



OPEN ACCESS

EDITED BY

Antonio Andriella,
Spanish National Research Council
(CSIC), Spain

REVIEWED BY

Luca Raggioli,
University of Naples Federico II, Italy
Alessandra Sorrentino,
University of Florence, Italy

*CORRESPONDENCE

Xixin Shen,
yixishen1@student.unimelb.edu.au

RECEIVED 17 December 2024

ACCEPTED 14 March 2025

PUBLISHED 08 April 2025

CITATION

Shen Y and Johal W (2025) TED-culture: culturally inclusive co-speech gesture generation for embodied social agents. *Front. Robot. AI* 12:1546765.
doi: 10.3389/frobt.2025.1546765

COPYRIGHT

© 2025 Shen and Johal. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

TED-culture: culturally inclusive co-speech gesture generation for embodied social agents

Xixin Shen* and Wafa Johal

School of Computer Science and IT, FEIT, University of Melbourne, Melbourne, VIC, Australia

Generating natural and expressive co-speech gestures for conversational virtual agents and social robots is crucial for enhancing their acceptability and usability in real-world contexts. However, this task is complicated by strong cultural and linguistic influences on gesture patterns, exacerbated by the limited availability of cross-cultural co-speech gesture datasets. To address this gap, we introduce the TED-Culture Dataset, a novel dataset derived from TED talks, designed to enable cross-cultural gesture generation based on linguistic cues. We propose a generative model based on the Stable Diffusion architecture, which we evaluate on both the TED-Expressive Dataset and the TED-Culture Dataset. The model is further implemented on the NAO robot to assess real-time performance. Our model surpasses state-of-the-art baselines in gesture naturalness and exhibits rapid convergence across languages, specifically Indonesian, Japanese, and Italian. Objective and subjective evaluations confirm improvements in communicative effectiveness. Notably, results reveal that individuals are more critical of gestures in their native language, expecting higher generative performance in familiar linguistic contexts. By releasing the TED-Culture Dataset, we facilitate future research on multilingual gesture generation for embodied agents. The study underscores the importance of cultural and linguistic adaptation in co-speech gesture synthesis, with implications for human-robot interaction design.

KEYWORDS

co-speech gesture generation, human-robot interaction, social agents, virtual avatar, humanoid robot

1 Introduction

As virtual agents and robots are becoming more popular, optimizing the interaction between humans and these technologies is becoming increasingly important. According to a human evaluation instrument, the ESI (Evaluation of Social Interaction) ([Fisher and Griswold, 2010](#)), some essential social skills are identified, such as approaching, speaking, turn-taking, gazing, and gesturing. These social interaction skills are equally applicable to the interaction between humans with virtual agents or robots. A significant portion of human interaction occurs through non-verbal means, frequently involving gestures made alongside spoken language ([Knapp et al., 2013](#)). Since the gestures contain rich non-verbal information, these movements play an important role in human communication ([Studdert-Kennedy, 1994](#)). The rise of telepresence in virtual/augmented

reality, 3D animation, and social games highlights the importance of real-time gesture generation in dialogues and conversations (Lee et al., 2019a). Therefore, accompanying the natural co-speech gestures to virtual agents and robots is extremely desired.

Two main approaches to gesture generation are rule-based and data-driven methodologies (Liu et al., 2021). In addition, combining both approaches, hybrid systems have been introduced in some recent studies to generate natural and semantically meaningful gestures (Zhou et al., 2022). Rule-based systems can be repetitive and monotonous, while data-driven approaches leverage deep neural networks, from CNN (Habibie et al., 2021) to GANs (Goodfellow et al., 2014); (Liu et al., 2023). Despite GANs' state-of-the-art performance, they face challenges like mode collapse and unstable training. Inspired by the success of the Stable Diffusion (Ho et al., 2020) in image creation, exploring its application in gesture generation is promising.

Gesture generation research is limited by data scarcity. Capturing finger motion accurately remains challenging. Studies have demonstrated a close relationship between speech and gestures across various cultures (Kita, 2009). However, existing models and datasets have not adequately addressed cultural or linguistic impacts on gesture generation. The BEAT Dataset (Liu H. et al., 2022) includes four languages and diverse cultural backgrounds but has limited speaker diversity and high data collection costs. Similarly, Gjaci et al. (2022) introduced a dataset featuring Indian and English speakers, however, both groups speak only in English, limiting its utility for investigating cultural factors. To address these gaps, a multimodal dataset that captures cultural differences, ensures speaker diversity, and includes detailed finger motion data would significantly advance the field.

Gesturing predominantly involves upper body and finger movements, with precise finger motion being particularly challenging. Improving finger motion quality could significantly enhance the authenticity and appropriateness of distal finger movements in social agents such as robots. Only two social robots, BERTI Bremner et al. (2009) and Erica Ishi et al. (2018), can render complete finger motions. Research on gesture generation for these robots is limited. Yoon et al. converted 2D poses to 3D poses and retargeted them to the NAO robot but did not address finger movements due to the absence of individual actuators of NAO's fingers (Yoon et al., 2019). Therefore, the challenge of rendering finger motions on robots such as NAO using a dataset containing finger motion remains an unaddressed gap in the field.

To address the aforementioned challenges, our main contributions are as follows: (1) Introducing the TED-Culture Dataset for cross-cultural gesture generation based on linguistic cues; (2) Developing a novel generative model based on the DiffGesture framework (Zhu et al., 2023), achieving state-of-the-art performance on the TED-Expressive Dataset and rapid convergence across several languages in the TED-Culture Dataset; (3) Analyzing cultural factors using the TED-Culture Dataset, showing that people are more critical of outputs in familiar languages; and (4) Developing a robot prototype that maps our model's gestures onto

an NAO robot, enabling it to speak six languages with corresponding gestures.

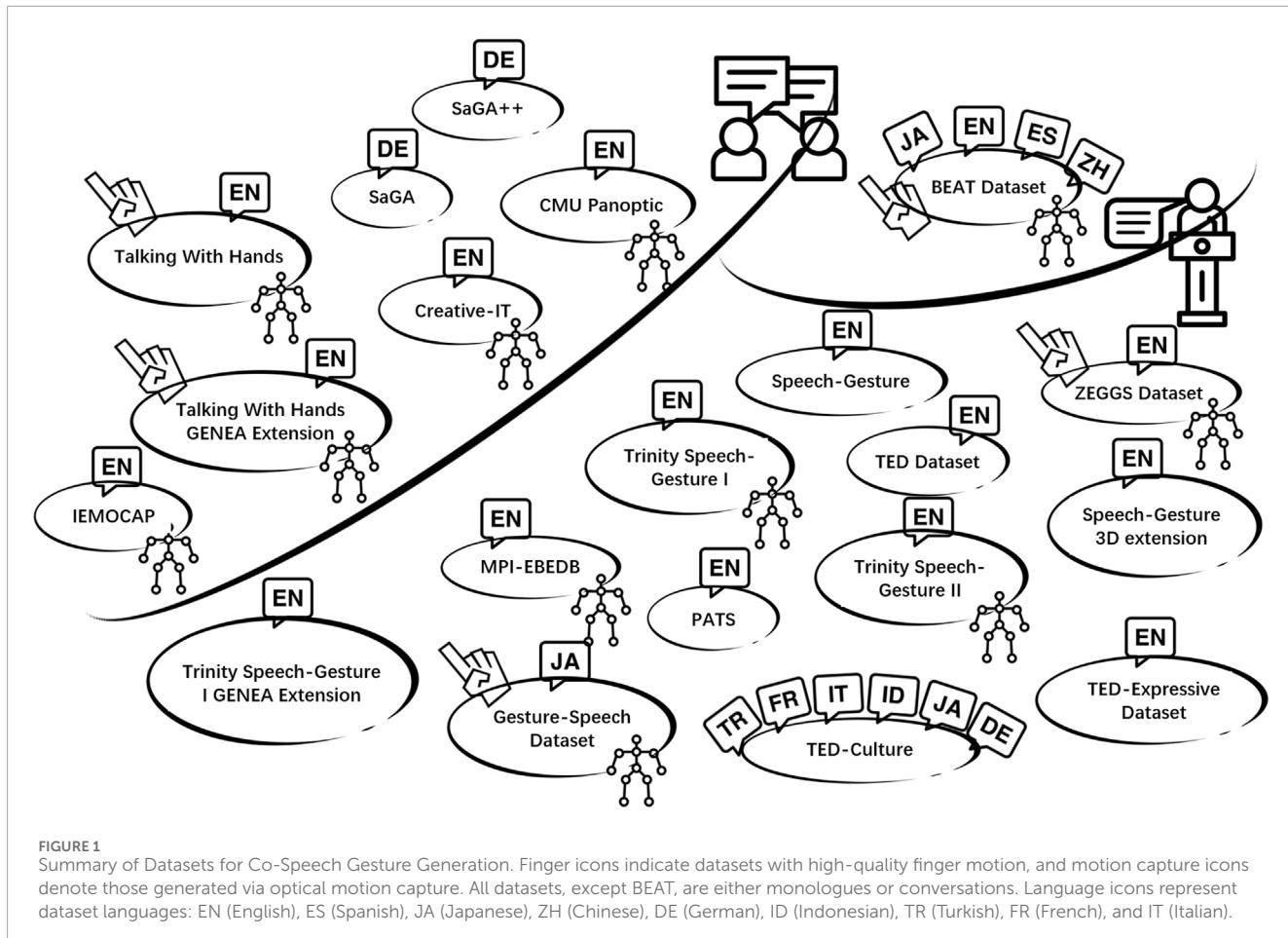
2 Related works

2.1 Co-speech gesture generation

Data-driven approaches have gained popularity due to their ability to reduce manual efforts in designing rules, unlike rule-based methods. Nyatsanga et al. (2023) provide a comprehensive survey of these methods. Data-driven approaches are categorized into statistical and learning-based methods. Statistical approaches use models to derive rules from data, mapping input to gesture units for generation, as seen in Levine et al. (2010). With advancements in deep neural networks and increased human gesture data, end-to-end gesture generation systems have emerged. These systems include deterministic models like MLP (Kucherenko et al., 2020), CNN (Habibie et al., 2021), and RNN (Yoon et al., 2019); Yoon et al., 2020; (Liu X. et al., 2022), as well as non-deterministic models like VAEs (Ghorbani et al., 2023) and GANs (Yoon et al., 2020). Recent hybrid systems (Zhou et al., 2022) combine rule-based and data-driven methods to address issues like regression to the mean, showing superior performance. GAN-based systems have achieved cutting-edge results but face challenges like mode collapse and training instability. Inspired by Stable Diffusion (Ho et al., 2020) in image creation, new studies (Zhu et al., 2023) have adapted this framework for gesture generation, achieving state-of-the-art performance, though cross-cultural testing is still needed.

2.2 Human motion capture datasets

Creating a dataset for human gesture motion involves two primary methods: optical motion capture (Takeuchi et al., 2017); (Lee et al., 2019b) and pose estimation from monocular videos (Yoon et al., 2019; Habibie et al., 2021). While the quality of optical motion capture is relatively higher than monocular video, it tends to yield smaller datasets due to its higher cost and the labor-intensive nature of the data collection. In contrast, pose estimation from monocular videos offers the advantage of generating larger datasets since a plethora of online videos are available for analysis. Despite the growing availability and sizes of multi-modal datasets, those encompassing high-quality finger motion remain scarce. Figure 1 provides an overview of datasets used for co-speech gesture generation. Five speech-gesture datasets incorporate high-quality finger motion, (namely (Takeuchi et al., 2017), (Lee et al., 2019b), (Yoon et al., 2022), (Ghorbani et al., 2023), and (Liu H. et al., 2022)), and they are all collected using optical motion capture devices. Among these, the Gesture-Speech Dataset (Takeuchi et al., 2017) and ZEGGS Dataset (Ghorbani et al., 2023) are presented in a monologue context, where only one person conducts an interview or talks and is then recorded. The Talking With Hands Dataset (Lee et al., 2019b), on the other hand, comprises multi-modal recordings of face-to-face spontaneous conversations involving two individuals, making it the largest motion capture and audio dataset for natural conversations to date. Other publicly



available datasets lack high-quality finger motion, such as the TED Dataset (Yoon et al., 2019), which relies on pose estimation from monocular videos but is limited to 2D motion format, making it unsuitable for mapping onto social robots or 3D virtual agents and lacking finger motion data.

Fortunately, recent work from Habibie et al. (2021) has extended 2D skeletons into the three-dimensional realm. Yoon et al. (2020) expanded the TED Dataset from Yoon et al. (2019) by incorporating more TED videos into it, then converted all human poses into a 3D format using the 3D pose estimator (Pavllo et al., 2019). Additionally, Liu et al. introduced the TED-Expressive Dataset (Liu X. et al., 2022), addressing a limitation of the original TED Dataset Yoon et al. (2020) by including expressive co-speech finger movements alongside upper body key points. In the realm of culture-related datasets, the BEAT Dataset (Liu H. et al., 2022) places its main emphasis on capturing the emotions expressed by actors, and it offers multi-cultural and multi-language gesture data since the thirty participants are from ten countries. However, it is important to recognize that this dataset requires significant resources and has a limited variety of speaker identities, largely attributable to its reliance on optical motion capture technology. Consequently, among the existing 3D datasets, none have addressed the influence of culture or language on gesture generation while simultaneously ensuring efficient data collection and encompassing a diverse range of speaker identities.

3 TED-culture dataset

To address the scarcity of the culture-aware co-speech gesture dataset, we developed a new dataset called *TED-Culture Dataset*, featuring six different languages: Indonesian, Japanese, German, Italian, French, and Turkish. Inspired by the TED Dataset built by Yoon et al. (2019), Yoon et al. (2020), we choose the TEDx Talks channel on YouTube¹ as the original source of our dataset. While TED Talks have certain limitations, such as the lack of representativeness of TED speakers in reflecting real-world diversity and the tendency of these professional and trained speakers to overuse gestures, the TED Talk video source has some advantages including ample data, diversity, and well-prepared gestures. We follow the dataset collection pipeline of Liu X. et al., (2022), excluding word-level alignments due to the use of auto-generated subtitles. Following the production process, we consolidated the individual language datasets into a unified “Merged” dataset.

Regarding the source videos of the TED-Culture Dataset, the average video length is 15 minutes. The total length of the source video is 60.1 h, with each language having a relatively even distribution of approximately 10 h each. The final dataset format represents 3D coordinates, encompassing multimodal aspects

¹ <https://www.youtube.com/@TEDx>

TABLE 1 Statistics information about the TED-Culture Dataset.

Language	Speakers	Valid clip number	Seconds	Hours
Turkish	21	306	11,729	3.3
French	29	457	9,701	2.7
Italian	35	580	11,651	3.2
Indonesian	29	351	8,343	2.3
German	39	901	14,142	3.9
Japanese	30	202	7,407	2.1
Total	183	2,797	62,974	17.5

including Gesture, Audio, and Text. All poses are spine-centered, with 43 key points (13 upper body joints and 30 finger joints) defined in the dataset. The statistical information for the TED-Culture Dataset is presented in Table 1.

In the table, we note a total of 183 speakers in our dataset, surpassing speaker counts in datasets captured by motion capture devices (typically involving fewer than or equal to 50 speakers), as well as most pose estimation datasets like Speech-Gesture (Ginosar et al., 2019) and Speech-Gesture 3D extension (Habibie et al., 2021). The TED-Culture Dataset comprises 2,797 valid clips, totaling 17.5 h with an average clip length of 23 s. The distribution of our dataset demonstrates a relatively even distribution across all languages. German holds the largest share at 22.3%, while Japanese has the smallest share at 12%. The distribution percentages for the other languages are as follows: Indonesian 13.1%, Italian 18.3%, French 15.4%, and Turkish 18.9%.

4 The proposed approach

Expanding upon insights from speech and its accompanying gestures, we have developed a deep learning approach focused on uncovering the intrinsic relationship between these modalities. Figure 2 provides an overview of our proposed *DiffCulture* framework, which is based on Zhu et al. (2023) and aims to enhance the fidelity of co-speech gesture generation. Unlike the original DiffGesture model (Zhu et al., 2023), our approach updates the objective function and modifies the architecture of the audio encoder. Despite testing other similar audio encoders in previous work (Zhu et al., 2023) (Zhi et al., 2023), no performance gains were observed. For details on the individual components of the framework, please refer to Zhu et al. (2023).

4.1 Problem formulation

For this gesture generation problem, we utilize the large-scale co-speech gesture training corpus introduced in Section 3, which focuses on videos featuring distinct and prominent upper body movements synchronized with speech to conduct the model

learning process. Specifically, we assume every video clip has N frames, then we can define the co-speech audio sequence as $a = \{a_1, \dots, a_N\}$ and annotate the per-frame human skeletons as $x = \{p_1, \dots, p_N\}$, where p_i denotes the i^{th} pose. These skeletal representations are pre-processed as the concatenation of unit direction vectors using the same method in the baselines as in Yoon et al. (2020); Liu X. et al. (2022). The unit direction vectors are represented as $p_i = [d_{i,1}, d_{i,2}, \dots, d_{i,J}]$ where p_i means the pose description coordinates of the i^{th} frame, J is the total joint number and $d_{i,j}$ stands for the j^{th} unit direction vector among the J joints in the i^{th} image frame. The diffusion model's reverse denoising process G parameterized by θ is optimized to generate the human skeleton sequence x , taking into account the speech audio sequence a and initial poses p_1, \dots, p_M from the first M frames as conditioning factors. The primary objective of the framework is to generate a pose sequence that closely approximates the ground truth x . This objective can be mathematically formulated as $\arg \min_{\theta} \|x - G_{\theta}(a, p_1, \dots, p_M)\|$. In the context of deep learning, the system's training phase involves providing a gesture sequence (x), an audio sequence (a), and seed poses (p) as inputs. In contrast, during the testing phase, only the audio sequence (a) and seed poses (p) are supplied, while the gesture sequence (x) is predicted.

4.2 Proposed model architecture

Figure 2 shows the DiffCulture Framework we developed, which is mostly based on the DiffGesture (Zhu et al., 2023). The orange area shows the forward diffusion process given the gesture sequence x_0 . After adding Gaussian noise to the gesture sequence x_0 based on the formulation $x_t = \sqrt{a_t}x_0 + \sqrt{1 - a_t}\epsilon$, we will get a corrupted gesture x_t . The blue area indicates the context information part, which includes the initial poses $p_{(1:M)}$, speech audio a , and time embedding t . It is worth noting that we add another text modal to the context information when we do the ablation experiments. Then the given gesture sequence x_0 and context information were input into the green part together to indicate the conditional denoising process. The two neural network blocks consist of Linear Embedding, Transformer Blocks, and MLP, which receive multiple modalities in the context information and corrupted gestures with

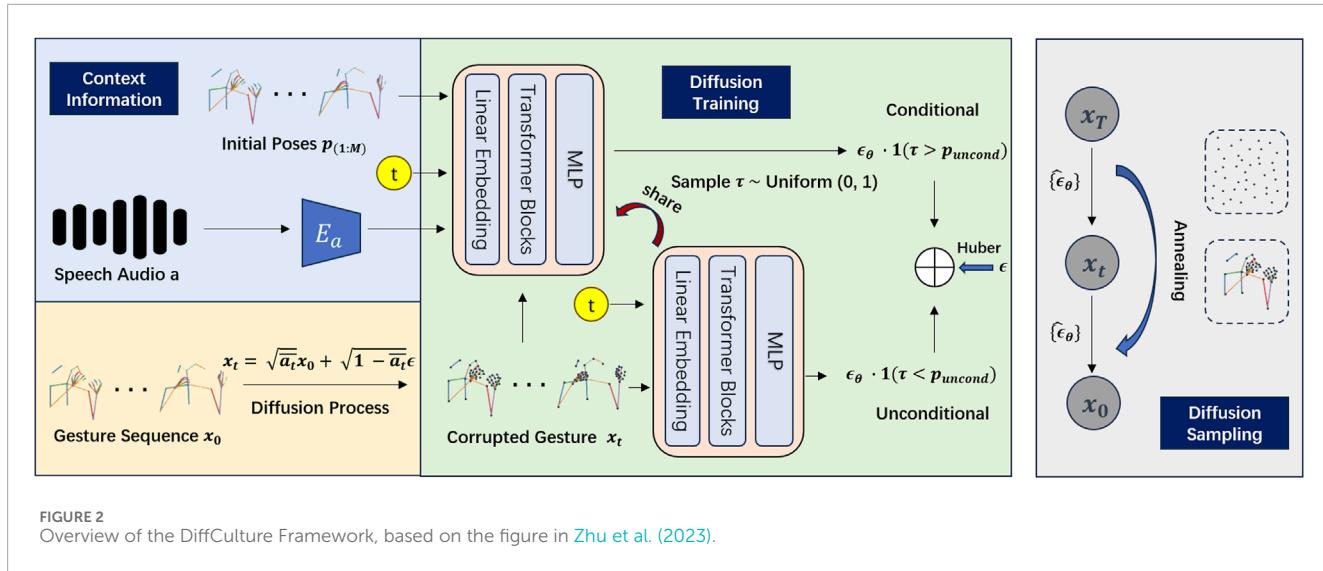


FIGURE 2
Overview of the DiffCulture Framework, based on the figure in Zhu et al. (2023).

time embedding. Instead of the MSE loss utilized in Zhu et al. (2023), we use the Huber loss as our learning objective and do the element-wise plus conditional and unconditional embedding features together for further diffusion sampling. Huber loss is a robust loss function used in regression tasks that is less sensitive to outliers than the Mean Squared Error (MSE) loss. Given that we employ the Huber loss as our learning objective, we can simplify the training objective through parameterization to an ensemble of Huber loss formulated as:

$$l_n = \begin{cases} 0.5(x_n - y_n)^2, & \text{if } |x_n - y_n| < \text{delta} \\ \text{delta}(|x_n - y_n| - 0.5 \text{ delta}), & \text{otherwise} \end{cases} \quad (1)$$

where $(x_n - y_n)^2 = L(\theta) = \mathbb{E}_q[\|\epsilon - \epsilon_\theta(\sqrt{a_t}x_0 + \sqrt{1 - a_t}\epsilon, c, t)\|^2]$. Here t is uniformly chosen from 1 to T . As we concurrently train the model in both conditional and unconditional settings, a trainable masked embedding with probability p_{uncond} replaces the context c , and the diffusion model predicts the noise in the unconditional settings.

Lastly, the grey area highlights the diffusion sampling phase, where we introduce the Diffusion Gesture Stabilizer, which employs an annealed noise sampling strategy to address temporal inconsistencies. Additionally, to integrate implicit classifier-free guidance, we jointly train conditional ($1 - p_{uncond}$) and unconditional (p_{uncond}) models. This approach enables us to balance between diversity and quality when performing inference.

5 Experiments

5.1 Co-speech gesture datasets

5.1.1 TED-expressive

In contrast to the TED Dataset (Yoon et al., 2019; Yoon et al., 2020), which includes only 10 upper body key points and lacks detailed finger movements, the TED-Expressive Dataset (Liu X. et al., 2022) offers a more comprehensive representation of both finger and body movements. This enhancement is achieved

using the state-of-the-art 3D pose estimator ExPose (Choutas et al., 2020), which captures detailed pose information. Consequently, the TED-Expressive Dataset annotates the 3D coordinates of 43 key points, encompassing 13 upper body joints and 30 finger joints.

5.1.2 TED-culture

The data collection pipeline for the TED-Culture Dataset follows the same methodology as the TED-Expressive Dataset, resulting in an identical representation format. In this work, we focus on the TED-Culture Merged dataset, with experimental results for specific languages provided in the Project Website².

5.2 Experimental settings

5.2.1 Baselines

We evaluate our method on two benchmark datasets, comparing it with several state-of-the-art methods developed in recent years: 1) Attention Seq2Seq (Yoon et al., 2019) elaborates on the attention mechanism to generate pose sequences from speech text; 2) Speech2Gesture (Ginosar et al., 2019) employs speech audio spectrums as input to adversarially generate speech gestures; 3) Joint Embedding (Ahuja and Morency, 2019) maps text and motion to the same embedding space to generate outputs from motion description text; 4) Trimodal (Yoon et al., 2020) serves as a robust baseline that learns from text, audio, and speaker identity to generate gestures, significantly outperforming previous methods; 5) HA2G (Liu X. et al., 2022) introduces a hierarchical audio learner that captures information across different semantic granularities, surpassing former methods; and 6) DiffGesture (Zhu et al., 2023) leverages the stable diffusion model (Ho et al., 2020) and Transformer architecture (Vaswani et al., 2017), achieving state-of-the-art performance. We also present evaluations directly on the pseudo Ground Truth annotated in the dataset.

² https://yixin-shen-1218.github.io/TED_Culture

5.2.2 Implementation details

In our experiments, we utilize two datasets: TED-Expressive (Liu X. et al., 2022) and TED-Culture. We preprocess these datasets following the method outlined in Yoon et al. (2020), setting the length of each pose sequence N to 34 frames. Additionally, the length of the seed gesture is set to $M = 4$, representing the gestures of the first four frames used for reference during inference. To eliminate the effect of joint lengths and root motion, we follow (Yoon et al., 2019) and represent joint positions using $J - 1$ 3D directional unit vectors. For audio processing, we employ an audio encoder consisting of three convolutional layers, each followed by a Rectified Linear Unit (ReLU) activation layer and ending with a one-dimensional AdaptiveAvgPool layer. This configuration extracts features directly from raw audio clips, encoding them into N audio feature vectors, each with 32 dimensions. These audio features are concatenated with the initial poses to form the conditional context for the diffusion model. In the diffusion process, we use $T = 500$ timesteps, with variances increasing linearly from $\beta_1 = 1 \times 10^{-4}$ to $\beta_T = 0.02$. However, for specific cases like the Japanese dataset, additional experiments were conducted with increased timesteps of 1,000 and 1,500 to compare their effects as experimental results indicate that 500 epochs are insufficient for the model to converge fully on these datasets. For the Stabilizer (Zhu et al., 2023), t_0 can be adjusted between 20 and 30 for thresholding, and a quadratic non-increasing function $\sigma_a(t)$ is applied for smooth sampling. The hidden dimension of the transformer blocks is set to 512 for both the TED-Expressive and TED-Culture. We use eight transformer blocks, each comprising a multi-head self-attention block and a feed-forward network. The Adam optimizer is used with a learning rate of 5×10^{-4} . The threshold for the Huber loss is set to 0.1. Training the model using each separate language dataset TED-Culture takes approximately 1 h (6 h for Merged) and 44 h for TED-Expressive on a single NVIDIA A100 Tensor Core GPU on HPC.

5.3 Quantitative evaluation

For quantitative analysis, we employ evaluation metrics previously used in co-speech gesture generation (Liu X. et al. 2022); Zhu et al., 2023) and related tasks such as music-to-dance (Sun et al., 2020).

5.3.1 Quantitative metrics

5.3.1.1 Fréchet gesture distance (FGD)

Like the commonly used Fréchet Inception Distance (FID) metric in image generation research, the FGD metric serves the purpose of quantifying the dissimilarity between the distribution of synthesized gestures and that of real data. Yoon et al. (2020) introduce the FGD metric by training a skeleton sequence auto-encoder to extract features from both real gesture sequences X and generated gesture sequences

$$\hat{X}: FGD(X, \hat{X}) = \|\mu_r - \mu_g\|^2 + T_r \left(\sum_r + \sum_g - 2 \left(\sum_r \sum_g \right)^{1/2} \right),$$

where μ_r and \sum_r represent the first and second moments of the latent feature distribution of the real gestures X , while μ_g and \sum_g correspond to the first and second moments of the latent feature distribution of the generated gestures \hat{X} .

5.3.1.2 Beat Consistency Score (BC)

The Beat Consistency Score (BC) Li et al. (2022), Li et al. (2021) is designed to gauge the correlation between motion and speech beats. Recognizing the diversity in kinematic velocities among different joints, Liu X. et al. (2022) propose employing changes in the included angle between bones to identify motion beats. To initiate this process, they compute the Mean Absolute Angle Change (MAAC) for angle θ_i between consecutive frames using the following equation:

$$MAAC(\theta_j) = \frac{\sum_{s=1}^S \sum_{t=1}^{T-1} \|\theta_{j,s,t+1} - \theta_{j,s,t}\|_1}{S * (T-1)},$$

where S represents the total number of clips within the dataset, T signifies the number of frames contained in each clip, and $\theta_{j,s,t}$ corresponds to the included angle between the j^{th} and $(j+1)^{\text{th}}$ bone of the s^{th} clip at time-step t . The angle change rate for frame t within the s^{th} clip can be computed as

$$\frac{1}{J-1} \sum_{j=1}^{J-1} (\|\theta_{j,s,t+1} - \theta_{j,s,t}\|_1 / MAAC(\theta_j)),$$

Subsequently, kinematic beats are identified as local optima whose first-order difference exceeds a predefined threshold. To detect audio beats, we follow the methodology outlined in Li et al. (2022), utilizing the onset strength (Ellis, 2007). The Beat Consistency score is then determined as the average distance between each audio beat and its closest motion beat:

$$BC = \frac{1}{n} \sum_{i=1}^n \exp \left(- \frac{\min \forall t_j^x \in B^x \|t_i^x - t_j^y\|^2}{2\sigma^2} \right),$$

where t_i^x represents the i -th audio beats, $B^y = t_i^y$ denotes the set of kinematic beats, and σ is a parameter used for sequence normalization.

5.3.1.3 Diversity

This metric assesses the disparities in generated gestures that correspond to different inputs, as detailed in Lee H.-Y. et al. (2019). When calculating FGD, we employ the same feature extractor to map synthesized gestures into latent feature vectors and determine the mean feature distance. Specifically, in some studies (Liu X. et al., 2022); (Zhu et al., 2023) 500 randomly selected generated samples are used to compute the mean absolute error between the features and shuffled features.

5.3.2 Evaluation results

Table 2 presents the objective evaluation results for the TED-Expressive and TED-Culture Merged datasets. Baseline results for TED-Expressive are sourced from Zhu et al. (2023). Our model, DiffCulture, surpasses all baselines and achieves state-of-the-art performance on the TED-Expressive Dataset. Although the performance improvement over DiffGesture is minimal, DiffCulture still demonstrates superior overall results. Specifically, for the BC and Diversity metrics, our model, despite being slightly weaker than DiffGesture, outperforms all other models. On the TED-Culture Merged dataset, DiffGesture achieves the best FGD score, improving by nearly 30% compared to HA2G and DiffCulture. Additionally, since BC and Diversity measure motion-audio beat correlation and variation, these metrics for Ground Truth should not be treated as upper bounds. Notably, our results are on par with

TABLE 2 The Quantitative Results on TED-Expressive [Liu X. et al. \(2022\)](#) and TED-Culture Merged.

Methods	TED-expressive Liu et al. (2022b)			TED-culture		
	FGD ↓	BC ↑	Diversity ↑	FGD ↓	BC ↑	Diversity ↑
Ground Truth	0	0.703	178.827	0	0.702	181.900
Attention Seq2Seq (Yoon et al., 2019)	54.920	0.152	122.693	27.858	0.205	150.985
Speech2Gesture (Ginosar et al., 2019)	54.650	0.679	142.489	53.676	0.567	136.512
Joint Embedding (Ahuja and Morency, 2019)	64.555	0.130	120.627	52.993	0.135	120.380
Trimodal (Yoon et al. 2020)	12.613	0.563	154.088	12.026	0.396	146.988
HA2G (Liu et al., 2022b)	5.306	0.641	173.899	5.919	0.310	160.225
DiffGesture (Zhu et al., 2023)	2.600	0.718	182.757	4.216	0.728	175.025
DiffCulture (Ours)	2.398	0.715	177.814	5.532	0.722	160.603

We compare the proposed diffusion-based method against recent state-of-the-art (SOTA) methods ([Yoon et al., 2019](#)); ([Ginosar et al., 2019](#)); ([Ahuja and Morency, 2019](#)); ([Yoon et al., 2020](#)); ([Liu X. et al., 2022](#)); ([Zhu et al., 2023](#)) and ground truth. Lower values are better for FGD, while higher values are better for the other metrics. Bold values denote the best performance for each respective metric in the table.

Ground Truth, indicating high-quality generated gestures. Results for specific languages in the TED-Culture Dataset are listed in the Project Website³, demonstrating both SOTA performance and faster convergence across languages compared with [Zhu et al., \(2023\)](#), especially in Indonesian, Japanese, and Italian. It is worth noting that during the evaluation process, the BC and Diversity metrics exhibit significant fluctuations, highlighting the need for further refinement of the quantitative metrics.

5.4 User evaluation

Given that the generated gestures will ultimately be used in interactions with virtual agents or social robots, involving real individuals or users in the evaluation process is the ideal approach for assessing the quality and effectiveness of these gestures. We will analyze the result from the cultural perspective since the particularity of our dataset is multilingual and multicultural. [Figure 3](#) displays two cases, one in Japanese and the other in Turkish. It is noteworthy that translation is subject to a phenomenon known as the “word order change phenomenon”, wherein the order of morphemes (words or parts of words) in the target language differs from that in the source language after translation. Consequently, the word order in the English translation may not directly correspond to the order of the gesture sequence due to changes in the timing of morphemes. Take the Japanese case illustrated in the figure as an example: in the Japanese sentence, “60” is positioned at the beginning, whereas in the English sentence, “60” appears in the middle. When interpreting semantic information conveyed through gestures, the focus should be on the original subtitles rather than the translated version. The translated version is intended to facilitate understanding for individuals who do not comprehend the original language

but understand English. For example, in the Japanese case, the framework generates iconic gestures depicting “become so popular”, constituting the third key frame where the gesture extends the left arm to indicate emphasis. In the Turkish case, a similar phenomenon is observed with phrases like “and for 8 years in a row” and “our conversations were long”, corresponding to the fourth and fifth key frames in the figure.

5.4.1 Case study

To better validate the qualitative performance, we conducted a user case study on the generated co-speech gestures. The study involved 42 participants aged between 18 and 45 years, with 11 men, 30 women, and 1 non-binary person. From the language perspective, English was the predominant language spoken among participants, with 17 individuals, followed closely by Chinese and Indonesian, spoken by 14 and 7 individuals, respectively. The remaining 4 participants spoke Vietnamese or other less common languages. Each participant was required to assess the quality and coherence of the motion, with all clips presented without labels. A total of 136 cases⁴ were selected, comprising 17 (3 for Indonesian, 2 for Japanese, 3 for German, 3 for Italian, 3 for French, and 3 for Turkish) for each baseline (seven different frameworks and one ground truth). When distributing the questionnaires offline via campus bulletin boards or online through email, we use Qualtrics’ randomization function to select three testing videos for each method, resulting in a total of 24 videos per questionnaire. The Mean Opinion Scores (MOS) rating protocol was adopted, where each participant rated three aspects of the generated motions: naturalness, smoothness, and synchrony with speech. The results are presented in [Table 3](#). Interestingly, the Attention Seq2Seq model achieved the highest subjective evaluation scores, in contrast to

³ https://yixin-shen-1218.github.io/TED_Culture

⁴ <https://www.youtube.com/playlist?list=PLTTEUXLjZvB9-Vil1ZkmpZdphsQb92cAt>



TABLE 3 User Study Results.

Methods	Naturalness	Smoothness	Synchrony
Ground Truth	2.21 (0.94)	1.96 (0.74)	2.21 (0.86)
Attention Seq2Seq (Yoon et al., 2019)	3.15 (0.69)	3.57 (0.52)	3.05 (0.64)
Speech2Gesture (Ginosar et al., 2019)	2.91 (0.73)	2.56 (0.65)	3.03 (0.80)
Joint Embedding (Ahuja and Morency, 2019)	2.66 (0.65)	3.37 (0.44)	2.41 (0.64)
Trimodal (Yoon et al., 2020)	2.74 (1.01)	2.27 (0.82)	2.74 (1.04)
HA2G (Liu et al., 2022b)	1.98 (0.69)	1.86 (0.64)	1.91 (0.64)
DiffGesture (Zhu et al. 2023)	2.29 (0.81)	2.11 (0.59)	2.43 (0.63)
DiffCulture (Ours)	2.59 (0.83)	2.27 (0.70)	2.52 (0.80)

The ratings for motion naturalness, smoothness, and synchrony are assessed on a scale of 1–5, where 5 indicates the highest performance. All the results in the table are presented in the format of Average (SD).

Bold values denote the best performance for each respective metric in the table.

the results obtained in objective evaluation, where the Attention Seq2Seq framework performed worse. Qualitative analysis revealed that gestures generated by this model exhibited a slow and rigid behavior, which intuitively might not lead to high subjective evaluation scores. However, the slower gestures made by this model contributed to smoother and more coherent movements for virtual agents driven by gestures. In contrast, other models such as DiffGesture were perceived to produce overly jerky gestures due to pronounced changes between each frame, leading to less favorable and poorer subjective evaluation results. Overall, our DiffCulture model performed intermediately compared to the other models, surpassing the Ground Truth and some of the other models such as HA2G.

From a language acquisition perspective, we exclusively examine the subjective evaluation results from seven participants who are proficient in Indonesian as either their first or second language. The study includes two conditions: one focusing on subjective

evaluation results for Indonesian videos and the other for non-Indonesian videos. Table 4 presents the perception of Indonesian participants in these two scenarios. According to the table, participants assigned similar scores for smoothness to videos in both familiar and unfamiliar languages. However, participants tended to assign lower scores to gestures performed in their native language, with Indonesian videos receiving significantly lower scores in naturalness and synchrony compared to non-Indonesian videos. This indicates that individuals are more critical of co-speech gestures in their native language and expect higher performance from generative models.

5.5 Ablation studies

We conducted text-embedding experiments to investigate whether using word embeddings from different languages would

TABLE 4 The correlation between language acquisition and the subjective perception of the participants.

Video Type	Naturalness	Smoothness	Synchrony
Indonesian Videos	1.94 (1.07)	2.18 (1.16)	1.80 (1.16)
Non-Indonesian Videos	2.34 (1.05)	2.18 (1.14)	2.34 (1.01)

Bold values denote the best performance for each respective metric in the table.

influence the training process and results. Unlike the typical ablation study that usually tests how system performance changes when components are removed, we extended the proposed framework to include text modality handling and performed three sets of experiments by varying the language of the FastText embeddings, using TED-Culture French. Aside from the language of the FastText embeddings, we also adjusted the number of training epochs, increasing them from 500 to 1,000 when using the French FastText to embed the subtitles. Table 5 presents the results of these ablation experiments on text embedding. From the table, it is evident that the English FastText Word Embedding outperforms the other two conditions, even though in the last case we trained the model for 1,000 epochs to ensure complete convergence. Conversely, when using French FastText Word Embedding, the results indicate that 500 epochs are insufficient for the framework with text modality to converge to an optimal state. Even with 1,000 epochs, the model using French FastText still performs worse than its English counterpart, despite the longer training time. We can conclude that the language of word embeddings based on FastText does not significantly impact the final performance of the framework, but it can affect the convergence time.

6 Robot prototype

Unlike mapping gestures to virtual agents, some constraints exist when mapping gestures to robots because the joints of the robots are motor-driven and not as flexible as those of virtual agents. Since the output of our framework consists of 3D directional vectors, we first need to calculate the radian values for each vector. These radian values often exceed the joints' angle range, making post-processing an essential step. We applied different post-processing methods to the different joints to make it look as natural as possible and consistent with the gesture motion trends generated by the model. To mitigate the jittering issue in the generated gestures, Bézier interpolation is employed during the retargeting process. Since the robot has difficulty walking and the dataset focuses only on the upper body, not all joints are used in this project. We concentrate on the head and arms, as well as hands which only have open and close functions. Specifically, we utilize 12 degrees of freedom (DoF) in the upper body, namely HeadYaw, HeadPitch, RShoulderPitch, RShoulderRoll, RElbowYaw, RElbowRoll, RWristYaw, LShoulderPitch, LShoulderRoll, LEElbowYaw, LEElbowRoll, and LWristYaw, plus the open and close functions of the hands. All the robot prototype codes are available

TABLE 5 Ablation study on the impact of using corresponding text embeddings in the DiffCulture model after incorporating the text modality on TED-Culture French.

Methods	FGD ↓	BC ↑	Diversity ↑
Without Text Modality, 500	5.053	0.747	91.817
English Text Embedding, 500	5.003	0.749	96.225
French Text Embedding, 500	10.691	0.750	101.947
French Text Embedding, 1,000	5.205	0.745	96.936

Bold values denote the best performance for each respective metric in the table.

at this repository⁵, and the playlist⁶ includes both the outputs of the generative model visualized as skeletons and the robot prototype demonstrations for better comparison.

7 Conclusion

In this paper, we present a large-scale monologue dataset for cross-cultural gesture generation grounded in language and refine the gesture generation model based on Zhu et al. (2023). Additionally, we devise a state-of-the-art co-speech gesture framework and implement the generated gestures on the NAO robot, enabling synchronized speech and gesture performance. We conduct culturally subjective evaluations and an ablation study, validating the cultural relevance of co-speech gestures and demonstrating that word embeddings may not need to correspond to the language of the text. However, certain limitations are evident. Firstly, low subjective evaluation scores for Ground Truth indicate the dataset's relative lack of quality. Additionally, during the dataset creation process, the filtering of valid clips lacks manual filtering. Furthermore, the experiments were conducted exclusively on the NAO robot, which has limited or no finger mobility, thus restricting the generalizability of the findings. To address this, future work could explore testing on more advanced platforms with greater dexterity to validate and extend the results.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/[Supplementary Material](#).

Author contributions

YS: Writing – original draft, Data curation, Investigation, Methodology, Visualization. WJ: Conceptualization, Funding acquisition, Supervision, Writing – review and editing.

5 https://github.com/Yixin-Shen-1218/NAO_Gesture_Generation

6 https://www.youtube.com/playlist?list=PLTTEUXLjZvB8qEKfwHjQlgOJ3ahP_gOPS

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the Australian Research Council (DE210100858).

Acknowledgments

We would like to thank the authors for kindly sharing their articles in closed access.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Ahuja, C., and Morency, L.-P. (2019). "Language2pose: natural language grounded pose forecasting," in *2019 international conference on 3D vision (3DV)* (IEEE), 719–728.
- Bremner, P., Pipe, A. G., Fraser, M., Subramanian, S., and Melhuish, C. (2009). "Beat gesture generation rules for human-robot interaction," in *RO-MAN 2009-the 18th IEEE international Symposium on Robot and human interactive communication (IEEE)*, 1029–1034.
- Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., and Black, M. J. (2020). "Monocular expressive body regression through body-driven attention," in *Computer vision-ECCV 2020: 16th European conference, glasgow, UK, august 23–28, 2020, proceedings, Part X 16* (Springer), 20–40.
- Ellis, D. P. (2007). Beat tracking by dynamic programming. *J. New Music Res.* 36, 51–60. doi:10.1080/09298210701653344
- Fisher, A., and Griswold, L. (2010). Evaluation of social interaction (ESI) (Fort Collins, CO: Three Star Press)
- Ghorbani, S., Ferstl, Y., Holden, D., Troje, N. F., and Carbonneau, M.-A. (2023). Zeroeggs: zero-shot example-based gesture generation from speech. In *Comput. Graph. Forum*, 42, 206–216. doi:10.1111/cgf.14734
- Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., and Malik, J. (2019). "Learning individual styles of conversational gesture," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3497–3506.
- Gjaci, A., Recchiuto, C. T., and Sgorbissa, A. (2022). Towards culture-aware co-speech gestures for social robots. *Int. J. Soc. robotics* 14, 1493–1506. doi:10.1007/s12369-022-00893-y
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial networks
- Habibie, I., Xu, W., Mehta, D., Liu, L., Seidel, H.-P., Pons-Moll, G., et al. (2021). "Learning speech-driven 3d conversational gestures from video," in *Proceedings of the 21st ACM international conference on intelligent virtual agents*, 101–108.
- Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Adv. neural Inf. Process. Syst.* 33, 6840–6851. doi:10.48550/arXiv.2006.11239
- Ishi, C. T., Machiyashiki, D., Mikata, R., and Ishiguro, H. (2018). A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics Automation Lett.* 3, 3757–3764. doi:10.1109/lra.2018.2856281
- Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: a review. *Lang. cognitive Process.* 24, 145–167. doi:10.1080/09507379831059783-1
- Knapp, M. L., Hall, J. A., and Horgan, T. G. (2013). *Nonverbal communication in human interaction* (Cengage Learning)
- Kucherenko, T., Jonell, P., Van Waveren, S., Henter, G. E., Alexandersson, S., Leite, I., et al. (2020). "Gesticulator: a framework for semantically-aware speech-driven gesture generation," in *Proceedings of the 2020 international conference on multimodal interaction*, 242–250. doi:10.1145/3382507.3418815
- Lee, G., Deng, Z., Ma, S., Shiratori, T., Srinivasa, S., and Sheikh, Y. (2019a). "Talking with hands 16.2m: a large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 763–772. doi:10.1109/ICCV.2019.00085
- Lee, G., Deng, Z., Ma, S., Shiratori, T., Srinivasa, S. S., and Sheikh, Y. (2019b). "Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 763–772. doi:10.1109/iccv.2019.00085
- Lee, H.-Y., Yang, X., Liu, M.-Y., Wang, T.-C., Lu, Y.-D., Yang, M.-H., et al. (2019c). Dancing to music. *Adv. neural Inf. Process. Syst.* 32. doi:10.48550/arXiv.1911.02001
- Levine, S., Krähenbühl, P., Thrun, S., and Koltun, V. (2010). "Gesture controllers," in *AcM siggraph 2010 papers*, 1–11.
- Li, B., Zhao, Y., Zhelun, S., and Sheng, L. (2022). Danceformer: music conditioned 3d dance generation with parametric motion transformer. *Proc. AAAI Conf. Artif. Intell.* 36, 1272–1279. doi:10.1609/aaai.v36i2.20014
- Li, R., Yang, S., Ross, D. A., and Kanazawa, A. (2021). "Ai choreographer: music conditioned 3d dance generation with aist++," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13401–13412.
- Liu, C. Y., Mohammadi, G., Song, Y., and Johal, W. (2023). "Speech-gesture gan: gesture generation for robots and embodied agents," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 405–412. doi:10.1109/RO-MAN57019.2023.10309493
- Liu, H., Zhu, Z., Iwamoto, N., Peng, Y., Li, Z., Zhou, Y., et al. (2022a). "Beat: a large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis," in *Computer Vision – ECCV 2022. Editors Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., Hassner* (Cham: Springer Nature Switzerland) T612–630. doi:10.1007/978-3-031-20071-7_36
- Liu, X., Wu, Q., Zhou, H., Xu, Y., Qian, R., Lin, X., et al. (2022b). "Learning hierarchical cross-modal association for co-speech gesture generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10462–10472.
- Liu, Y., Mohammadi, G., Song, Y., and Johal, W. (2021). "Speech-based gesture generation for robots and embodied agents: a scoping review," in *Proceedings of the 9th International Conference on Human-Agent Interaction*, 31–38. doi:10.1145/3472307.3484167
- Nyatsanga, S., Kucherenko, T., Ahuja, C., Henter, G. E., and Neff, M. (2023). A comprehensive review of data-driven co-speech gesture generation. In *Comput. Graph. Forum*, 42, 569–596. doi:10.1111/cgf.14776
- Pavlo, D., Feichtenhofer, C., Grangier, D., and Auli, M. (2019). "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7753–7762.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frobt.2025.1546765/full#supplementary-material>

- Studdert-Kennedy, M. (1994). Hand and mind: what gestures reveal about thought. *Lang. Speech* 37, 203–209. doi:10.1177/002383099403700208
- Sun, G., Wong, Y., Cheng, Z., Kankanhalli, M. S., Geng, W., and Li, X. (2020). Deepdance: music-to-dance motion choreography with adversarial learning. *IEEE Trans. Multimedia* 23, 497–509. doi:10.1109/tmm.2020.2981989
- Takeuchi, K., Kubota, S., Suzuki, K., Hasegawa, D., and Sakuta, H. (2017). “Creating a gesture-speech dataset for speech-based automatic gesture generation,” in *HCI international 2017-posters' extended abstracts: 19th international conference, HCI international 2017* (Vancouver, BC, Canada: Springer), 198–202.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. neural Inf. Process. Syst.* 30. doi:10.48550/arXiv.1706.03762
- Yoon, Y., Cha, B., Lee, J.-H., Jang, M., Lee, J., Kim, J., et al. (2020). Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Trans. Graph. (TOG)* 39, 1–16. doi:10.1145/3414685.3417838
- 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) (IEEE), 4303–4309.
- Yoon, Y., Wolfert, P., Kucherenko, T., Viegas, C., Nikolov, T., Tsakov, M., et al. (2022). “The genea challenge 2022: a large evaluation of data-driven co-speech gesture generation,” in Proceedings of the 2022 International Conference on Multimodal Interaction, 736–747. doi:10.1145/3536221.3558058
- Zhi, Y., Cun, X., Chen, X., Shen, X., Guo, W., Huang, S., et al. (2023). “Livelyspeaker: towards semantic-aware co-speech gesture generation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 20807–20817.
- Zhou, C., Bian, T., and Chen, K. (2022). “Gesturemaster: graph-based speech-driven gesture generation,” in Proceedings of the 2022 International Conference on Multimodal Interaction, 764–770. doi:10.1145/3536221.3558063
- 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, 10544–10553. doi:10.1109/cvpr52729.2023.01016