Speech-driven Personalized Gesture Synthetics: Harnessing Automatic Fuzzy Feature Inference

Fan Zhang, Zhaohan Wang, Xin Lyu, Siyuan Zhao, Mengjian Li, Weidong Geng, Naye Ji (✉), Hui Du, Fuxing Gao, Hao Wu, Shunman Li

***Abstract*—Speech-driven gesture generation is an emerging field within virtual human creation. However, a significant challenge lies in accurately determining and processing the multitude of input features (such as acoustic, semantic, emotional, personality, and even subtle unknown features). Traditional approaches, reliant on various explicit feature inputs and complex multimodal processing, constrain the expressiveness of resulting gestures and limit their applicability. To address these challenges, we present *Persona-Gestor*, a novel end-to-end generative model designed to generate highly personalized 3D full-body gestures solely relying on raw speech audio. The model combines a fuzzy feature extractor and a non-autoregressive Adaptive Layer Normalization (AdaLN) transformer diffusion architecture. The fuzzy feature extractor harnesses a fuzzy inference strategy that automatically infers implicit, continuous fuzzy features. These fuzzy features, represented as a unified latent feature, are fed into the AdaLN transformer. The AdaLN transformer introduces a conditional mechanism that applies a uniform function across all tokens, thereby effectively modeling the correlation between the fuzzy features and the gesture sequence. This module ensures a high level of gesture-speech synchronization while preserving naturalness. Finally, we employ the diffusion model to train and infer various gestures. Extensive subjective and objective evaluations on the Trinity, ZEGGS, and BEAT datasets confirm our model's superior performance to the current state-of-the-art approaches. *Persona-Gestor* improves the system's usability and generalization capabilities, setting a new benchmark in speech-driven gesture synthesis and broadening the horizon for virtual human technology. Supplementary videos and code can be accessed at https://zf223669.github.io/Diffmotion-v2-website/.**

***Index Terms*—Speech-driven, Gesture synthesis, Fuzzy inference, AdaLN, Diffusion, Transformer.**

Fan Zhang is with the Faculty of Humanities and Arts, Macau University of Science and Technology, Macau, China; The College of Media Engineering, Communication University of Zhejiang, China; Research Center for Artificial Intelligence and Fine Arts, Zhejiang Lab, Zhejiang, China (e-mail: fanzhang@cuz.edu.cn)

Zhaohan Wang, Xin Lyu are with the School of Animation and Digital Arts Communication University of China, Beijing, China (e-mail: 2022201305j6018@cuc.edu.cn; lvxinlx@cuc.edu.cn)

Mengjian Li is with the Research Center for Artificial Intelligence and Fine Arts, Zhejiang Lab, Zhejiang, China(e-mail: limengjian@zhejianglab.com)

Weidong Geng is with the College of Computer Science and Technology, Zhejiang University, the Research Center for Artificial Intelligence and Fine Arts, Zhejiang Lab, Zhejiang, China (e-mail: gengwd@zju.edu.cn)

Naye Ji, Hui Du, Fuxing Gao, Hao Wu are with the College of Media Engineering, Communication University of Zhejiang, China (e-mail: jinaye@cuz.edu.cn; fuxing@cuz.edu.cn; duhui@cuz.edu.cn; 210207140@stu.cuz.edu.cn)

Siyuan Zhao is with the Faculty of Humanities and Arts, Macau University of Science and Technology, Macau, China(e-mail: 2109853jai30001@student.must.edu.mo)

Shunman Li is with Zhejiang Institute of Economics and Trade, Zhejiang, China (e-mail: 2017000018@zjiet.edu.cn)

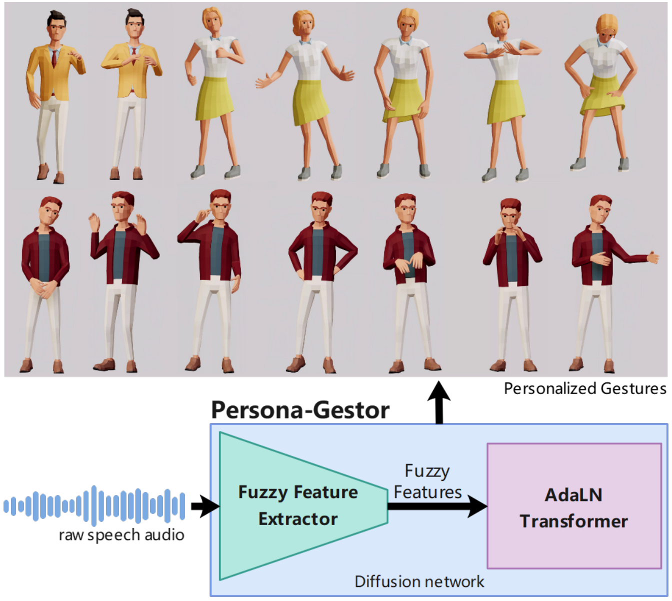


Fig. 1: Each pose depicted is personalized gestures generated solely relying on raw speech audio. Persona-Gestor offers a versatile solution, bypassing complex multimodal processing and thereby enhancing user-friendliness.

I. Introduction

RECENT advancements have significantly expanded the use of 3D virtual human technology. Its growing appeal spans numerous applications, including animation, gaming, digital receptionists, and human-computer interaction. A major task in this research area is to create credible, personalized co-speech gestures. Speech-driven gesture generation through deep learning provides a cost-effective solution, eliminating the need for manual intervention associated with conventional motion capture systems.

However, the primary challenges in speech-driven gesture generation face precisely identifying the vast array of input conditions necessary for driving gesture synthesis. This complexity arises because co-speech gestures are shaped by an extensive range of factors, including acoustics, semantics, emotions, personality traits, and demographic variables like gender, age, etc.

Previous approaches [1]–[7] have explored the use of man-

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3393236

ual labels and diverse feature inputs to facilitate the synthesis of personalized gestures. Nonetheless, these methods depend heavily on various unstructured feature inputs and require complex multimodal processing. These approaches present a significant barrier to the practical application and broader adoption of virtual human technologies.

The fuzzy inference strategy, which pertains to the concept of fuzzy logic [8], is particularly useful in the field for dealing with uncertain or imprecise information. The fuzzy inference strategy is known for its effectiveness in speech-emotion recognition [9] and audio classification [10]. These methods do not necessarily require explicit classification outputs but instead provide fuzzy feature information, which broadens the explicit discrete space into an expansive implicit continuous fuzzy space. The information in fuzzy space better aligns with the actual scenario. Relevantly, research in psychology highlights the significance of various factors in speech [11]–[14]. These factors, called fuzzy features, are intricately intertwined with co-speech gestures. These studies present novel opportunities for synthesizing personalized gestures based solely on speech audio, thereby simplifying the feature inputs and reducing the complexity of multimodal processing.

Another challenge in this field is ensuring a high level of gesture-speech synchronization while preserving naturalness. Recent developments have focused on the application of Transformer and Diffusion-based models. This methodological shift has led to substantial progress in the efficiency and flexibility of gesture-generation technologies. Key examples of such innovative efforts include Taming [15], Diffuse Style Gesture [2], Diffuse Style Gesture+ [3], GestureDiffuClip [16], and LDA [4]. Yet, these approaches encounter challenges with either insufficient or excessive correlation between gesture and speech, reducing the naturalness of the generated gestures.

The success of the Diffusion Transformers(DiTs) in text2image [17] and text2video generation tasks, such as Sora [1], which incorporates AdaLN, marks a significant advancement. This framework introduces a conditional mechanism that applies a uniform function across all tokens, enhancing the model's ability to represent conditional and output features. This conditional mechanism also holds promise for effectively enhancing the ability to model the intricate mapping between speech and gestures. While the original DiTs take discrete text prompts as conditional inputs, its adaptability for sequence-to-sequence tasks, such as speech-driven gesture generation, presents an area of exploration.

In this study, we propose *Persona-Gestor*, a novel approach aimed at synthesizing personalized gestures solely from raw speech audio. This model innovatively introduces a fuzzy feature inference strategy within its condition extractor and incorporates AdaLN in a diffusion-based transformer module. *Persona-Gestor* transitions from explicit conditions to a nuanced, continuous representation of fuzzy features by employing fuzzy inference, which captures a broad spectrum of stylistic nuances and specific audio details. These features are integrated into a unified latent representation, synthesizing intricate 3D full-body gestures. Adopting AdaLN significantly enhances the model's capability to depict the nuanced relationship between speech and gestures. Leveraging a diffusion process, the framework can generate diverse gesture outputs, showcasing the potential for high fidelity in gesture synthesis.

For clarity, our contributions are summarized as follows:

- **We pioneering introduce the fuzzy feature inference strategy that enables driving a wider range of personalized gesture synthesis from speech audio alone, removing the need for style labels or extra inputs.** This fuzzy feature extractor improves the usability and the generalization capabilities of the system. To the best of our knowledge, it is the first approach that uses fuzzy features to generate co-speech personalized gestures.
- **We combined AdaLN transformer architecture within the diffusion model to enhance the Modeling of the gesture-speech interplay.** We demonstrate that this architecture can generate gestures that achieve an optimal balance of natural and speech synchronization.
- **Extensive subjective and objective evaluations reveal our model superior outperform to the current state-of-the-art approaches.** These results show the remarkable capability of our method in generating credible, speech-appropriateness, and personalized gestures.

II. Related Work

The present discussion offers a succinct overview of the conditional extraction mechanism and generative models within speech-driven gesture generation.

A. Condition Extraction Mechanism

Recent advancements in co-speech gesture generation systems have incorporated various unstructured conditional information as input.

Selecting optimal representations for conditional input is a crucial research challenge in creating virtual human motions [18] [19]. For accurate reflection of gestures that match the auditory perception, prevalent research [7], [20]–[22] utilizes preprocessed audio features, such as MFCCs, log amplitude spectrogram, etc. Li et al. [6] develop a model for direct audio-to-gesture mapping. Despite these methods capturing acoustic nuances, the quest for richer feature sets continues. This has prompted investigations into the WavLM model, a refined, pre-trained wav2vec framework, for enhanced speech extraction, showcasing in ReprGesture [23], QPGesture [24], and DiffuseStyleGesture [2].

Text-based co-speech gesture synthesis has seen significant contributions, such as Yoon et al.'s [25] recurrent neural network approach and Taras et al.'s [26] system, which merges acoustic and semantic speech features, employing BERT for semantic analysis [27]. Additionally, Uttaran et al. [28] utilize GloVe embeddings [29] to surpass models of similar dimensions, like Word2Vec [30] and FastText [31]. Merging acoustic with semantic data offers a valuable path to enhance the relevance and context of generated gestures. Nonetheless, these modalities' manual alignment and integration pose a challenge in effectively superior gesture synthesis.

[1] https://openai.com/sora

For creating style-specific gestures, ReprGesture [23] and QPGesture [24] integrate textual data with audio features, whereas DiffuseStyleGesture [2] employs discrete labels to influence the stylistic aspects of the gestures produced. LDA [4] enables the system to generate style gestures with classifier-free guidance. Additionally, recent research has explored using textual prompts to generate stylized gestures [16]. Given that human emotions are more accurately represented on a continuous spectrum [32] [33] and emerge from a complex interplay of fuzzy factors, depending on discrete emotion labels can overly simplify the gesture generation process. This could limit the expressiveness and subtlety of the produced gestures. To address these limitations, Ghrobani et al. [7] introduced ZeroEGGS, a model that utilizes example motion clips to guide the style of gestures. Although achieving zero-shot is feasible, it still necessitates sample animation clips.

B. Generative approaches

DiffMotion [22], is the pioneering application of diffusion models integrating an LSTM for the synthesis of diverse gestures. UnifiedGesture [5] presents a retargeting network to learn latent homeomorphic graphs to homeomorphic graphs for various gesture representations. Maximizing the transformer architecture's potential, Alexanderson et al. [4] enhanced DiffWave by replacing dilated convolutions. Conformers [34] implementing classifier-free guidance to improve style expression. GestureDiffuCLIP [16] propose a network based on the transformer and AdaIN layers to incorporate style guidance into the diffusion model. LivelySpeaker [35] depends on contrastive learning to create a joint embedding space between gestures and transcripts. DiffuseStyleGesture (DSG) [2] and DSG+ [3], integrating cross-local attention and layer normalization within transformers. Conversely, these methodologies face difficulties in achieving an optimal balance between gesture and speech synchronization, resulting in gestures that may appear either underrepresented or overly matched.

In this study, we employ a fuzzy feature inference strategy to implicitly capture fuzzy features in speech audio, synthesizing natural, personalized co-speech gestures solely relying on raw speech audio without additional modalities. Furthermore, we employ an AdaLN transformer architecture to enhance the model's capacity to capture the intricate relationship between speech and gestures.

III. System Overview

Persona-Gestor, as an end-to-end architecture, processes raw speech audio as its sole input, synthesizing personalized gestures that adeptly balance naturalness with synchronized alignment to speech.

A. Problem Formulation

We introduce the challenge of co-speech gesture generation by framing it as a sequence-to-sequence problem, where the objective is to translate a sequence of speech audio features into a corresponding sequence of gestures. We denote the sequence of full-body gesture features and the sequence of the audio signal as $g^0 = g^0_{1:T} \in [g^0_1, ..., g^0_t, ..., g^0_T] \in \mathbb{R}^{T\times(D+3+3)}$ and $a = a_{1:T} \in [a_1, ..., a_t, ..., a_T] \in \mathbb{R}^T$. $g^0_t = \mathbb{R}^{(D+3+3)}$ symbolizes the representation of 3D joint angles, along with the root positional and rotational velocity at frame t, where D denoting the number of channels for these joints. The superscript indicates the diffusion time step n. Here, a_t refers to the current subsequence audio waveform signal at frame t, while T denotes the sequence length. Let us define $p_\theta(\cdot)$ as the Probability Density Function (PDF), which aims to approximate the actual distribution of gesture data $p(\cdot)$ and enables easy sampling. The objective is to generate a non-autoregressive whole pose sequence (g^0) from its conditional probability distribution given audio signal (a) as covariate:

$$g^0 \sim p_\theta\left(g^0|a\right) \approx p(\cdot) := p\left(g^0|a\right) \tag{1}$$

where the $p_\theta(\cdot)$ aims to approximate $p(\cdot)$ trained by the denoising diffusion model.

B. Model Architecture

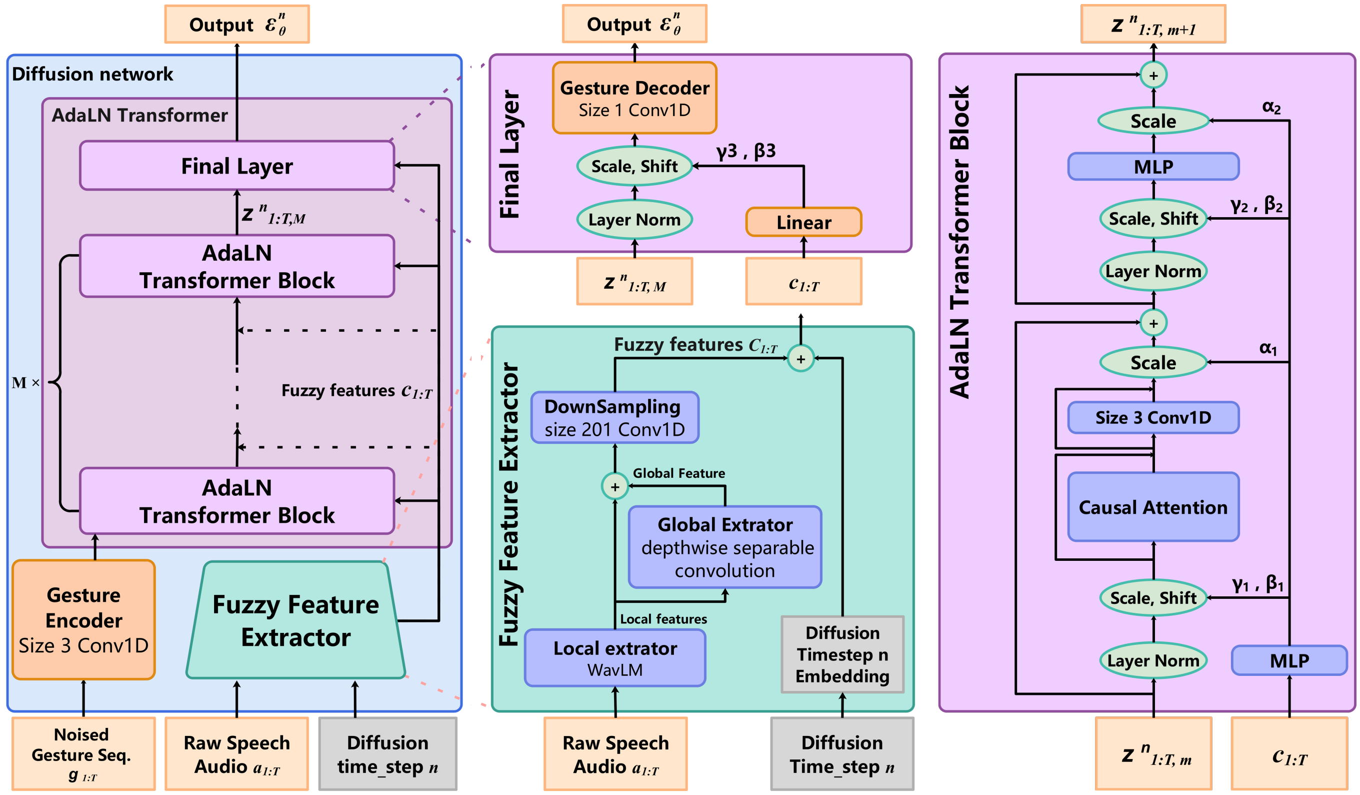
The architecture of Persona-Gestor is depicted in Figure 2. It comprises four primary components: (1) a Fuzzy Feature Extractor, (2) an AdaLN Transformer, (3) a Gesture Encoder and Decoder, and (4) a diffusion network.

1) Fuzzy Feature Extractor: This module utilizes a fuzzy inference strategy, meaning it does not generate explicit classification outputs. Instead, it offers implicit, continuous, fuzzy feature information, automatically learning and inferring the global style and details directly from raw speech audio. The module, showcased in Figure 2b and Figure 3, is a dual-component extractor that integrates both global and local extractors. The local extractor leverages the WavLM large-scale pre-trained model [36] to convert the audio sequence into tokens. We chose WavLM for its adeptness at extracting the complex features of speech audio to capture universal audio latent representations, denoted as z_a.

We observe that the local extractor alone falls short of fully capturing the array of stylistic features and ensuring style consistency across sequences. To overcome this, we integrate a global style extractor, employing a depthwise separable convolution 1D layer [37] across the z_a. This global extractor is designed to automatically capture and embed global fuzzy style information from z_a into a token $z_s \in \mathbb{R}^{1\times D'}$. This token is then broadcasted and combined with the universal audio latent representations $z_a \in \mathbb{R}^{T'\times D'}$ to form a unified latent representation $z_l \in \mathbb{R}^{T\times D''}$. We enhance the sequence's overall representational fidelity by merging local and global insights for co-speech gesture generation. Subsequently, the unified latent representation is directed to the downsampling module for further processing.

The downsampling module is integrated into the condition extractor to ensure alignment between each latent representation and its corresponding sequence of encoded gestures. In our exploration, we experimented with linear alignment like DSG [2] and DSG+ [3], but noted an issue of foot-skating arising from these methods. On the contrary, We adopt a Conv1D layer with a kernel size of 201 for this module

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3393236



(a) Overall schematic

(b) Fuzzy Feature Extractor (below), Final Layer in AdaLN Transformer (above)

(c) AdaLN Transformer Block

Fig. 2: The Architecture of Persona-Gestor mainly integrates a fuzzy feature extractor and an adaptive layer normalization (AdaLN) transformer diffusion architecture. The fuzzy feature extractor comprises a dual-component framework to comprehensively capture the fuzzy style and detail-oriented audio features. These features, as unified latent features, are subsequently fed into the AdaLN transformer to model the relationship with the accompanist gesture, facilitating the estimation of diffusion noise for the diffusion model. (a) Overall Schematic. (b) Fuzzy Feature Extractor. (c) AdaLN Transformer Block.

that maps every 201-length target token output from WavLM to one gesture frame. Finally, the fuzzy feature extractor outputs $c_{1:T}$, representing a unified latent representation that combines encoded audio features and diffusion time step n. The condition extractor can be formalized by:

$$\begin{aligned}
z_a &= LE(a) & z_a &\in \mathbb{R}^{T' \times D'} \\
z_s &= GE(z_a) & z_s &\in \mathbb{R}^{1 \times D'} \\
z_l &= DS(z_a + z_s) & z_l &\in \mathbb{R}^{T \times D''} \\
n' &= DTE(n) & n &\in \mathbb{R}, \quad n' \in \mathbb{R}^{1 \times D''} \\
c_{1:T} &= z_l + n' & c_{1:T} &\in \ \mathbb{R}^{T \times D''}
\end{aligned} \tag{2}$$

Where $LE(\cdot)$ and $GE(\cdot)$ denote the local extractor (WavLM) and the global extractor. $DS(\cdot)$ represents the down sampling process. $DTE(\cdot)$ signifies the diffusion time step embedding. The final output of the fuzzy feature extractor is denoted as $c_{1:T}$. Here, T', D', and D'' refer to the WavLM output token length, feature dimensionality of WavLM's output token, and the feature (h) dimensionality of the proposed model's hidden state, respectively. a is the input raw speech audio waveform. Z_a and Z_s are extracted by the local extractor(WavLM model) and the global extractor. Z_l is the unified latent representation. n is the diffusion time step, n' is the embedded diffusion time step feature.

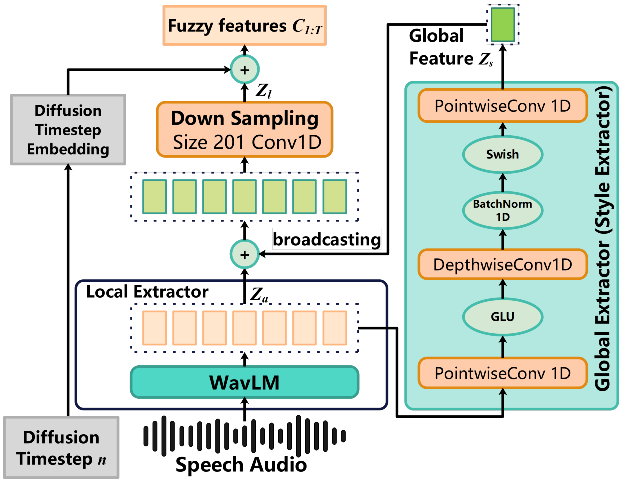


Fig. 3: An overview of the fuzzy inference condition extractor.

2) AdaLN Transformer: The AdaLN's fundamental purpose is to incorporate a conditional mechanism that uniformly applies a specific function across all tokens, thereby significantly improving the model's capacity for representing both conditional and output features with enhanced efficiency. It offers a more sophisticated and nuanced approach to modeling, enabling the system to capture and articulate the complex dynamics between various input conditions and their corresponding outputs. Consequently, this leads to an improvement in the model's predictive accuracy and its ability to generate outputs that are more aligned with the given conditions.

Diffusion Transformers (DiTs) [17] represent an advanced transformer-based backbone for diffusion models, surpassing previous U-Net models in performance. By incorporating AdaLN within transformer blocks for text-to-image synthesis, DiTs achieve lower Fréchet Inception Distance (FID) [38] scores, indicating superior image quality. Recently, this framework has been used for text-conditional video generation. Despite Diffusion Transformers (DiTs) success in handling discrete text prompts conditional inputs, their effectiveness in speech-driven gesture generation, a sequence-to-sequence task, necessitates a thorough investigation.

Distinctively with the DiTs, our approach utilizes continuous fuzzy features as conditional input tokens. Further, it is without any patchy for spatial input, resulting in the output being the latent feature of a sequence of gestures.

The module involves regressing the dimensionwise scale and shift parameters (γ and β), which are derived from the fuzzy feature extractor output $c_{1:T}$, instead of directly learning γ and β, as depicted in Figure 2c. In each AdaLN transformer, a latent feature denoted as $z^n_{1:T,m}$ is generated by fusing condition information and gesture using AdaLN and causal self-attention. Here, $1 \leq m \leq M$, where M represents the total number of AdaLN transformer stacks. In addition, the final layer, as illustrated in Figure 2b, fed the same fuzzy features but with additional scale and shift operation.

This method facilitates the creation of detailed gesture sequences solely from speech audio, eliminating the requirement for discrete style labels or supplementary inputs. Consequently, it significantly improves the model's capacity to generate personalized and closely aligned gestures with the context of the speech, offering a more refined and context-sensitive gesture synthesis capability.

3) Gesture Encoder and Decoder: The architecture of the gesture encoder and decoder is designed to encode and decode the gesture sequence, as illustrated in Fig.2a and Fig.2b. The gesture encoder comprises a Convolution1D with a kernel size of 3. It encodes the initial sequence of gestures g into a hidden state $h \in \mathbb{R}^{T \times D''}$. Our experimental results revealed that employing a kernel size of 1 resulted in animation jitter. Conversely, a kernel size of 3 is instrumental in mitigating this issue by effectively capturing the spatial-temporal relationships inherent in gesture sequences.

The gesture decoder reduces the feature dimension of the output from the transformer D'' to the original dimension D, corresponding to the number of channels representing skeleton joints. Result in outputting the predicted noise (ϵ_θ). We utilize a size of 1 convolution1D By convolving a 1D kernel with each position in the input sequence, our model can effectively extract meaningful features and relationships between adjacent joint channels.

C. Training and Inferencing with Denoising Diffusion Probabilistic Model

The diffusion process in this architecture aims to reconstruct the conditional probability distribution between gestures and fuzzy features. This entails employing a systematic approach to sample from this restored distribution, thereby enabling the generation of diverse gestures.

Following our previous work, Diffmotion [22], incorporating the Denoising Diffusion Probabilistic Model (DDPM) into our approach. However, we employ a non-autoregressive transformer to generate the entire sequence of gestures instead of frame-by-frame. The form is represented by $p_\theta := \int p_\theta\left(g^{0:N}\right) dg^{1:N}$, where $g^1, ..., g^N$ are latent of the same dimensionality as the data g^n at the n-th diffusion time stage.

The model contains two processes: the diffusion process and the generation process. At training time, the diffusion process gradually converts the original gesture data(g^0) to white noise(g^N) by optimizing a variational bound on the data likelihood. At inference time, the generation process recovers the data by reversing this noising process through the Markov chain using Langevin sampling [39]. The Markov chains in the diffusion process and the generation process are:

$$\begin{aligned} &p\left(g^n|g^0\right) = \mathcal{N}\left(g^n; \sqrt{\overline{\alpha}^n} g^0, \left(1-\overline{\alpha}^n\right) I\right) \quad and \\ &p_\theta\left(g^{n-1}|g^n, g^0\right) = \mathcal{N}\left(g^{n-1}; \tilde{\mu}^n\left(g^n, g^0\right), \tilde{\beta}^n I\right), \end{aligned} \tag{3}$$

where $\alpha^n := 1 - \beta^n$ and $\overline{\alpha}^n := \prod_{i=1}^n \alpha^i$. As shown by [40], β^n is a increasing variance schedule $\beta^1, ..., \beta^N$ with $\beta^n \in (0, 1)$, and $\tilde{\beta}^n := \frac{1-\overline{\alpha}^{n-1}}{1-\overline{\alpha}^n}\beta^n$.

The training objective is to optimize the parameters θ that minimizes the Negative Log-Likelihood (NLL) via Mean Squared Error (MSE) loss between the true noise $\epsilon \sim \mathcal{N}(0, I)$ and the predicted noise ϵ_θ:

$$\mathbb{E}_{g^0_{1:T},\epsilon,n}[||\epsilon - \epsilon_\theta\left(\sqrt{\overline{\alpha}^n} g^0 + \sqrt{1-\overline{\alpha}^n}\epsilon, a_{1:T}, n\right)||^2], \tag{4}$$

Here ϵ_θ is a neural network (see figure 2a), which uses input g^0_t , a_{t-1} and n that to predict the ϵ, and contains the similar architecture employed in [41]. The complete training procedure is outlined in Algorithm 1.

Algorithm 1: Training for the whole sequence gesture

```
Input: data $g^0_{1:T} \sim p\left(g^0|a_{1:T}\right)$ and $a_{1:T}$
repeat
    Initialize $n \sim$ Uniform$(1, ..., N)$ and $\epsilon \sim \mathcal{N}(0, I)$
    Take the gradient step on
        $\nabla_\theta||\epsilon - \epsilon_\theta\left(\sqrt{\overline{\alpha}_n} g^0_{1:T} + \sqrt{1-\overline{\alpha}_n}\epsilon, a_{1:T}, n\right)||^2$
until converged;
```

After training, we utilize variational inference to generate the whole sequence of new gestures matching the original data distribution($g^0_t \sim p_\theta\left(g^0_t, a_t\right)$). We followed the sampling procedure in Algorithm 2 to obtain a sample g_t^0 of the current frame. The σ_θ is the standard deviation of the $p_\theta\left(g^{n-1}|g^n\right)$. We choose $\sigma_\theta := \tilde{\beta}^n$.

Algorithm 2: Sampling $g^0_{1:T}$ via annealed Langevin dynamics

```
Input: noise $g^N_{1:T} \sim \mathcal{N}(0, I)$ and raw audio waveform
        $a_{1:T}$
for $n = N$ to 1 do
    if $n > 1$ then
        $z \sim \mathcal{N}(0, I)$
    else
        $z = 0$
    end if
    $g^{n-1}_{1:T} = \frac{1}{\sqrt{\alpha^n}}\left(g^n_{1:T} - \frac{\beta^n}{\sqrt{1-\bar{\alpha}^n}}\epsilon_\theta\left(g^n_{1:T}, a_{1:T}, n\right)\right) + \sqrt{\sigma_\theta}z$
end for
Return: $g^0_{1:T}$
```

During inferencing, we send the whole sequence of the raw audio to the condition extractor component. Then, the component output is fed to the Diffusion Model to generate the whole sequence of the accompanying gesture(g^0).

IV. Experiments

To validate our approach, we utilized three co-speech gesture datasets (Trinity [42], ZEGGS [7], and BEAT [43]). Our experiments concentrated on producing full 3D body gestures (including finger motions and locomotion). This choice presented a greater challenge than merely focusing on upper body motions due to the expanded output dimensionality and the need to overcome visual complexities, such as foot-skating, the naturalness of finger movements, and locomotion.

A. Dataset and Data Processing

1) Datasets: The Trinity dataset focuses on individual spontaneous speech, the ZEGGS dataset encompasses a wide range of emotional expressions, and the BEAT dataset consists of personalized movements exhibited by various individuals. Further details are elaborated in TableII found in Appendix A.

2) Speech Audio Data Process: In the Trinity dataset, the audio was recorded at a sampling rate of 44 kHz, while 48 kHz in ZEGGS and BEAT. However, due to the pre-training of the WavLM large model on speech audio sampled at 16 kHz, we uniformly resample all audio to match this frequency.

3) Gesture Data Process: We focus solely on full-body gestures, adopting the data processing techniques outlined by Alexanderson et al. [20]. Given the variability in data quality and structure across motion datasets, we tailor our approach by selecting specific joints for analysis in each dataset. We omit hand skeleton data for the Trinity Gesture Dataset due to its inferior quality. For ZEGGS and BEAT datasets, our analysis includes finger joints and the same set of joints considered in the Trinity dataset. All data capture translational and rotational velocities to detail the root's trajectory and orientation. The datasets are uniformly downsampled to a frame rate of 20 fps. To ensure accurate and continuous representation of joint angles, we apply the exponential map technique [44]. All data are segmented into 20-second clips for training and validation purposes. As for the user evaluation, we segment the generated gesture sequence into 10 seconds to improve the efficiency of the evaluation.

B. Model Settings

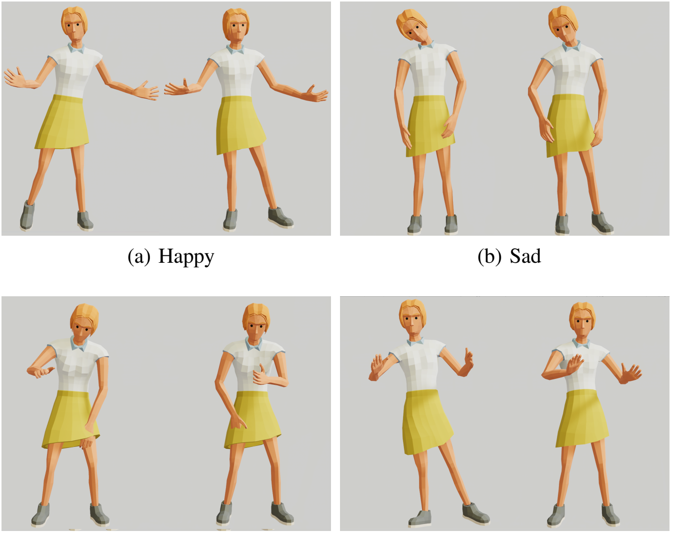
Our experiments employ 12 causal attention blocks, each comprising 16 attention heads (as depicted in Figure 2a). The encoding process transforms each frame of the gesture sequence into hidden states $h \in \mathbb{R}^{1280}$. For the WavLM model, we utilize the pre-trained WavLM Large model [2]. To ensure temporal translation invariance, we employ a translation-invariant self-attention (TISA) mechanism [45].

The quaternary variance schedule of diffusion model starts from $\beta_1 = 1 \times 10^{-4}$ till $\beta_N = 5 \times 10^{-5}$ with linear beat schedule. The number of diffusion steps is $N = 1000$. The training batch size is 32 per GPU.

The model was developed using the Torch Lightning framework and tested on an Intel i9 processor with an A100 GPU. Training durations were approximately 4 hours for Trinity and ZeroEGGS and 21 hours for BEAT.

C. Visualization Results

Our system excels in creating personalized gestures that align with the speech context, leveraging the fuzzy inference strategy to autonomously derive fuzzy features directly from speech audio. Furthermore, it showcases remarkable generalization and robustness by utilizing in-the-wild speech.



(a) Happy

(b) Sad

(c) Old

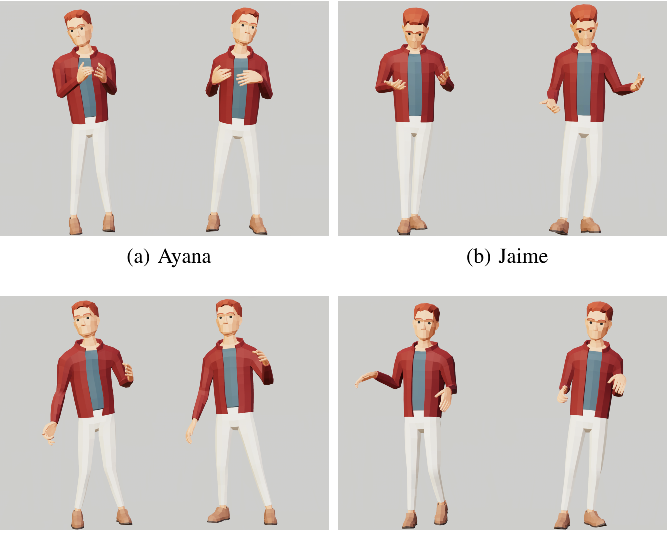
(d) Speech

Fig. 4: Samples of gestures corresponding to different emotions. The left side of the subfigure displays ground truth gestures, while the right side showcases gestures generated by our architecture.

[2] https://github.com/microsoft/unilm/tree/master/wavlm

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2024.3393236

Figure 4 depicts the visual outcomes of gestures aligned with the emotional valence conveyed by the audio. For instance, the system produces gestures of joy in response to happy audio cues (refer to Figure 4a) and gestures of sadness for sorrowful audio (as depicted in Figure 4b). The system can also infer age-related characteristics or other nuanced states from the speech audio (as illustrated in 4c and 4d).



(a) Ayana

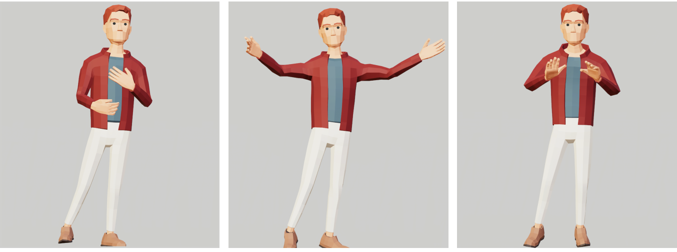
(b) Jaime

(c) Luqi

(d) Zhao

Fig. 5: Samples of gestures corresponding to different personalities. The left side of the subfigure displays ground truth gestures, while the right side showcases gestures generated by our architecture.

Figure 5 shows the system's ability to generate personalized gestures Predicated upon individuals' unique speech traits. For example, Ayana's gestures, with hands together and palms facing, denote reserved expressiveness. In contrast, Jaime's "palm up" gestures imply openness, and Luqi's alternating hand movements add dynamic variability. These results highlight the system's adeptness at depicting a wide range of personality-specific gestures.



(a) ...with my mother...

(b) ...1400 miles away...

(c) ...I´m not saying...

Fig. 6: Samples of gestures corresponding to semantic.

Interestingly, as shown in Figure 6, the system can produce gestures with certain semantic relevance even in the absence of explicit semantic constraints. For instance, Carla's remark about her mother is matched with a self-referential gesture. Likewise, Lawrence's reference to distance is visually enhanced by a gesture that emphasizes the semantic essence of his speech.

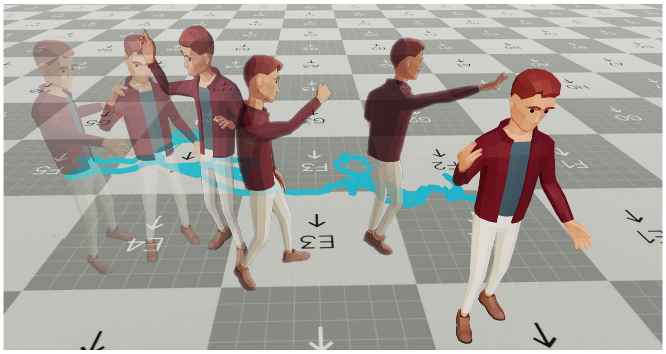
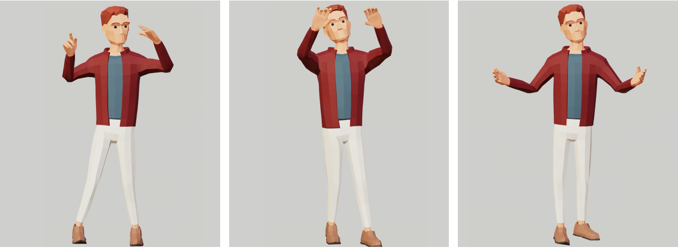


Fig. 7: Sample of gesture including finger movements and locomotion.[3]

Further, finger movements and locomotion are included, as shown in Figure 7, which highlight the system's proficiency in creating realistic, character-specific animations, thereby increasing the virtual interactions' believability and immersion.



(a) CamenAgraDeedy

(b) JinhaLee

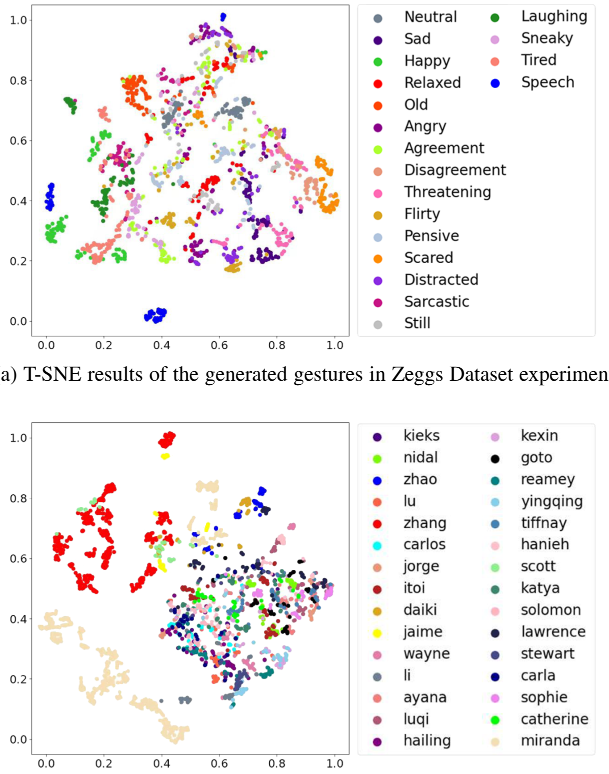
(c) SakiMafunikwa

Fig. 8: Samples of gestures corresponding to in-the-wild speech audio collected from TED Talks.

The Figure 8 showcase gesture outcomes generated from in-the-wild speech audio, like TED talks, to demonstrate the system's ability to create lifelike and style movements directly from unstructured real-world audio, without additional prompts or labels. This highlights the system's robust generalization capabilities. Testing in noisy environments with background music, applause, and urban sounds further revealed the system's strong anti-interference performance, emphasizing its resilience. This efficiency simplifies the input process, enabling effortless generation of dynamic character animations from raw audio, thus enhancing user experience and system accessibility.

Finally, we represent the visualizes (Figure 9) of the distribution of generated gestures corresponding to different emotional states (Fig. 9a) and personalities(Fig. 9b) using the t-SNE method. The figure illustrates distinct separations between certain states, while others exhibit a degree of similarity yet remain distinguishable. These findings demonstrate the capability of our proposed method to generate nuanced and discernible gestures solely from raw speech audio without relying on labels or manual annotations.

[3] Due to the inherent challenges in retargeting finger motion to the avatar, please refer to the support video for more details.



(a) T-SNE results of the generated gestures in Zeggs Dataset experiment.

(b) T-SNE results of the generated gestures in BEAT Dataset experiment.

Fig. 9: The T-SNE clustering visualization displays a variety of gestures distinguished by color-coded styles, revealing distinct regions for gestures tied to specific emotions or speakers, albeit with some boundary overlaps. This highlights our approach's capability to produce distinct style gestures through a fuzzy feature inference strategy, relying solely on speech audio.

D. Subjective and Objective Evaluation

Consistent with the prevailing practices in gesture generation research, we conducted a series of subjective and objective evaluations to evaluate the co-speech gestures generated by our proposed Persona-Gestor (PG) model.

We adopted slightly varied baselines for different datasets. For the Trinity dataset, we employed LDA [4] and Taming [15]. In addition to LDA and Taming, for the ZEGGS dataset, we also incorporated DiffuseStyleGesture (DSG) [2] and ZeroEGGS [7] Furthermore, for the BEAT dataset, we utilized the same baseline models as in ZEGGS but replaced DSG with DSG+ [3] and introduced GestureDiffuCLIP (GDC) [16] as an additional baseline model.

In our experiments with the ZEGGS and BEAT datasets, we extended the original LDA, DSG, and DSG+[4] models to cover all styles within these datasets. Originally, the Taming model, trained exclusively on the TED dataset, focused on upper-body gestures. We have since augmented it to support full-body gestures across the three datasets. Efforts to adapt LDA to include finger motions were met with challenges, leading to unsatisfactory outcomes in gesture generation. Consequently, we utilized LDA-generated gestures, excluding finger movements, for our analysis. For more implementation details of these baselines, please refer to Appendix C.

1) Subjective Evaluation: The goal of speech-driven gesture generation is to produce gestures that are both natural and convincing. However, exclusive reliance on objective metrics may not adequately reflect human subjective quality assessments [20], [46], [47]. This study prioritizes subjective evaluations to gauge human perception, complemented by objective evaluations detailed in Section IV-D2.

For thorough subjective evaluations, we utilize three metrics: human likeness, appropriateness, and style appropriateness. Human likeness gauges the naturalness and resemblance of gestures to real human movements independent of speech. Appropriateness examines the temporal alignment of gestures with speech rhythm, intonation, and semantics, ensuring natural fluidity. Style-appropriateness evaluates the similarity between generated and original gestures.

We conducted a user study with pairwise comparisons, as recommended by [48]. In each trial, participants were shown two 10-second video clips generated by different models (including the Ground Truth (GT)) side by side for direct comparison. The videos were accompanied by instructions for participants to select their preferred clip based on their evaluations. Preferences were quantified on a 0 to 2 scale, with the unselected clip in each pair receiving an inverse score (e.g., a -2 score for the non-chosen clip if the chosen one received 2). A score of zero indicated no preference. Attention checks were included in the study to ensure engagement. Further details are available in Appendix B.

Considering the extensive range of styles in ZEGGS (19) and BEAT (30), individual evaluations for each style were deemed impractical. Consequently, we utilized a random selection method to assign a subset of 5 styles from ZEGGS and 6 characters from BEAT to each participant. For the Trinity dataset, we chose Record_008 and Record_015. The training or validation sets include none of the selected audio clips.

A total of thirty volunteer participants, 17 males and 13 females aged between 19 and 31, were recruited for this study. Among them, 22 participants were Chinese nationals, while the remaining eight were international students from the USA and UK. Notably, all participants in this study exhibited a high level of English proficiency.

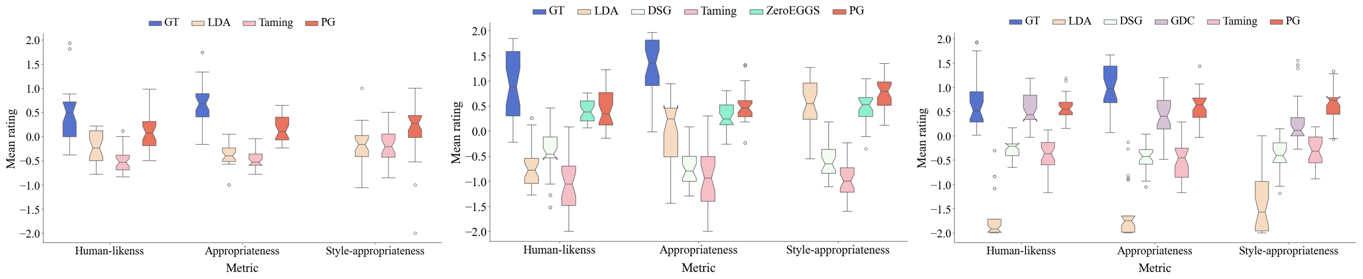
One-way ANOVA and post-hot Tukey multiple comparison tests were conducted to determine if the models' scores differed on the three evaluation aspects. The results are shown in Table I and Figure 10. The post-hoc analysis information is provided in the Appendix B.

The results indicate that the GT achieves the highest scores (0.51 ± 1.73 and 0.95 ± 1.13) in the Trinity and ZEGGS datasets, exhibiting statistically significant differences ($p < 0.001$) in human-likeness evaluations when compared to model-generated gestures. The GT is characterized by a diverse yet limited array of gestures, each with distinct traits that enhance movement realism. However, these gestures belong to the dataset's long-tail distribution, challenging the models' learning capabilities. Additionally, these unique gestures

[4] The authors have expanded their coverage to include all types in the BEAT dataset, as originally presented in the project of that study.

TABLE I: The subject mean perceptual rating score. Bold fonts were utilized to emphasize the best results for each metric among the different methods, except for the GT.

Methods			Subject Evaluation Metric			Objective Evaluation Metric		
Dataset	Model	With Fingers	Human↑ likeness	appropriateness↑	Style↑ appropriateness	FGD↓ on feature space	FGD↓ on raw data space	BeatAlign↑
Trinity	GT	*Y*	0.51±1.73	0.66±1.24	/	/	/	/
	LDA [4]	*N*	-0.22±0.98	-0.39±1.08	-0.18±1.10	349.53	6008.37	0.68
	Taming [15]	*N*	-0.48±0.96	-0.47±1.07	-0.23±1.05	3970.14	52196.87	0.68
	(Proposed)PG	N	**0.12±1.09**	**0.19±1.12**	**0.20±1.06**	**289.42**	**5080.57**	**0.69**
	(Ours)PGNSE	N	0.08±1.06	0.13±1.18	0.15±1.03	8002.65	70617.28	0.68
	(Ours)PGCA	N	-0.17±1.02	-0.01±1.11	-0.03±1.02	2566.65	26124.39	0.68
	(Ours)PGCF	N	-0.08±1.06	-0.13±1.05	0±1.02	12540.17	156895.12	0.68
	(Ours)PGICC	N	0.07±1.01	0.06±1.12	0.08±1.01	125921.53	1940124.44	0.67
ZEGGS	GT	Y	0.95±1.13	1.19±1.03	/	/	/	/
	LDA [4]	N	-0.73±1.12	0.02±1.31	0.53±1.41	124.55	50996.33	0.66
	DSG [2]	Y	-0.47±1.14	-0.71±1.11	-0.61±1.08	66.77	33297.50	0.63
	Taming [15]	Y	-1.08±1.01	-0.91±1.09	-0.98±0.97	1419.76	293245.12	0.67
	ZeroEGGS [7]	Y	0.38±1.11	0.29±1.35	0.49±1.31	37.19	26666.85	0.66
	(Proposed)PG	Y	**0.42±1.17**	**0.48±1.29**	**0.76±1.34**	**28.13**	**26193.92**	**0.68**
	(Ours)PGNSE	Y	0.33±1.15	0.35±1.29	0.59±1.38	125.40	49081.55	0.66
	(Ours)PGOnehot	Y	0.25±1.19	0.33±1.31	0.51±1.28	122.56	50259.95	0.63
	(Ours)PGCA	Y	-0.36±1.22	-0.62±1.14	-0.66±1.09	807.12	156686.01	0.67
	(Ours)PGCF	Y	0.27±1.20	-0.03±1.28	-0.23±1.26	97.57	39256.55	0.67
	(Ours)PGICC	Y	0.04±1.20	-0.39±1.23	-0.40±1.21	407.89	89893.96	0.67
BEAT	GT	Y	0.65±1.16	0.96±1.04	/	/	/	/
	LDA [4]	*N*	-1.65±0.73	-1.59±0.74	-1.35±1.05	276.25	3584.95	0.66
	DSG+ [3]	Y	-0.28±1.17	-0.49±1.15	-0.40±1.24	23811.46	2384465.64	0.43
	GDC [16]	*N*	0.54±1.12	0.47±1.25	0.30±1.27	432.15	93215.56	**0.69**
	Taming [15]	Y	-0.42±1.14	-0.52±1.14	-0.32±1.24	1251.56	46828.23	0.66
	(Proposed)PG	Y	**0.56±1.14**	**0.63±1.10**	**0.66±1.16**	**264.06**	**3471.26**	0.68
	(Ours)PGNSE	Y	0.09±1.16	0.27±1.23	0.46±1.31	1514.94	51077.98	0.66
	(Ours)PGOnehot	Y	-0.01±1.16	0.18±1.31	0.32±1.36	1863.69	63872.78	0.63
	(Ours)PGCA	Y	0.35±1.09	0.17±1.26	0.28±1.33	703.83	18990.56	0.66
	(Ours)PGCF	Y	0.14±1.01	0.15±1.22	0.30±1.31	1160.63	48899.63	0.66
	(Ours)PGICC	Y	0.02±1.10	-0.24±1.17	-0.25±1.25	2057.31	78754.92	0.66



(a) Trinity (b) ZEGGS (c) BEAT

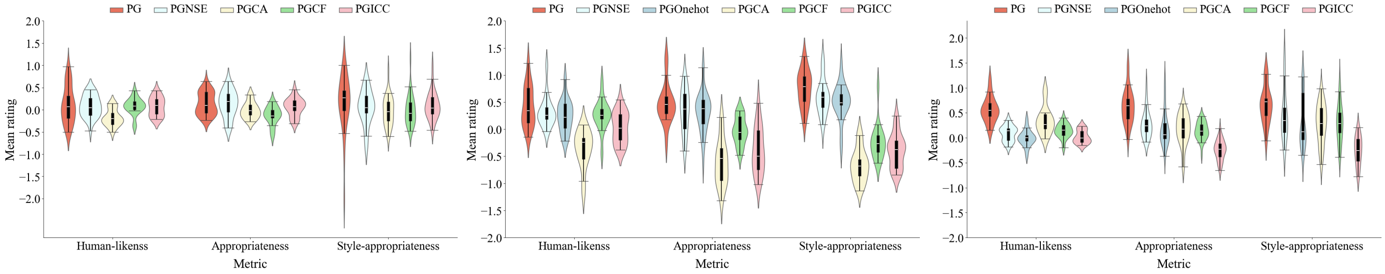
Fig. 10: The mean rating of each metric for each approach across the three datasets in comparative experiments.

impact the appropriateness and style-appropriateness scores. Conversely, while the GT achieves higher scores (0.65 ± 1.16), no significant differences were observed compared with the PG (0.56 ± 1.14) and GDC (0.54 ± 1.12) in the BEAT dataset analysis. This suggests that these models are more closely aligned with GT benchmarks in this dataset.

The experiments on the Trinity dataset show our proposed model (0.12 ± 1.09, 0.138 ± 1.12, and 0.203 ± 1.06) outperforming both LDA (-0.22 ± 0.98, -0.39 ± 1.08, and -0.18 ± 1.10) and Taming (-0.48 ± 0.96, -0.47 ± 1.07, and -0.23 ± 1.05) architectures significantly ($p < 0.001$) across all metrics. This superior performance is due to the more natural and relaxed gestures produced by our model, PG, enhancing its effectiveness compared to the LDA and Taming models, which fall short in accurately capturing the acoustic rhythm.

Evaluation of the ZEGGS dataset showed statistically significant differences ($p < 0.001$) between our method (0.42 ± 1.17, 0.48 ± 1.29, and 0.76 ± 1.34) and others across all three metrics. However, there was no statistically significant difference ($p > 0.05$) between our method and ZeroEGGS (0.38 ± 1.11) in terms of human likeness, though our method achieved a slightly higher score. These findings suggest that both our proposed model and ZeroEGGS can generate vivid gestures. Our advantage lies in the ability to synthesize emotional gestures solely through audio input, without relying on any reference example animations or labels.

In the BEAT dataset experiments, our PG model exhibited significant improvements (0.56 ± 1.14, 0.63 ± 1.10, and 0.66 ± 1.16) in three metrics compared to DSG+ (-0.28 ± 1.17, -0.49 ± 1.15, and -0.40 ± 1.24), LDA (-1.65 ± 0.73, -1.59 ± 0.74, and -1.35 ± 1.05), and Taming (-0.41 ± 1.14, -0.52 ± 1.14, and -0.32 ± 1.24) , reflecting the degradation



(a) Trinity (b) ZEGGS (c) BEAT

Fig. 11: The mean rating of each metric for each approach across the three datasets in ablation experiments.

in synthesis quality observed in DSG+ and LDA when incorporating all styles. While human-likeness metrics were comparable to GDC, the appropriateness metric of PG achieved higher scores than GDC (0.47 ± 1.25) despite the GDC's ability to better align with speech rhythm. Users reported that gestures generated by GDC overly emphasized prosodic cues, resulting in unnatural and frequent gestures. Furthermore, these gestures often displayed repetitive patterns with limited stylistic diversity, clearly indicating their origin from the GDC model and leading to a lower score (0.30 ± 1.27) in the style-appropriateness metric. This finding implies that the quality of gestures is not solely determined by accurately matching the audio rhythm.

2) Objective Evaluation: We introduce three objective evaluation metrics: Fréchet Gesture Distance (FGD) in both feature and raw data spaces [49], and BeatAlign [50]. FGD, inspired by the Fréchet Inception Distance (FID) [38], assesses the quality of generated gestures and demonstrates a moderate correlation with human-likeness ratings, outperforming other objective metrics [47]. Additionally, BeatAlign measures gesture-audio synchrony by calculating the Chamfer Distance between audio and gesture beats, providing insights into the temporal alignment of generated gestures with speech rhythms.

Table I displays our results, highlighting the state-of-the-art performance of our method in objective evaluations using FGD and BeatAlign metrics. Our model outperforms (289.42 for Trinity, 28.13 for ZEGGS, and 264.06 for BEAT) other architectures in FGD, effectively generating gestures that align closely with the Ground Truth (GT). It also achieves superior BeatAlign scores (0.69 for Trinity, 0.68 for ZEGGS, and 0.68 for BEAT) compared to other models, except for GDC (0.69 for BEAT), demonstrating its efficacy in producing co-speech gestures that synchronize accurately with speech rhythms. Although GDC scores highest in BeatAlign, corroborating user feedback, its overemphasis on prosodic cues leads to frequent high-frequency gestures. While technically accurate, this diminishes gesture naturalness.

E. Ablation Studies

Ablation studies were performed to evaluate the impact of key components on our model's efficacy, specifically targeting the global fuzzy feature extractor and Adaptive Layer Normalization (AdaLN).

1) Ablation of Global Fuzzy Feature Extractor: For the global fuzzy feature extractor, we explored the outcomes of removing this component (we call it: No Style Encoding, PGNSE) and replacing it with One-hot embedding (PGOnehot) for discrete feature extraction. PGOnehot was not applied to the Trinity dataset due to its limited style variability.

Our analysis of the global fuzzy feature extractor shows no significant differences ($p > 0.05$) between PG and PGNSE on the Trinity dataset in three subjective metrics, likely due to its limited range of styles. However, the ZEGGS dataset reveals significant variances in three metrics between PG (0.42 ± 1.17, 0.48 ± 1.29, and 0.76 ± 1.34) and PGOnehot (0.25 ± 1.19, 0.33 ± 1.31, and 0.51 ± 1.28), while no notable differences in human-likeness and appropriateness metrics are observed between PG and PGNSE (0.35 ± 1.29). PG (0.76 ± 1.34) outperforms PGNSE (0.59 ± 1.38) in style-appropriateness, likely because PGNSE cannot ensure a consistent style throughout the sequence. Conversely, the BEAT dataset exhibits significant differences ($p < 0.001$) between PG (0.56 ± 1.14, 0.63 ± 1.10, and 0.66 ± 1.16) and the other methods, indicating the superior capability of the global fuzzy feature inference mechanism in capturing stylistic nuances. Moreover, while PGOnehot is capable of capturing various logo gesture styles, it may compromise the naturalness of the movements.

2) Ablation of AdaLN: We integrated Cross-Attention (PGCA), In-Context Conditions (PGICC), and Concatenation of Features (PGCF) into our analysis to evaluate AdaLN's effectiveness. This structured approach enabled a comprehensive assessment of each component's contribution to the model's overall performance and its role in audio-based gesture generation. The implementations of Cross-Attention and In-Context Conditions follow the designs in [17], while Feature Concatenation combines gesture and encoded audio features along the feature axis, a technique proven effective in related studies [51]. Separate user studies were conducted for each component, with findings presented in Table I and Figure 11.

In the ablation studies concerning AdaLN, the replacement of the AdaLN module with alternative architectural frameworks precipitated a significant degradation in performance across all metrics. This reduction in efficacy can be ascribed to the deficiency of alternative architectures in synchronizing speech rhythm and capturing stylistic nuances with precision.

This outcome underscores the pivotal role of a uniform mechanism that applies an identical function across all attention layers throughout the sequence.

F. Generalization and Robustness

In addition, we test our method's generalization capabilities. We utilized in-the-wild speech audio collected from TED talks. Our system adeptly generates consistent gestures from dataset types and seamlessly produces gestures from untagged, in-the-wild audio. It also showcases remarkable robustness against various auditory disturbances, such as background music, applause, urban noise, and decorative sounds. This adaptability highlights the system's ability to handle a broad spectrum of audio inputs, ensuring the creation of naturalistic gestures despite significant noise interference. Such resilience emphasizes the system's suitability for real-world applications. Yet, we encountered certain inherent challenges in assessing the generalizability and robustness of alternative models during our experimentation. More details can be found in Appendix D, and supporting videos.

V. DISSCUSTION AND CONCLUSION

In this work, we introduce *Persona-Gestor*, a novel network architecture designed for the generation of personality gestures, leveraging solely raw speech audio. At its core, *Persona-Gestor* combines a fuzzy feature extractor and an AdaLN transformer diffusion architecture.

The fuzzy feature extractor utilizes a fuzzy feature inference strategy in the dual-component module to implicitly infer both fuzzy stylistic features and specific details embedded within the audio data autonomously. These elements are combined into a unified latent representation, facilitating the generation of speaker-aware personalized 3D full-body gestures. This approach incorporates a highly influential feature into the capability to synthesize personality gestures through automatically inferred fuzzy features, removing the necessity for explicit style labels or additional features. This advancement facilitates the end-to-end generation of gestures that resonate with the speaker's unique characteristics, directly from raw speech audio. Thereby, integrating fuzzy feature inference ensures a seamless and intuitive creation process that enhances generalization and user accessibility.

The AdaLN mechanism is a conditional mechanism that uniformly applies a specific function across all sequence tokens. This strategic incorporation significantly augments the model's proficiency in accurately capturing and representing both conditional dependencies and output characteristics with greater efficiency. We demonstrate that AdaLN also facilitates a refined understanding and processing of the complex interplay between the continuous fuzzy features conditional input and the resultant gesture synthesis, leading to enhanced model performance and output fidelity. Ultimately, *Persona-Gestor* utilizes diffusion mechanism for producing a diverse spectrum of gesture outputs.

Our approach presents multiple benefits: 1) It exclusively uses raw speech audio to synthesize speaker-aware personalized gestures, bypassing the requirement for extra inputs, which enhances user-friendliness. 2) It achieves the full-body (including finger motions and locomotion) gestures' superior synchronization with speech, capturing rhythm, intonation, and certain semantics without compromising naturalness. 3) It showcases improved generalization and robustness, adapting effectively across varied conditions.

Our study highlights key areas for enhancement: Firstly, the model's sole dependence on speech audio may limit its effectiveness in capturing style features within segments of minimal speech. Secondly, the lack of control over the movement path and orientation of the digital human could lead to unintended gestures. Thirdly, our model may not effectively replicate certain gestures, which are crucial for expressing specific states. These observations underscore the necessity for improvements to broaden the model's ability to accurately convey a wide range of human gestures.

VI. ACKNOWLEDGMENTS

This work was partially supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No.2023 C01212), the National Key Research and Development Program of China (No.2022YFF 0902305), the Public Welfare Technology Application Research Project of Zhejiang (No.LGF21F020002, No.LGF22F020008), the Key Program and development projects of Zhejiang Province of China (No.2021C03137), and the Key Lab of Film and TV Media Technology of Zhejiang Province (No.2020E10015).

REFERENCES

[1] U. Bhattacharya, E. Childs, N. Rewkowski, and D. Manocha, "Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning," in *Proc. of the 29th ACM International Conf. on Multimedia*, 2021, pp. 2027–2036.

[2] S. Yang, Z. Wu, M. Li, Z. Zhang, L. Hao, W. Bao, M. Cheng, and L. Xiao, "DiffuseStyleGesture: Stylized audio-driven co-speech gesture generation with diffusion models," *arXiv preprint arXiv:2305.04919*, 2023.

[3] S. Yang, H. Xue, Z. Zhang, M. Li, Z. Wu, X. Wu, S. Xu, and Z. Dai, "The diffusestylegesture+ entry to the genea challenge 2023," *arXiv preprint arXiv:2308.13879*, 2023.

[4] S. Alexanderson, R. Nagy, J. Beskow, and G. E. Henter, "Listen, denoise, action! audio-driven motion synthesis with diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–20, 2023.

[5] S. Yang, Z. Wang, Z. Wu, M. Li, Z. Zhang, Q. Huang, L. Hao, S. Xu, X. Wu, and C. Yang, "Unifiedgesture: A unified gesture synthesis model for multiple skeletons," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1033–1044.

[6] J. Li, D. Kang, W. Pei, X. Zhe, Y. Zhang, L. Bao, and Z. He, "Audio2gestures: Generating diverse gestures from audio," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–15, 2023.

[7] S. Ghorbani, Y. Ferstl, D. Holden, N. F. Troje, and M.-A. Carbonneau, "Zeroeggs: Zero-shot example-based gesture generation from speech," in *Computer Graphics Forum*, vol. 42, no. 1. Wiley Online Library, 2023, pp. 206–216.

[8] L. A. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, no. 3, pp. 338–353, 1965.

[9] S. Vashishtha and S. Susan, "Unsupervised fuzzy inference system for speech emotion recognition using audio and text cues (workshop paper)," in *2020 IEEE sixth international conference on multimedia big data (BigMM)*. IEEE, 2020, pp. 394–403.

[10] N. M. Patil and M. U. Nemade, "Content-based audio classification and retrieval using segmentation, feature extraction and neural network approach," in *Advances in computer communication and computational sciences: Proceedings of IC4S 2018*. Springer, 2019, pp. 263–281.

[11] R. A. Calvo, S. D'Mello, J. M. Gratch, and A. Kappas, *The Oxford handbook of affective computing*. Oxford Library of Psychology, 2015.

[12] M. Goudbeek and K. Scherer, "Beyond arousal: Valence and potency/control cues in the vocal expression of emotion," *The Journal of the Acoustical Society of America*, vol. 128, no. 3, pp. 1322–1336, 2010.

[13] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, 2015.

[14] K. Campbell-Kibler, "Intersecting variables and perceived sexual orientation in men," *American Speech*, vol. 86, no. 1, pp. 52–68, 2011.

[15] L. Zhu, X. Liu, X. Liu, R. Qian, Z. Liu, and L. Yu, "Taming diffusion models for audio-driven co-speech gesture generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 544–10 553.

[16] T. Ao, Z. Zhang, and L. Liu, "Gesturediffuclip: Gesture diffusion model with clip latents," *arXiv preprint arXiv:2303.14613*, 2023.

[17] W. Peebles and S. Xie, "Scalable diffusion models with transformers," *arXiv preprint arXiv:2212.09748*, 2022.

[18] J. Windle, D. Greenwood, and S. Taylor, "Uea digital humans entry to the genea challenge 2022," in *GENEA: Generation and Evaluation of Non-Verbal Behaviour for Embodied Agents Challenge*, 2022.

[19] L. Yu, H. Xie, and Y. Zhang, "Multimodal learning for temporally coherent talking face generation with articulator synergy," *IEEE Transactions on Multimedia*, vol. 24, pp. 2950–2962, 2021.

[20] A. Simon, H. G. Eje, K. Taras, and B. Jonas, "Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows," in *Computer Graphics Forum*, vol. 39. Wiley Online Library, 2020, pp. 487–496, issue: 2.

[21] Taylor Sarah, Windle Jonathan, Greenwood David, and Matthews Iain, "Speech-driven conversational agents using conditional flow-vaes," in *European Conf. on Visual Media Production*, 2021, pp. 1–9.

[22] F. Zhang, N. Ji, F. Gao, and Y. Li, "Diffmotion: Speech-driven gesture synthesis using denoising diffusion model," in *MultiMedia Modeling: 29th International Conf., MMM 2023, Bergen, Norway, January 9–12, 2023, Proc., Part I*. Springer, 2023, pp. 231–242.

[23] S. Yang, Z. Wu, M. Li, M. Zhao, J. Lin, L. Chen, and W. Bao, "The reprgesture entry to the genea challenge 2022," in *Proc. of the 2022 International Conf. on Multimodal Interaction*, 2022, pp. 758–763.

[24] S. Yang, Z. Wu, M. Li, Z. Zhang, L. Hao, W. Bao, and H. Zhuang, "Qpgesture: Quantization-based and phase-guided motion matching for natural speech-driven gesture generation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2023, pp. 2321–2330.

[25] Yoon Youngwoo, Ko Woo-Ri, Jang Minsu, Lee Jaeyeon, Kim Jaehong, and Lee Geehyuk, "Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots," in *2019 International Conf. on Robotics and Automation (Icra)*, 2019, pp. 4303–4309.

[26] Kucherenko Taras, Jonell Patrik, van Waveren Sanne, Henter Gustav Eje, Alexandersson Simon, Leite Iolanda, and Kjellström Hedvig, "Gesticulator: A framework for semantically-aware speech-driven gesture generation," in *Proc. of the 2020 International Conf. on Multimodal Interaction*, 2020, pp. 242–250.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *2019 Conf. of The North American Chapter of The Association for Computational Linguistics: Human Language Technologies, vol. 1*, 2019, pp. 4171–4186.

[28] Bhattacharya Uttaran, Rewkowski Nicholas, Banerjee Abhishek, Guhan Pooja, Bera Aniket, and Manocha Dinesh, "Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents," in *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2021, pp. 1–10.

[29] Pennington Jeffrey, Socher Richard, and Manning Christopher D., "Glove: Global vectors for word representation," in *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[30] Mikolov Tomas, Sutskever Ilya, Chen Kai, Corrado Greg S., and Dean Jeff, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[31] Bojanowski Piotr, Grave Edouard, Joulin Armand, and Mikolov Tomas, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[32] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 12 1980.

[33] E. Cambria, A. Livingstone, and A. Hussain, "The hourglass of emotions," in *Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers*. Springer, 2012, pp. 144–157.

[34] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *INTERSPEECH 2020*, ser. Interspeech, 2020, pp. 5036–5040, interspeech Conf., Shanghai, PEOPLES R CHINA, OCT 25-29, 2020.

[35] Y. Zhi, X. Cun, X. Chen, X. Shen, W. Guo, S. Huang, and S. Gao, "Livelyspeaker: Towards semantic-aware co-speech gesture generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 807–20 817.

[36] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, and X. Xiao, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[37] L. Kaiser, A. N. Gomez, and F. Chollet, "Depthwise separable convolutions for neural machine translation," *arXiv preprint arXiv:1706.03059*, 2017.

[38] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[39] L. Paul, "sur la théorie du mouvement brownien," *C. R. Acad. Sci.*, vol. 65, no. 11, pp. 146,530–533, 1908, publisher: American Association of Physics Teachers.

[40] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[41] K. Rasul, C. Seward, I. Schuster, and R. Vollgraf, "Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting," in *International Conf. on Machine Learning*, 2021, pp. 8857–8868.

[42] Ferstl Ylva and McDonnell Rachel, "Investigating the use of recurrent motion modelling for speech gesture generation," in *Proc. of the 18th International Conf. on Intelligent Virtual Agents*, 2018, pp. 93–98.

[43] H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, E. Bozkurt, and B. Zheng, "Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis," in *COMPUTER VISION, ECCV 2022, PT VII*, vol. 13667, 2022, pp. 612–630.

[44] Grassia F. Sebastian, "Practical parameterization of rotations using the exponential map," *Journal of graphics tools*, vol. 3, no. 3, pp. 29–48, 1998.

[45] U. Wennberg and G. E. Henter, "The case for translation-invariant self-attention in transformer-based language models," in *ACL-IJCNLP 2021: The 59th Annual Meeting of The Association for Computational Linguistics and The 11th International Joint Conf. on Natural Language Processing, vol 2*, 2021, pp. 130–140.

[46] P. Wolfert, N. Robinson, and T. Belpaeme, "A review of evaluation practices of gesture generation in embodied conversational agents," *IEEE Transactions on Human-Machine Systems*, 2022.

[47] T. Kucherenko, P. Wolfert, Y. Yoon, C. Viegas, T. Nikolov, M. Tsakov, and G. E. Henter, "Evaluating gesture-generation in a large-scale open challenge: The genea challenge 2022," *arXiv preprint arXiv:2303.08737*, 2023.

[48] P. Wolfert, J. M. Girard, T. Kucherenko, and T. Belpaeme, "To rate or not to rate: Investigating evaluation methods for generated co-speech gestures," in *Proceedings of the 2021 International Conference on Multimodal Interaction*. ACM, 2021-10-18, pp. 494–502.

[49] Yoon Youngwoo, Cha Bok, Lee Joo-Haeng, Jang Minsu, Lee Jaeyeon, Kim Jaehong, and Lee Geehyuk, "Speech gesture generation from the trimodal context of text, audio, and speaker identity," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–16, 2020.

[50] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "Ai choreographer: Music conditioned 3d dance generation with aist++," in *Proc. of the IEEE/CVF International Conf. on Computer Vision*, 2021, pp. 13 401–13 412.

[51] M. Zhao, J. Ning, J. Hu, and T. Li, "Attention-driven dual feature guidance for hyperspectral super resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2023, publisher: IEEE.

Appendix A
Detail of Datasets

Table II refers to an overview of the three datasets (Trinity, ZEGGS, and BEAT).

TABLE II: Overview of the three datasets.

Dataset	Total Time	fps	Audio Sample Rate	Character	Content
Trinity	244 min	60	44 kHz	1 male	spontaneous speech on different topics.
ZEGGS	135 min	60	48 kHz	1 female	cover 19 different motion styles.
BEAT	35h	120	48 kHz	30 speakers	speak on diverse content.

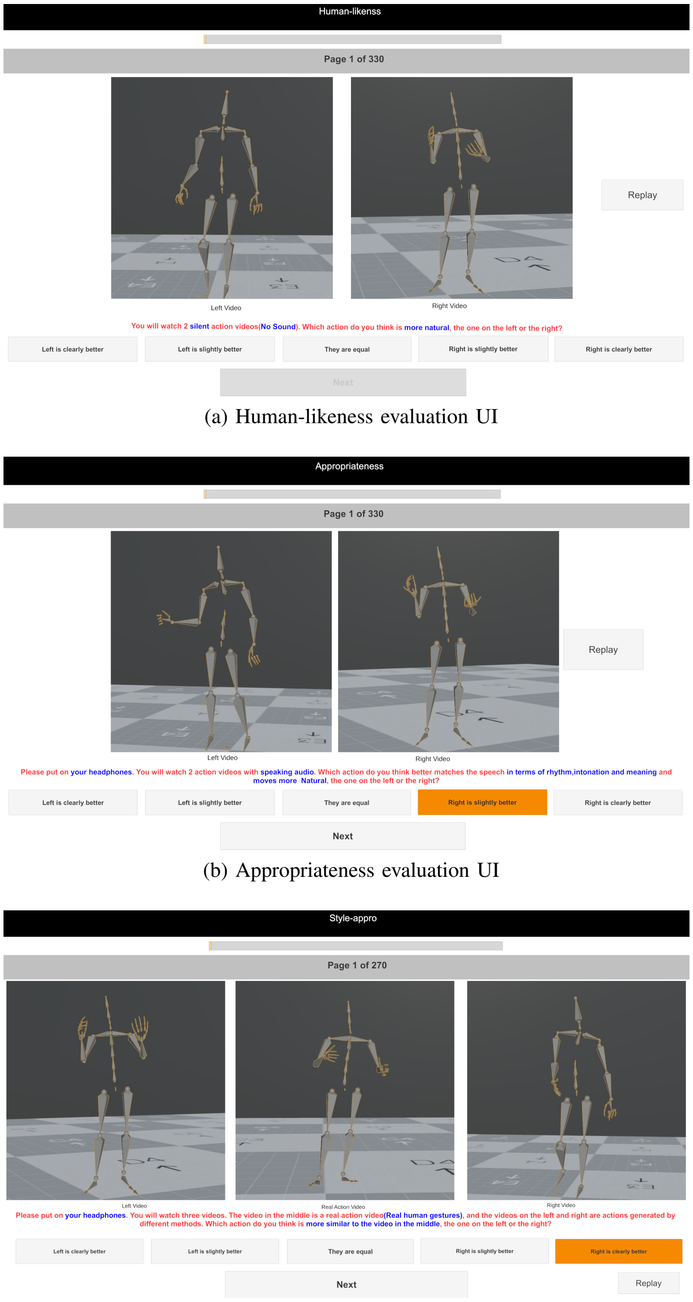
Appendix B
Details of User Study

1) Processing of User Study: Pairwise comparisons, found to be quicker and slightly more reliable inter-rater wise, offer a different perspective from the rating scale method, which illuminates both absolute and comparative qualities, thus better accommodating the simultaneous assessment of multiple stimuli [48]. In our study, we implemented pairwise comparisons, showing participants two 10-second clips from various models (including ground truth) for the same speech excerpt. To streamline and clarify the evaluation, clips were played side by side, with an arrangement of three clips for style-appropriateness assessments, featuring the ground truth in the middle. This setup allowed for the concurrent viewing of two (three) clips, enhancing the directness of comparisons.

We opted for the original skeletal display gestures instead of avatars, as accurately retargeting skeletal movements to avatars, particularly the finger parts, poses significant challenges. Relevant details can be found in Figure 12.

Participants were instructed to select their preferred clip by five response options below the videos: ”Left is clearly better”,”Left is slightly better”, ”They are equal”, ”Right is clearly better”, and ”Right is clearly better”. The platform assigns a score to each video based on the user’s selection, using a scale ranging from 0 to 2, where 0 indicates no preference. In cases where a video is not chosen within a pair, it automatically receives an inverse score (e.g., if the participants select the left video as ”Left is clearly better”, then the unselected video (the right video) receives a score of -2). Participants were given the opportunity to rate the videos only after both had been presented, and they were provided with the option to replay them at their discretion. For each metric experiment, we provide explicit evaluation directions and detailed instructions:

- **Human-likeness experiment:** You will watch 2 silent action videos (No sound). Which action do you think is more natural, the one on the left or the right?
- **Appropriateness experiment:** Please put on your headphones. You will watch 2 videos with speaking audio. Which action do you think better matches the speech in terms of rhythm, intonation, meaning and moves more naturally, the one on the left or the right?
- **Style-appropriateness experiment:** Please put on your headphones. You will watch three videos. The video in the middle is a real action video (Real Human gestures), and the videos on the left and right are actions generated by different methods. Which action do you think is more similar to the video in the middle, the one on the left or the right?



You will watch 2 silent action videos(No Sound). Which action do you think is more natural, the one on the left or the right?

(a) Human-likeness evaluation UI

Please put on your headphones. You will watch 2 action videos with speaking audio. Which action do you think better matches the speech in terms of rhythm,intonation and meaning and moves more Natural, the one on the left or the right?

(b) Appropriateness evaluation UI

Please put on your headphones. You will watch three videos. The video in the middle is a real action video(Real human gestures), and the videos on the left and right are actions generated by different methods. Which action do you think is more similar to the video in the middle, the one on the left or the right?

(c) Style-appropriateness evaluation UI

Fig. 12: The screenshot depicting the user evaluation interface.

After conducting the user evaluation, we administered a user questionnaire survey to gather feedback from participants, primarily focusing on their overall perceptions of the obtained results.

A total of thirty volunteer participants, comprising 17 males and 13 females aged between 19 and 31, were recruited for this study. Among them, 22 participants were from China while the remaining eight were international students hailing from countries such as the USA and UK. It is worth noting that all participants in this study demonstrated proficiency in English.

To initiate the formal experiment, we provided all participants with a comprehensive introduction to the methodology and presented them with illustrative clips that were not included in the evaluation set. Subsequently, participants were instructed to don headphones and situate themselves in a tranquil environment devoid of any disturbances while facing a computer screen. It is noteworthy that throughout the experiment, participants remained uninformed about which method each video corresponded to. The sequence of videos was randomized and scored by participants upon presentation. All assessments incorporated attention checks.

Given the substantial number of styles in ZEGGS (19) and BEAT (30), conducting individual evaluations for each style would be impractical. Therefore, we employed a random selection process to choose a subset of 5 styles from ZEGGS. Regarding the Trinity dataset specifically, we specifically selected Record_008 and Record_015. Given that the GDC model is contingent upon extensive data annotations and due to the framework and dataset being proprietary and subject to licensing restrictions, our evaluation was limited to selecting only 6 clips for testing purposes.

Trinity Dataset: A total of 6 speech clips were selected from Record_008 and Record_015 for gesture synthesis. In evaluating human-likeness and appropriateness, 7 methods along with a ground truth (GT) were employed, resulting in the creation of 48 videos. For this evaluation, users needed to perform $(8\times7)/2\times6 = 168$ tests. For the style-appropriateness evaluation, without needing to assess the Ground Truth (GT), participants were required to complete $(7 \times 6)/2 \times 6 = 126$ tests.

ZEGGS Dataset: To distribute the variety of styles from the ZEGGS dataset, a random selection method was utilized, assigning 5 distinct styles to each participant. This approach guaranteed that participants would experience a diverse set of styles, thereby comprehensively encompassing the entire spectrum of styles within the ZEGGS dataset. For the evaluation of human-likeness and appropriateness, 11 different architectures, including the Ground Truth (GT), were tested, resulting in the generation of 55 videos per participant. Consequently, users were required to conduct $(11\times10)/2\times5 = 275$ comparisons. For the style-appropriateness evaluation, the procedure necessitated users to complete $(10 \times 9)/2 \times 5 = 225$ comparative assessments.

BEAT Dataset: In assessing human-likeness and appropriateness within the BEAT dataset, 10 distinct architectural models, inclusive of the Ground Truth (GT), underwent testing, culminating in the creation of 66 videos for each participant's evaluation. This setup mandated users to undertake $(11 \times 10)/2 \times 6 = 330$ comparison tasks. For evaluating style-appropriateness, participants were required to execute $(10 \times 9)/2 \times 6 = 270$ comparative analyses.

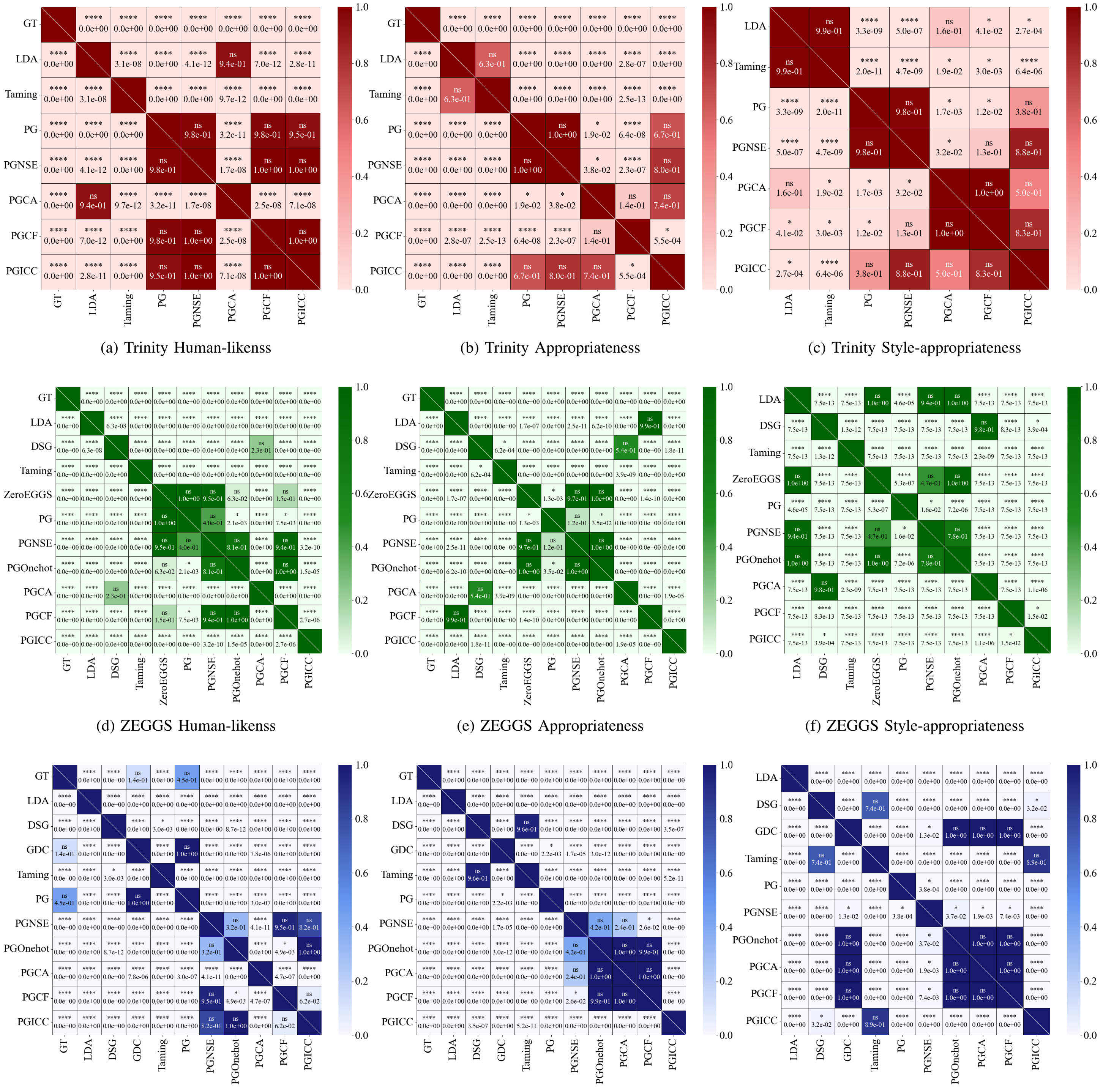
The experiment for the three datasets required a total of 2 days for each participant to complete. Participants were allowed to pause and take breaks after each experiment. The evaluation information is summarized in Table III

TABLE III: The summaries of evaluation information

Dataset	Metric	Test Count	Duration per experiment
Trinity	Human likeness	168	0.85±0.13h
	Appropriateness	168	0.92±0.12h
	Style appropriateness	126	1.13±0.32h
ZEGGS	Human likeness	275	1.53±0.21h
	Appropriateness	275	1.66±0.28h
	Style appropriateness	225	1.67±0.37h
BEAT	Human likeness	330	2.08±0.41h
	Appropriateness	330	2.12±0.53h
	Style appropriateness	270	1.82±0.28h

2) Process of User Study Data: By conducting a statistical analysis of user evaluation feedback data, we aim to investigate the existence of a significant correlation between the average scores obtained from different gesture generation methods. Initially, to ensure the applicability of the analytical approach, a normality test was conducted on the data. Given the limited sample size of 30, we opted for the Kolmogorov-Smirnov (K-S) test to assess data normality. The test results revealed significant relationships ($p < 0.05$) between the various gesture generation methods and action scores, leading to the rejection of the null hypothesis of normality. Therefore, it was concluded that the test data did not exhibit normal distribution characteristics.

In light of the non-normality of the data, we employed the Brown-Forsythe and Welch analysis of variance methods, suitable for non-normally distributed data. Our analysis of the differences in Method on Score indicated significant differences ($p < 0.05$) among the various Method samples, suggesting that different gesture generation methods had a significant impact on gesture scores. To further elucidate the significance of differences among the gesture generation methods, we conducted a post hoc analysis using the Tukey HSD method. This analysis revealed notable differences in mean scores between multiple groups. Given the complexity of the methods involved, we present in Figure 13 heatmaps illustrating the disparity in mean significance across different methods for each metric experiment conducted on the three datasets.



(a) Trinity Human-likenss

(b) Trinity Appropriateness

(c) Trinity Style-appropriateness

(d) ZEGGS Human-likenss

(e) ZEGGS Appropriateness

(f) ZEGGS Style-appropriateness

(g) BEAT Human-likenss

(h) BEAT Appropriateness

(i) BEAT Style-appropriateness

Fig. 13: Heatmaps of the mean ratings of user studies significant differences across all methods for each metric and datasets. Asterisks indicated significant effects (*: $p < 0.05$,****:$p < 1.00e-04$, ns: no significant difference). We use distinct colors to represent each dataset: red for Trinity, green for ZEGGS, and blue for BEAT, with lighter shades indicating greater significance differences.

APPENDIX C
CHALLENGES OF AUGMENTING STYLE ADAPTABILITY IN LDA AND DSG

In our evaluations, we expanded the capabilities of the LDA and DSG models on the ZEGGS dataset to accommodate all 19 styles, significantly surpassing the 6 styles that were initially supported. This extension entailed adjusting the dimensionality of the models' style one-hot embeddings from $\mathbb{R}^6$ to $\mathbb{R}^{19}$. For the BEAT dataset, we applied the full set of personality IDs, comprising 30 IDs, to the LDA model to gauge its capacity for capturing a broad spectrum of personality expressions. Notably, the DSG+ model's creators have similarly expanded its ID range to 30, as indicated in their open-source code. These adjustments evaluated the models' flexibility and efficacy in generating gestures across various stylistic expressions.

The modifications to the DSG model exposed its constraints, particularly in the generation of gestures, which were limited in their range of motion and exhibited a constrained motion. This limitation hindered the model's ability to produce gestures corresponding to specific states, indicating a shortfall in the model's adaptability and expressiveness. Conversely, the LDA model retained the pose characteristics inherent in the original data; it faced challenges in generating diverse gestures and displayed a notable lack of stability in its movements. These findings suggest a fundamental limitation of One-hot encoding in accommodating an expanded style spectrum. Contrarily, by employing a fuzzy inference strategy, our proposed model, *Persona-Gestor*, demonstrated remarkable stability and diversity across various styles, evidencing its superior adaptability to an extensive range of gesture styles. An ablation study that replaced the global feature extractor with One-hot embedding further supports this conclusion.

Meanwhile, in our efforts to enhance the LDA model's capabilities, we attempted to extend its functionality to include finger motion synthesis. Unfortunately, this endeavor did not yield successful outcomes, as we encountered difficulties in accurately generating the intended gestures. As shown in Table IV

TABLE IV: Unveiling the challenges of extended LDA and DSG training and testing on the ZEGGS and BEAT datasets. The symbol ✓ denotes the generation of proper gestures, while × indicates poor or crash gesture generation, ✓✗ signifies gestures generated with fewer bodily movements.

Dataset	Model	Style	Finger	g	Training	Train	Quality
				dim.	Steps	Loss	of gesture
ZEGGS	LDA	6	N	70	12260	0.0089	✓
		19	N	70	25410	0.0079	✓✗
		6	Y	103	12260	0.246	×
		19	Y	103	25410	0.171	×
	DSG	6	Y	1141	450000	0.014	✓
		19	Y	1141	450000	0.046	✓✗
	PG(Ours)	19	Y	103	24319	0.0142	✓
BEAT	LDA	30	N	70	222240	0.0112	✓
		30	Y	103	222240	0.116	×
	DSG	30	Y	2232	215638	0.0128	✓
	PG(Ours)	30	Y	159	118899	0.0226	✓

APPENDIX D
GENERALIZATION AND ROBUSTNESS

1) Generalization: In evaluating the generalization of gesture generation models using in-the-wild audio data from TED Talks, we encountered varied challenges across different models, as summarized in Table V. The LDA model's performance was contingent on inputting a style label for gesture generation, showing limitations in its ability to produce gestures accurately without explicit style guidance. Conversely, the DSG model required not only Style type information but also the corresponding text file of the audio and alignment data between the text and audio, significantly complicating the model's scalability. Furthermore, when trained on the BEAT dataset, the LDA model was ineffective in generating appropriate gestures. While responsive to speech rhythm, the GDC model tended to generate excessive and unnecessary limb movements due to its oversensitivity. In contrast, the PG model showcased remarkable scalability and efficiency by autonomously converting discrete styles into fuzzy style features, bypassing manual style labeling, and simplifying the gesture generation process.

2) Robustness: To assess the robustness of gesture generation models against auditory disturbances, we infused test audio inputs with various noise types, including applause and music, to simulate real-world scenarios. The outcomes, as detailed in Table V, illustrate distinct responses from each model to the presence of noise. The LDA model struggled significantly, failing to generate relevant gestures when exposed to noisy audio from TED Talks. The GDC model, while able to produce gestures, showed a tendency towards jittery and erratic movements under noisy conditions, indicating sensitivity to audio quality. Conversely, the PG model exhibited commendable stability and efficacy in gesture generation, unaffected by the introduced noise. This performance underscores the PG model's superior robustness, highlighting its potential for practical applications in environments with variable audio quality.

TABLE V: Testing Results of in-the-wild audios (TED Talk)

Input	LDA	DSG	GDC	PG(Ours)
Origin	generated improperly	Complex text processing	unnecessary movements	Good
Noisy	generated improperly	Complex text processing	with slight jittery	Good