LLM Knows Body Language, Too: Translating Speech Voices into Human Gestures

Chenghao Xu[1], Guangtao Lyu[1], Jiexi Yan[2*], Muli Yang[3], Cheng Deng[1*]
[1] School of Electronic Engineering, Xidian University, Xi'an, Shaanxi, China,
[2] School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, China,
[3] Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
{chx,guangtaolyu}@stu.xidian.edu.cn, {jxyan1995,muliyang.xd,chdeng.xd}@gmail.com

Abstract

In response to the escalating demand for digital human representations, progress has been made in the generation of realistic human gestures from given speeches. Despite the remarkable achievements of recent research, the generation process frequently includes unintended, meaningless, or non-realistic gestures. To address this challenge, we propose a gesture translation paradigm, GesTran, which leverages large language models (LLMs) to deepen the understanding of the connection between speech and gesture and sequentially generates human gestures by interpreting gestures as a unique form of body language. The primary stage of the proposed framework employs a transformer-based auto-encoder network to encode human gestures into discrete symbols. Following this, the subsequent stage utilizes a pre-trained LLM to decipher the relationship between speech and gesture, translating the speech into gesture by interpreting the gesture as unique language tokens within the LLM. Our method has demonstrated state-of-the-art performance improvement through extensive and impartial experiments conducted on public TED and TED-Expressive datasets.

1 Introduction

The synthesis of human motions and gestures that correspond with concurrent speech, a process known as Co-Speech gesture generation, is instrumental in conveying messages during human communication and augmenting self-expression (Kucherenko et al., 2021). Therefore, generating realistic and controllable gesture motions that are both plausible and synchronous with the corresponding speech input can significantly bolster the acceptance of social robots by human users (Cassell et al., 1999; Wagner et al., 2014). This holds substantial promise for various applications, including but not limited to, education, training, and medical contexts. Additionally, the potential for this pursuit extends to the development of digital humans in emerging virtual environments, non-player game characters, robotic assistants (Salem et al., 2012, 2011), and embodied artificial intelligence.

In practical scenarios, a single speech input could correspond to a variety of gestures or motions. For example, a speech input corresponding to a beat motion could be performed using the left hand, right hand, or both hands; all these variations are plausible and would be considered appropriate by human users(Yan et al., 2022b; Xu et al., 2023). However, prior methodologies (Ao et al., 2022; Li et al., 2021a) often frame co-speech gesture generation as a regression problem, resulting in a model that is more likely to learn an average of all plausible gestures rather than distinct ones, thereby generating excessively smoothed and unrealistic average gestures. Consequently, these methods tend to yield more restrained gesture motions, which are less engaging from a human perception standpoint. It remains unclear how existing methods could capture such one-to-many variability. Furthermore, such methods tend to exhibit instability in practical applications and are susceptible to regression towards nonstandard poses beyond the gesture subspace, such as freezing or meaningless swaying(Yan et al., 2024, 2022a).

A primary inherent issue contributing to the aforementioned problems is that previous methods fail to adequately model the semantic relation between speech and gesture, specifically, their diverse temporal correspondence. To enable the generative model to more comprehensively understand and encapsulate this relationship, we propose treating the gesture as a distinct form of body language that can be seamlessly translated into and out of speech. Recent research suggests that large language models (LLMs) (Touvron et al., 2023; Radford et al., 2018, 2019) can process multimodal inputs, such as

*Corresponding author

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5004–5013
August 11-16, 2024

images and videos, through a lightweight adapter. Consequently, we anticipate that LLMs, with a suitable adapter, can also comprehend gesture sequences. The integration of gesture and speech (audio and its corresponding text) data, encoded within a unified vocabulary, makes the relationship between motion and language more discernible. This would enable the gesture generator, which is fine-tuned from LLMs, to produce gestures with diverse patterns and flexible sequences.

In this paper, we propose a new LLM-driven co-speech gesture translation method, namely GesTran, which emulates the procedure of bilingual translation in humans and has the capability to comprehend and translate human gestures that are concomitantly associated with speech and its corresponding text. In order to equip GesTran with the capability to comprehend and generate gestures akin to humans, an initial step involves training a gesture-specific Vector Quantized Variational Autoencoder (VQ-VAE) (Van Den Oord et al., 2017) model. The objective of this step is to compile a "gesture lexicon" (Chiu et al., 2015), analogous to a natural language vocabulary, which subsequently allows for the conversion of unprocessed gesture data into a series of gesture tokens. These tokens are subsequently processed by a pre-trained language model, which has been trained to understand the inherent grammar and syntax of the gesture language, as well as its correlation with the corresponding audio and text of human speeches. To efficiently amalgamate speech and gesture, we fine-tune the pre-trained language model on a multi-modal co-speech gesture dataset, which is instrumental in learning the correlation and conversion between speech and gesture. In this way, we can easily translate the speech to desirable body language, *i.e.*, human gesture, in the unified LLM. Extensive experimental results demonstrate that GesTran attains a performance that surpasses current benchmarks in the task of co-speech generation.

We summarize our contributions as follows:

- We present a novel gesture translation framework, **GesTran**, for co-speech gesture generation by incorporating a pre-trained LLM. Regarding human gesture as a specific body language, our method can better comprehend the correlation between speech and gesture and effectively translate them in the unified pre-trained LLM.
- By leveraging the strong language generation and zero-shot transfer abilities of pre-trained language models, our gesture generation model can synthesize diverse human gestures and have better generalization ability.
- The proposed GesTran consistently outperforms state-of-the-art co-speech gesture generation methods across benchmark datasets and metrics.

2 Related Work

Co-speech Gesture Generation. The synthesis of co-speech gestures holds significant importance across various applications. Conventional approaches (Cassell et al., 1994; Huang and Mutlu, 2012) typically employ rule-based pipelines, wherein linguistic experts define speech-gesture pairs and refine transitions between different motions. Additionally, motion-matching-based models (Yang et al., 2023; Büttner and Clavet, 2015), if appropriately designed, exhibit greater effectiveness compared to neural network-based counterparts. Moreover, researchers are delving into comprehending the influence of input modalities, investigating the relationships between co-speech gestures and speech audio, text transcripts, speaking styles, and speaker identity (Yoon et al., 2020). Previous studies aim to augment model capacity through a range of architectural choices, including Convolutional Neural Networks (CNN) (Xu et al., 2023; Ao et al., 2022), Recurrent Neural Networks (RNN) (Yoon et al., 2019), Transformer models (Pang et al., 2023), Generative Adversarial Networks (GANs) (Yoon et al., 2020; Liu et al., 2022), and Diffusion models (Zhu et al., 2023; Zhi et al., 2023; Ao et al., 2023).

Large Language Model. Large-scale language models (LLMs) (Touvron et al., 2023; Du et al., 2021; Team et al., 2023; Minaee et al., 2024), facilitated by extensive datasets and large model sizes, have showcased remarkable comprehension and generation capabilities, significantly advancing the field of natural language processing. BERT (Devlin et al., 2018), for instance, pre-trains deep bidirectional language representations capable of supporting downstream tasks. T5 (Raffel et al., 2020) introduced a unified framework that transforms all text-based language tasks into a text-to-text format. Recent studies have demonstrated that fine-tuning pre-trained models using input-output pairs comprising instructions and corresponding responses can fur-

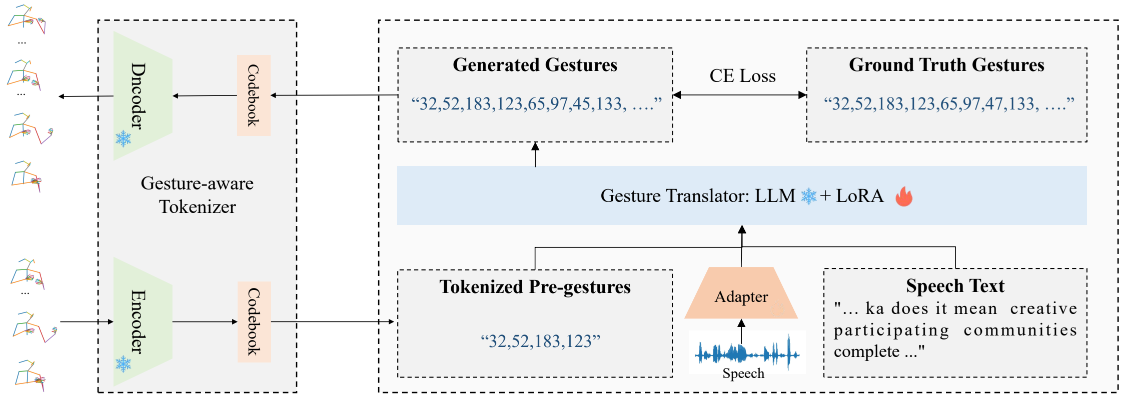

Figure 1: The overall framework of our gesture translator.

ther enhance their performance. FLAN (Chung et al., 2022) introduces an instruction-tuning technique that outperforms non-tuned models on unseen tasks. LLaMA (Touvron et al., 2023)is a collection of open-source and efficient foundation large language models ranging from 7B to 65B parameters. Moreover, the emergence of multimodal models, which process text along with other modalities such as images, audio, and videos, has garnered considerable attention. Despite the success of language models in various vision-language tasks, the development of multi-modal language models capable of interpreting human gestures remains relatively limited.

3 Method

Inspired by MotionGPT (Zhang et al., 2024; Jiang et al., 2023; Ribeiro-Gomes et al., 2024), we introduce LLMs into the task of co-speech gesture generation. Capitalizing on the exceptional ability of LLMs to comprehend and translate multilingual data, we propose a co-speech gesture translator (GesTran) governed by multimodal conditions, namely speeches (audio and corresponding text) and human gestures captured in video frames. We intend to frame human gestures as a particular form of body language, thereby enabling the Large Language Model to translate desired human gestures in accordance with corresponding prompts and control conditions. The overall framework of our GesTran is shown in Figure 1. Specifically, we first quantize raw gesture data into discrete tokens using VQ-VAE (Van Den Oord et al., 2017).

3.1 Gesture-wise Token Quantization

To effectively conceptualize gesture as a language, thereby facilitating the integration and translation of gesture and speech, we pre-train a human gesture tokenizer. This is accomplished by utilizing the Vector Quantized Variational Autoencoders (VQ-VAE) architecture, which enables the attainment of discrete representations of gesture data with discrete tokens. Our gesture-wise tokenizer consists of a gesture encoder $\boldsymbol{E}$ and gesture decoder $\boldsymbol{D}$.

Specifically, given a gesture sequence $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T]$, where T is the number of frames, our gesture-wise tokenizer aims to recover the gesture sequence with a learnable codebook $\mathcal{C} = \{\boldsymbol{c}_k\}_{k=1}^{N} \subset \mathbb{R}^d$ containing N codes, each of dimension d. With the gesture encoder $\boldsymbol{E}$, the latent feature $\boldsymbol{V} = [\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \cdots, \boldsymbol{\nu}_T]$ can be computed as $\boldsymbol{V} = \boldsymbol{E}(\boldsymbol{X})$. We can train the gesture-wise tokenizer by the combination of the reconstruction loss, the embedding loss, and the commitment loss as follows:

$$\mathcal{L}_r = \underbrace{||\boldsymbol{D}(\boldsymbol{E}(\boldsymbol{x}_i)) - \boldsymbol{x}_i||^2}_{reconstruction\, loss} + \underbrace{||\mathrm{sg}[\boldsymbol{E}(\boldsymbol{x}_i)] - \hat{\boldsymbol{\nu}}_i||_2^2}_{embedding\, loss} + \beta \underbrace{||\boldsymbol{E}(\boldsymbol{x}_i) - \mathrm{sg}[\hat{\boldsymbol{\nu}}_i]||_2^2}_{committing\, loss}. \tag{1}$$

Here, for the i-th latent feature $\boldsymbol{\nu}_i$, the estimated embedding $\hat{\boldsymbol{\nu}}_i$ can be found by searching the nearest embedding in the codebook $\mathcal{C}$ through the quantization process $Q(\cdot)$:

$$\hat{\boldsymbol{\nu}}_i = Q(\boldsymbol{x}_i) := \underset{\boldsymbol{c}_k \in \mathcal{C}}{\arg\min} \|\boldsymbol{\nu}_i - \boldsymbol{c}_k\|_2. \tag{2}$$

Based on the estimation latent representation $\hat{\boldsymbol{V}} = [\hat{\boldsymbol{\nu}}_1, \hat{\boldsymbol{\nu}}_2, \cdots, \hat{\boldsymbol{\nu}}_T]$, the reconstructed human gesture can be produced by the decoder $\boldsymbol{D}(\cdot)$, *i.e.*, $\tilde{\boldsymbol{X}} = \boldsymbol{D}(\hat{\boldsymbol{V}})$.

3.2 Gesture-aware Translation

With the utilization of our learned gesture-wise tokenizer, a gesture sequence denoted as $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T]$ can be mapped into a sequence of gesture tokens, represented as $\hat{\boldsymbol{V}} = [\hat{\boldsymbol{\nu}}_1, \hat{\boldsymbol{\nu}}_2, \cdots, \hat{\boldsymbol{\nu}}_T]$. This interpretation facilitates the joint representation and translation with audio and text embeddings of speech in LLMs. Specifically, we first represent the motion sequence $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T]$ to a sequence of indices $\mathcal{I} = \{\text{</sog>}\} \cup \{s_i\}_{i=1}^{T} \cup \{\text{</eog>}\}$ with $s_i = [1, 2, \cdots, T]$. Note that the special </sog> and </eog> tokens are added to indicate the start and stop of the gesture. By projecting $\mathcal{I}$ back to their corresponding codebook entries, we can reconstruct the gesture through decoding $\hat{\boldsymbol{\nu}}_i = \boldsymbol{c}_{s_i}$ with the learned gesture decoder $\boldsymbol{D}(\cdot)$.

In a bid to ingeniously frame the speech-to-gesture autoregressive prediction as a comprehensible language translation paradigm, we establish a bridge between gesture and speech. This allows LLMs to comprehend human gesture concepts by fine-tuning the pre-trained LLMs with the widely utilized and efficient Low-Rank Adaptation (LoRA) (Hu et al., 2021). Specifically, we unify the audio and text of the speech and human gestures within a single LLM. For the audio data of the speech, we incorporate an adapter to extract the sequence of audio embeddings, denoted as $\boldsymbol{A}$. Simultaneously, the text embeddings, represented as $\boldsymbol{T}$, can be directly derived through the LLM. Treating the audio and text embeddings of the speech as the source language, we aim to translate them into a diverse and meaningful target body language, namely human gesture, on a frame-by-frame basis.

Given the source language pair $\{\boldsymbol{A}, \boldsymbol{T}\}$ and previous $i-1$ predicted indices $[s_1, s_2, \cdots, s_{i-1}]$, the LLM is enforced to translate the subsequent gesture index s_i. The final translation output of LLM, denoted as $\tilde{\mathcal{V}}$, constitutes a series of generated gesture tokens, which can be decoded to human gesture using our learned gesture-wise tokenizer. Analogous to the majority of language models, we employ cross-entropy loss, which constrains the similarity between estimated and ground-truth tokens, to fine-tune LLMs using LoRA, which can be represented as

$$\mathcal{L}_t = \text{CE}(\tilde{\mathcal{V}}, \tilde{\mathcal{V}}^*), \tag{3}$$

where $\tilde{\mathcal{V}}^*$ is the gesture tokens of ground-truth gestures calculated by Eq.(2) and $\tilde{\mathcal{V}}$ is the translated gesture tokens by the LLM.

3.3 Zero-shot Generalized Extension Analysis

Present co-speech generation methodologies lack the capacity to directly synthesize corresponding gestures in response to speeches encapsulating unseen sentences. This poses a significant challenge in practical applications as it is implausible to guarantee that the speech requiring translation has been previously exposed to our model during its training phase.

LLMs have also proven to be instrumental in advancing zero-shot learning. LLMs are trained on a vast corpus of text from the internet, learning a wealth of linguistic patterns, facts about the world, and to some extent, reasoning abilities. This extensive training enables LLMs to leverage their learned knowledge when presented with new tasks, making them highly versatile tools for zero-shot learning. It's not explicitly trained on the specific task, but it uses its general understanding of language and world knowledge to generate a meaningful response. The development and application of LLMs in zero-shot learning continue to be an active area of research, with potential impacts across various fields, including natural language processing, computer vision, and more. Due to the superior zero-shot generalization ability of the LLM, our method can also deal with unseen speeches and well translate them into diverse gestures.

4 Experiments

4.1 Co-Speech Gesture Datasets

TED Gesture: Serving as a significant dataset for gesture generation research, the TED Gesture dataset (Yoon et al., 2019, 2020) comprises 1,766 TED videos featuring different narrators discussing various topics. The data processing methodology from previous works is adopted (Yoon et al., 2020; Liu et al., 2022), where poses are resampled at 15 FPS, and frame segments of length 34 are obtained with a stride of 10.

TED Expressive: In contrast to TED Gesture, which includes poses with only 10 upper body key points without detailed finger movements, the TED Expressive dataset (Liu et al., 2022) goes further by capturing expressive finger and body movements. The state-of-the-art 3D pose estimator, ExPose (Choutas et al., 2020), is employed to fully

Methods	TED Gesture			TED Expressive		
	FGD ↓	BC ↑	Diversity ↑	FGD ↓	BC ↑	Diversity ↑
Ground Truth	0	0.698	108.525	0	0.703	178.827
Gesture VQ-VAE	0.205	0.698	108.501	0.190	0.728	184.595
Attention Seq2Seq (Yoon et al., 2019)	18.154	0.196	82.776	54.920	0.152	122.693
Speech2Gesture (Ginosar et al., 2019)	19.254	0.668	93.802	54.650	0.679	142.489
Joint Embedding (Ahuja and Morency, 2019)	22.083	0.200	90.138	64.555	0.130	120.627
Trimodal (Yoon et al., 2020)	3.729	0.667	101.247	12.613	0.563	154.088
HA2G (Liu et al., 2022)	3.072	0.672	104.322	5.306	0.641	173.899
DiffGesture (Zhu et al., 2023)	1.506	**0.699**	106.722	2.600	0.718	182.757
GesTran (Ours)	**1.087**	0.697	**108.190**	**1.836**	**0.720**	**182.295**

Table 1: **The Quantitative Results on TED Gesture and TED Expressive** . We compare the proposed GesTran against recent methods and ground truth. For FGD, the lower, the better; for other metrics, the higher, the better.

	# Train	# Test	# New
TED Gesture	22662	7992	1240
TED Gesture Ext	14459	18209	9443
TED Express	24016	7586	901
TED Express Ext	16826	16780	8091

Table 2: The statistics of speech word distribution.

capture pose information in the data. Consequently, TED Expressive annotates the 3D coordinates of 43 keypoints, including 13 upper body joints and 30 finger joints.

TED Gesture Ext & TED Expressive Ext: In order to better verify the zero-shot generalization ability of the model, we re-separate the training/testing split for the TED Gesture dataset and TED Expressive in a different way. We first count the frequency of different words and filter out parts of low-frequency words to form a testing split. After this operation, many words in the testing split have never appeared in the training split. This zero-shot way of segmenting the dataset can better describe the situations that occur in reality, and can also better verify the generalization of our model. Detailed dataset details are provided in Table 2.

4.2 Experimental Settings

Comparison Methods: Our method is compared against recent state-of-the-art techniques on two benchmark datasets. **1) Attention Seq2Seq** (Yoon et al., 2019) elaborates on the attention mechanism for generating pose sequences from speech text. **2) Speech2Gesture** (Ginosar et al., 2019) utilizes spectrums of speech audio segments as input, generating speech gestures adversarially. **3) Joint Embedding** (Ahuja and Morency, 2019) maps text and motion to a shared embedding space, generating outputs from motion description text. **4) Trimodal** (Yoon et al., 2020) serves as a robust baseline learning from text, audio, and speaker identity to generate gestures, outperforming prior methods significantly. **5) HA2G** (Liu et al., 2022) introduces a hierarchical audio learner capturing information across different semantic granularities. **6) DiffGesture** (Zhu et al., 2023) introduces a novel diffusion audio-gesture transformer with a diffusion gesture stabilizer to eliminate temporal inconsistency.

Implementation Details. For a fair comparison, we maintained consistency in our experimental setup with that of previous methods. For all the methods in both datasets, we set $N = 34$ and $M = 4$ to get N-frame pose sequences where the first M frames are used for reference, termed as initial poses. Following (Yoon et al., 2020), to eliminate the effect of the joint lengths and root motion, we represent the joints' positions using directional vectors normalized to the unit vectors and train the model to learn the directional vectors. We use standard transformer blocks for gesture Gesture VQ-VAE. The size of the codebook is set to length, groups, and dims. It is set to 1024, 2, and 512 for both datasets. We use an Adam optimizer, and the learning rate is 0.0001. All experiments are produced on two NVIDIA A6000 GPUs.

4.3 Evaluation Metrics

In accordance with previously established methodologies, we employ three distinct metrics: Fréchet Gesture Distance (FGD), Beat Consistency Score (BC), and Diversity.

Fréchet Gesture Distance (FGD). FGD is employed to quantify the divergence between the distribution of synthesized gesture and the actual data distribution. As delineated by (Yoon et al., 2020),

Methods	TED Gesture Ext			TED Expressive Ext		
	FGD ↓	BC ↑	Diversity ↑	FGD ↓	BC ↑	Diversity ↑
Ground Truth	0	0.695	107.214	0	0.711	184.641
Gesture VQ-VAE	0.186	0.695	107.188	0.174	0.727	188.548
Attention Seq2Seq (Yoon et al., 2019)	19.989	0.196	80.542	75.341	0.145	120.142
Speech2Gesture (Ginosar et al., 2019)	21.603	0.654	87.067	74.227	0.615	145.623
Joint Embedding (Ahuja and Morency, 2019)	26.771	0.213	81.561	79.523	0.149	118.324
Trimodal (Yoon et al., 2020)	8.374	0.653	101.667	18.744	0.510	148.624
HA2G (Liu et al., 2022)	5.595	0.660	103.303	6.85	0.621	169.352
DiffGesture (Zhu et al., 2023)	2.902	0.681	106.738	4.491	0.697	171.639
GesTran (Ours)	**1.874**	**0.692**	**107.207**	**2.854**	**0.714**	**183.188**

Table 3: The Quantitative Results on TED Gesture Ext and TED Expressive Ext. We compare the proposed GesTran against recent methods and ground truth.

Methods	GT	Seq2Seq.	Speech2Gesture	Joint.	Trimodal	HA2G	DiffGesture	**GesTran(Ours)**
Naturalness	4.35	1.32	1.56	2.73	3.22	3.51	3.72	**4.23**
Smoothness	4.11	3.37	2.61	3.14	3.27	3.59	3.71	**3.98**
Synchrony	4.23	2.17	1.82	3.19	3.28	3.54	3.87	**4.11**

Table 4: User Study Results. The ratings of motion naturalness, smoothness, and synchrony are 1-5, with 5 being the best.

the FGD is conceptualized through the development of an auto-encoder for the gesture sequence, designed to abstract the attributes of authentic gesture sequences, denoted as X, in addition to the characteristics of the artificially generated gesture sequences, referred to as $\hat{X}$.

$$\begin{aligned}\mathrm{FGD}(X, \hat{X}) = &\|\mu_r - \mu_g\|^2 \\ &+ \mathrm{Tr}(\Sigma_\mathrm{r} + \Sigma_\mathrm{g} - 2(\Sigma_\mathrm{r}\Sigma_\mathrm{g})^{1/2}),\end{aligned} \tag{4}$$

where μ_r and Σ_r are the first and the second moments of the latent feature distribution of the real gestures X, and μ_g and Σ_g are the first and the second moments of the latent feature distribution of the generated gestures $\hat{X}$.

Beat Consistency Score (BC). Proposed in (Li et al., 2021b, 2022), BC measures motion-audio beat correlation.

$$\mathrm{BC} = \frac{1}{n}\sum_{i=1}^{n}\exp(-\frac{\min_{\forall t_j^y \in B^y} \|t_i^x - t_j^y\|^2}{2\sigma^2}), \tag{5}$$

where t_i^x is the i-th audio beats, $B^y = \{t_i^y\}$ is the set of the kinematic beats, and σ is a parameter to normalize sequences, set to 0.1 empirically.

Diversity. This metric evaluates the variations among generated gestures (Lee et al., 2019). In detail, we randomly pick 500 generated samples and compute the mean absolute error between the features and the shuffled features.

4.4 Evaluation Results

Quantitative Results. We conducted a comprehensive comparison between our proposed method and all baseline approaches, evaluating their performance across three metrics on both TED Gesture and TED Expressive datasets. The results, presented in Table 1, highlight that our **GestureGPT** attains state-of-the-art performance across most metrics on both datasets, particularly showcasing a substantial superiority over existing methods in the case of TED Expressive.

To substantiate the robustness and generalization capability of our proposed technique, supplementary evaluations were executed on two distinct datasets, namely, TED Gesture Ext and TED Expressive Ext. Empirical findings reveal a notable deterioration in the performance of alternate methods when confronted with a substantial influx of samples in the test set that were absent during the training phase, while our methodology consistently maintains superior performance. This observation underscores the efficacy of the LLMs enhancement in fortifying the generalization aptitude of our method, rendering it markedly superior to its counterparts. Furthermore, it underscores our method's capacity to effectively retain and exploit the inherent knowledge encapsulated within the LLMs, thereby facilitating the generation of accurate and vivid gestures.

Qualitative Results. The qualitative results are

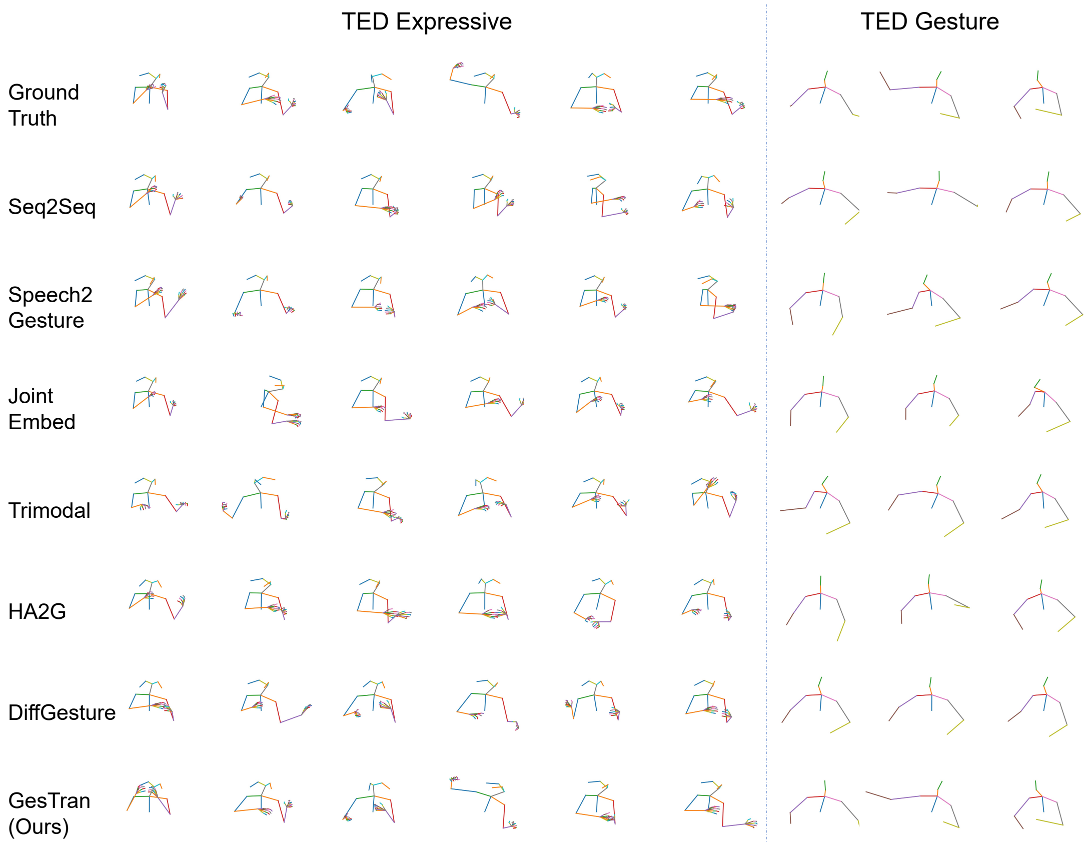

Figure 2: Visualization Results of Our GesTran on Two Benchmarks. Best view in color and zoom in.

		TED Gesture Ext			TED Expressive Ext		
Parameter	Pretrained	FGD ↓	BC ↑	Diversity ↑	FGD ↓	BC ↑	Diversity ↑
LLaMA-7B	✔	1.874	0.692	107.207	2.854	0.714	183.188
LLaMA-7B	✖	19.254	0.668	93.802	54.650	0.679	142.489
LLaMA-13B	✔	1.803	0.694	108.957	2.771	0.706	185.431
LLaMA-13B	✖	27.425	0.612	108.345	84.452	0.609	120.652

Table 5: Evaluation of different pre-trained LLaMA on TED Gesture Ext and TED Expressive Ext datasets.

illustrated in Figure 2. We can see that the results of GesTran are the most similar to GT. We also visualized in Figure 3 under some words with clear reference to the gesture. In the top of Figure 3, when the speaker says "five minutes", the gesture generated by GesTran is really consistent with the semantics. Gestran can understand the correct semantics and generate the corresponding gesture, which shows that GesTran indeed extracts the knowledge in the LLM to generate vivid gestures.

User Study. To meticulously authenticate the qualitative outcomes, a user study was conducted, emphasizing the synthesized co-speech gestures, and was steered in accordance with well-established methods (Liu et al., 2022; Zhu et al., 2023). This empirical investigation involved 28 respondents, an equal distribution of 14 males and 14 females, all within the demographic age bracket of 18-25 years. Responsibilities assigned to the participants included the adjudication of the quality and consistency of the generated movements, in scenarios devoid of labels. A total of 30 cases were procured for evaluation, of which 20 were dedicated to TED-Expressive and the remaining 10 to TED Gesture. Each case was represented through eight videos, which were rendered in a randomized sequence, inclusive of the ground truth. The mean opinion scores rating protocol was utilized, obligating participants to assess three distinct facets of the generated movements: *Naturalness*; *Smoothness*; *Synchrony between speech and generated gestures*. The outcomes, as delineated in Table 4, epitomize

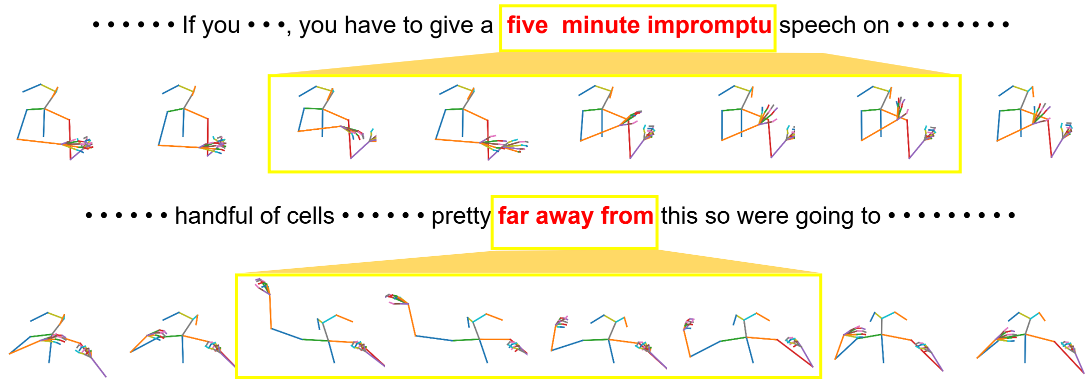

Figure 3: Visualization Results of Our GesTran on Two Benchmarks.

LLM architecture	TED Gesture Ext			TED Expressive Ext		
	FGD ↓	BC ↑	Diversity ↑	FGD ↓	BC ↑	Diversity ↑
LLaMA 7B (Touvron et al., 2023)	1.874	0.692	107.207	2.854	0.714	183.188
LLaMA 13B (Touvron et al., 2023)	1.803	0.694	108.957	2.771	0.706	185.431
T5 (Raffel et al., 2020)	2.386	0.681	105.08	3.550	0.689	177.199

Table 6: Evaluation of co-speech gesture generation using different backbone architectures.

ratings on a scale of 1 to 5, with 5 signifying an optimal rating. The empirical evidence suggests a predominant consensus among the participants establishing that our methodology is capable of delivering high-fidelity results.

4.5 Abtion Study

Effect of the numbers of model parameters. To further explore the impact of LLM capabilities on generated results, we conduct experiments using the LLaMA model with different numbers of parameters. The results are shown in Table 5. We can see that our method can achieve consistently superior performance when using different LLMs.

Effect of the pre-training. Pre-trained LLMs can provide robust priors about human motion from texts. In this context, we experiment with base models pre-trained to varying degrees, *i.e.*, LLaMA-7B, LLaMA-13B, and LLaMA without pre-training. For the un-pretrained LLaMA, we adopt LLaMA-7B without loading the pre-trained weights. The randomly initialized LLaMA is tuned by LoRA as well, fixing weights during training. As shown in Table 5, there exists a strong correlation between the level of pre-training in LLMs and the performance of our model. This highlights the significant influence of gesture prior extracted from LLM.

Effect of different model architecture. In order to explore the impact of different LLM architectures on the results, we also conducted experiments with T5 (Raffel et al., 2020) as the backbone. The experimental results are shown in Table 6. We can see that our method can achieve consistently superior performance when using different LLMs.

5 Conclusion

In this paper, we introduce an innovative gesture translation technique, termed GesTran, that capitalizes on the capabilities of Large Language Models (LLMs) to enhance the comprehension of the intricate relationship between verbal and non-verbal communication. This is accomplished by sequentially generating human gestures, thereby interpreting them as a distinct mode of body language. The initial phase of the proposed architecture incorporates a transformer-based auto-encoder network to transcribe human gestures into discrete symbolic representations. Subsequently, the succeeding phase exploits a pre-trained LLM, aiming to decipher the interplay between speech and gesture. This is achieved by transforming the verbal input into corresponding gestures, thereby interpreting the gestures as unique language tokens within the LLM's context. Through a series of rigorous experiments, conducted on two universally acknowledged datasets, consistent evidence of the superior performance of our proposed approach was observed across almost all evaluative metrics. Consequently, this strongly corroborates the efficacy of the method introduced in this study.

Limitations

Our methodology, having been trained solely on English data, is currently limited to generating gestures pertaining to English speakers and lacks the capacity to adapt to a broader spectrum of languages.

Acknowledgments

Our work is supported in part by the National Key R&D Program of China (No. 2023YFC3305600), National Natural Science Foundation of China (62132016 and 62302372), and Fundamental Research Funds for the Central Universities (ZDRC2102 and XJSJ23036).

References

Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2pose: Natural language grounded pose forecasting. In *3DV*, pages 719–728. IEEE.

Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. 2022. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Trans. Graph.*, 41(6):1–19.

Tenglong Ao, Zeyi Zhang, and Libin Liu. 2023. Gesturediffuclip: Gesture diffusion model with clip latents. *arXiv*.

Michael Büttner and Simon Clavet. 2015. Motion matching-the road to next gen animation. *Proc. of Nucl. ai*, 1(2015):2.

Justine Cassell, David McNeill, and Karl-Erik McCullough. 1999. Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & cognition*, 7(1):1–34.

Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420.

Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predicting co-verbal gestures: A deep and temporal modeling approach. In *IVA*, pages 152–166. Springer.

Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. 2020. Monocular expressive body regression through body-driven attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 20–40. Springer.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv*.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *CVPR*, pages 3497–3506.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv*.

Chien-Ming Huang and Bilge Mutlu. 2012. Robot behavior toolkit: generating effective social behaviors for robots. In *HRI*, pages 25–32.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. Motiongpt: Human motion as a foreign language. *NeurIPS*, 36.

Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021. A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020. In *26th international conference on intelligent user interfaces*, pages 11–21.

Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. 2019. Dancing to music. *NeurIPS*, 32.

Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. 2022. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *AAAI*, pages 1272–1279.

Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 2021a. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *ICCV*, pages 11293–11302.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021b. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, pages 13401–13412.

Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. 2022. Learning hierarchical cross-modal association for co-speech gesture generation. In *CVPR*, pages 10462–10472.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv*.

Kunkun Pang, Dafei Qin, Yingruo Fan, Julian Habekost, Takaaki Shiratori, Junichi Yamagishi, and Taku Komura. 2023. Bodyformer: Semantics-guided 3d body gesture synthesis with transformer. *ACM Trans. Graph.*, 42(4):1–12.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(1):5485–5551.

Jose Ribeiro-Gomes, Tianhui Cai, Zoltán A Milacski, Chen Wu, Aayush Prakash, Shingo Takagi, Amaury Aubel, Daeil Kim, Alexandre Bernardino, and Fernando De La Torre. 2024. Motiongpt: Human motion synthesis with improved diversity and realism via gpt-3 prompting. In *WACV*, pages 5070–5080.

Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. 2012. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4:201–217.

Maha Salem, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2011. A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In *2011 ro-man*, pages 247–252. IEEE.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv*.

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

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *NeurIPS*, 30.

Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview.

Chenghao Xu, Jiexi Yan, Yanhua Yang, and Cheng Deng. 2023. Implicit compositional generative network for length-variable co-speech gesture synthesis. *IEEE Trans. Multimed.*

Jiexi Yan, Cheng Deng, Heng Huang, and Wei Liu. 2024. Causality-invariant interactive mining for cross-modal similarity learning. *IEEE Trans. Pattern Anal. Mach. Intell.*

Jiexi Yan, Lei Luo, Chenghao Xu, Cheng Deng, and Heng Huang. 2022a. Noise is also useful: Negative correlation-steered latent contrastive learning. In *CVPR*, pages 31–40.

Jiexi Yan, Erkun Yang, Cheng Deng, and Heng Huang. 2022b. Metricformer: A unified perspective of correlation exploring in similarity learning. *NeurIPS*, 35:33414–33427.

Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, and Haolin Zhuang. 2023. Qpgesture: Quantization-based and phase-guided motion matching for natural speech-driven gesture generation. In *CVPR*, pages 2321–2330.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Trans. Graph.*, 39(6):1–16.

Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *ICRA*, pages 4303–4309. IEEE.

Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. 2024. Motiongpt: Finetuned llms are general-purpose motion generators. In *AAAI*, volume 38, pages 7368–7376.

Yihao Zhi, Xiaodong Cun, Xuelin Chen, Xi Shen, Wen Guo, Shaoli Huang, and Shenghua Gao. 2023. Livelyspeaker: Towards semantic-aware co-speech gesture generation. In *ICCV*, pages 20807–20817.

Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. 2023. Taming diffusion models for audio-driven co-speech gesture generation. In *CVPR*, pages 10544–10553.