DialoGen: Towards Dialog Gesture Generation via Identity-Decoupled Style Guidance in Interactive Diffusion Model

Weiyu Zhao[1], Chenyang Wang[1]*, Liangxiao Hu[1], Zonglin Li[1], Wei Yu[2], Shengping Zhang[1]

[1]School of Computer Science and Technology, Harbin Institute of Technology
[2]School of Information Science and Technology, Tsinghua University

Abstract

We propose DialoGen, a novel framework for generating realistic gestures for both interlocutors in dialog scenarios, conditioned on conversational audios. Unlike most existing methods that focus solely on a single speaker, DialoGen simultaneously generates synchronized gestures for both participants while also embedding identity-decoupled style into generated gestures that enhance realism and expressiveness. To ensure precise synchronization between interlocutors, DialoGen adopts an interactive dual-diffusion model with mutual interaction estimation, which integrates interaction correlation into the diffusion process. More importantly, by leveraging supervised contrastive learning, we develop the identity-decoupled style guidance to adaptively decompose the identity-specific style of interlocutors into latent space, enabling multi-style dialog gesture generation. Extensive experimental results demonstrate that our model significantly outperforms existing methods in generating realistic, speech-aligned, identity-specific gestures, offering a high-quality solution for various dialog scenarios.

Figure 1: Our proposed DialoGen delves into *Dialog Gesture Generation*, a defined task that produces coherent, life-like gestures for both interlocutors based on the speech context while preserving each speaker's identity style.

Introduction

Dialog gesture synthesis aims to generate realistic body movements in conversational settings, loyal to the audio of both the speaker and the interlocutor. This task has broad applications in 3D gaming, augmented reality (AR), virtual reality (VR), and virtual social platforms (Liang et al. 2024). Although significant efforts (Ng et al. 2024; Mughal et al. 2024) have been devoted to developing effective dialog gesture synthesis, it is still a challenging task due to the variety of body gestures and intricate human interaction.

Previous approaches mainly concentrate on generating co-speech gestures tailored to individual speakers by employing a variety of computational techniques, including Transformer architectures (Qi et al. 2024b; Ghorbani et al. 2023), Variational Autoencoder models (VAE) (Yi et al. 2023; Li et al. 2021), Diffusion models (Ao, Zhang, and Liu 2023; Zhu et al. 2023; Yang et al. 2023a), Mamba frameworks (Xu et al. 2024; Fu et al. 2024), and Generative Pre-trained Transformers (GPT) (Zhang et al. 2024; Cheng, Li, and Fu 2024). These advancements have significantly improved the realism and applicability of virtual agents in various single-speaker scenarios. Unfortunately, they usually fail to generate pleasant gesture results casting to diverse conversational contexts due to the neglect of complex interactive factors such as dialog content, emotional states, and speaker identity. Recently, researchers have shifted attention towards two-person gesture generation in conversational contexts, notably initiated by the GENEA Challenge 2023 (Kucherenko et al. 2023). Some studies primarily focus on generating gestures for the primary speakers (Yang et al. 2023b; Zhao, Hu, and Zhang 2023), while others (Ng et al. 2024; Mughal et al. 2024) generate the gestures of both interlocutors sequentially according to their conversational turns. Although these works have improved the performance of conversational gesture synthesis, two key factors still hinder the generation of highly realistic interactive effects in real-world scenarios. First, these methods typically generate gestures for a single participant based solely on their own audio rather than perceiving the mutual dynamics between participants, limiting the realism and interactivity of their results. On the other hand, generating gestures with participant-specific styles is essential for producing vivid and realistic dialog gestures. However, existing methods mostly employ one-hot encoding to identify each

*Corresponding author (c.wang@stu.hit.edu.cn).

participant, but the sparse and discrete characteristics of this representation prevents it from capturing the latent stylistic differences among individuals, making them unable to generate high-quality, multi-style dialog gestures.

To overcome these limitations, we present DialoGen, a novel diffusion-based framework (Ho, Jain, and Abbeel 2020a) dedicated to generating synchronized interaction gestures for both participants in dialog scenarios. Unlike existing methods that focus on single-speaker gesture generation, DialoGen introduces an interactive dual-diffusion model that simultaneously generates temporally coherent gestures for both participants. Specifically, we employ two weight-sharing Transformer models as denoisers within the diffusion process, ensuring that the gestures of both interlocutors are produced in a harmonized and temporally aligned manner. To further capture the mutual dynamics between speakers, we augment the Transformer-based denoisers with a mutual interaction estimation module. By integrating the estimated interaction weights into the diffusion process, our model produces gestures that are not only well-aligned with speech but also exhibit realistic and responsive interpersonal coordination.

Furthermore, DialoGen employs a supervised contrastive learning approach (Khosla et al. 2020) to decouple identity-specific style representation for each participant, facilitating controllable personalization in gesture generation. Specifically, we introduce an identity feature extractor to extract the identity-specific style representations from gestures. Unlike simplistic one-hot encoding techniques (Liu et al. 2022; Yang et al. 2023b), the proposed supervised contrastive learning approach is able to maximize the distinction between individuals with different identities, as described in Section . The decoupled style representation is seamlessly integrated into the denoiser modules to enable multi-style gesture generation, thereby enhancing both realism and expressiveness.

In summary, we claim three following contributions:

- We propose a novel framework, DialoGen, which for the first time achieves simultaneous generation of identity-specific co-speech gestures for both interlocutors in dialog scenarios.
- We introduce an interactive dual-diffusion model to synchronize dialog gesture generation. Furthermore, we augment the model with mutual interaction estimation to enhance the responsiveness and coordination of mutual speaking gestures.
- We propose to learn the identity-decoupled style guidance via a supervised contrastive learning approach, enabling our method to effectively preserve the distinctive stylistic traits of each individual.

Related Work

Co-Speech Gesture Modeling. Co-speech gesture generation has been a widely studied topic, with early work focusing on rule-based methods (Cassell, Vilhjálmsson, and Bickmore 2001; Kopp and Wachsmuth 2004) and statistical models (Kipp et al. 2007; Levine et al. 2010). These approaches, while interpretable, require extensive manual effort and are limited by handcrafted features. With the advent of deep learning, models such as MLPs (Kucherenko et al. 2020), CNNs (Habibie et al. 2021), RNNs (Yoon et al. 2020), and Transformers (Qi et al. 2024a) have significantly advanced co-speech gesture generation. Recent works have leveraged generative models, such as VAEs (Li et al. 2021), VQ-VAEs (Ao et al. 2022), and diffusion-based models (Ao, Zhang, and Liu 2023), to improve diversity in generated gestures. To address dataset sparsity, Semantic Gesticulator (Zhang et al. 2024) uses GPT-2 (Radford et al. 2019) for semantic gesture retrieval, and SIGGesture (Cheng, Li, and Fu 2024) generates gestures from speech using large language models. While these methods focus on semantic accuracy, they often overlook individual gesture styles. Our approach explicitly integrates speaker identity, preserving stylistic nuances throughout the generation process.

Dyadic Interaction Modeling. Modeling dyadic interactions is crucial for generating realistic co-speech gestures in conversational agents. Recent work includes ComMDM (Shafir et al. 2023) and InterGen (Liang et al. 2024), which use collaborative transformer-based models to generate interactions between two participants. Approaches like ReMoS (Ghosh et al. 2025) and FreeMotion (Fan et al. 2025) further extend these methods to handle multi-character interactions. The GENEA Challenge 2023 (Kucherenko et al. 2023) spurred advancements in generating co-speech gestures within dyadic interactions, including diffusion models (Yang et al. 2023b), VQ-VAE (Korzun, Beloborodova, and Ilin 2023), and autoregressive models (Harz, Voß, and Kopp 2023). Audio2Photoreal (Ng et al. 2024) generates 3D movements for both speaker and listener in dyadic conversations, while ConvoFusion (Mughal et al. 2024) introduces semantic guidance for motion generation. Despite these advances, existing methods often fail to account for the complex, dynamic non-verbal interactions between participants. Our approach addresses this by modeling both participants' gestures and ensuring the preservation of identity style information throughout the generation process.

Method

Our goal is to simultaneously generate personalized gestures for both participants in dialog scenarios. We first introduce the interactive diffusion model to synchronize dialog gesture generation. We then augment the model with mutual interaction estimation to capture the mutual dynamics. To enable multi-style gesture generation, we leverage a supervised contrastive learning approach to learn the latent stylistic representation. In the following, we provide a comprehensive explanation of the technical details.

Interactive Dual-Diffusion Model

Diffusion models (Ho, Jain, and Abbeel 2020b) have demonstrated significant advances in co-speech gesture generation (Ao, Zhang, and Liu 2023; Chen et al. 2024; Yang et al. 2023a). Building on these developments, we extend the framework to an interaction dual-diffusion model for dialog gesture generation. Unlike previous methods that predominantly focus on single-speaker gesture generation, we propose two weight-sharing Transformer models as denoisers

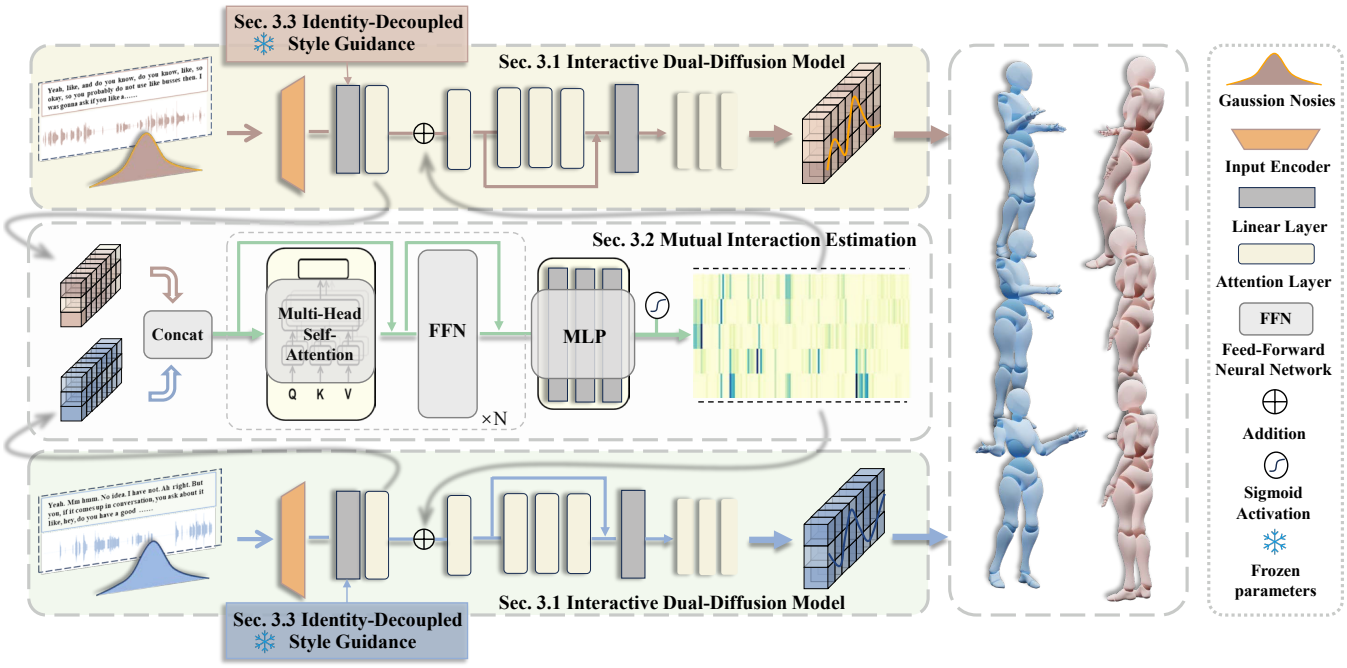

Figure 2: Overview of DialoGen. Given the dialog speech context, we propose the interactive dual-diffusion model with mutual interaction estimation to predict co-speech gestures for both participants by leveraging the extracted content features and decoupled identity-style guidance. Specifically, the mutual interaction estimation module within the diffusion model is trained to generate interaction weights that guide the coordination of gestures between interlocutors. Note that the decoupled identity-style guidance injects person-specific stylistic traits into the model, ensuring individualized gesture generation.

to accommodate dual-stream audio inputs. Each component of our framework is detailed in the following.

Feature Extraction. We adopt the feature encoder presented in (Yang et al. 2023a; Chang, Zhang, and Kapadia 2022) to extract both speech and text features. Specifically, the extracted speech features include MFCC, Mel Spectrum, Pitch, Energy, and WavLM (Chen et al. 2022), capturing essential audio characteristics (such as rhythm, emotion, and beat) to balance both statistical and latent-space representations. Meanwhile, we utilize FastText(Bojanowski et al. 2017) to extract 300-dimensional word embeddings from text. In addition, we introduce a binary indicator to denote laughter within the speech data. The identity-decoupled style embeddings ensure that the generated style features align with the speaker identity.

Gesture Representation. We denote the gesture clip in the diffusion model as $x_0 \in \mathbb{R}^{(N_{seed}+N)\times J\times L}$. The first N_{seed} frames of the gesture clip serve as seed gestures and the remaining N frames are what the model needs to predict. This enables the gesture sequences generation at arbitrary length (Zhao, Hu, and Zhang 2023; Yang et al. 2023b). We use $J = 62$ joints including fingers and represent each frame with $L = 36$ dimensional features comprising joint position, velocity, acceleration, rotation matrix, rotational angular velocity, and rotational angular acceleration.

Transformer Denoising for Gesture Generation We propose two weight-sharing Transformer models as denoisers in the diffusion process, guided by input conditions $\mathcal{C} = (\mathcal{C}_c, \mathcal{C}_s)$, where $\mathcal{C}_c$ consists of speech features, word embeddings, and a laughter indicator, and $\mathcal{C}_s$ is formed by broadcasting style embeddings $\mathbf{W}_{\text{style}}$ across the temporal dimension of $\mathcal{C}_c$. Following the Denoising Diffusion Probabilistic Model (DDPM) (Ho, Jain, and Abbeel 2020b), the forward process adds Gaussian noise to the clip x_0:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, \tag{1}$$

with $\varepsilon \sim \mathcal{N}(0, I)$. The reverse process is modeled as a conditional Gaussian distribution, where the model learns the mean μ_θ and the variance Σ_θ of the distribution:

$$p_\theta(x_{t-1} \mid x_t, \mathcal{C}) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t, \mathcal{C}), \Sigma_\theta\right), \tag{2}$$

where the mean is:

$$\mu_\theta(x_t, t, \mathcal{C}) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(x_t, t, \mathcal{C})\right). \tag{3}$$

This process iteratively reconstructs the clean data x_0.

Training Objective. To ensure the generated gestures accurately reflect the expected behavior, we train the transformer denoising network using the Huber Loss (Huber 1992) for both participants. The training objective is defined as:

$$\mathcal{L} = \sum_{id=1}^{2}\sum_{t=1}^{T} HuberLoss(x_{0,id} - \hat{x}_{t,id}), \tag{4}$$

where id indicates two individuals.

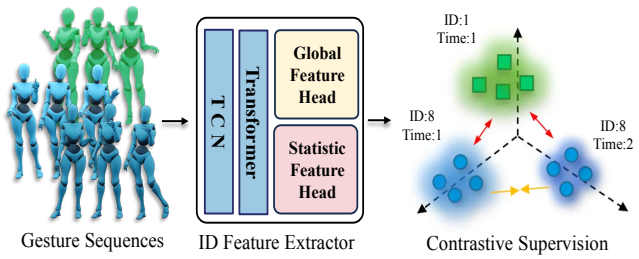

Figure 3: Contrastive supervision on the identity style representation. Blue points represent positive samples, while green points indicate negative samples.

Mutual Interaction Estimation

In this section, we aim to explore how interlocutors interact with each other. To capture these intricate dynamics, we augment the proposed dual-diffusion model with a mutual interaction estimation module. As shown in Fig. 2, this module is designed to model the interaction weights between the latent gesture features of interlocutors during the diffusion process. We extract the gesture features from the second layer following the injection of identity-style guidance. The features of the two interlocutors at time step t are denoted as $f_{t,1}$ and $f_{t,2}$. The module can be formulated as:

$$z_t = G_\theta\left([f_{t,1};\ f_{t,2}]\right). \tag{5}$$

The module G_θ consists of N blocks, each with multi-head self-attention mechanism and a feed-forward network (FFN), followed by an MLP head and a Sigmoid function.

With the Sigmoid function ensuring that the interaction weights z_t lie within the interval $[0, 1]$, these weights serve as a quantitative measure of the degree to which the two participants influence each other. Specifically, values of z_t closer to 1 indicate a stronger mutual influence, whereas values near 0 reflect a weaker interaction.

Identity-Decoupled Style Guidance

Individuals often exhibit distinctive gestural behaviors during speech, reflecting personal communication styles. In this section, we focus on disentangling identity-specific styles from speech gestures. Unlike previous methods that employ one-hot encoding to identify each participants style, we leverage a supervised contrastive learning approach to learn latent stylistic representation.

As depicted in Fig. 3, we introduce an feature extractor $\mathcal{F}$ to map the the input gesture sequence $\mathbf{x} \in \mathbb{R}^{N^{id} \times J \times L}$ to two latent features as follows:

$$\begin{aligned} \langle \mathbf{W}_{\text{global}}, \mathbf{W}_{\text{statistic}} \rangle &= \mathcal{F}(\mathbf{x}), \\ \mathbf{W}_{\text{style}} &= \mathbf{W}_{\text{global}} \oplus \mathbf{W}_{\text{statistic}}, \end{aligned} \tag{6}$$

where $\mathbf{W}_{\text{style}} \in \mathbb{R}^{D_{\text{style}}}$, $\oplus$ stands for feature concatenation. We set $N^{id} = 700$, which is significantly longer than the sequence length used in diffusion prediction, allowing the model to capture richer style information. The feature extractor $\mathcal{F}$ is composed of a Temporal Convolutional Network (TCN) and a Transformer network, followed by global and statistic feature heads. The global feature head applies mean pooling along the time dimension to compute the global feature vector $\mathbf{W}_{\text{global}}$. Meanwhile, the statistic feature head calculates the mean and standard deviation of inter-frame differences to derive the statistic feature vector $\mathbf{W}_{\text{statistic}}$. We then concatenate $\mathbf{W}_{\text{global}}$ and $\mathbf{W}_{\text{statistic}}$ to obtain the final style representation $\mathbf{W}_{\text{style}}$.

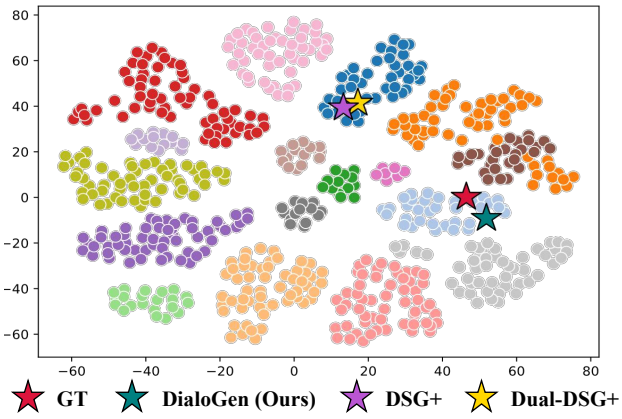

Figure 4: Visualization of style representation clusters using U-MAP. Each color represents a unique ID from the TWH dataset. ☆ means the predicted gestures. GT stands for Ground Truth, DSG+ stands for DiffustyleGesture+.

After obtaining the personal style representation, our goal is to amplify the distinctions among individual representations in the latent space, thereby promoting more discernible clustering. As illustrated in Fig. 3, we apply a supervised contrastive learning approach on the style representation $\mathbf{W}_{\text{style}}$. Specifically, we designate gestures performed by the same speaker at different times as positive samples, while gestures produced by different speakers serve as negative samples. We employ a supervised contrastive loss $\mathcal{L}_{\text{identity}}$ (Khosla et al. 2020) to cluster latent style representations, which is defined as follows:

$$\mathcal{L}_{\text{id}} = -\sum_{i \in I} \mathbb{E}_{p \in P(i)} \log \frac{\exp(w_i \cdot w_p / \tau)}{\sum_{a \in A(i)} \exp(w_i \cdot w_a / \tau)}, \tag{7}$$

where $i \in I \equiv \{1...2N\}$ is the index of an arbitrary augmented sample within a multiviewed batch. $A(i) \equiv I \setminus \{i\}$ denotes the set of indices of all samples excluding the anchor i. $P(i) \equiv \{p \in A(i) : \tilde{y}_p = \tilde{y}_i\}$ represents the set of indices of all positive samples in the multiviewed batch that are distinct from i. $\mathbb{E}_{p \in P(i)}[\cdot]$ denotes the expectation over positive samples of i. The temperature parameter τ is a positive scalar that controls the smoothness of the contrastive loss. The supervised contrastive loss function is designed to maximize the similarity between positive samples while minimizing it between negative samples, enhancing the encoder's capacity to learn discriminative features across diverse identities.

As shown in Fig. 4, we visualize the clusters of the style representation $\mathbf{W}_{\text{style}}$ using U-MAP (McInnes, Healy, and Melville 2018). Each cluster exhibits clear and compact dis-

Methods	FD_g↓	FD_k↓	SRGR↑	DSRGR↑	Div_{sample}↑	Div_g	Div_k
DG	1.74	3.48	2.85	3.68	<u>410.13</u>	138.02 ◇11.3%	189.55 ◇9.28%
DG^{Dual}	1.69	3.61	2.28	3.43	**460.53**	<u>117.38</u> ◇5.26%	149.90 ◇13.58%
DSG+	1.36	2.81	3.32	3.87	266.95	91.84 ◇25.87%	128.57 ◇25.88%
$DSG+^{Dual}$	1.41	2.73	3.30	3.75	273.68	109.59 ◇11.6%	<u>157.27</u> ◇9.34%
A2P	<u>1.29</u>	<u>2.41</u>	<u>3.37</u>	<u>3.90</u>	231.38	86.17 ◇30.45%	119.91. ◇30.87%
DialoGen (Ours)	**1.18**	**2.33**	**3.61**	**4.19**	293.47	**123.94** ◇0.03%	**175.23** ◇1.01%
Ground Truth	0	0	–	–	–	123.90	173.47

Table 1: Quantitative comparison of our proposed method with sota methods trained on the TWH dataset. **Bold** denotes the best performance, while <u>underline</u> indicates the second-best. ◇ represents the percentage difference from the ground truth. DG stands for Diffgesture, DSG+ stands for DiffustyleGesture+ and A2P stands for Audio2photoreal.

tribution boundaries, indicating well-separated and distinctive style representations. We also validate the style distribution of the predicted gestures and surprisingly find that our predicted gestures are located within the same latent style clusters as the ground truth. In contrast, the other two methods that rely on one-hot encoding fail to capture the correct style, resulting in misaligned cluster assignments.

Experiments

Experiment Setup

Datasets. We perform training and evaluation on the Talking With Hands (TWH) dataset, originally proposed in (Lee et al. 2019) and further refined in (Kucherenko et al. 2023). The TWH dataset comprises over 18 hours of conversational data involving 18 participants, including both male and female subjects. Each conversation has an average duration of 183 seconds. The dataset is split into 86% for training, 7% for validation, and 7% for testing.

Compared Methods. We compare our method with representative state-of-the-art co-speech generation methods, including DiffGesture (Zhu et al. 2023), DiffuseStyleGesture+ (Yang et al. 2023b), and Audio2Photoreal (Ng et al. 2024) on the TWH dataset. For fair comparison, we retrained these three baselines on the TWH, applying the same smoothing techniques. Additionally, to ensure that DiffGesture and DiffuseStyleGesture+ receive the same contextual information as our method and Audio2Photoreal, we provided the audio from both speakers in the dialogs to adapt to the dual-person dialog scenario. In Tab. 1, the results of these changes to these baselines are denoted as "xxx^{Dual} ".

Evaluation Metrics. We evaluate our model using the following metrics, as referenced in (Ng et al. 2024; Liu et al. 2022): For Fréchet distance-based metrics, we measure (a) **FD_g** as the Fréchet distance between generated and ground truth static gestures, and (b) **FD_k** as the Fréchet distance on gesture motion velocities. For semantic alignment, we propose (c) **SRGR** to capture gesture–text relevance, and (d) **DSRGR** to handle interactions in the dual-speaker setting. For diversity metrics, we assess (e) **Div_g** using the average L2 distance between sampled gesture pairs, (f) **Div_k** to capture gesture variation within a sequence, and (g) **Div_{sample}** as the variance across samples from the same audio.

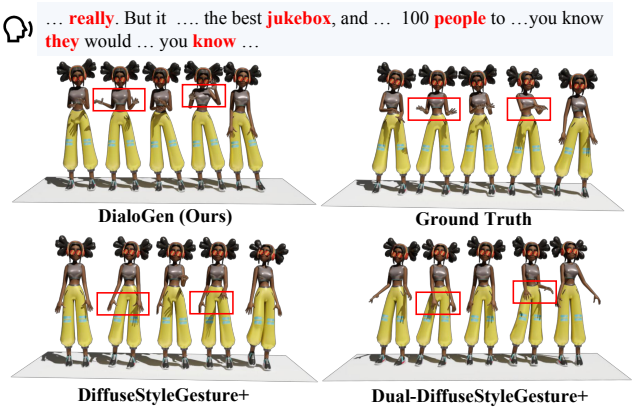

Figure 5: Gesture generation with identity style preservation. Our method generates motions that more closely align with the ground truth style, in contrast to DiffuseStyleGesture+ (Yang et al. 2023b), which uses one-hot encoding.

Quantitative Evaluation

The results in Tab. 1 demonstrate that our method achieves state-of-the-art performance on the TWH datasets, consistently outperforming other baselines, particularly in terms of Fréchet distance-based metrics (FD_g and FD_k) and semantic metrics(SRGR and DSRGR). These results highlight the strong distribution-matching capability of our approach, especially in dialog scenarios. It is worth noting that higher diversity scores do not always correlate with improved generation quality. The Div metric only counts when the synthesized motion is smooth and natural (Chen et al. 2024). Our method achieves the most comparable diversity scores (Div_g and Div_k) to real data, indicating that the generated outputs within the TWH dataset are both diverse and realistic.

Further, we observe that using paired audio as input typically leads to improved performance across all baselines. Audio2Photoreal achieves the suboptimal performance in the Fréchet distance metric thanks to its guidance mechanism and extended generation sequence, yet it exhibits notable shortcomings in diversity. DiffGesture achieves seemingly superior quantification on Div_{sample}. We argue that the reason for this observation stems from the inherent random jitter in the generation results, which can be confirmed by the inferior performance on the other metrics.

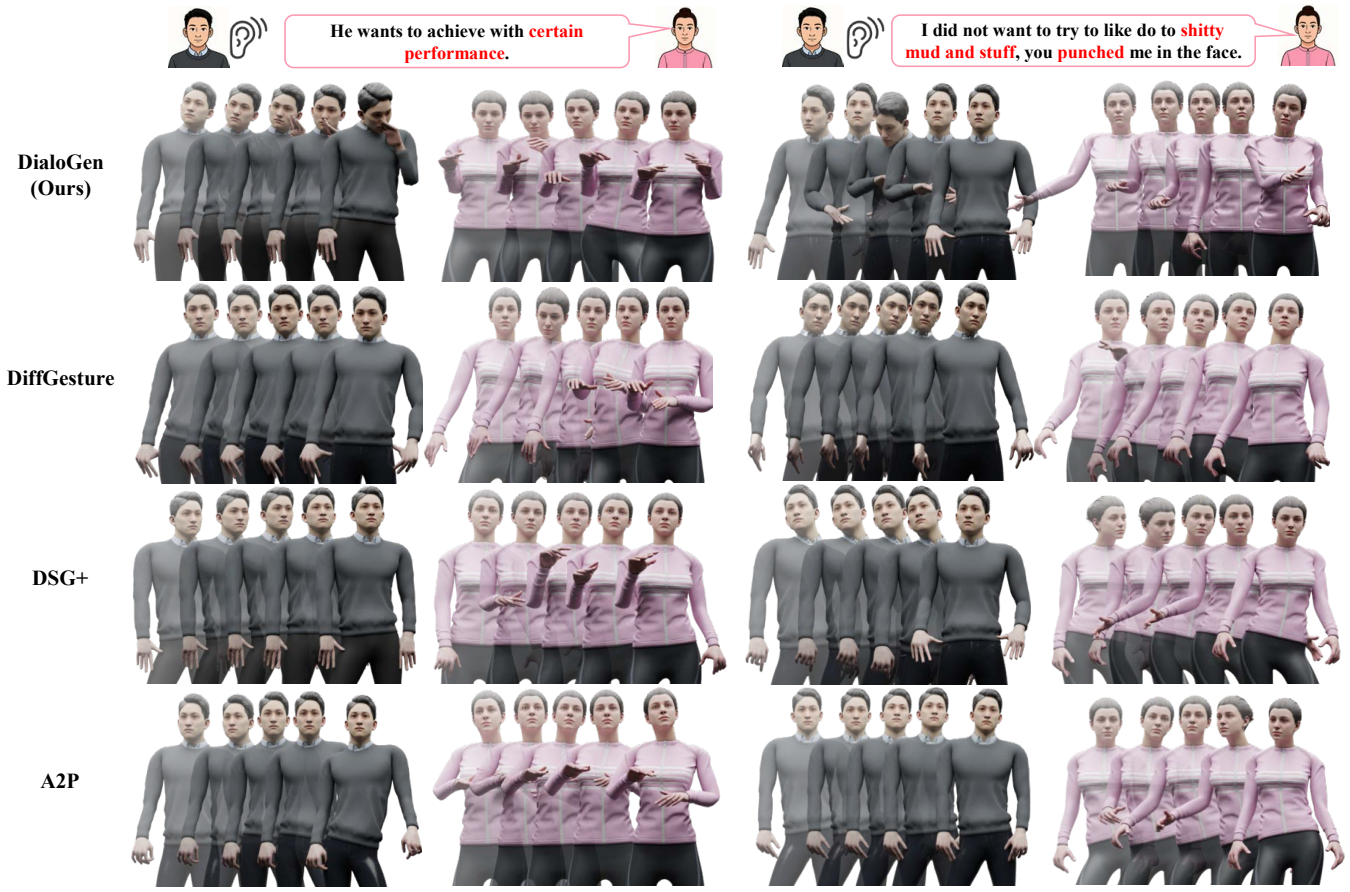

Figure 6: Qualitative Comparison on TWH Dataset. DiffuseGesture, DiffuseStyleGesture+, Audio2Photoreal, and Ground Truth are abbreviated as DG, DSG+, A2P, and GT respectively. Compared to state-of-the-art methods, our approach generates a wider range of natural, agile, and diverse speaking and listening gesture motions.

Qualitative Evaluation

As shown in Fig. 6, our approach generates visually striking and human-like outcomes that outperform baseline methods. While the baseline methods produce reasonably coherent sequences, DiffGesture sometimes exhibits subtle jitters, and DiffuseStyleGesture+ occasionally produces stiff and unnatural gestures. Audio2Photoreal generates co-speech gestures with slower gesture transitions. Our method, however, is capable of generating realistic, diverse, and identity-preserving outputs when processing speech inputs. In listening mode, all three baseline methods are limited to generating only minimal body and head movements, struggling to produce contextually appropriate responses. By leveraging the design of our dual-stream diffusion network and mutual interaction estimation, our approach can generate contextually appropriate gestures, such as a "face-covering motion indicating shame" or a "wide-body sway reflecting laughter", enabling more expressive non-verbal communication.

As demonstrated in Fig. 5, the gesture styles produced by our method accurately match the ground truth, owing to the incorporation of our ID style representation module. Our identity-decoupled representation approach, in contrast to the traditional one-hot encoding method, proves to be more effective in preserving identity during gesture generation. Moreover, we are surprised to discover that by modifying the style input, our method enables the transfer of ID styles.

Ablation Study

As shown in Tab. 2, we conduct an ablation study to evaluate the effectiveness of each component. We conduct ablation experiments from five aspects. Firstly, we evaluate the impact of the weight-sharing mechanism (w/o Weights Sharing) by training two networks with independent weights. The results show a significant decline across all four evaluation metrics, underscoring the importance of the weight-sharing mechanism in capturing the complex interactions between dual-person gestures while maintaining a balance between motion quality and diversity. Additionally, without this mechanism, the generated gestures exhibit noticeable drift at the root node, resulting in inconsistencies in motion across consecutive frames. After removing the mutual interaction estimation (w/o MIENet) mechanism, both FD_g and Div_g witness a significant decline, highlighting the crucial role of information transfer between the dual-stream diffusion networks in gesture generation.

Next, we evaluate the impact of the Identity Decoupled

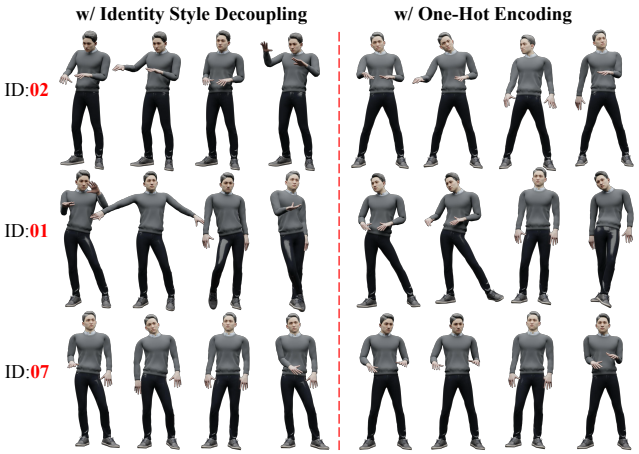

Figure 7: Examples of identity style transfer results. Identity-decoupled style guidance (left), one-hot encoding style guidance (right).

Methods	$FD_g\downarrow$	$FD_k\downarrow$	$Div_g\rightarrow$	$Div_k\rightarrow$
DialoGen (Ours)	**1.179**	**2.326**	**123.94**	**175.23**
w/o Weights Sharing	1.895	2.861	95.72	136.47
w/o MIENet	1.493	2.863	100.81	143.70
w/o Identity Style	1.705	3.361	131.32	188.68
w/o Identity Style†	1.749	2.887	139.48	194.68
w/o Text Input	1.325	2.594	107.62	155.01

Table 2: Ablation study results. † denotes the use of one-hot encoding for injecting identity style information.

Representation module. We find that removing the identity decoupled representation input (w/o Identity Style) led to a significant increase in both diversity metrics, even surpassing the performance of all baseline methods. However, the two Fréchet distance metrics deteriorated considerably, and the method lost its ability to capture identity-specific styles. Combined with the user study, our identity style representation effectively preserves personal style information. We then replace the identity information with one-hot encoding (w/o Identity Style†) and observe that the changes in the four metrics were similar to those when identity information is completely removed. As shown in Fig. 7, when using identity style representation, modifying the input identity style allows for a change in the overall motion style. For example, transitioning from ID 02, which represents a "gentleman with hands crossed in front of the chest" to ID 01 results in more exaggerated body language, resembling the "energetic gestures of a lively young woman". The one-hot encoding approach struggles to generate gestures that are consistent with the intended style. Finally, we evaluate the necessity of text inputs. The results indicate that the semantic information from text better guides the gesture generation

User Study

User studies are considered the gold standard for evaluating co-speech gesture generation (Qi et al. 2024b). We randomly

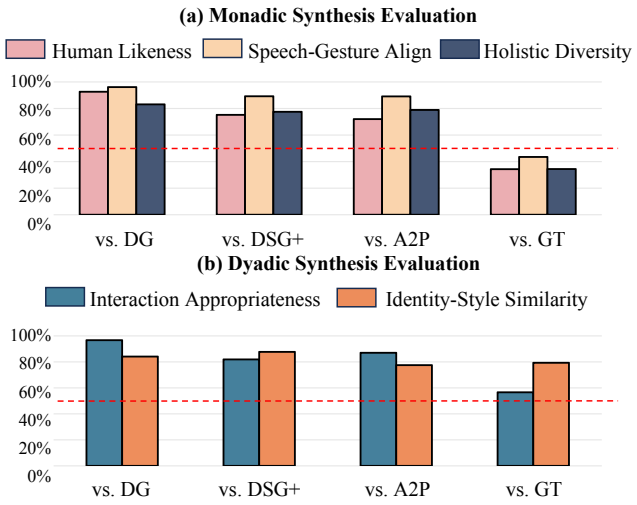

Figure 8: Results of the user study. DiffuseGesture (Zhu et al. 2023), DiffuseStyleGesture+ (Yang et al. 2023b) Audio2Photoreal (Ng et al. 2024), and Ground Truth are abbreviated as DG, DSG+, A2P, and GT respectively.

sampled 30-minute dialog segments from the test set and recruited 26 participants from diverse backgrounds. Each evaluation involved 10-second video clips and was conducted in two phases (Fig. 8). In Phase 1, participants compared pairs of videos generated from the same audio and rated Human Likeness, Speech-Gesture Alignment, and Holistic Diversity. In Phase 2, we assess the Interaction Appropriateness and Identity-style Similarity within the context of Dyadic Synthesis in dialog scenarios. In this phase, real data was provided as a reference, and participants were asked to evaluate the appropriateness of the non-verbal gesture interactions and their similarity to the reference style. This demonstrates that our proposed DialoGen effectively generates high-quality, realistic, and diverse outputs.

Conclusion

We introduce DialoGen, a novel and comprehensive framework designed to generate realistic, synchronized and context-aware gestures for both interlocutors in dialog scenarios, conditioned on conversational audio. Unlike existing methods that focus primarily on generating gestures for a single speaker, DialoGen innovatively augments the proposed interactive dual-diffusion model with mutual interaction estimation, allowing for the simultaneous generation of high-quality gestures for both participants. Moreover, we leverage a supervised contrastive learning approach to decouple identity style guidance for preserving individual identity styles. Through extensive experiments across various dialog contexts, we demonstrate that DialoGen outperforms current state-of-the-art methods by generating more realistic and expressive dialog gestures. Overall, we believe our method represents a significant milestone in advancing dialog gesture generation.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62572152).

References

Ao, T.; Gao, Q.; Lou, Y.; Chen, B.; and Liu, L. 2022. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)*, 41(6): 1–19.

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

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5: 135–146.

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*, 477–486.

Chang, C.-J.; Zhang, S.; and Kapadia, M. 2022. The IVI Lab entry to the GENEA Challenge 2022–A Tacotron2 based method for co-speech gesture generation with locality-constraint attention mechanism. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, 784–789.

Chen, J.; Liu, Y.; Wang, J.; Zeng, A.; Li, Y.; and Chen, Q. 2024. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7352–7361.

Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518.

Cheng, Q.; Li, X.; and Fu, X. 2024. SIGGesture: Generalized Co-Speech Gesture Synthesis via Semantic Injection with Large-Scale Pre-Training Diffusion Models. *arXiv preprint arXiv:2405.13336*.

Fan, K.; Tang, J.; Cao, W.; Yi, R.; Li, M.; Gong, J.; Zhang, J.; Wang, Y.; Wang, C.; and Ma, L. 2025. Freemotion: A unified framework for number-free text-to-motion synthesis. In *European Conference on Computer Vision*, 93–109. Springer.

Fu, C.; Wang, Y.; Zhang, J.; Jiang, Z.; Mao, X.; Wu, J.; Cao, W.; Wang, C.; Ge, Y.; and Liu, Y. 2024. MambaGesture: Enhancing Co-Speech Gesture Generation with Mamba and Disentangled Multi-Modality Fusion. *arXiv preprint arXiv:2407.19976*.

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, 206–216. Wiley Online Library.

Ghosh, A.; Dabral, R.; Golyanik, V.; Theobalt, C.; and Slusallek, P. 2025. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *European Conference on Computer Vision*, 418–437. Springer.

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*, 101–108.

Harz, L.; Voß, H.; and Kopp, S. 2023. FEIN-Z: Autoregressive Behavior Cloning for Speech-Driven Gesture Generation. In *Proceedings of the 25th International Conference on Multimodal Interaction*, 763–771.

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

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

Huber, P. J. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, 492–518. Springer.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33: 18661–18673.

Kipp, M.; Neff, M.; Kipp, K. H.; and Albrecht, I. 2007. Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. In *Intelligent Virtual Agents: 7th International Conference, IVA 2007 Paris, France, September 17-19, 2007 Proceedings 7*, 15–28. Springer.

Kopp, S.; and Wachsmuth, I. 2004. Synthesizing multimodal utterances for conversational agents. *Computer animation and virtual worlds*, 15(1): 39–52.

Korzun, V.; Beloborodova, A.; and Ilin, A. 2023. The FineMotion entry to the GENEA Challenge 2023: DeepPhase for conversational gestures generation. In *Proceedings of the 25th International Conference on Multimodal Interaction*, 786–791.

Kucherenko, T.; Jonell, P.; Van Waveren, S.; Henter, G. E.; Alexandersson, S.; Leite, I.; and Kjellström, H. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 international conference on multimodal interaction*, 242–250.

Kucherenko, T.; Nagy, R.; Yoon, Y.; Woo, J.; Nikolov, T.; Tsakov, M.; and Henter, G. E. 2023. The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the 25th International Conference on Multimodal Interaction*, 792–801.

Lee, G.; Deng, Z.; Ma, S.; Shiratori, T.; Srinivasa, S. S.; and Sheikh, Y. 2019. 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.

Levine, S.; Krähenbühl, P.; Thrun, S.; and Koltun, V. 2010. Gesture controllers. In *Acm siggraph 2010 papers*, 1–11.

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*, 11293–11302.

Liang, H.; Zhang, W.; Li, W.; Yu, J.; and Xu, L. 2024. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 1–21.

Liu, H.; Zhu, Z.; Iwamoto, N.; Peng, Y.; Li, Z.; Zhou, Y.; Bozkurt, E.; and Zheng, B. 2022. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European Conference on Computer Vision*, 612–630. Springer.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mughal, M. H.; Dabral, R.; Habibie, I.; Donatelli, L.; Habermann, M.; and Theobalt, C. 2024. ConvoFusion: Multi-Modal Conversational Diffusion for Co-Speech Gesture Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1388–1398.

Ng, E.; Romero, J.; Bagautdinov, T.; Bai, S.; Darrell, T.; Kanazawa, A.; and Richard, A. 2024. From audio to photoreal embodiment: Synthesizing humans in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1001–1010.

Qi, X.; Liu, C.; Li, L.; Hou, J.; Xin, H.; and Yu, X. 2024a. Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation. *IEEE Transactions on Multimedia*.

Qi, X.; Pan, J.; Li, P.; Yuan, R.; Chi, X.; Li, M.; Luo, W.; Xue, W.; Zhang, S.; Liu, Q.; et al. 2024b. Weakly-Supervised Emotion Transition Learning for Diverse 3D Co-speech Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10424–10434.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Shafir, Y.; Tevet, G.; Kapon, R.; and Bermano, A. H. 2023. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*.

Xu, Z.; Lin, Y.; Han, H.; Yang, S.; Li, R.; Zhang, Y.; and Li, X. 2024. Mambatalk: Efficient holistic gesture synthesis with selective state space models. *arXiv preprint arXiv:2403.09471*.

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

Yang, S.; Xue, H.; Zhang, Z.; Li, M.; Wu, Z.; Wu, X.; Xu, S.; and Dai, Z. 2023b. The diffusestylegesture+ entry to the genea challenge 2023. In *Proceedings of the 25th International Conference on Multimodal Interaction*, 779–785.

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*, 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.

Zhang, Z.; Ao, T.; Zhang, Y.; Gao, Q.; Lin, C.; Chen, B.; and Liu, L. 2024. Semantic Gesticulator: Semantics-Aware Co-Speech Gesture Synthesis. *ACM Transactions on Graphics (TOG)*, 43(4): 1–17.

Zhao, W.; Hu, L.; and Zhang, S. 2023. Diffugesture: Generating human gesture from two-person dialogue with diffusion models. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, 179–185.

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.