A Multimodal Large Language Model Framework for Gesture Generation in Social Robots

Le Thien Doanh[1,2], Nguyen Tan Viet Tuyen[3], Le Duy Tan[1,2] and Sarvapali D. Ramchurn[3]

[1]Vietnam National University Ho Chi Minh City, Vietnam
[2]School of Computer Science and Engineering, Vietnam International University Ho Chi Minh City, Vietnam
[3]School of Electronics and Computer Science, University of Southampton, United Kingdom

Abstract

Non-verbal gestures play a crucial role in social robots, enabling them to signal their intentions to users during human–robot interaction (HRI). While recent research in this domain has primarily focused on robot gesture generation, there remains a limited number of studies on multimodal generation frameworks, where generated gestures are harmonized with other generated modalities, to better convey the robot's intention to users via a wider range of communication channels. Inspired by recent advancements in multimodal large language models (MLLMs), we propose a novel framework that integrates motion generation models with existing MLLMs to produce high-quality 3D motions without the need for extensive multimodal training. Our framework comprises three key components: a Denoising Diffusion Motion Generation (DDMG) module that maps text descriptions to motion sequences using a diffusion-based approach; a Motion Decoding Alignment (MDA) module that refines motion representations by incorporating signal embeddings generated by an LLM; and a Fusion Module (FM) that integrates motion features trained from previous phases to enhance coherence and realism. We conducted a series of experiments on a publicly available dataset to evaluate the efficiency of the proposed framework in terms of motion quality, diversity, and semantic alignment. The results suggest that our multimodal approach can serve as a powerful controller for robot gesture generation, offering a more scalable and effective solution, particularly for social HRI.

Keywords

Non-verbal Gestures, Social Robots, Human-Robot Interaction, Large Language Model

1. Introduction

Non-verbal gestures are essential for social robots to signal their intentions to users during human–robot interaction (HRI). By integrating gestures with other visual and auditory modalities (e.g., speech, images, video), robots can communicate with users more effectively through a wider range of channels. Multimodal communication has been shown to be a crucial strategy for social robots, as users can better perceive the robot's communicative messages through complementary modalities, towards enhancing their understanding of the robot behaviors [1]. In various situations, such as service environments with background noise [2] or elderly care settings where users may have sensory impairments [3], multimodal communication has become a key for social robots to deliver messages to users effectively. However, there remains a limited number of studies on multimodal generation frameworks in which gesture features are harmonized with other modalities to effectively convey intentions through diverse communication channels. One possible reason is that training a powerful model capable of handling multiple modalities is highly resource-intensive. Inspired by recent advancements in multimodal large language models (MLLMs), this paper contributes a novel framework that integrates a diffusion-base motion generation model with an MLLM to address multimodal generation tasks, particularly for social HRI. These models leverage strong encoders trained on large-scale datasets to process multiple modalities without the need for additional training. By employing an MLLM as a reasoning controller and integrating it with a robust motion generation model, our approach aims to reduce the reliance on extensive multimodal datasets while maintaining high-quality gesture generation.

BEAR 2025 - Workshop on Benefits of pErsonalization and behAvioral adaptation in assistive Robots, within IEEE RO-MAN 2025, August 25-29 2025, Eindhoven, The Netherlands
ITCSIU22237@student.hcmiu.edu.vn (L. T. Doanh); tuyen.nguyen@soton.ac.uk (N. T. V. Tuyen); ldtan@hcmiu.edu.vn (L. D. Tan); sdr1@soton.ac.uk (S. D. Ramchurn)


The rest of this paper is organized as follows. In section 2, we review studies on human-inspired gesture generation framework for social robots and virtual agents. Then, we highlight recent works on MLLMs and their limitations in the context of motion generation for social robots. In section 3, the proposed approach consisting of three training phases will be explained in detail. In section 4, we evaluate the efficiency of the proposed framework using a wide range of evaluation metrics and under different configurations. As a proof of concept, we demonstrate how the designed framework on a social robot. Finally, the experimental results and future work are summarized in section 5.

2. Related Works

2.1. Non-verbal Gesture Generation in Social Robots

Studies on generating robots' non-verbal gestures are categorized into two groups: rule-based approach and data-driven approach. However, rule-based approaches often struggle to adapt to dynamic interactions, motivating researchers to explore data-driven methods for more natural and realistic motion generation. Regression-based models has been used in previous studies to generate human motion by encoding input features. For example, the Joint Language-to-Pose (JL2P) model learns a shared embedding space for language and motion, facilitating natural language-driven animations [4]. However, these models often struggle with capturing long-range dependencies, leading to overly smooth and less realistic predictions. Generative frameworks such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion have been used in recent studies, enabling more diverse and realistic sequences of the generated motion. Previous GAN-based approach, such as ActFormer, integrate Transformers for temporal and interaction modeling in motion generation [5]. VAEs has been applied in [6] to improve generative modeling by leveraging latent variables, as seen in text-to-motion applications [7]. MoMask extends VAEs with hierarchical quantization for high-quality motion synthesis [8]. Normalizing flows [9] construct complex distributions through invertible transformations, enhancing the realism of the motion. Recently, Denoising Diffusion Probabilistic Models (DDPMs) have demonstrated state-of-the-art performance. MotionDiffuse [10], [11] use the diffusion model to generate high quality of motion based on the contextual. The study [12] extends diffusion models to open-vocabulary motion synthesis, enabling zero-shot generation from text.

While recent data-driven frameworks have extensively addressed motion generation task, enabling robots or virtual agents to convey their intention and the verbal content of their speech through synthesized gestures, there remains a limited number studies on multimodal frameworks, particularly inspired from recent advancement of MLLM. In such frameworks, the motion modality could be harmonized with others in the latent space to produce coherent multimodal communication outputs. Notably, current motion generation approaches mostly rely on text-based condition [4, 11, 13], which limits their ability to fully interpret and align with the tokens representations provided by MLLM. In the context of social HRI, however, non-verbal communication extends beyond robot gestures. The ability to engage users through other communication channels, such as audio, images, or other visual cues, is essential for enhancing interaction across different HRI scenarios [2, 3].

2.2. Multimodal Large Language Models

Recent advancements in Multimodal Large Language Models (MLLMs) have significantly enhanced the integration of diverse modalities. Mini-Gemini [14] employs a dual-encoder approach, notably improving the interpretation of high-resolution visuals. Similarly, the work in [15] introduces a flexible vision encoding technique, optimizing the processing of high-resolution images. In addition, OneLLM [16] integrates diverse data types using universal projection modules, improving scalability. NExT-GPT [17] extends LLM capabilities to handle a broader range of modalities, including text, image, video, and audio. However, those solutions typically rely on domain-specific training, resulting in substantial computational costs and limited generalization capabilities. Importantly, existing models in this domain

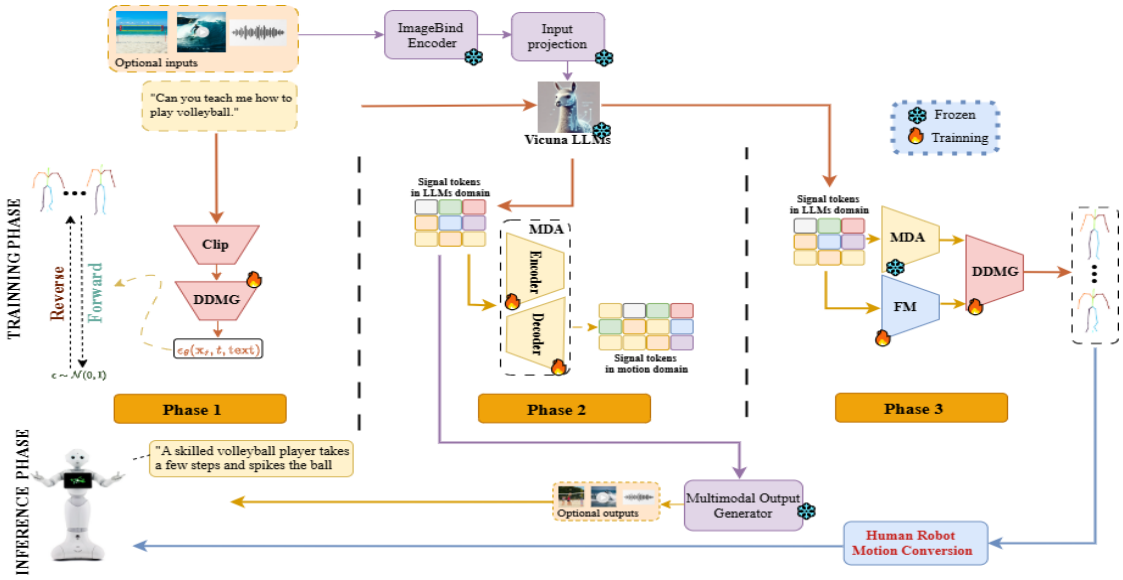


Figure 1: The proposed framework consists of three training phases. Phase 1 uses the Denoising Diffusion Motion Generation module (DDMG) to map textual descriptions into initial motion representations. Phase 2 employs the Motion Decoding Alignment module(MDA) module to refine motion representations within the motion domain. Phase 3 combines the outputs of MDA with the Fusion module (FM) to optimal the accuracy of motion sequence output.

primarily focus on audio, visual, and textual modalities, often neglecting motion modality, an essential channel for robots to convey semantic contents of their speech during HRI.

This paper addresses this gap by extending the multimodal LLM framework introduced in [17] to incorporate motion as an additional modality, alongside audio, image, and video, making it more practical for HRI settings. The proposed unified multimodal framework is capable of synthesizing high-quality motion without requiring extensive domain-specific training. In this paper, we utilize Vicuna-7B-v0 to effectively synthesize textual descriptions and bridge text-based prompts with motion generation. Our approach aims to advance the deployment of MLLM in the context of HRI where the synthesis of multimodality communication, particularly the nonverbal gesture channel, is the key for the robot to interact with users efficiently.

3. Methodology

Overview of the Proposed Framework

Fig. 1 illustrates an overview of the proposed framework, including three Phases. In 3.1, we illustrate the Denoising Diffusion Motion Generation module (DDMG), the core mechanism for generating motion sequences, trained during Phase 1. In 3.2, we introduce the Motion Decoding Alignment module (MDA) trained in Phase 2. In 3.3, we explain our novel refinement method for motion diffusion, trained in the final Phase with fusion module (FM).

3.1. Phase 1: Motion Diffusion Generation Module

The **Motion Diffusion Generation Module (DDMG)** plays a pivotal role in the framework, as outlined in the first phase of the training process. In Phase 1, the module is trained using a traditional text-to-motion model to establish basic text-guided motion generation. The model is trained with a forward diffusion process, where a Markov chain adds Gaussian noise to the motion data over several steps, and the sequence of noisy samples is generated according to a variance schedule. This forward process is modeled using the equation:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{1}$$

where $\mathbf{x}_0$ represents the original motion sequence, and $\bar{\alpha}_t$ is the cumulative product of $\alpha_t = 1 - \beta_t$, the variance schedule. To reconstruct the motion sequence, the reverse motion diffusion process is used, where the noise addition is reversed. The reverse process is modeled as a Gaussian distribution conditioned on the text embedding. The mean for this process is calculated as:

$$\mu_\theta(\mathbf{x}_t, t, \mathbf{text}) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t, \mathbf{text}) \right), \tag{2}$$

where $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{text})$ is the predicted noise component. The model is trained to minimize the mean squared error (MSE) loss, which ensures the accuracy of the noise prediction:

$$L = \mathbb{E}_{t, \mathbf{x}_0, \boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{x}_t, t, \mathbf{text})\|^2 \right], \tag{3}$$

where the true noise $\boldsymbol{\epsilon}$ is sampled from a Gaussian distribution. The pretrained weights from Phase 1 are then utilized in Phase 3, where they are integrated into the overall framework, enabling coherent motion synthesis throughout the system.

To implement the reverse denoising function $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{text})$, we adopt a Transformer-based architecture composed of three core components: a Self-Attention (S-Attention) layer for modeling intra-motion temporal dependencies, a Cross-Attention (C-Attention) mechanism for conditioning on text or signal embeddings, and a Feed-Forward Network (FFN) for non-linear transformation.

3.2. Phase 2: Motion Decoding Alignment Module

The MLLM in our framework, as shown in Fig.1, generates multiple outputs, including a text response and signal tokens containing motion-related information. These tokens are passed to the transformer-based MDA, which models temporal dependencies and refines motion sequence representations. The Learning Queries Output Projection component adapts the multimodal LLM output into a format suitable for motion generation, ensuring coherent responses across text, image, and motion modalities.

Unlike traditional motion diffusion models, which rely on text-based conditioning, our approach uses signal tokens as conditional inputs. These tokens pass through transformer-based projection layers to better align with LLM outputs, resulting in more coherent motion sequences.

To ensure alignment between signal tokens and CLIP text embeddings, we introduce a loss function combining consistency loss ($\mathcal{L}_\text{c}$) and discrepancy loss ($\mathcal{L}_\text{d}$). The total alignment loss is given by:

$$\mathcal{L}_\text{align} = \alpha \cdot \mathcal{L}_\text{c} + \beta \cdot \mathcal{L}_\text{d} \tag{4}$$

$$\mathcal{L}_\text{c} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{a}_i - \mathbf{b}_i\|^2 \quad \mathcal{L}_\text{d} = 1 + \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{a}_i \cdot \mathbf{b}_i}{\|\mathbf{a}_i\| \|\mathbf{b}_i\|} \tag{5}$$

This combination ensures stable training, better semantic alignment, and improved motion generation quality. The aligned signal tokens guide the motion synthesis process in the diffusion model.

3.3. Phase 3: Fusion Module

Directly pushing the signal token embedding from Phase 2 to the DDMG module resulted in suboptimal motion quality. To address this, we designed a Fusion module (FM) integrated into the diffusion pipeline, which enhances both fusion quality and motion performance. The FM acts as a controller for the function $\epsilon_\theta(\mathbf{x}_t, t, \text{text})$, integrating diverse inputs to produce tailored motion sequences. Fig. 2 illustrates the Phase 3 pipeline, which trains the FM and DDMG components. The FM integrates the signal token embedding with other motion-related features, capturing semantic and temporal relations for improved motion representations. The pre-trained DDMG from Phase 1 is trained alongside the FM module.

Our FM, based on the MoE architecture [18], enables dynamic feature fusion and adaptive learning. The Gate function dynamically assigns weights to different experts based on input data. The top-k

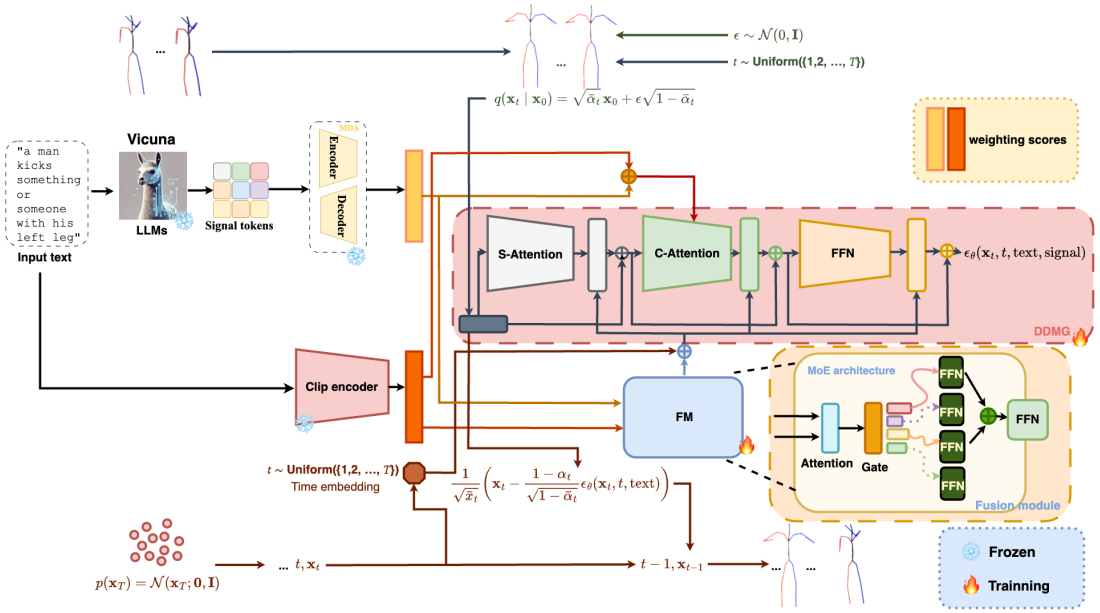


Figure 2: Overall framework of the pipeline on Phase 3. The input text is processed by a LLM to generate a semantic signal. The Clip encoder processes the input to align motion features. The MDA module bridges textual and motion embeddings. The DDMG module, consisting of self-attention (S-Attention), cross-attention (C-Attention) mechanisms along with Feed-Forward Networks (FFNs), utilizes a probabilistic diffusion process to model motion data. FM integrates text and motion features via attention gating and outputs sequential motion frames, which reverse the diffusion process to produce a synthesis motion. Time embeddings are incorporated to encode temporal dynamics of the motion.

gating mechanism selects the most relevant experts via a similarity measure. The gating function is defined as Eq. 6 with softmax normalization $\mathbf{w}_{\text{gating}} = \text{softmax}(\mathbf{g}_{\text{top-k}})$.

$$\mathbf{g} = -\|\mathbf{W}_g - \mathbf{E}_{\text{concat}}\|_2, \quad \mathbf{g}_{\text{top-k}}, \text{indices} = \text{top-k}(\mathbf{g}, k) \tag{6}$$

The final aggregated output followed by an output projection is presented in Eq. 7 where $\mathbf{W}_{\text{out}}$ and $\mathbf{b}_{\text{out}}$ are trainable projection parameters.

$$\mathbf{F}_{\text{agg}} = \sum_{i=1}^{k} w_{\text{gating},i} \cdot \mathbf{E}_i, \quad \mathbf{F}_{\text{final}} = \mathbf{W}_{\text{out}}\mathbf{F}_{\text{agg}} + \mathbf{b}_{\text{out}} \tag{7}$$

To enforce consistency between signal and text embeddings, we incorporate L2 regularization as below. This regularization aligns multimodal feature distributions, stabilizes training, and enhances representation consistency.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \lambda \left\|\mathbf{E}_{\text{signal}} - \mathbf{E}_{\text{text}}\right\|_2^2 \tag{8}$$

4. Experimental Results and Dicussion

4.1. Dataset and Evaluation Metrics

We evaluated on the HumanML3D dataset [19], which includes 14,616 motion samples from AMASS [20] and HumanAct12 [21], with a total of 44,970 annotations. Each motion sequence consists of N joints, where joint j_i at time t is represented by position $\mathbf{p}_i(t)$ and rotation $\mathbf{R}_i(t)$. The position follows $\mathbf{p}_i(t) = \mathbf{T}_i + \mathbf{R}_i(t) \cdot \mathbf{l}_i$, with $\mathbf{T}_i$ as the joint root translation, $\mathbf{R}_i(t)$ as the rotation matrix, and $\mathbf{l}_i$ as the local offset. The motion is represented as a time series $\mathcal{P}(t) = \{\mathbf{p}_1(t), \ldots, \mathbf{p}_N(t)\}$, mapped from textual descriptions $\mathcal{D} = \{w_1, \ldots, w_m\}$ via $\mathcal{D} \xrightarrow{f} \mathcal{P}(t)$.

To comprehensively evaluate the quality of the generated motions, we adopted eight key metrics validating semantic, spatial, temporal, and distributional aspects. Specifically, **R Precision** measures

Table 1
Quantitative results on the HumanML3D test set. All methods use the real motion length from the ground truth. ± indicates 95% confidence interval, and → means the closer to real motions the better. Bold face indicates the best result.

Methods	R Precision↑			FID↓	MultiModal Dist.↓	Diversity→	MultiModality↑	APE↓	Velocity↓	Jerk↓
	Top 1	Top 2	Top 3							
Real motions	$0.500^{\pm.005}$	$0.695^{\pm.006}$	$0.795^{\pm.005}$	$0.002^{\pm.004}$	$2.788^{\pm.012}$	$8.425^{\pm.109}$	-	-	-	-
4 signal tokens	$0.297^{\pm.005}$	$0.502^{\pm.003}$	$0.598^{\pm.004}$	$4.318^{\pm.022}$	$5.147^{\pm.030}$	$7.426^{\pm.109}$	3.014	$0.652^{\pm.005}$	$0.022^{\pm.002}$	$0.016^{\pm.002}$
8 signal tokens	$0.342^{\pm.004}$	$0.509^{\pm.004}$	$0.626^{\pm.005}$	$7.918^{\pm.022}$	$2.980^{\pm.029}$	$7.655^{\pm.079}$	4.124	$0.547^{\pm.008}$	$0.017^{\pm.006}$	$0.012^{\pm.003}$
24 signal tokens	$\mathbf{0.371}^{\pm.005}$	$\mathbf{0.558}^{\pm.004}$	$\mathbf{0.676}^{\pm.005}$	$\mathbf{3.428}^{\pm.183}$	$\mathbf{2.970}^{\pm.030}$	$\mathbf{7.918}^{\pm.109}$	**4.516**	$\mathbf{0.532}^{\pm.005}$	$\mathbf{0.012}^{\pm.002}$	$\mathbf{0.006}^{\pm.002}$

Table 2
Quantitative results on the HumanML3D test set. All methods use the real motion length from the ground truth. ± indicates the 95% confidence interval, and → means the closer to real motions, the better. Bold face indicates the best result. **Experts** refer to the number of specialized sub-networks used within the mixture of experts (MoE), where each expert captures different motion characteristics. **Top-k** denotes the number of experts dynamically selected by the gating function for each input instance, ensuring that only the most relevant experts contribute to the final motion generation.

Methods	R Precision↑			FID↓	MultiModal Dist.↓	Diversity→	MultiModality↑	APE↓	Velocity↓	Jerk↓
	Top 1	Top 2	Top 3							
Real motions	$0.501^{\pm.005}$	$0.695^{\pm.006}$	$0.795^{\pm.005}$	$0.002^{\pm.004}$	$2.788^{\pm.012}$	$8.425^{\pm.109}$	-	-	-	-
With 4 experts, $top_k = 2$	$0.371^{\pm.005}$	$0.558^{\pm.004}$	$0.676^{\pm.005}$	$7.918^{\pm.183}$	$3.014^{\pm.029}$	$7.455^{\pm.079}$	3.014	$0.552^{\pm.008}$	$0.012^{\pm.006}$	$0.006^{\pm.003}$
With 6 experts, $top_k = 2$	$0.385^{\pm.005}$	$\mathbf{0.568}^{\pm.003}$	$0.662^{\pm.003}$	$5.818^{\pm.021}$	$5.276^{\pm.030}$	$\mathbf{7.932}^{\pm.109}$	3.023	$\mathbf{0.539}^{\pm.005}$	$0.013^{\pm.002}$	$0.007^{\pm.004}$
With 8 experts, $top_k = 2$	$\mathbf{0.387}^{\pm.004}$	$0.564^{\pm.004}$	$\mathbf{0.678}^{\pm.005}$	$\mathbf{2.411}^{\pm.183}$	$2.963^{\pm.029}$	$7.725^{\pm.079}$	2.963	$0.576^{\pm.008}$	$\mathbf{0.012}^{\pm.006}$	$\mathbf{0.005}^{\pm.003}$
With 8 experts, $top_k = 4$	$0.345^{\pm.004}$	$0.523^{\pm.004}$	$0.641^{\pm.005}$	$3.456^{\pm.126}$	$\mathbf{2.959}^{\pm.029}$	$7.782^{\pm.079}$	**3.062**	$0.541^{\pm.008}$	$0.012^{\pm.003}$	$0.006^{\pm.002}$

semantic alignment by retrieving the top-k captions closest to each generated motion in an embedding space, with the score defined as R-Precision@$k = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[c_i \in \text{Top-}k(m_i)]$. **Fréchet Inception Distance (FID)** evaluates distributional similarity between real and generated motion features via $\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right)$. **Diversity** quantifies motion variation using the average pairwise distance across samples: $\text{Diversity} = \frac{2}{N(N-1)}\sum_{i<j}\|m_i - m_j\|_2$. **Multimodality** measures intra-prompt diversity through $\text{Multimodality} = \frac{1}{N}\sum_{i=1}^{N}\frac{2}{K(K-1)}\sum_{j<l}\|m_{ij} - m_{il}\|_2$. **Multimodal Distance** assesses cross-modal alignment via the average embedding distance between generated motions and their paired captions: $\text{MD} = \frac{1}{N}\sum_{i=1}^{N} d(E_m(m_i), E_t(c_i))$. **Average Positional Error (APE)** evaluates spatial accuracy using $\text{APE} = \frac{1}{NTJ}\sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{J}\|\hat{x}_{ntj} - x_{ntj}\|_2$. **Velocity** and **Jerk** assess temporal consistency and smoothness, defined respectively as $\text{Velocity} = \frac{1}{NTJ}\sum_{n,t,j}\|x_{ntj} - x_{n(t-1)j}\|_2$ and $\text{Jerk} = \frac{1}{NTJ}\sum_{n,t,j}\left\|x_{ntj} - 3x_{n(t-1)j} + 3x_{n(t-2)j} - x_{n(t-3)j}\right\|_2$. We additionally report **Energy** as an auxiliary metric without further elaboration.

4.2. Number of Signal Tokens

Our experiment was firstly conducted using different numbers of signal tokens, as illustrated in Fig.2, to assess their impact on motion generation. The experimental results presented in Table 1 demonstrate that using only 4 signal tokens resulted in the lowest R-Precision scores (0.297, 0.502, 0.598) and the highest MultiModal Distance (5.147), indicating weak alignment with real motions. Increasing to 8 tokens improved retrieval accuracy (Top-1 R-Precision: 0.342) and reduced MultiModal Distance to 2.980, though FID increased to 7.918. Our experiment noticed that with 24 tokens, the model achieved the best performance across all evaluation metrics, including the highest R-Precision (0.371, 0.558, 0.676), Diversity (7.918), and the lowest FID (3.428). This finding implies that increasing number of tokens enhances motion diversity, and retrieval accuracy. Additionally, lower APE (0.532), Velocity (0.012), and Jerk (0.006) indicate smoother and more natural motion sequences.

4.3. Mixture of Experts

As our designed FM utilizes the mixture of experts (MoE) approach, an experiment was conducted with different numbers of experts and top_k values to determine the optimal configuration. The results in Table 2 indicate that increasing the number of experts generally improves motion diversity but introduces trade-offs in retrieval accuracy. Expanding from 4 to 6 experts slightly enhances Top-2 R-Precision

Table 3
Quantitative results on the HumanML3D test set. All methods use the real motion length from the ground truth.± indicates 95% confidence interval, and $\rightarrow$ means the closer to Real motions the better. Bold face indicates the best result, while underscore refers to the second best. The Total loss is the sum of Original loss, Projection loss and Out loss

Methods	R Precision↑			FID↓	MultiModal Dist.↓	Diversity→	MultiModality↑	APE↓	Velocity↓	Jerk↓
	Top 1	Top 2	Top 3							
Real motions	$0.500^{\pm.005}$	$0.695^{\pm.006}$	$0.795^{\pm.005}$	$0.002^{\pm.004}$	$2.788^{\pm.012}$	$8.425^{\pm.109}$	-	-	-	-
Original loss	$0.371^{\pm.005}$	$0.558^{\pm.004}$	$0.676^{\pm.005}$	$7.918^{\pm.183}$	$3.014^{\pm.029}$	$7.455^{\pm.079}$	**3.014**	$0.552^{\pm.008}$	$0.012^{\pm.006}$	$0.006^{\pm.003}$
Original loss + Out loss	**0.385**$^{\pm.003}$	**0.571**$^{\pm.003}$	**0.686**$^{\pm.003}$	$3.085^{\pm.167}$	$2.980^{\pm.029}$	$7.509^{\pm.067}$	2.871	$0.532^{\pm.008}$	**0.012**$^{\pm.005}$	**0.006**$^{\pm.001}$
Original loss + Projection loss	$0.363^{\pm.004}$	$0.548^{\pm.004}$	$0.661^{\pm.005}$	**3.074**$^{\pm.183}$	**2.959**$^{\pm.029}$	**7.782**$^{\pm.079}$	2.959	**0.531**$^{\pm.008}$	$0.012^{\pm.006}$	$0.006^{\pm.003}$
Total loss	$0.156^{\pm.004}$	$0.255^{\pm.004}$	$0.338^{\pm.005}$	$12.194^{\pm.183}$	$6.964^{\pm.029}$	$9.343^{\pm.079}$	2.945	$0.591^{\pm.008}$	$0.214^{\pm.006}$	$0.147^{\pm.003}$

(from 0.558 to 0.568) and significantly reduces FID (from 7.918 to 5.818), suggesting better alignment with real motions. Further increasing to 8 experts leads to the lowest FID (2.411) and MultiModal Distance, demonstrating improved the realistic and diversity of generated motions. However, Top-1 R-Precision remains stable (0.385–0.387), indicating that additional experts do not necessarily improve single-instance retrieval accuracy. Adjusting top_k from 2 to 4 with 8 experts decreases Top-1 R-Precision (from 0.387 to 0.345) and increases FID again (from 2.411 to 3.456). This result suggests that selecting more experts introduces variability without consistently enhancing accuracy. Despite these trade-offs, diversity remains relatively stable, indicating that expert fusion preserves a broad range of motion styles. Overall, the configuration with 8 experts and $top_k = 2$ achieves the relatively balanced performance, offering the lowest FID (2.411) and reasonably low MultiModal Distance (2.963) while maintaining strong retrieval accuracy and motion diversity, making it the most effective choice for realistic motion generation.

4.4. Configuration of the Loss Function in Phase 3

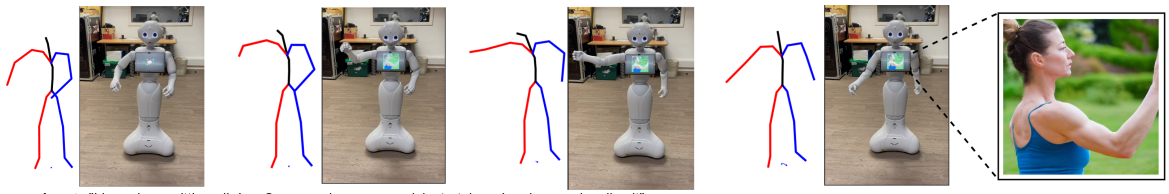
We examined the impact of the loss function in Phase 3 by comparing the model's performance using the original DDMG loss and an augmented loss with L2 regularization, as defined in Equation 8. The L2 loss consists of two components: the embedding from the MDA module and the text embedding from the CLIP encoder. These embeddings are processed through a transformer encoder to produce *tensor out*, which is projected to form *tensor projection*, used in the cross-attention layers of the DDMG module.

The loss function is evaluated using two approaches: Out Loss, which measures the L2 distance between *tensor out* and the CLIP text embedding, and Projection Loss, which compares the L2 distance between *tensor projection* representations.

We tested four configurations of the loss function within the same network architecture. As shown in Table 3, the loss function configuration significantly affects model performance. The original loss yields a Top-1 R-Precision of 0.371 with a high FID (7.918), indicating the generated motions deviate from real motions. Adding Out loss improves Top-1 R-Precision (0.385) and reduces FID (3.085), aligning the model's output closer to real motions. The projection loss decreases the accuracy of the search (0.363) but reduces the Multi-Modal Distance (2.959 vs. 3.014 in the original loss), improving motion diversity. The Total loss configuration, combining all components, results in a large drop in Top-1 R-Precision (0.156) and an increase in FID (12.194), suggesting excessive regularization harms model performance. These results highlight the need for a balanced loss formulation to optimize retrieval accuracy, realism, and motion diversity.

4.5. Demonstration of the Inference Phase on a Social Robot

We implemented the proposed framework on a social robot, namely Pepper. During the inference phase, as illustrated in Fig. 1, the framework receives a human textual prompt as a mandatory input, while other modalities are optional. The model generates a textual response and corresponding motion as mandatory outputs, with additional modalities that could be produced depending on the input prompt. The Motion Conversion module was deployed to convert gestured motions into the robot's motion



Human Input: *"I have been sitting all day. Can you show me a quick stretch and an image visualize it"*
Robot Response: *"Of course! I can definitely help you with that. Here's a quick stretch that you can do to warm up your muscles and get some fresh air. It's important to take breaks and stretch regularly to prevent stiffness and improve overall health."*

Figure 3: The robot responses to human input with a generated motion and a generated image of stretching with the right hand (displayed on the robot tablet) to explain verbal contents of it speech.

space, resulting in communicative gestures performed by the robot. The generated textual responses were converted to robot's speech, while other generated modalities were displayed on the robot's tablet, to support the robot speech, simultaneously. In Fig. 3, in response to a user request of showing a quick stretch illustrated with an image, the framework additionally generated an image of right-hand stretch, which was displayed on the robot's tablet. This enabled the robot to comprehend their communicative messages through a combination of speech, gesture, and an image.

5. Conclusion

This paper introduced a novel framework that leverages a multimodal LLM and a diffusion-based motion approach to generate high-quality motions that can be implemented into social robots, supporting for multimodal communication. By integrating a DDMG module with a MDA module and a novel FM module, our approach enables the network to capture both semantic and temporal information, ensuring diverse and contextually coherent motions. Experimental results on the HumanML3D dataset demonstrated that our framework yields competitive performance across a wide range of metrics. Ablation studies revealed the importance of carefully balancing signal tokens, mixture-of-experts configurations, and alignment losses, underscoring the effectiveness of the fusion mechanism in bridging the gap between textual instructions and the target motion domain. Our approach mitigates the need for extensive multimodal datasets by exploiting pre-trained components, such as ImageBind and well-established language models, thereby reducing computational costs without compromising generation quality. There are several remaining works that will be considered in our future works. First, we plan to expand the framework to a higher number of interaction modalities commonly seen in HRI (e.g., eye-contact, haptic feedback, etc.). We also aim to explore more advanced gating mechanisms and integration strategies, aiming to further improve the scalability and adaptability of our multimodal system.

Acknowledgments

This work was supported by Responsible Ai UK (EP/Y009800/1).

References

[1] H. Su, W. Qi, J. Chen, C. Yang, J. Sandoval, M. A. Laribi, Recent advancements in multimodal human–robot interaction, Frontiers in Neurorobotics 17 (2023) 1084000.

[2] N. T. V. Tuyen, S. Okazaki, O. Celiktutan, A study on customer's perception of robot nonverbal communication skills in a service environment, in: 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), IEEE, 2023, pp. 301–306.

[3] C. Papadopoulos, T. Hill, L. Battistuzzi, N. Castro, A. Nigath, G. Randhawa, L. Merton, S. Kanoria, H. Kamide, N.-Y. Chong, et al., The caresses study protocol: testing and evaluating culturally

competent socially assistive robots among older adults residing in long term care homes through a controlled experimental trial, Archives of Public Health 78 (2020) 1–10.

[4] C. Ahuja, L.-P. Morency, Language2pose: Natural language grounded pose forecasting, 2019. URL: https://arxiv.org/abs/1907.01108. arXiv:1907.01108.

[5] L. Xu, Z. Song, D. Wang, J. Su, Z. Fang, C. Ding, W. Gan, Y. Yan, X. Jin, X. Yang, W. Zeng, W. Wu, Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation, 2022. URL: https://arxiv.org/abs/2203.07706. arXiv:2203.07706.

[6] D. P. Kingma, M. Welling, Auto-encoding variational bayes, 2022. URL: https://arxiv.org/abs/1312.6114. arXiv:1312.6114.

[7] C. Guo, X. Zuo, S. Wang, L. Cheng, Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts, 2022. URL: https://arxiv.org/abs/2207.01696. arXiv:2207.01696.

[8] C. Guo, Y. Mu, M. G. Javed, S. Wang, L. Cheng, Momask: Generative masked modeling of 3d human motions, 2023. URL: https://arxiv.org/abs/2312.00063. arXiv:2312.00063.

[9] D. J. Rezende, S. Mohamed, Variational inference with normalizing flows, 2016. URL: https://arxiv.org/abs/1505.05770. arXiv:1505.05770.

[10] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, Z. Liu, Motiondiffuse: Text-driven human motion generation with diffusion model, 2022. URL: https://arxiv.org/abs/2208.15001. arXiv:2208.15001.

[11] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, A. H. Bermano, Human motion diffusion model, 2022. URL: https://arxiv.org/abs/2209.14916. arXiv:2209.14916.

[12] H. Liang, J. Bao, R. Zhang, S. Ren, Y. Xu, S. Yang, X. Chen, J. Yu, L. Xu, Omg: Towards open-vocabulary motion generation via mixture of controllers, 2024. URL: https://arxiv.org/abs/2312.08985. arXiv:2312.08985.

[13] N. T. V. Tuyen, A. Elibol, N. Y. Chong, Conditional generative adversarial network for generating communicative robot gestures, in: 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), IEEE, 2020, pp. 201–207.

[14] Y. Li, Y. Zhang, C. Wang, Z. Zhong, Y. Chen, R. Chu, S. Liu, J. Jia, Mini-gemini: Mining the potential of multi-modality vision language models, 2024. URL: https://arxiv.org/abs/2403.18814. arXiv:2403.18814.

[15] B. Li, P. Zhang, J. Yang, Y. Zhang, F. Pu, Z. Liu, Otterhd: A high-resolution multi-modality model, 2023. URL: https://arxiv.org/abs/2311.04219. arXiv:2311.04219.

[16] J. Han, K. Gong, Y. Zhang, J. Wang, K. Zhang, D. Lin, Y. Qiao, P. Gao, X. Yue, Onellm: One framework to align all modalities with language, 2023. URL: https://arxiv.org/abs/2312.03700. arXiv:2312.03700.

[17] S. Wu, H. Fei, L. Qu, W. Ji, T.-S. Chua, Next-gpt: Any-to-any multimodal llm, 2024. URL: https://arxiv.org/abs/2309.05519. arXiv:2309.05519.

[18] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, J. Huang, A survey on mixture of experts, 2024. URL: https://arxiv.org/abs/2407.06204. arXiv:2407.06204.

[19] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, L. Cheng, Generating diverse and natural 3d human motions from text, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5142–5151. doi:10.1109/CVPR52688.2022.00509.

[20] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, M. J. Black, Amass: Archive of motion capture as surface shapes, 2019. URL: https://arxiv.org/abs/1904.03278. arXiv:1904.03278.

[21] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, L. Cheng, Action2motion: Conditioned generation of 3d human motions, in: Proceedings of the 28th ACM International Conference on Multimedia, MM '20, ACM, 2020. URL: http://dx.doi.org/10.1145/3394171.3413635. doi:10.1145/3394171.3413635.