↑	FID ↓	Diversity ↑	Multimodality ↑
Baseline (re-implemented)	0.6805 ± 0.0040	0.1826 ± 0.0031	0.3160 ± 0.0048	0.4275 ± 0.0053	0.0532 ± 0.0021	1.2804 ± 0.0082	0.2593 ± 0.0064
Ours (Motion-X)	$\mathbf{0.6210 \pm 0.0030}$	$\mathbf{0.2326 \pm 0.0039}$	$\mathbf{0.3941 \pm 0.0052}$	$\mathbf{0.5187 \pm 0.0054}$	$\mathbf{0.0164 \pm 0.0011}$	$\mathbf{1.3016 \pm 0.0076}$	$\mathbf{0.4141 \pm 0.0058}$

The baseline model was re-implemented and trained under the same Motion-X experimental setting to ensure a fair comparison. Therefore, we primarily focus on a fair in-dataset evaluation on Motion-X. The baseline model used in the results table corresponds to the re-implemented CompV6 text-to-motion generation framework [17], which follows the original encoder-decoder latent motion generation architecture without the proposed latent-guided refinement modules. In addition, direct experimental comparison with many recent text-to-motion generation frameworks remains challenging because most existing approaches are developed and evaluated under HumanML3D-based settings using different motion representations, skeletal structures, pre-processing strategies, and evaluation protocols.

To provide broader contextual analysis, we additionally discuss several representative text-to-motion generation approaches that were originally evaluated on HumanML3D. However, these comparisons should not be interpreted as direct benchmark evaluations because the compared methods were developed under different dataset settings, skeletal structures, motion representations, pre-processing strategies, and evaluation protocols. In contrast, the proposed framework is specifically designed and evaluated under the Motion-X setting, which introduces significantly more complex full-body motion together with detailed hand and finger articulation. Therefore, the HumanML3D-related discussions are included only as contextual reference, while the primary quantitative evaluation in this study is based on controlled in-dataset Motion-X experiments. To provide the most controlled and reproducible evaluation possible under the Motion-X setting, we re-implemented and trained the baseline framework using the same Motion-X pre-processing pipeline, motion representation, training protocol, evaluator network, and metric computation procedure as the proposed method. This experimental design isolates the contribution of the proposed latent-guided refinement strategy while minimizing confounding factors arising from differences in datasets, skeletal structures, feature representations, and evaluation settings. Although direct comparisons with additional recent Motion-X-based and whole-body motion generation methods would further strengthen the experimental evaluation, such comparisons are currently limited by the availability of publicly released implementations and standardized Motion-X evaluation protocols. Therefore, the current study emphasizes a controlled in-dataset comparison under identical experimental conditions to ensure fairness and reproducibility.

As shown in Table 4, the proposed framework consistently improves both semantic alignment and motion generation quality across all evaluation metrics. Specifically, the pro-

posed method improves the Matching Score from 0.6805 to 0.6210, indicating stronger alignment between the generated motion sequences and the input textual descriptions. The retrieval-based R-precision metrics are also significantly improved, with R@1 increasing from 0.1826 to 0.2326, R@2 increasing from 0.3160 to 0.3941, and R@3 increasing from 0.4275 to 0.5187. These improvements demonstrate that the generated motions more accurately preserve the semantic content of the input text. In terms of motion quality, the proposed framework achieves a lower FID score of 0.0164 compared to 0.0532 for the baseline, suggesting that the generated motion distribution is closer to the real Motion-X motion distribution. The diversity and multi-2modality scores are also improved, indicating that the model can generate more varied and semantically consistent motion patterns under the same textual condition. Overall, these results confirm that the proposed latent-guided framework effectively balances semantic alignment, motion realism, and motion diversity on the challenging Motion-X dataset. To provide broader context for the proposed framework, we additionally compare our method with several representative text-to-motion generation approaches in Table 5, including MoCoGAN [21], Dance2Music [22], Language2Pose [23], Text2Gesture [24], and Seq2Seq [25]. Since these methods were originally evaluated on different datasets, primarily HumanML3D, using different motion representations and evaluation protocols, the comparison is intended only to provide a general performance reference rather than a direct benchmark ranking. It should be noted that the numerical scales of evaluation metrics, particularly FID, can vary significantly across different datasets and evaluation pipelines. Most existing methods reported in Table 3 were evaluated under HumanML3D-based settings using different skeletal structures, motion feature dimensionalities, embedding spaces, pre-processing strategies, and evaluator networks. In contrast, the proposed framework is evaluated under the Motion-X full-body setting, which contains substantially higher-dimensional motion representations including detailed hand and finger articulation. These differences directly affect the feature distribution used for FID computation and therefore lead to different numerical scales across studies. To ensure fair evaluation within our work, both the baseline and the proposed framework were evaluated using the same Motion-X pre-processing pipeline, embedding space, feature extraction procedure, and metric computation protocol. For all experiments in this work, the same evaluator network and identical metric computation settings were consistently applied to both the baseline and the proposed framework to ensure reproducibility and fair comparison under the Motion-X setting.

Table 5. Reference comparison with representative text-to-motion generation methods reported under different experimental settings.

Method	R@1 ↑	R@2 ↑	R@3 ↑	FID ↓	Diversity ↑
MoCoGAN [21]	0.035 ± 0.002	0.069 ± 0.003	0.107 ± 0.002	94.381 ± 0.018	0.458 ± 0.004
Dance2Music [22]	0.029 ± 0.001	0.066 ± 0.003	0.102 ± 0.003	67.032 ± 0.011	0.732 ± 0.008
Language2Pose [23]	0.247 ± 0.003	0.385 ± 0.008	0.489 ± 0.005	10.021 ± 0.085	7.681 ± 0.062
Text2Gesture [24]	0.172 ± 0.003	0.271 ± 0.005	0.351 ± 0.003	7.672 ± 0.028	6.410 ± 0.068
Seq2Seq [25]	0.180 ± 0.002	0.300 ± 0.002	0.396 ± 0.002	11.750 ± 0.035	6.223 ± 0.061
Ours (Motion-X)	$\mathbf{0.2326 \pm 0.0039}$	$\mathbf{0.3941 \pm 0.0052}$	$\mathbf{0.5187 \pm 0.0054}$	$\mathbf{0.0164 \pm 0.0011}$	$\mathbf{1.3016 \pm 0.0076}$

Although the numerical scales are not directly comparable due to differences in datasets and evaluation protocols, the proposed framework demonstrates competitive retrieval-based performance while maintaining strong motion realism and semantic consistency. In particular, the proposed latent-guided framework achieves stable text–motion alignment and coherent full-body motion synthesis on the challenging Motion-X dataset, which includes detailed hand and finger articulation. The diversity score is lower compared with some existing approaches, which can be attributed to the stronger semantic conditioning and latent regularization applied in the proposed framework. By constraining the latent motion space through the prior–posterior formulation and cross-modal alignment,

the model prioritizes semantically accurate and temporally stable motion generation over excessive motion variation. Despite this trade-off, the proposed method maintains sufficient diversity while producing coherent and realistic full-body motion sequences. For additional reference, we also evaluated the configuration on the HumanML3D dataset. This setting achieved a Matching Score of 3.1539, R@1 of 0.4572, R@2 of 0.6429, R@3 of 0.7494, FID of 0.8779, and Diversity of 9.3503. However, these values are not directly compared with the Motion-X results because HumanML3D and Motion-X differ substantially in motion representation, dimensionality, skeletal structure, and evaluation distribution.

Overall, the quantitative results demonstrate that the proposed latent-guided framework effectively improves text–motion alignment while producing realistic, coherent, and semantically consistent full-body motion sequences on the challenging Motion-X dataset. To further analyze the robustness and stability of the proposed framework, we report the mean values with 95% confidence intervals over repeated evaluation runs under the same experimental setting. These repeated runs use the same trained checkpoint; therefore, the reported statistics reflect inference and evaluation stability rather than independent training-seed robustness. In addition, we performed Welch's t-test based on the repeated evaluation statistics reported to assess the statistical significance of the observed improvements. As shown in Table 6, all reported metrics achieve p-values (measures the probability that the observed improvement occurred by chance) below 0.001, indicating statistically significant differences between the proposed method and the re-implemented baseline under the same evaluation protocol. However, we acknowledge that repeated independent training runs with different random seeds would provide stronger evidence regarding training-level robustness. Due to the substantial computational cost associated with large-scale Motion-X full-body motion training, conducting multiple complete training runs was beyond the scope of the current study. Therefore, while the reported results demonstrate evaluation-level stability and statistical consistency, repeated multi-seed training experiments remain an important direction for future work.

Table 6. Statistical significance analysis based on repeated evaluation runs.

Metric	p-Value
Matching Score ↓	5.20×10^{-23}
R-Precision@1 ↑	6.68×10^{-21}
R-Precision@2 ↑	7.12×10^{-23}
R-Precision@3 ↑	2.59×10^{-24}
FID ↓	2.18×10^{-23}
Diversity ↑	6.52×10^{-4}
Multimodality ↑	2.23×10^{-30}

To qualitatively evaluate the proposed method, we compare the generated motions with the corresponding ground-truth sequences under the same textual descriptions. As shown in Figure 4, the generated motions effectively preserve the core semantic content of the input text, including action type, body posture, and temporal dynamics. The proposed framework successfully synthesizes text-consistent motions across diverse categories, such as walking actions, seated gestures, and sports-like movements. Compared with the ground-truth sequences, the generated motions exhibit similar global movement patterns and smooth temporal transitions, indicating strong text–motion alignment. The model is also capable of capturing fine-grained motion characteristics, including upper-body coordination and articulated hand movements, demonstrating its effectiveness in handling complex full-body motion structures. Additional qualitative analysis demonstrates that the proposed framework performs reasonably well across various locomotion, standing gesture,

and sports/dynamic motion categories on Motion-X. We demonstrated these in Table 7. Nevertheless, several challenging scenarios remain for highly detailed hand and finger articulation under complex textual prompts. For motions involving subtle finger-level interactions, such as finger scrolling, finger pressing, crossing fingers, or snapping gestures, the generated motions generally preserve the overall hand and arm movement patterns, although some fine-grained articulations may not yet be fully reproduced. Similarly, long textual prompts containing multiple sequential or simultaneous actions may occasionally lead to simplified motion generation for certain detailed movements. These observations suggest that additional fine-tuning and refinement strategies focused on detailed hand articulation and complex sequential motion modeling could further improve the overall realism and generation quality of the proposed framework.

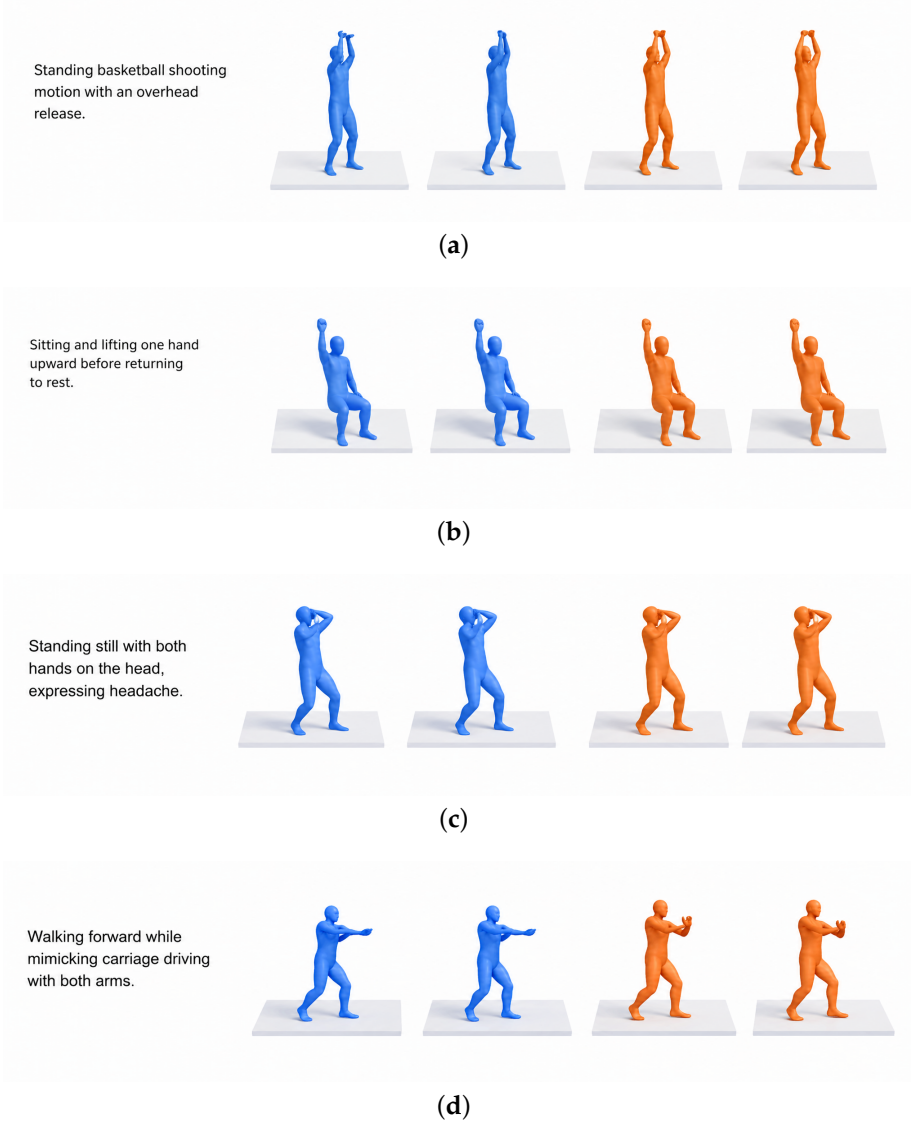


Figure 4. Qualitative comparison between ground-truth (GT) and generated motions under different text descriptions. The blue motion sequences represent the ground-truth motions, while the orange motion sequences denote the motions generated by the proposed framework. The results demonstrate that the proposed method effectively captures textual semantics and generates realistic full-body motions with coherent temporal dynamics, stable global structures, and fine-grained articulations.

Table 7. Category-wise evaluation of our experiment on the Motion-X dataset.

Category	Samples	Top 1 ↑	Top 2 ↑	Top 3 ↑	FID ↓	Diversity ↑	Multimodality ↑
Locomotion	173	0.1272	0.2081	0.3006	0.1251	1.1048	0.7383
Sitting actions	304	0.1382	0.1875	0.2533	0.0744	0.9011	0.5130
Standing gestures	284	0.1056	0.2430	0.3028	0.1320	1.1054	0.5492
Sports/dynamic	78	0.1795	0.2821	0.4359	0.0851	1.1713	0.6742

Overall, the qualitative comparisons demonstrate that the proposed framework can generate realistic, coherent, and semantically consistent human motions across a wide range of action categories while preserving both global motion dynamics and detailed articulated movements.

4.6. Category-Wise Analysis

Table 7 presents the performance of the proposed model across different motion categories. The model achieves the strongest semantic alignment on sports/dynamic motions, obtaining the highest R-precision scores across all levels. This indicates that motions with clear and distinctive global dynamics are effectively captured by the proposed framework. In terms of realism, sitting actions achieve the lowest FID score, suggesting that the generated motions closely follow the real motion distribution for relatively stable and structured actions. For locomotion, the model demonstrates the highest multi-modality, indicating its strong ability to generate diverse motion patterns for walking-related actions. Standing gestures also show consistent performance across metrics, demonstrating the model's capability to handle fine-grained and upper-body-dominant movements. Overall, the results indicate that the proposed method performs effectively across diverse motion categories, capturing both global motion dynamics and detailed articulation.

We further analyze the relationship between semantic consistency and motion generation quality in the proposed model. As shown in the quantitative results, the method achieves strong performance in retrieval-based metrics, particularly R-precision, indicating effective alignment between generated motions and input textual descriptions. At the same time, the model maintains competitive performance across generation quality metrics, including FID, Diversity, and Multi-modality. This behavior suggests that strong text conditioning guides the model toward semantically consistent motion generation while preserving a reasonable level of motion realism and variation. The structured latent representation, combined with the prior–posterior framework, enables the model to balance alignment and motion quality in a unified manner. In addition, the category-wise results demonstrate that the model effectively adapts to different motion types. Motions with clear global dynamics, such as sports and locomotion, show strong semantic alignment, while more structured actions, such as sitting, exhibit stable and realistic motion patterns. This indicates that the proposed framework is capable of handling both global motion structures and fine-grained articulation.

Overall, the results highlight the effectiveness of the proposed latent-guided framework in jointly modeling semantic alignment and motion generation quality, while maintaining stable performance across diverse motion categories. To further understand the behavior of the proposed method, we analyze representative challenging cases. The model consistently captures the overall action semantics and global motion structure across various scenarios. In some cases, fine-grained details such as subtle hand articulations and motion amplitude variations may be less pronounced, particularly for complex or interaction-heavy actions. Additionally, for longer and more complex text descriptions, the model maintains the main semantic structure while minor variations in temporal consistency can be observed. These observations are consistent with the quantitative results,

where strong semantic alignment is achieved alongside stable motion generation performance. Overall, the analysis indicates that the proposed method effectively models both global dynamics and semantic alignment, while future improvements may further enhance fine-grained articulation and long-range temporal consistency.

4.7. Failure-Case and Baseline Qualitative Analysis

To provide a more balanced qualitative evaluation, we further visualize challenging Motion-X examples and compare the proposed method with the re-implemented baseline. Figure 5 presents several representative failure cases involving fine-grained hand-centric motions, including finger scrolling, hand-hand interaction, and finger snapping actions. These examples were selected because they require subtle local articulation and precise contact relationships, which remain difficult for current text-to-motion generation frameworks. Compared with the baseline, the proposed method generally preserves the intended global body configuration and semantic motion structure more consistently. For example, in the finger scrolling example, both methods maintain the standing posture; however, the proposed method exhibits more stable arm placement and reduced body drift. Similarly, in the hand-hand interaction example, the proposed framework generates hand trajectories that remain closer to the intended interaction region, although precise contact is not always maintained. In the seated finger-snapping example, the overall seated posture and upper-body motion are preserved, but the detailed snapping articulation remains under-expressed.

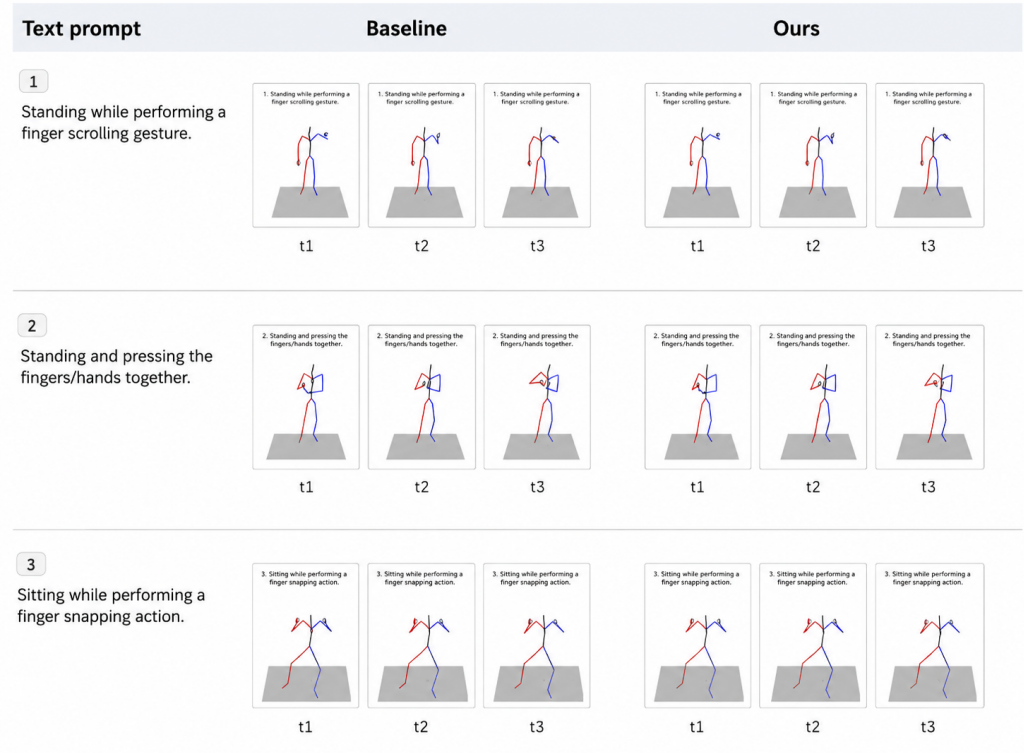


Figure 5. Qualitative comparison of the baseline and proposed framework on challenging Motion-X hand-centric motions. Examples include finger scrolling, hand-hand interaction, and finger snapping. Representative frames (t_1–t_3) are shown for each motion sequence.

These observations suggest that the proposed latent-guided framework effectively improves global text-motion alignment and full-body motion realism while retaining coherent motion dynamics. Nevertheless, several limitations remain. In particular, fine-grained finger articulation, precise hand-hand contact, and subtle hand-centric interactions

are still difficult to reproduce accurately. This limitation is likely related to the complexity of detailed hand motion modeling and the limited representation of contact-aware interactions in current motion generation pipelines. Overall, the qualitative comparisons indicate that the proposed framework achieves more semantically consistent and stable motion generation than the baseline, while also highlighting important challenges that remain for future research. Improving detailed finger articulation, contact-aware motion generation, and fine-grained hand-object interaction modeling represents a promising direction for future work.

4.8. Ablation Study

To evaluate the effectiveness of the proposed components, we conducted an ablation study by progressively removing key modules from the framework. Specifically, we compare the baseline model, the model without the text attention module, the model without the prior refiner, and the complete framework. The training dynamics of each ablation setting are illustrated in Figure 6, while the quantitative evaluation results are summarized in Table 8.

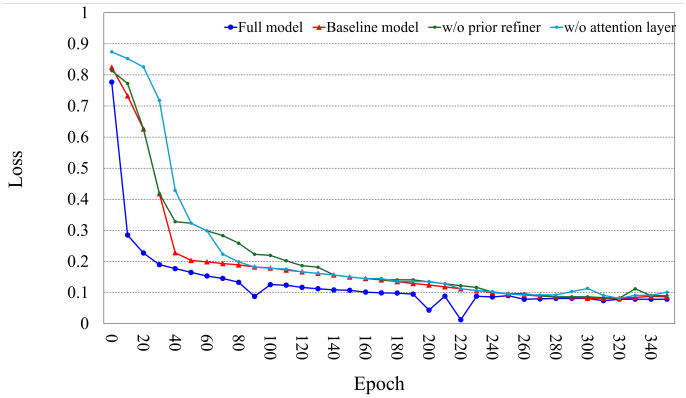


Figure 6. Training loss curves for the ablation study. The proposed full model achieves faster convergence and lower final loss compared with the baseline and ablated variants.

Table 8. Ablation study of the proposed framework on the Motion-X dataset.

Method	Matching ↓	R@1 ↑	R@2 ↑	R@3 ↑	FID ↓	Diversity ↑	Multimodality ↑
Baseline (re-implemented)	0.6805	0.1826	0.3160	0.4275	0.0532	1.2804	0.2593
W/O text attention	0.6310	0.2003	0.3420	0.4565	0.0567	1.2912	0.2468
W/O prior refiner	0.6285	0.2203	0.3820	0.4965	0.0497	1.3016	0.2347
Full model	**0.6210**	**0.2369**	**0.3848**	**0.5046**	**0.0489**	**1.2795**	**0.2477**

Each proposed component contributes to the overall performance improvement of the framework. The baseline model produces the weakest performance across most retrieval-based metrics, indicating limited semantic alignment between text and generated motion sequences. Removing the text attention module leads to a noticeable reduction in R-precision and overall motion quality. This demonstrates that the attention mechanism plays an important role in capturing fine-grained semantic relationships between textual descriptions and motion representations. Without attention-guided alignment, the generated motions become less semantically consistent with the input text. Similarly, removing the prior refiner degrades retrieval performance and slightly increases the FID score. This suggests that the prior refiner contributes to latent motion stabilization and improves the consistency of the generated motion distribution. The prior refinement mechanism helps regularize the latent space, leading to smoother and more coherent motion generation.

Among all configurations, the full model achieves the best overall balance between semantic alignment and motion realism. In particular, the proposed framework achieves the highest R@1, R@2, and R@3 scores while maintaining competitive FID and diversity values. These results indicate that the integration of text attention and prior refinement jointly improves the quality, coherence, and semantic consistency of generated full-body motions.

To further investigate the contribution of each optimization component, we analyzed the relative influence of the loss terms used in the proposed framework. As summarized in Table 9, reconstruction-related objectives dominate the optimization process, whereas sparsity regularization, smoothness regularization, KL-divergence, and prior refinement act as auxiliary constraints that enhance latent stability and semantic consistency. The sparsity and smoothness losses account for only 0.07% and 0.04% of the decomposition objective, respectively, indicating that they function as lightweight regularizers. In the composition stage, the latent reconstruction and KL-divergence losses contribute 20.14% and 10.60%, respectively, demonstrating their importance in maintaining latent consistency and distribution regularization. Overall, these objectives play complementary roles in improving the robustness and motion generation quality of the proposed latent-guided framework on the Motion-X dataset.

Table 9. Contribution analysis of optimization objectives in the proposed Motion-X framework.

Stage	Loss Component	Weighted Contribution	Share of Total Loss
Decomposition	Reconstruction loss	2.102228	99.89%
Decomposition	Sparsity loss	0.001488	0.07%
Decomposition	Smoothness loss	0.000790	0.04%
Composition	Motion reconstruction loss	0.109552	69.18%
Composition	Latent reconstruction loss	0.031892	20.14%
Composition	KL divergence loss	0.016780	10.60%
Composition	Prior refinement loss	0.000123	0.08%

The training curves shown in Figure 6 further support these observations. The proposed full model converges faster and achieves lower final loss values compared with the baseline and ablated variants. In contrast, the baseline model exhibits slower convergence and higher training loss, indicating reduced optimization stability. Overall, the ablation study confirms that each component of the proposed framework contributes positively to robust text-to-motion generation on the Motion-X dataset.

Although the current ablation study evaluates the contributions of the text attention module and prior refinement module, a more comprehensive component-wise analysis involving all architectural modules and optimization objectives was not conducted due to the substantial computational cost associated with large-scale Motion-X training and evaluation. Future work will investigate more extensive ablation studies to further quantify the contribution of individual components and loss terms.

5. Conclusions

In this paper, we presented a latent-guided text-to-motion generation framework for human motion synthesis. The proposed method improves text–motion consistency by learning a structured motion representation that effectively captures semantic information while preserving realistic motion dynamics. Experimental results on the Motion-X dataset demonstrate that the model achieves strong semantic alignment, particularly in retrieval-based metrics such as R-precision, while maintaining competitive motion realism and diversity. Qualitative results further confirm that the generated motions preserve key action semantics, body posture, and temporal structure across diverse motion categories. Through detailed analysis, we observe that the proposed framework effectively handles motions with

both global dynamics and articulated structures, demonstrating its capability in modeling complex full-body motion. In future work, we aim to further improve fine-grained motion modeling, particularly for subtle hand and upper-body articulation, and enhance long-range temporal consistency for extended motion sequences. Additionally, we plan to extend the evaluation with broader baseline comparisons and human perceptual studies to provide more comprehensive validation.

Author Contributions: J.N. conducted the conceptualization, methodology development, software implementation, experiments, validation, formal analysis, data curation, visualization, and preparation of the original manuscript draft. H.-B.L. contributed to validation, manuscript review and technical discussions. Y.-H.S. supervised the research, provided project administration and funding support, and contributed to manuscript review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was also supported by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency (KOCCA), funded by the Ministry of Culture, Sports and Tourism (MCST), Republic of Korea, in 2025 (Project Name: Development of Digital Twin and Simulation Technology for Optimization of Safety Management in Mobile Performance Hall; Project Number: RS-2025-02221168; Contribution Rate: 100%).

Data Availability Statement: The datasets generated and analysed during the current study are not publicly available due to copyright restrictions on the original sources. However, they are available from the corresponding author upon reasonable request.

Acknowledgments: This research was supported by the Excellent Researcher Support Project of Kwangwoon University in 2024.

Conflicts of Interest: Author Young-Ho Seo was employed by the company Omotion Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Petrovich, M.; Black, M.J.; Varol, G. TEMOS: Generating diverse human motions from textual descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2022.
2. Zhong, C.; Hu, L.; Zhang, Z.; Xia, S. AttT2M: Text-Driven Human Motion Generation with Multi-Perspective Attention Mechanism. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2023; pp. 509–519.
3. Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; Chen, T. MotionGPT: Human Motion as a Foreign Language. *arXiv* **2023**, arXiv:2306.14795. [CrossRef]
4. Wang, R.; He, Y.; Sun, T.; Li, X.; Shi, T. UniTMGE: Uniform Text-Motion Generation and Editing Model via Diffusion. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: Piscataway, NJ, USA, 2025; pp. 6104–6114. [CrossRef]
5. Chen, W.; Xiao, H.; Zhang, E.; Hu, L.; Wang, L.; Liu, M.; Chen, C. SATO: Stable Text-to-Motion Framework. *arXiv* **2024**. [CrossRef]
6. Petrovich, M.; Black, M.J.; Varol, G. TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2023.
7. Zhang, Y.; Lin, J.; Zeng, A.; Wu, G.; Lu, S.; Fu, Y.; Cai, Y.; Zhang, R.; Wang, H.; Zhang, L. Motion-X++: A Large-Scale Multimodal 3D Whole-body Human Motion Dataset. *arXiv* **2025**, arXiv:2501.05098.
8. Lin, J.; Zeng, A.; Lu, S.; Cai, Y.; Zhang, R.; Wang, H.; Zhang, L. Motion-X: A Large-scale 3D Expressive Whole-body Human Motion Dataset. *Adv. Neural Inf. Process. Syst.* **2023**, 36. [CrossRef]
9. Athanasiou, N.; Petrovich, M.; Black, M.J.; Varol, G. TEACH: Temporal Action Compositions for 3D Humans. In *2022 International Conference on 3D Vision (3DV)*; IEEE: Piscataway, NJ, USA, 2022.
10. Zhang, J.; Zhang, Y.; Cun, X.; Huang, S.; Zhang, Y.; Zhao, H.; Lu, H.; Shen, X. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2023.
11. Tevet, G.; Raab, S.; Cohan, S.; Reda, D.; Luo, Z.; Peng, X.B.; Bermano, A.H.; van de Panne, M. CLoSD: Closing the Loop between Simulation and Diffusion for multi-task character control. *arXiv* **2024**, arXiv:2410.03441.

12. Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-or, D.; Bermano, A.H. Human Motion Diffusion Model. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
13. Bensabath, L.; Petrovich, M.; Varol, G. TMR++: A Cross-Dataset Study for Text-based 3D Human Motion Retrieval. *arXiv* **2024**, arXiv:2405.16909.
14. Yin, K.; Zou, S.; Ge, Y.; Tian, Z. Tri-Modal Motion Retrieval by Learning a Joint Embedding Space. *arXiv* **2024**. [CrossRef]
15. Lu, S.; Chen, L.H.; Zeng, A.; Lin, J.; Zhang, R.; Zhang, L.; Shum, H.Y. HumanTOMATO: Text-aligned Whole-body Motion Generation. *arXiv* **2023**, arXiv:2310.12978.
16. Petrovich, M.; Litany, O.; Iqbal, U.; Black, M.J.; Varol, G.; Peng, X.B.; Rempe, D. Multi-Track Timeline Control for Text-Driven 3D Human Motion Generation. *arXiv* **2024**. [CrossRef]
17. Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; Cheng, L. Generating Diverse and Natural 3D Human Motions from Text. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2022 ; pp. 5152–5161.
18. Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A.A.A.; Tzionas, D.; Black, M.J. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2019.
19. Zhang, S.; Xue, Y.; Zhang, Y.; Wu, X.; Tuan, L.A.; Zhao, C. MRAG: A Modular Retrieval Framework for Time-Sensitive Question Answering. *arXiv* **2025**. [CrossRef]
20. Zhang, M.; Guo, X.; Pan, L.; Cai, Z.; Hong, F.; Li, H.; Yang, L.; Liu, Z. ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model. *arXiv* **2023**. [CrossRef]
21. Tulyakov, S.; Liu, M.Y.; Yang, X.; Kautz, J. MoCoGAN: Decomposing Motion and Content for Video Generation. *arXiv* **2017**. [CrossRef]
22. Aggarwal, G.; Parikh, D. Dance2Music: Automatic Dance-driven Music Generation. *arXiv* **2021**. [CrossRef]
23. Ahuja, C.; Morency, L.P. Language2Pose: Natural Language Grounded Pose Forecasting. *arXiv* **2019**. [CrossRef]
24. Asakawa, E.; Kaneko, N.; Hasegawa, D.; Shirakawa, S. Evaluation of text-to-gesture generation model using convolutional neural network. *Neural Netw.* **2022**, *151*, 365–375. [CrossRef] [PubMed]
25. Lin, A.S.; Wu, L.; Corona, R.; Tai, K.W.H.; Huang, Q.X.; Mooney, R.J. Generating Animated Videos of Human Activities from Natural Language Descriptions. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, QC, Canada, 3–8 December 2018.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.