Diffusion Transformer Framework for Speech-Driven Stylized Gesture Generation

Nada Elmasry[a], Yanbo Cheng[b] and Yingying Wang[c]

Department of Computing and Software, McMaster University, Hamilton, Canada

Keywords: Speech-Driven, Gesture Synthesis, Motion Stylization, Multimodal Coordination, Deep Learning, Diffusion Transformer.

Abstract: Gestures are a vital component of human expression, playing a pivotal role in conveying information and emotions. Generating co-speech gestures remains challenging in human-computer interaction due to the intricate relationship between speech and gestures. While recent advances in learning-based methodologies have shown some progress, they still encounter limitations, as a lack of diversity and a mismatch between generated gestures and the semantic and emotional context of speech, impacting the effectiveness of communication. In this work, we propose a novel gesture generation framework that takes speech audio and a target style gesture example as inputs, automatically synthesizing new gesture performances that align with the speech in the desired style. Specifically, our framework comprises four main components: a dual-stream audio encoder, a gesture-style encoder, a cross-attention modality fusion module, and a latent diffusion generation module. The dual-stream audio encoder and gesture style encoder extract diverse modality embeddings from audio and motion inputs; the cross-attention fusion module maps the multi-modal embeddings into a unified latent space, and the diffusion module produces expressive and stylized gestures. The results demonstrate the exceptional performance of our method in generating natural and diversified gestures that accurately and coherently convey the intended information, surpassing the benchmarks established by traditional methods. Finally, we discuss future directions for our research.

1 INTRODUCTION

Gestures are integral components of human communication, functioning as co-expressive elements complementing speech (David, 1992; McNeill, 2019). They consist of non-verbal hand and arm movements that enhance communication by synchronizing with speech emphasis in time and matching speech content in semantics. Thus, gestures play a pivotal role in conveying information, emotion and personality.

Co-speech gesture synthesis is crucial in developing lifelike conversational virtual characters, for human-computer interaction, computer graphics, and social robotics applications. However, the multimodal and multi-functional nature of gestures causes great challenges in their automatic generation. Unlike generic human motions, gesture performance is not standalone, but part of multimodal conversational behaviors dependent on speech and prosody. McNeill (David, 1992) categorizes gestures into four types: beat, iconic, metaphoric, and deictic, each of which correlates to different prosodic or semantic aspect in speech. Thus automatically generating realistic gestures of all categories that well synchronize with speech emphasis and match the spoken content is a hard multi-modal coordination problem to solve. Another challenge is that gesture motions are free form expressive motions that do not follow a regular pattern like locomotion, and identities, personalities and emotions can all cast significant impact on gesture performance styles. Recent research has utilized deep learning approaches to predict and produce gestures, however, the generated motion quality and diversity is still restricted by the data and network design, due to the spatiotemporal complexity of gesture performance.

In this work, we propose an example-based stylistic co-speech gesture generation framework that solves the aforementioned challenges. Our framework takes speech audio and an example of target style gesture as input, and outputs novel gesture performance in the specified style that matches the speech. The gesture synthesis task is achieved by four major components: a dual-stream audio encoder, a gesture style encoder, a multimodal cross-attention

[a] https://orcid.org/0009-0006-3694-1453
[b] https://orcid.org/0009-0009-1684-0585
[c] https://orcid.org/0000-0002-5680-1929

Elmasry, N., Cheng, Y. and Wang, Y.
Diffusion Transformer Framework for Speech-Driven Stylized Gesture Generation.
DOI: 10.5220/0013318400003912
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 1: GRAPP, HUCAPP and IVAPP*, pages 355-362
ISBN: 978-989-758-728-3; ISSN: 2184-4321

fusion module, and a diffusion-based gesture generator. First, the dual-stream audio encoder extracts effective acoustic embeddings from the speech input, and the gesture style encoder extracts motion style embeddings from the example gesture. The multimodal cross-attention module aligns the audio and motion embeddings extracted from different modalities through attention mechanisms, and fuses them to a unified latent space. The diffusion model takes the fused latent embedding and outputs diverse gesture performance.

Compared to existing gesture synthesis research, our proposed framework has many advantages. Single-shot example gesture is an efficient and feasible way for specifying the desired style. Incorporating style features addresses the challenge of capturing the vast combination of motion content and stylistic variations in human movement. Our dual-stream audio encoder extracts effective acoustic features from speech, and the subsequent cross-attention module is capable of capturing the corresponding between gesture style features and the speech audio features from different modalities. Our diffusion-based gesture generator ensures the stylistic diversity in the synthesized gestures. Preliminary results demonstrate that our framework outperforms existing gesture synthesis work under similar training conditions, generating expressive and context-appropriate gestures that align with the given speech. We summarize the contribution of our work as follows:

- We propose a novel framework that takes single-shot style example for synthesizing expressive gestures in desired styles matching speech input;
- We introduce a dual-stream audio encoder that effectively extracts acoustic features from speech;
- We demonstrate the multi-modal cross-attention module for fusing the correlated features between speech and style;
- We present the latent diffusion-based gesture generator, capable of synthesizing diverse stylistic gesture performances.

2 RELATED WORK

Rule-based Methods for Gesture Synthesis

Early gesture generation relied on rule-based systems with manual speech-gesture mappings. Cassell et al.'s Animated Conversation (Cassell et al., 1994) pioneered the automatic production of context-appropriate gestures, facial expressions, and intonation by integrating dialogue generation, text-to-speech, and symbolic representations. Thórrison's Ymir (Wei et al., 2022) enhanced this approach by incorporating multimodal inputs—speech, gaze, gesture, and intonation—through perception, dialogue, decision-making, and action scheduling modules, enabling more interactive animations. Further advancements included Cassell et al.'s Behaviour Expression Animation Toolkit (BEAT) (Cassell et al., 2001), which synthesized nonverbal cues with customizable personalities; Kopp et al.'s Max (Kopp and Wachsmuth, 2002; Kopp et al., 2003), generating complex gestures from XML specifications using non-uniform cubic B-Splines; and Pelachaud et al.'s Greta (Pelachaud et al., 2002), a 3D virtual agent expressing emotions through a Belief-Desire-Intention framework. The development of domain-specific languages (DSLs) such as MURML (Kopp et al., 2003), APML (De Carolis et al., 2004), and RRL (Piwek et al., 2004) followed, although they were primarily XML-based and incompatible. To resolve this, the Behavior Markup Language (BML) (Kopp et al., 2006; Vilhjálmsson et al., 2007) was created as a comprehensive framework for intent and behavior planning, becoming the standard for rule-based systems and integrating into platforms like SmartBody and humanoid robots. Despite their ability to produce synchronized gestures, rule-based systems are limited by finite handcrafted rules and pre-recorded motions, resulting in restricted motion diversity, scalability challenges due to manual effort, and reliance on explicit speech-gesture mappings based on text or acoustic features.

Data-driven Statistical Gesture Generation

Researchers developed data-driven statistical models for gesture synthesis to address the limitations of rule-based methods, but these often relied on curated gesture libraries and manual annotations, limiting scalability and adaptability. Kipp used ANVIL (Kipp, 2001) to annotate co-speech gestures, modeling them based on features like handedness, timing, and communicative function. Neff et al. (Neff et al., 2008) created an animation lexicon to generate gestures from text. Bergmann and Kopp introduced Bayesian networks for transforming speech into gestures (Bergmann and Kopp, 2009), enhancing them with probabilistic and rule-based components. Levine et al. employed hidden Markov model (HMM) and conditional random field (CRF) (Levine et al., 2009) to select motion clips based on prosodic features and reduce overfitting. Chiu et al. developed Hierarchical Factored Conditional Restricted Boltzmann Machine (HFCRBM) for audio-based smooth gesture generation, and Yang et al. (Yang et al., 2020) implemented statistical motion

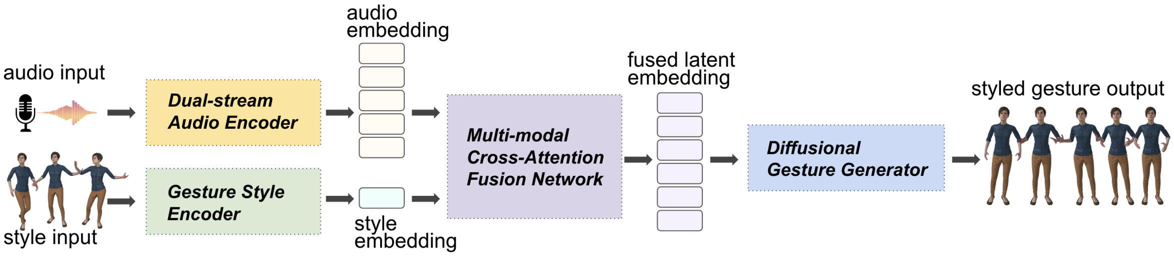

Figure 1: Framework overview, where four major components are illustrated, i.e. dual-stream audio encoder, gesture style encoder, multi-modal cross-attention module and diffusion-based gesture generator.

graphs for synchronized body motions, enhancing diversity with stochastic search algorithms.

Deep Learning for Gesture Synthesis

Deep learning has significantly advanced co-speech gesture generation by enabling the synthesis of natural and diverse gestures from large datasets, eliminating the need for manually designed lexicons and mapping rules. Early approaches utilized deterministic models such as Convolutional Neural Networks (CNNs) (Habibie et al., 2021) and Recurrent Neural Networks (RNNs) (Liu et al., 2022; Yoon et al., 2019; Yoon et al., 2020) to map speech inputs directly to gesture sequences. While these models improved the perceived naturalness and appropriateness of generated gestures, they often produced more averaged and less diverse outputs.

Generative models have emerged as a superior alternative by introducing stochasticity into the generation process, leading to more diverse and human-like gestures. These approaches include Normalizing Flows, Generative Adversarial Networks (GANs), and diffusion-based models. Generative models, such as Normalizing Flows, Variational Autoencoders (VAEs), and Vector Quantized VAEs (VQ-VAEs), have been employed to learn diverse and realistic gesture distributions. For instance, Ahuja et al. (Ahuja et al., 2020) developed a Temporal Convolutional Network (TCN) to create stylized gestures, enhancing motion expressiveness, Yoon et al. (Yoon et al., 2020) utilized adversarial networks with multimodal information for gesture generation, and Li et al. (Li et al., 2021) employed VAEs to train generators with shared and motion-specific latent spaces for coherent gesture sequences. Despite these advancements, generative models often suffer from low semantic alignment with speech input due to the inherent many-to-many relationship between speech and gestures. Recent approaches aim to improve intent alignment with gesture prediction and incorporate gesture styles for personalized synthesis.

Diffusion Models for Gesture Generation

Recently, diffusion-based models have advanced gesture generation by leveraging stochastic diffusion processes to learn data distributions, enhancing flexibility and diversity. These models produce gestures that are semantically or emotionally aligned with input speech. Notable approaches include DiffGesture, which employs a transformer-based diffusion pipeline with annealed noise sampling for temporal consistency (Zhu et al., 2023); GestureDiffuCLIP, which integrates latent-diffusion models and CLIP-based conditioning for better control (Ao et al., 2023); TalkSHOW, utilizing VQ-VAEs for body and hand motions (Yi et al., 2023); and LDA, which provides style control using classifier-free guidance for diffusion models in both music-to-dance (Alexanderson et al., 2023). Additionally, models have been developed for predicting the movement of multiple speakers in social settings (Tanke et al., 2023), multi-modal diffusion for video and audio generation (Ruan et al., 2023), and efficient omni-modal representation learning paradigms (Lei et al., 2023).

Despite these advancements, diffusion-based generative models still struggle to maintain semantic alignment with speech due to the many-to-many relationship between speech and gestures. Recent research aims to improve intent alignment and incorporate personalized gesture styles.

3 METHOD

In this work, we propose a novel framework for generating diverse and stylized co-speech gestures through diffusion. As illustrated in Figure 1, our framework mainly consists of four components: a dual-stream audio encoder (Sec. 3.1), a gesture style encoder (Sec. 3.2), a multi-modal cross-attention fusion module (Sec. 3.3) and a diffusion-based gesture generator (Sec. 3.4). Given a speech input, the framework allows users to provide a one-short gesture example to specify their desired target style. The dual-stream audio encoder and the gesture style encoder take the speech and the style example as input, and project them to audio embeddings and style embeddings respectively. Instead of naively piecewising the audio and style embeddings together, our multimodal cross-attention module correlates the audio emphasis and speech elements with the salient gesture style features, and aligns embeddings from the two modalities in one unified latent space. Lastly, the unified embeddings are passed to diffusion-based gesture generator, which synthesizes co-speech gesture performance in the target style frame by frame in an auto-regressive fashion. We discuss the details of each component of our gesture generation framework in the following sections.

3.1 Speech Encoding

Given a speech input, our framework employs a dual-stream audio encoder to project the speech audio into latent embeddings. Specifically, the audio input is a sequence of T-frame total length. At each frame, a window of N neighboring frames is cut into an audio segment and fed to the audio encoder to extract its features. We propose to extract the audio features from two streams: raw audio input in time domain, and mel-spectrogram in frequency domain. Outputs from the two streams are then fused into the sequence of audio embedding vectors denoted by $A = [a_0, a_1, \ldots, a_{T-1}]$ where $A \in \mathbb{R}^{T \times D_a}$, and D_a denotes the dimension of the audio embedding vector for each frame.

Time Domain Stream: A convolutional neural network (CNN) designed to process raw audio features. It consists of four 1D convolutional layers with progressively increasing channel sizes (64, 128, 256, 512) and kernel sizes (1, 3, 5, 7). Each convolutional layer is followed by a GELU activation function and dropout layers for regularization. The convolutional layers are succeeded by three fully connected layers, reducing the dimensionality to the target embedding size.

Frequency Domain Stream: Based on the Audio Spectrogram Transformer (AST) architecture (Gong et al., 2021), this encoder transforms the input spectrogram into embedded patches using a patch embedding layer. Positional embeddings are added to provide spatial context. The patches are processed through multiple self-attention layers and feedforward networks, producing a refined spectrogram encoding.

Stream Fusion Block: The Fusion block integrates the outputs from the Speech Encoder and Audio Spectrogram Encoder by concatenating their outputs. The combined vector is then normalized using layer normalization and processed through a multilayer perceptron network (MLP) with SiLU activation. This fusion allows the model to jointly learn from both time-domain (raw audio) and frequency-domain (spectrogram) representations, resulting in a comprehensive audio embedding.

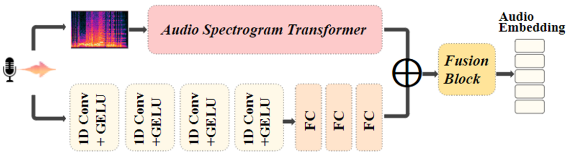

Figure 2: Dual-stream Audio Encoder.

3.2 Style Encoding

The style input processing module encodes desired gesture characteristics, including motion type, posture, and expressiveness. It utilizes detailed animation and joint data, encompassing joint local translations and rotations, translational and rotational velocities, and joint movements relative to the character's root transform (Ghorbani et al., 2023). These features capture both static and dynamic properties of gestures, ensuring that the generated gestures are realistic and stylistically accurate.

Each frame of the animation clip is represented by a feature vector $\mathbf{a} = [\rho_p, \rho_r, \dot{\rho}_p, \dot{\rho}_r, \dot{r}_p, \dot{r}_r]$, where $\rho_p \in \mathbb{R}^{3j}$ and $\rho_r \in \mathbb{R}^{6j}$ represent the joint local translations and rotations, $\dot{\rho}_p$ and $\dot{\rho}_r$ represent the joint local translational and rotational velocities, and $\dot{r}_p$ and $\dot{r}_r$ represent the character root translational and rotational velocity local to the character root transform. $\mathbf{j}$ corresponds to the number of joints in the kinematic tree.

Inspired by attention mechanisms and variational autoencoder (VAE) (Vaswani, 2017; Kingma, 2013), the Style Encoder transforms a reference style animation clip into a low-dimensional embedding vector that encodes the stylistic properties of the gestures. A Variational Auto-Encoder (VAE) samples the style embeddings from multivariate Gaussian distribution. The extracted style sequence is then processed through convolutional layers and an Attention-based Feed Forward Transformer network to produce the style embedding vector $\mathbf{e}$.

3.3 Multimodal Cross-Attention Fusion

The Cross-Attention Fusion Network integrates the audio and style embeddings to enable the generator to generate semantically and stylistically coherent gestures. Leveraging multi-head attention mechanisms,

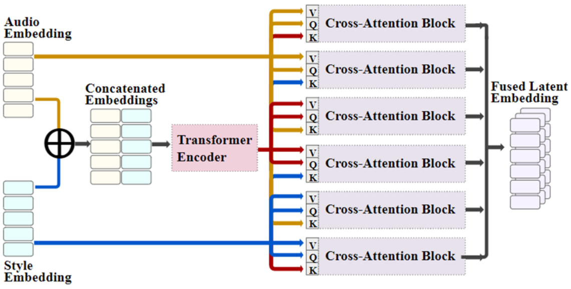

Figure 3: Multimodal Cross-Attention Fusion Network.

the cross-attention network captures the intricate relationships between speech and style. The audio and style embeddings are first concatenated. Multi-head attention layers are then applied to capture interactions between these modalities as shown in Figure 3. A feedforward network further processes the fused embeddings to the target embedding size, enabling the model to produce high-quality, context-aware latent embeddings for gesture generation.

Let F_A represent the deep features extracted from the speech encoder, and F_S represent the deep features from the style encoder. The joint feature representation is obtained by concatenating F_S and F_A, followed by a transformer encoder with self-attention mechanisms. Cross-attention layers are subsequently applied to share context between audio and style features, producing a mixed encoding that informs the final gesture output.

3.4 Gesture Diffusion

Our gesture generation method leverages a diffusion transformer model operating within a pose feature space to synthesize realistic and contextually appropriate gestures. During training, we employ a forward diffusion process that incrementally adds Gaussian noise to the initial pose sequence representation x_0, resulting in a sequence of progressively noisier pose representations $\{x_t\}_{t=1}^{T}$ that approximate a standard normal distribution $\mathcal{N}(0,I)$. This process is defined by eq.(1), where β_t is a predefined variance schedule.

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}\,x_{t-1}, \beta_t I\right), \tag{1}$$

The cumulative effect over t timesteps can be expressed directly in terms of x_0 in eq.2 with $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$.

$$q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}\,x_0, (1-\bar{\alpha}_t)I\right), \tag{2}$$

In the reverse diffusion process, our model learns to recover the original pose sequence from the noisy input by estimating the noise added at each timestep. The denoising model ε_θ predicts the noise given the noisy pose x_t, the timestep t, and the conditioning information c:

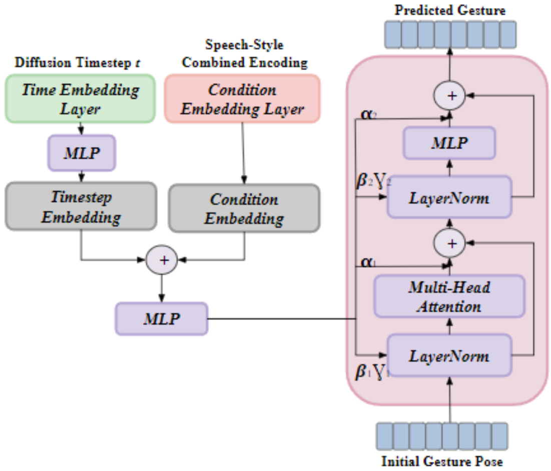

Figure 4: Our Diffusion Transformer architecture based on the adaLN-Zero architecture introduced by (Peebles and Xie, 2023).

$$x_0 = \varepsilon_\theta(x_t, t, c). \tag{3}$$

The conditioning information c is represented by the output embedding of our miltomodal corss-attention fusion network. The model is trained by minimizing the mean squared error between the predicted noise and the actual noise added during the forward process:

$$\mathcal{L}_{\text{LD}} = \mathbb{E}_{x_0,t,\varepsilon}\left[\|\varepsilon - \varepsilon_\theta(x_t,t,c)\|^2\right], \tag{4}$$

where $\varepsilon \sim \mathcal{N}(0,I)$ and $x_t = \sqrt{\bar{\alpha}_t}\,x_0 + \sqrt{1-\bar{\alpha}_t}\,\varepsilon$.

During inference, we generate gestures by starting from random noise $x_T \sim \mathcal{N}(0,I)$ and iteratively applying the reverse diffusion steps using the Denoising Diffusion Probabilistic Model (DDPM) sampling algorithm (Ho et al., 2020) to obtain the denoised pose sequence x_0. At each timestep t, the model predicts the noise to be removed, guided by the conditioning information c.

4 IMPLEMENTATION

4.1 Dataset and Data Preprocessing

We train and evaluate our system using the ZeroEGGS dataset (Ghorbani et al., 2023), which comprises full-body motion capture and synchronized audio recordings from a single English-speaking female actor performing 67 monologues across 19 distinct gesture styles. These styles range from posture-focused categories like ”Tired” and ”Oration” to intricate hand and head movements, ensuring a wide variety of gesture types for training context-appropriate

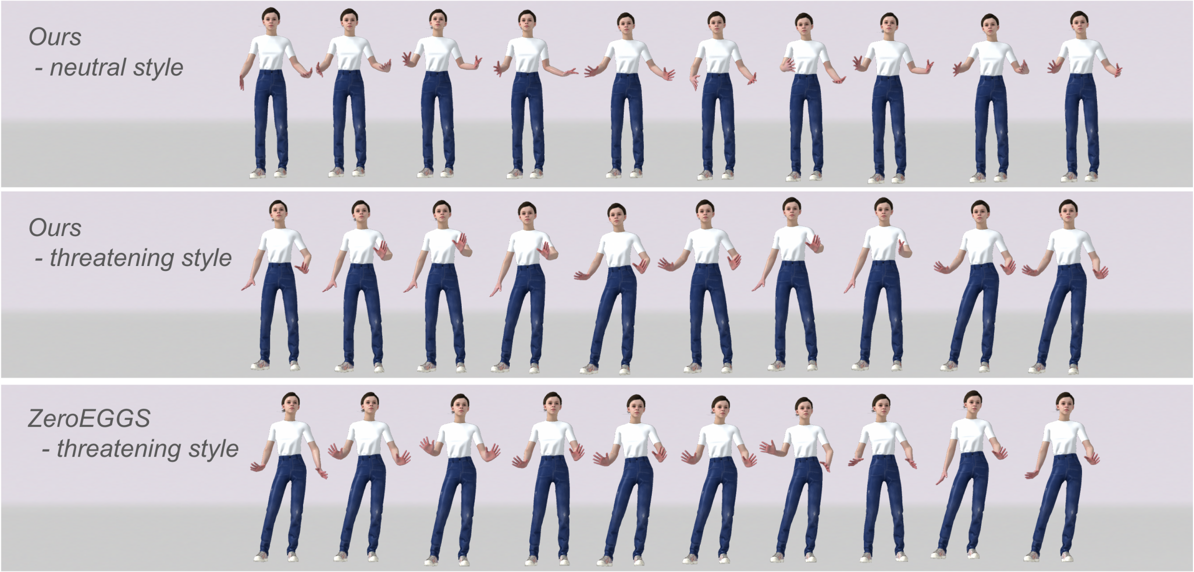

Figure 5: Visual result comparison.

models. The dataset includes 135 minutes of data recorded at 60 frames per second, represented by a 75-joint skeletal model that captures detailed hand and finger movements, providing high fidelity for gesture generation tasks.

To augment the data, we mirrored all animation sequences, effectively doubling the training data. Head orientation was processed by projecting the head z-axis direction onto the ground plane and computing the median to establish a global target facing direction for each sequence, which is set to the global z-axis during runtime. The style labels, based on actor instructions, may differ from external annotations, introducing subjectivity that is considered during training and analysis. Preprocessing steps include normalizing skeleton data to ensure consistent joint positioning, downsampling audio to 16kHz for compatibility with the speech encoder, and extracting Mel-frequency cepstral coefficients (MFCCs) and energy per frame to represent speech content.

4.2 Implementation Details

Our gesture generation model is implemented using PyTorch and trained on an NVIDIA RTX 3060 GPU with a batch size of 32 and an initial learning rate of 0.0001. We utilize the RAdam optimizer for its adaptive learning rate properties and an exponential learning rate scheduler to promote faster convergence and better generalization.

For audio feature extraction, our speech encoder comprises a custom CNN-based encoder that processes the energy and log-amplitude of mel-spectrograms, alongside a pretrained Audio Spectrogram Transformer (AST) that extracts additional mel-spectrogram features. These features are fused to form a single feature vector representing each speech segment. Gesture style data is encoded using an attention-based style encoder, which captures general features from a reference animation style sample clip with a dynamic window length between 256 and 512 frames, sampled from the same animation clip as the target sequence. Our diffusion-based gesture generator employs 1000 diffusion timesteps with a linear variance schedule ranging from $\beta_1 = 1 \times 10^{-4}$ to 0.1, and the hidden dimension of all transformer layers in the Diffusion Transformer (DiT) is set to 1024.

5 RESULTS

5.1 Training Loss

The training process maximizes the Evidence Lower Bound (ELBO) of the gesture motion's log-likelihood given a speech sequence by minimizing the negative ELBO, which serves as the training loss. The total loss is defined as:

$$\begin{aligned} L &= \mathbb{E}_{q(z|e)}\left[-\log p(Y \mid S, z)\right] + D_{\mathrm{KL}}(q(z \mid e) \parallel p(z)) \\ &= L_{\mathrm{recon}} + D_{\mathrm{KL}}(q(z \mid e) \parallel p(z)) + L_{\mathrm{LD}} \end{aligned} \tag{5}$$

Reconstruction Loss. L_{recon} evaluates how accurately the model reconstructs the target gesture sequence from speech and style embeddings. It is composed of:

$$L_{\text{recon}} = \lambda_p L_p + \lambda_r L_r + \lambda_{vp} L_{vp} + \lambda_{vr} L_{vr} + \lambda_{dp} L_{dp} + \lambda_{dr} L_{dr} + \lambda_f L_f \quad (6)$$

where:

- L_p and L_r: Mean Absolute Error (MAE) for joint positions and rotations, ensuring pose accuracy.
- L_{vp} and L_{vr}: MAE for joint translational and rotational velocities, promoting smooth motion.
- L_{dp} and L_{dr}: MAE of velocities computed via finite differences, enhancing motion smoothness.
- L_f: MAE for the facing direction in world space, preventing rotational drift.

The weights λ_p, λ_r, λ_{vp}, λ_{vr}, λ_{dp}, λ_{dr}, and λ_f balance each loss component and are empirically determined during training.

Regularization Term. $D_{\text{KL}}(q(z \mid e) \parallel p(z))$ measures the Kullback–Leibler divergence between the posterior distribution $q(z \mid e)$ from the style encoder and the prior $p(z)$, a standard Gaussian. This encourages the latent space to resemble the prior, preventing overfitting and enhancing generalization.

Cost Annealing. gradually increases the weight of the regularization term during training, stabilizing the learning process and promoting a meaningful latent space.

Diffusion Loss. $\mathcal{L}_{\text{LD}}$ is the standard noise estimation loss used in diffusion models (Ho et al., 2020):

$$\mathcal{L}_{\text{LD}} = \mathbb{E}_{x_0,t,\varepsilon}\left[\|\varepsilon - \varepsilon_\theta(x_t,t,c)\|^2\right] \quad (7)$$

5.2 Qualitative Results

Figure 5 displays the gesture generation results for a sample from the ZeroEGGS dataset. The top figures illustrate gestures produced by our framework, while the bottom figure shows those generated by the ZeroEGGS (Ghorbani et al., 2023) model. Our model effectively captures emotional and semantic cues, demonstrating the desired style with consistent motion and appropriate emphasis in threatening gestures. In contrast, ZeroEGGS fails to accurately capture the style, resulting in average gestures with noticeable repetitiveness. More qualitative results are available in the video submission for our approach with different styles.

6 CONCLUSIONS

In this study, we introduced an example-based stylistic co-speech gesture generation framework that effectively produces expressive gestures aligned with speech and desired styles. The framework combines a dual-stream audio encoder, a gesture style encoder, a multimodal cross-attention fusion module, and a diffusion-based gesture generator to create high-quality and diverse gesture performances. Qualitative results show that the model outperforms benchmark systems by generating gestures that are both contextually appropriate, coherent, and realistic. Future work aims to enhance the framework's robustness and scalability by evaluating it on larger, more diverse datasets and benchmarking against state-of-the-art methods. We also plan to address limitations in generating specific styles—such as laughter and elderly gestures—to improve the model's generalization capabilities, conduct user studies to validate the naturalness of the gestures, and optimize the inference process by reducing the number of required seed frames for faster, near real-time gesture generation.

REFERENCES

Ahuja, C., Lee, D. W., Nakano, Y. I., and Morency, L.-P. (2020). Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 248–265. Springer.

Alexanderson, S., Nagy, R., Beskow, J., and Henter, G. E. (2023). Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–20.

Ao, T., Zhang, Z., and Liu, L. (2023). Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics (TOG)*, 42(4):1–18.

Bergmann, K. and Kopp, S. (2009). Increasing the expressiveness of virtual agents: autonomous generation of speech and gesture for spatial description tasks. In *AAMAS (1)*, pages 361–368.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (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.

Cassell, J., Vilhjálmsson, H. H., and Bickmore, T. (2001). Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 477–486.

David, M. (1992). Hand and mind: What gestures reveal about thought. *University of Chicago press.[Google Scholar].*

De Carolis, B., Pelachaud, C., Poggi, I., and Steedman, M. (2004). Apml, a markup language for believable behavior generation. *Life-like characters: tools, affective functions, and applications*, pages 65–85.

Ghorbani, S., Ferstl, Y., Holden, D., Troje, N. F., and Carbonneau, M.-A. (2023). Zeroeggs: Zero-shot example-based gesture generation from speech. In *Computer Graphics Forum*, volume 42, pages 206–216. Wiley Online Library.

Gong, Y., Chung, Y.-A., and Glass, J. (2021). Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

Habibie, I., Xu, W., Mehta, D., Liu, L., Seidel, H.-P., Pons-Moll, G., Elgharib, M., and Theobalt, C. (2021). Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 101–108.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Kingma, D. P. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kipp, M. (2001). Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European conference on speech communication and technology*. Citeseer.

Kopp, S., Jung, B., Lessmann, N., and Wachsmuth, I. (2003). Max-a multimodal assistant in virtual reality construction. *KI*, 17(4):11.

Kopp, S., Krenn, B., Marsella, S., Marshall, A. N., Pelachaud, C., Pirker, H., Thórisson, K. R., and Vilhjálmsson, H. (2006). Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent Virtual Agents: 6th International Conference, IVA 2006, Marina Del Rey, CA, USA, August 21-23, 2006. Proceedings 6*, pages 205–217. Springer.

Kopp, S. and Wachsmuth, I. (2002). Model-based animation of co-verbal gesture. In *Proceedings of Computer Animation 2002 (CA 2002)*, pages 252–257. IEEE.

Lei, W., Ge, Y., Yi, K., Zhang, J., Gao, D., Sun, D., Ge, Y., Shan, Y., and Shou, M. Z. (2023). Vit-lens-2: Gateway to omni-modal intelligence. *arXiv preprint arXiv:2311.16081*.

Levine, S., Theobalt, C., and Koltun, V. (2009). Real-time prosody-driven synthesis of body language. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–10.

Li, J., Kang, D., Pei, W., Zhe, X., Zhang, Y., He, Z., and Bao, L. (2021). Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11293–11302.

Liu, X., Wu, Q., Zhou, H., Xu, Y., Qian, R., Lin, X., Zhou, X., Wu, W., Dai, B., and Zhou, B. (2022). Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10462–10472.

McNeill, D. (2019). *Gesture and thought*. University of Chicago press.

Neff, M., Kipp, M., Albrecht, I., and Seidel, H.-P. (2008). Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions On Graphics (TOG)*, 27(1):1–24.

Peebles, W. and Xie, S. (2023). Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.

Pelachaud, C., Carofiglio, V., De Carolis, B., de Rosis, F., and Poggi, I. (2002). Embodied contextual agent in information delivering application. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*, pages 758–765.

Piwek, P., Krenn, B., Schröder, M., Grice, M., Baumann, S., and Pirker, H. (2004). Rrl: A rich representation language for the description of agent behaviour in neca. *arXiv preprint cs/0410022*.

Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N. J., Jin, Q., and Guo, B. (2023). Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228.

Tanke, J., Zhang, L., Zhao, A., Tang, C., Cai, Y., Wang, L., Wu, P.-C., Gall, J., and Keskin, C. (2023). Social diffusion: Long-term multiple human motion anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9601–9611.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Vilhjálmsson, H., Cantelmo, N., Cassell, J., E. Chafai, N., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A. N., Pelachaud, C., et al. (2007). The behavior markup language: Recent developments and challenges. In *Intelligent Virtual Agents: 7th International Conference, IVA 2007 Paris, France, September 17-19, 2007 Proceedings 7*, pages 99–111. Springer.

Wei, Y., Hu, D., Tian, Y., and Li, X. (2022). Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*.

Yang, Y., Yang, J., and Hodgins, J. (2020). Statistics-based motion synthesis for social conversations. In *Computer Graphics Forum*, volume 39, pages 201–212. Wiley Online Library.

Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., and Black, M. J. (2023). Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480.

Yoon, Y., Cha, B., Lee, J.-H., Jang, M., Lee, J., Kim, J., and Lee, G. (2020). Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16.

Yoon, Y., Ko, W.-R., Jang, M., Lee, J., Kim, J., and Lee, G. (2019). Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE.

Zhu, L., Liu, X., Liu, X., Qian, R., Liu, Z., and Yu, L. (2023). Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10553.