DOI: 10.1111/cgf.70412
EUROGRAPHICS 2026 / B. Masia and J. Thies
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 45 (2026), Number 2

Conversational Gesture Model (CGM): Extending Speaker-Centric Audio-Driven Motion Generation to Full Conversation Gestures

T. Koren[1] A. Rosenthal[1] D. Friedman[2] A. Shamir[1]

[1]School of Computer Science, Reichman University, Israel
[2]School of Communication, Reichman University, Israel

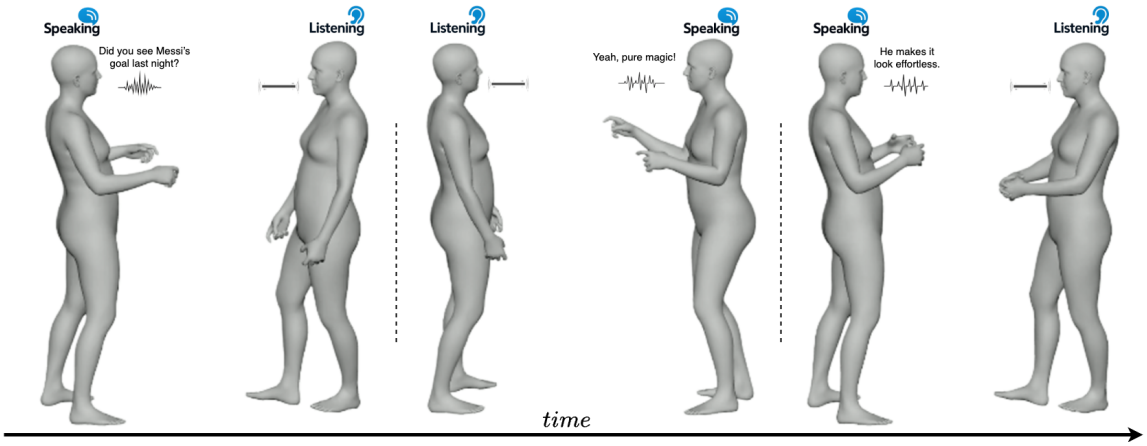

Figure 1: *Illustration of a dyadic conversation generated with the proposed* Conversational Gesture Model (CGM). *The model jointly synthesizes both* ***speaking*** *gestures and* ***listening*** *behaviors, and seamlessly alternates roles across time. This unified framework captures the fluid dynamics of natural conversation by conditioning gestures on both Character and Interlocutor inputs.*

Abstract

In this work we extend speaker-centric audio-driven gesture synthesis toward a unified conversational model that jointly captures both speaking and listening behaviors. Existing speaker-centric models effectively generate gestures aligned with speech but overlook the bidirectional dynamics that characterize natural dialogue. To address this limitation, we propose the Conversational Gesture Model (CGM), a cross-attention-based model capable of synthesizing gestures conditioned on interlocutor conversational cues such as gestures, tone, and textual semantics. By leveraging cross-attention mechanisms, the model fuses interlocutor audio and text features with character gesture encodings, enabling a single system to seamlessly alternate between speaking and listening roles of the same character. Hence, our model enables a single system to act as both speaker and listener, capturing the fluid role shifts and mutual influence inherent in conversation. Experiments demonstrate that this approach preserves the quality of speaker-driven gestures while significantly improving the realism, coherence, and responsiveness of full conversational interactions.

1. Introduction

Recent advances in audio-driven motion generation predominantly focus on speaker-centric scenarios, where generated gestures are aligned with a speaker's audio signals. While these approaches achieve impressive results, they remain limited to a single role and often under-represent the dynamics of natural conversations. In a real dialogue, interactions unfold between at least two participants, where behaviors from the interlocutor—such as nods, posture shifts, and attentive gazes—play a crucial role in shaping communication effectiveness and realism.

Several recent works have started to move beyond purely speaker-only pipelines [YXZ*23; ZCL*24; NWX*24], incorporat-

ing aspects of conversational motion. Yet most existing models either separate speaking and listening behaviors, depend on specialized settings, or train only on relatively small conversational datasets rather than exploiting the much broader availability of speaker-only data. This limits scalability and richness of motion that can be learned.

In this work, we move beyond purely speaker-driven pipelines and propose the *Conversational Gesture Model (CGM)*, a unified model that generates character gestures conditioned on conversational interactions with an interlocutor, and includes both speaking and listening gestures in one model (Fig. 1). CGM incorporates interlocutor-aware modules that interpret multimodal cues and dynamically respond to conversational inputs. By leveraging cross-attention mechanisms, the model fuses interlocutor audio and text features with character gesture encodings, enabling a single system to seamlessly alternate between speaking and listening roles of the same character.

Beyond the technical contribution, CGM moves toward embodied agents that interact in lifelike ways. By unifying speaking and listening roles, it enables avatars, robots, and digital assistants to respond naturally during conversation. Such motion is key for immersive communication, entertainment, and human–robot interaction, where realism fosters trust and engagement.

We explicitly distinguish between two conversational roles: the *Character*, whose body motion is synthesized by the model, and the *Interlocutor*, whose multimodal cues guide this synthesis. Both roles may speak and listen during a dialogue; however, CGM generates motion only for the Character, conditioning it on audio and textual cues from the Interlocutor. For two-person motion generation, CGM is applied twice by swapping the Character and Interlocutor roles, allowing both participants to be animated without modifying the model architecture.

CGM augments a pretrained speaker-centric diffusion transformer with an interlocutor-aware pathway (see Fig. 3 for an overview). By leveraging a pretrained speaker-centric model, our approach preserves the personalization encoded in speaker gestures while extending the system to generate responsive, role-aware behaviors across full conversational interactions. For both the *Character* and the *Interlocutor*, audio features A and text features T are used. A is represented using raw waveforms encoded by a six-block residual 1D CNN audio encoder, and transcripts are embedded with a fastText text encoder [BGJM17] yielding T. These features are concatenated and projected through a linear layer to form fused embeddings of the main Character F_c and the Interlocutor F_i. On the Character side, F_c is combined with a seed pose, temporal/positional embeddings, and latent motion tokens provided by a residual VQ-VAE (RVQ-VAE) trained per body region (upper body, hands, lower body (see Fig. 2). Inside the backbone transformer, the Interlocutor embedding F_i conditions the Character stream via lightweight cross-attention blocks, modulating generation without retraining the entire model. Finally, three parallel decoders reconstruct full-body motion of the main Character from the latent sequence.

Through experimental evaluations on combined conversational datasets, CGM demonstrates enhanced realism and coherence in full conversational interactions. Importantly, it preserves the quality of speaker-driven gestures while significantly improving interlocutor responsiveness, thus bridging the gap toward fully interactive, audio-driven conversational agents (see Fig. 1 and supplemental video).

2. Related Work

2.1. Speaker-Centric Generation from Speech

Recent advances in speaker-centric motion generation have increasingly adopted diffusion-based approaches [DN21] to address the limitations of earlier methods. For example, DiffGesture [ZLL*23] applies a diffusion model to capture the speech–gesture relationship. DiffuseStyleGesture [YWL*23] expands this by incorporating emotional control, while AMUSE [CDA*24] and EMOTE [DCT*23] disentangle emotion to provide stronger guidance. DiffSHEG [CLW*24] extends this line of work to real-time, holistic 3D expression and gesture generation. UnifiedGesture [YWW*23] further uses reinforcement learning to refine gesture quality, and GestureDiffuCLIP [AZL23] utilizes contrastive learning [TGH*22] for prompt-based style control. In parallel, retrieval-augmented and semantics-aligned gesture generation has been explored to better ground motion in linguistic content, as exemplified by RAG-based and semantic alignment approaches [ZAZ*24]. Listen, Denoise, Action! [ANBH23] further demonstrates diffusion as a strong generative prior for audio-driven motion, while EMAGE [LZB*24] targets holistic co-speech gesture generation (including face, hands, and body) via masked audio–gesture modeling. More recently, additional work has emerged in this space, such as SynTalker [CLD*24], which continues to explore diffusion-based architectures for enhanced semantic alignment and stylistic variation in gesture generation. Despite these advances, limited motion diversity in co-speech datasets remains a key bottleneck, making it difficult to accommodate a wide range of real-world scenarios, such as generating gestures for conversation.

2.2. Listener-Centric Generation from Speech

Listener-centric motion generation centers on how listeners nonverbally respond to a speaker's behavior. Early efforts primarily addressed the facial reactions of listeners. Gillies et al. [GPSS08] introduced a pioneering data-driven approach, which relied solely on audio clips as input. More recent work by Zhou et al. [ZBZ*22] incorporates both speaker facial cues and audio signals using deep neural networks to synthesize listener faces. Subsequent studies expand this scope to include head motion and additional generative techniques, such as a VQ-VAE (Ng et al., [NJH*22]) or a diffusion model (Liu et al.[LWF*23]). Song et al. [SYJ*23] further enhances the realism of listener head motion by integrating emotional information, while the REACT challenge (Song et al. [SSL*23]) focuses on real-time, multi-listener facial reactions to the speaker's live output.

Despite these advancements, most research has remained focused on facial expressions, neglecting full-body motion. Even less research has investigated the movements of the body of the interlocutor. Early approaches (Joo et al. [JSCS19]; Jonell et al. [JKHB20]) predict listener motion directly from the speaker's behavior; Ahuja et al. [AMMS19] extend this by incorporating

the listener's previous pose. More recently, GAN-based methods have been proposed for listener body pose generation (Tuyen and Celiktutan, [TC22], [TC23]), and challenges such as GENEA (Kucherenko et al. [KNY*23]) and DYAD (Palmero et al. [PBC*22]) explore holistic and upper-body reaction forecasting, respectively. Recent work [NRB*24] introduces a framework for generating full-bodied, photorealistic avatars that capture both facial nuances and body gestures for conversations. Furthermore, Zhang et al. [ZFT*24] present a contrastive learning-based model to align speaker-listener motion semantics. Recent dyadic and multi-party gesture generation further includes ConvoFusion [MDH*24] and concurrent two-person co-speech generation with interactive diffusion (Co^{3}Gesture) [QWZ*25], alongside the GENEA2023 competition and its submitted systems [KNY*23; DMAB23]. Several recent works also introduce new conversational datasets to support these models, including ConvoFusion [MDH*24], Co^{3}Gesture [QWZ*25], and Audio2Photoreal [NRB*24], reflecting a growing emphasis on high-quality dyadic motion capture. More recently, large language models have been employed to capture high-level conversational semantics and interaction structure, as demonstrated by Echo [XFWW25]. In parallel, some approaches focus on interaction with people and objects rather than conversational exchange, such as InterAct [XLZ*25] and It Takes Two [TC25], which study coordinated motion in interactive but non-conversational settings.

3. Our Approach

In this section, we introduce our novel framework for co-speech gesture synthesis that accommodates both single-speaker and interactive two-person scenarios.

We first provide a high-level overview of the method (Sec. 3.1). We then describe the Motion Representation Module (Secs. 3.2–3.3), which encodes/decodes upper body, hands, and lower body with separate RVQ-VAEs. Next, we detail the Conversational Gesture Model (Sec. 3.4) that injects interlocutor cues into the pretrained transformer and aligns them via a final cross-attention layer. We formulate generation as conditional denoising diffusion (Sec. 3.5), specify training objectives including a listening-aware loss (Sec. 3.6), and conclude with inference and a two-model pipeline for synthesizing full conversations (Sec. 3.8).

3.1. Overview

Our method extends speaker-only co-speech gesture generation to a dyadic setting, enabling the synthesis of gesture sequences for both speakers and listeners based on speech inputs.

To achieve this, we integrate a pre-trained speaker-centric model — originally capable of generating gestures for a single speaker based on speech — by integrating interlocutor-aware modules. These modules interpret the Character's gestures and produce corresponding listener behaviors, effectively extending the system to address both sides of an interaction. The architecture of the overall model is depicted in Figure 3.

3.2. Motion Encoding

Recent advances in motion generation have demonstrated that vector-quantized autoencoders (VQ-VAE) [vdOVK17] are highly effective at compressing motion information. In our approach, we also adopt vector quantization for motion encoding by employing a residual VQ-VAE (RVQ-VAE) [CLD*24; NRB*24; ZZC*23] as the quantization layer. To reduce the coupling between different body parts, we partition the body into three segments—upper body, fingers, and lower body—as in [CLD*24; LZB*24], and train a separate RVQ-VAE for each segment (see Figure 2).

Concretely, let the motion sequence be denoted by

$$x \in \mathbb{R}^{N\times J\times 3}$$

with N frames of motion and J three-dimensional joints.

This sequence is first encoded by a 1D convolutional encoder E into a latent representation

$$z_1 \in \mathbb{R}^{n\times d},$$

where the downsampling ratio is n/N, and d is the latent dimensionality. The latent vectors z_1 are then fed into the first quantization layer Q_1. In this layer, each latent vector is mapped to its nearest code in the codebook

$$C_1 = \{c_k^1\}_{k=1}^K \subset \mathbb{R}^d,$$

yielding the quantized output $\hat{z}^1_{1:n}$. The corresponding quantization residual is computed as

$$r^1_{1:n} = \hat{z}^1_{1:n} - z_1.$$

This residual is passed to the second quantization layer Q_2, which uses its own codebook

$$C_2 = \{c_k^2\}_{k=1}^K \subset \mathbb{R}^d,$$

to produce the second quantized code $\hat{z}^2_{1:n}$. This procedure can be iterated to obtain additional codes $\hat{z}^3_{1:n}, \hat{z}^4_{1:n}, \ldots$. Finally, the complete motion representation is obtained by summing all the quantized codes:

$$\hat{z} = \sum_{q=1}^{Q} \hat{z}^q.$$

In practice, we use $Q = 6$ quantization layers, each with a code dimension of 512 and a codebook size of 512 (see Section 4.2).

3.3. Motion Decoding

Motion decoding is performed using three parallel 1D convolutional decoders, each dedicated to a body segment (upper body, lower body, and hands). This design ensures that localized motion dynamics can be reconstructed independently while still contributing to full-body coherence (see Figure 2).

During training, the model learns from ground truth motion sequences, which are encoded and used as supervision to optimize the multimodal fusion process. The motion decoder is not involved during this stage.

During inference, the motion encoder is bypassed, and the decoders are applied to the model's output representations to generate the final *Character* gestures. These gestures are derived jointly from the *Character*'s and *Interlocutor*'s speech and text inputs.

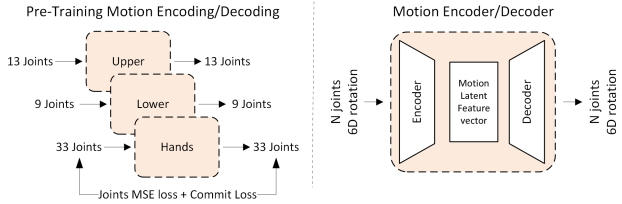

Figure 2: *Pretraining of the motion encoder/decoder.* ***Left:*** *Region-specific reconstruction of upper body (13 joints), lower body (9 joints), and hands (33 joints) using separate RVQ-VAE modules, optimized with joint-level MSE and commitment losses.* ***Right:*** *General motion encoder/decoder that maps N joints in 6D rotation to a compact latent feature vector and back. This pretraining step is performed prior to training the full Conversational Gesture Model (CGM).*

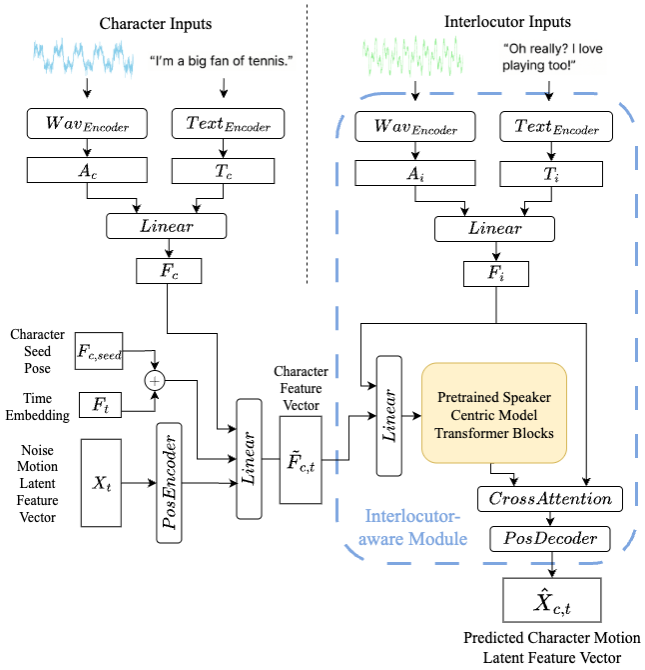

Figure 3: *Overview of the Conversational Gesture Model (CGM). Character inputs (left) provide the core speech-driven representation, which combines audio/text embeddings with seed pose, time, and latent motion features. The Interlocutor-aware Module (blue, right) encodes interlocutor audio/text features, injects them into pretrained speaker-centric transformer blocks, and fuses them via cross-attention. The final decoder outputs Character gestures that are both speech-aligned and contextually responsive.*

3.4. Conversational Gesture Model (CGM)

Unlike purely speaker-centric pipelines, the proposed *Conversational Gesture Model (CGM)* incorporates a dedicated *interlocutor-aware module* (Figure 3, blue box), that enables the model to function beyond speech-driven gesture synthesis, allowing CGM to capture conversational cues and dynamically modulate the generation of gestures. By modeling both roles within a single system, CGM supports more realistic bidirectional interaction.

We assume access to two synchronized input channels: (i) the **Character**'s audio and text, and (ii) the **Interlocutor**'s audio and text. Character inputs drive the primary gesture generation process, while interlocutor cues modulate these outputs, ensuring bidirectional responsiveness. For details on dataset preparation, see Section 4.1.

Conversational gesture generation is treated as an asymmetric conditioning task: the model synthesizes motion for the Character, while the Interlocutor provides contextual cues that guide this synthesis. This asymmetry motivates the design of our interlocutor-aware conditioning and cross-attention mechanisms.

We begin by extracting feature representations from both the Character's and the Interlocutor's speech and text inputs. For textual features, we employ a fastText-based encoder [BGJM17], which provides subword-level embeddings that capture semantic and morphological information.

For audio, we design a raw waveform encoder composed of six stacked residual convolutional blocks, inspired by ResNet-style architectures [HZRS16] and implemented following the `timm` library [Wig19]. This encoder hierarchically processes waveform signals, enabling the extraction of rich temporal and spectral features directly from raw audio.

3.4.1. Character Stream

The Character stream is initialized using two auxiliary embeddings: a seed pose embedding F_{seed}, which represents the initial body configuration of the Character, and a temporal embedding F_t, which encodes the current diffusion timestep.

Let A and T denote audio and text feature vectors, respectively, obtained from the encoders. The Character streams

$$A_c = \text{WavEnc}(\text{CharacterAudio}), \quad T_c = \text{TextEnc}(\text{CharacterText}),$$

are fused into a multimodal embedding:

$$F_c = \text{Linear}([A_c \,\|\, T_c]).$$

The Character representation is further conditioned on the seed pose F_{seed}, temporal embedding F_t, and a latent motion vector X_t. After positional encoding and projection, we obtain:

$$\tilde{F}_{c,t} = \text{Linear}(\text{PosEnc}(X_t), F_{\text{seed}} + F_t, F_c).$$

3.4.2. Interlocutor-aware Module

The Interlocutor stream is processed analogously:

$$A_i = \text{WavEnc}(\text{InterlocutorAudio}), \quad T_i = \text{TextEnc}(\text{InterlocutorText}),$$

$$F_i = \text{Linear}([A_i \,\|\, T_i]).$$

The Interlocutor embedding F_i is injected into the pretrained *Speaker-Centric Model* transformer blocks, conditioning the Character feature representation:

$$H = \text{PreTrainedSpeakerTransformers}(\text{Linear}(\tilde{F}_{c,t}, F_i)).$$

To explicitly model asymmetric conversational conditioning, we apply a cross-attention layer that allows the Character representation to selectively attend to Interlocutor cues. Intuitively, this mechanism lets the model decide *which* aspects of the interlocutor's behavior are relevant at each moment of gesture generation, reflecting the fact that Character motion is generated in response to the interlocutor.

The output of the transformer is passed through a cross-attention layer, where the Character states serve as the reference signals that select relevant information from the Interlocutor embedding, which provides the contextual content to be integrated:

$$H' = \text{CrossAttn}(Q = H, K = F_i, V = F_i).$$

This two-stage conditioning ensures that interlocutor information is (i) integrated throughout the transformer updates and (ii) explicitly aligned with Character motion features via cross-attention. Finally, a positional decoder maps this enriched representation into motion space:

$$\hat{X}_{c,t} = \text{PosDecoder}(H').$$

3.5. Diffusion Process

Gesture generation is formulated as a conditional denoising diffusion process, following the paradigm introduced in Motion Diffusion Models (MDM) [TRG*23]. Starting from Gaussian noise $F_{\text{motion}} \sim \mathcal{N}(0, I)$, the model iteratively refines a latent motion representation across T denoising steps. At each step t, the Conversational Gesture Model (CGM) predicts the clean motion feature x_0 conditioned on (i) the **Character**'s audio–text embedding F_c, (ii) the **Interlocutor** embedding F_i, and (iii) auxiliary conditioning signals such as the seed pose $F_{\text{c,seed}}$ and temporal embedding F_t.

Formally, the forward process gradually perturbs the ground-truth motion feature vector x_0 into noise:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\big(\sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I\big),$$

while the reverse process denoises step by step:

$$x_{t-1} = \text{Denoise}_\theta(x_t, F_c, F_i, F_{\text{c,seed}}, F_t).$$

The process is repeated for $T \to 1$ iterations until the final motion feature F_{motion} is obtained (Figure 4).

3.6. Model Training

Pretraining of Motion Encoder/Decoder. Before training the full conversational framework, we pretrain the motion encoder/decoder modules (Figure 2) to learn a compact latent representation of body motion. The body is partitioned into three regions—upper body (13 joints), lower body (9 joints), and hands (33 joints)—each with an RVQ-VAE that encodes 6D joint rotations into a latent vector and reconstructs them back. Training is guided by a combination of joint-level mean squared error (MSE) and a commitment loss that stabilizes codebook usage.

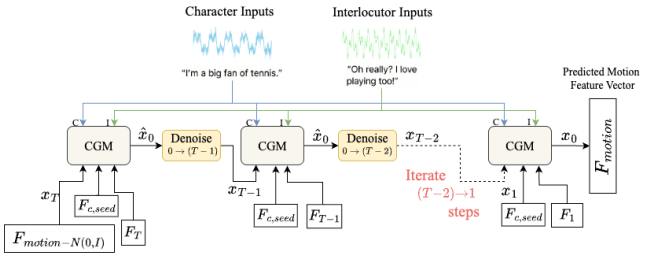

Figure 4: *Illustration of the diffusion-based motion generation process. Starting from Gaussian noise, the system iteratively denoises the motion representation using CGM, conditioned on Character inputs (blue) and Interlocutor inputs (green). At each step, seed pose $F_{c,seed}$ and temporal features F_t are injected, producing the final predicted motion feature vector F_{motion}.*

CGM Training. During training, only the pretrained encoder is used to obtain latent supervision targets, while the decoders are not involved in optimization (Figure 6, left). The diffusion model generates a predicted motion feature vector $F_{P,\text{motion}}$ from noisy latent inputs, conditioned on Character and Interlocutor features (see Section 3.5 for details). In parallel, the ground-truth motion sequence is passed through segment-specific motion encoders (upper body, lower body, and hands) to obtain the target representation $F_{GT,\text{motion}}$.

We train the model with a combination of a reconstruction loss and a listening-aware loss.

Huber Loss. To ensure stable and robust motion feature prediction, we employ the Smooth L1 loss [Gir15], also known as the Huber loss, defined as

$$\mathcal{L}_{\text{SmoothL1}}(x,y) = \begin{cases} 0.5(x-y)^2, & \text{if } |x-y| < 1 \\ |x-y| - 0.5, & \text{otherwise} \end{cases}$$

where x and y denote the predicted and target latent representations, respectively. This serves as the baseline reconstruction loss:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{SmoothL1}}(F_{P,\text{motion}}, F_{GT,\text{motion}}).$$

Listening-Aware Loss. To enhance the model's ability to learn listener-specific gestures, we introduce an auxiliary loss that weights the contribution of training samples based on the **frame-wise speech activity (turn-taking state)** of both participants. We estimate this state directly from each participant's isolated audio channel using a soft voice-activity score. This ensures that the model learns natural listener behaviors by prioritizing training frames where listening gestures are most relevant.

To enhance the model's ability to learn listener-specific gestures, we introduce an auxiliary loss that weights the contribution of training samples based on the **speaking dynamics** of both participants.

This ensures that the model learns natural listener behaviors by prioritizing training frames where listening gestures are most relevant.

Formally, let $A_c(t)$ and $A_i(t)$ denote the normalized audio amplitudes of the *Character* and the *Interlocutor* at time t, respectively. We define a soft speech-activity probability for each participant using a sigmoid-based classification:

$$S(t) = \sigma(\lambda_c(A_c(t) - \tau)), \quad L(t) = \sigma(\lambda_i(A_i(t) - \tau))$$

where $\sigma(x)$ is the sigmoid function, λ_c and λ_i are sensitivity parameters, and τ is a threshold that determines when a participant is considered active.

We then define the **importance weight function** $w(t)$ as:

$$w(t) = (1 - S(t))L(t) + (1 - L(t))S(t) + (1 - S(t))(1 - L(t)).$$

This function assigns:

- **High importance** when one participant is speaking and the other is listening: $(S = 1, L = 0)$ or $(S = 0, L = 1)$.
- **High importance** when both are silent: $(S = 0, L = 0)$, capturing moments of non-verbal attentiveness.
- **Low importance** when both are speaking: $(S = 1, L = 1)$, as overlapping speech is less relevant for listener modeling.

Figure 5 illustrates this process: the top panel shows raw amplitude waveforms, the middle panel displays speaking activity scores derived from sigmoid functions, and the bottom panel presents the resulting computed importance weights used to modulate the loss.

Total Loss The complete training objective combines the reconstruction and listening-aware terms:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{listen}}$$

In practice, we set the threshold and sensitivity parameters to $\tau = 0.02$, $\lambda_c = 7$, and $\lambda_i = 10$, values chosen to balance responsiveness to speech activity while avoiding over-weighting noisy fluctuations (see Section 4.2 for details).

3.7. Model Inference

During inference, motion encoders are not used, and only the decoders are active (Figure 6, right). The trained diffusion model produces the final motion feature $F_{P,\text{motion}}$, which is then passed through the segment-specific motion decoders (upper body, lower body, and hands) to reconstruct the full-body gesture sequence. For details on how the diffusion model is conditioned, please refer to Section 3.5.

3.8. Conversation Synthesis Pipeline

A full two-person conversation is generated by running two instances of CGM, one for each character. Each CGM defines a single character and treats the other participant as its interlocutor, so inputs are swapped across the two models. Let the participants be A and B with aligned audio and transcripts $(A_{\text{audio}}, A_{\text{text}})$ and $(B_{\text{audio}}, B_{\text{text}})$. Then

$$\hat{X}_A = \text{CGM}_A(A_{\text{audio}}, A_{\text{text}} \mid B_{\text{audio}}, B_{\text{text}}),$$
$$\hat{X}_B = \text{CGM}_B(B_{\text{audio}}, B_{\text{text}} \mid A_{\text{audio}}, A_{\text{text}}).$$

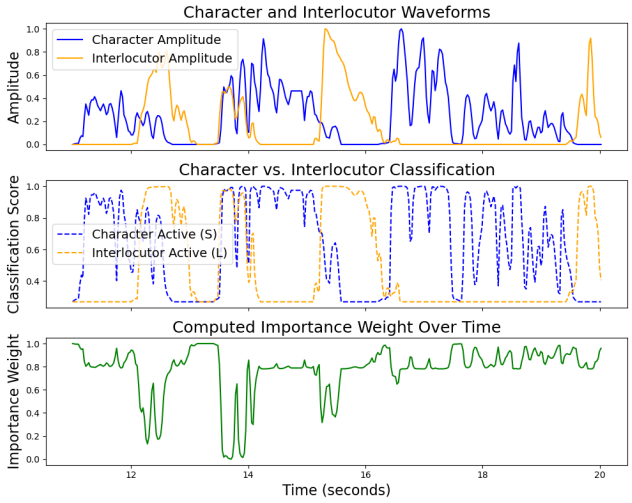

Figure 5: *Visualization of the Listening-Aware Loss components.* ***Top****: Audio amplitude waveforms of the* Character *(blue) and* Interlocutor *(orange) over time, illustrating speech activity (turn-taking)*

. **Middle**: Binary classification scores showing the speaking activity of each participant, computed using sigmoid functions. **Bottom**: The computed importance weight $w(t)$ over time, based on the classification scores.

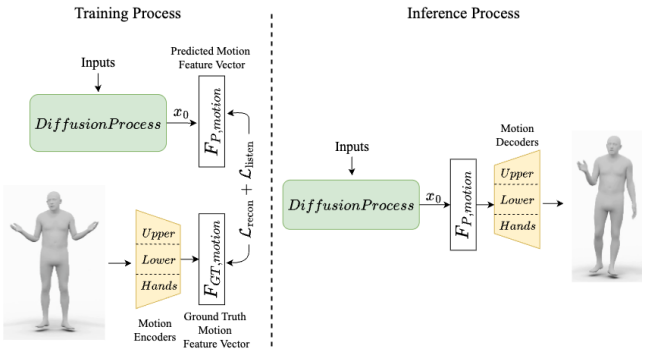

Figure 6: *Overview of the training and inference pipeline.* ***Left:*** *During training, the diffusion model predicts motion features $F_{P,motion}$, which are aligned with ground-truth features $F_{GT,motion}$ obtained from motion encoders, optimized using $\mathcal{L}_{recon} + \mathcal{L}_{listen}$.* ***Right:*** *At inference, the trained diffusion model produces motion features that are decoded by body-part-specific decoders (upper body, lower body, and hands) into full-body gestures.*

Each model uses its own inputs as *Character* features and the opposite participant's inputs as *Interlocutor* features.

4. Experiments

4.1. Dataset Construction and Alignment

At the time of conducting this work, only a small dyadic dataset was available. Therefore, our experiments build on two primary datasets:

- **BEAT2** - a large-scale **speaker only** motion dataset that utilizes SMPL-X for representing 3D human motion [LZB*24]
- **Talking With Hands** - **dyadic conversational** interactions dataset in BVH format [LDM*19]

To achieve a unified representation suitable for our approach, we conduct several preprocessing steps.

Motion Format Conversion. While BEAT2 uses SMPL-X [PCG*19], Talking With Hands relies on BVH skeleton data. SMPL-X is a parametric human body model that includes expressive hands and facial features, enabling detailed full-body motion representation. In contrast, BVH (Biovision Hierarchy) is a motion capture format that encodes joint angles for a predefined skeletal structure in a hierarchical manner. We convert the BVH sequences to SMPL-X format via inverse kinematics, following the procedure in [SGF*22]. This conversion involves:

1. Extracting 3D joint positions from the BVH motion.
2. Mapping the extracted joints to the SMPL-X anatomical structure.
3. Performing inverse kinematics to ensure proper bone lengths and joint rotations within SMPL-X.

By combining these steps, we obtain a *unified, aligned dataset* that contains synchronized audio streams (*Character* and *Interlocutor*) and SMPL-X motion sequences for both individuals. This merged dataset enables our extended system to learn from large-scale monadic and dyadic clips, ensuring robust co-speech gesture generation for single- and two-person scenarios.

Audio Channel Separation. Since the Talking With Hands dataset captures dyadic interactions, its raw audio often contains overlapping *Character-Interlocutor* segments. For each channel, we isolate the main voice from the background conversation voices, ensuring that each audio channel captures only one participant's voice. The effect of this separation can be seen in Figure 7, where the original mixed signals (top row) are transformed into individual *Character* and *Interlocutor* tracks (bottom row).

Audio Normalization. To align the two audio streams (*Character* and *Interlocutor*) across different clips, we normalize silences to zero amplitude. This step simplifies synchronization between the datasets, making it easier to align timestamps and correlate gestures with each participant's speech cues. As shown in Figure 7, silence regions are set to zero, improving the temporal alignment of the two audio channels.

4.2. Implementation Details

For our speaker-centric pre-trained model, we leverage the SynTalker model [CLD*24], and its residual vector-quantized autoencoder (RVQ-VAE) for motion representation. The RVQ-VAE includes residual blocks in both encoder and decoder, with a downscale factor of 4, and employs 6 quantization layers, each with a code dimension of 512 and a codebook size of 512.

The CGM is built upon a core diffusion-based transformer architecture, which consists of 8 transformer layers operating in a latent space of dimension 512. The seed pose embedding F_{seed} has

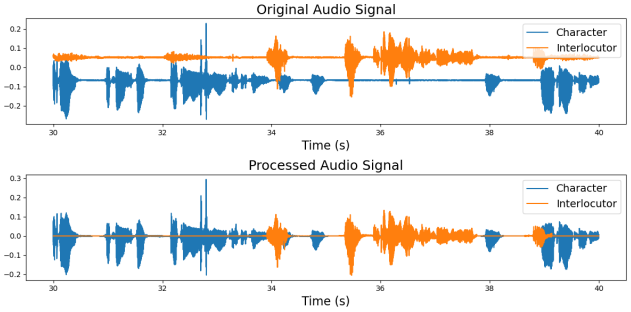

Figure 7: *Audio pre-processing pipeline example. The top row shows the original mixed audio signals of the* Character *and the* Interlocutor. *The bottom row shows the normalized signals after applying channel isolation and setting silence regions to zero, ensuring better alignment.*

dimensionality 512, while both audio and text embeddings are projected into a 256-dimensional space before being fused with the motion features. For motion representation, we use the SMPL-X body model with 55 joints, excluding facial expressions.

We trained the model using 1000 denoising steps and a batch size of 40, and the combined loss $\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{listen}}$ with $\tau = 0.02$, $\lambda_c = 7$ and $\lambda_i = 10$. Different weighting factors are adopted for the speaker and interlocutor losses, as the interlocutor audio in TWH typically exhibits lower amplitude. Using identical weights for both roles was found to reduce classifier sensitivity. Through empirical evaluation, we observed that setting $\lambda_c = 7$ and $\lambda_i = 10$ yields the most stable segmentation performance.

Each identity-specific model is trained separately using a merge of BEAT2[LZB*24] and TWH[LDM*19]. For each model, TWH contributes 3,338 motion sequences of 128 frames‘, while BEAT2 contributes 7,449 (ID 1), 7,189 (ID 2), or 7,003 (ID 6) sequences of the same length, resulting in a consistent BEAT2–TWH[LZB*24] ratio of approximately 2:1 across all training setups. For details on dataset pre-processing, see Section 4.1.

Since the BEAT2 dataset contains highly imbalanced amounts of data across speakers, we randomly selected identities with more than three hours of recordings to ensure that each model learns a stable and representative motion profile. The same training procedure is applied uniformly across all selected identities.

All components were trained on a single NVIDIA RTX 3090 GPU. On this hardware configuration, completing 500 training epochs requires approximately 3 hours. For inference, the model generates a 1-minute motion sequence from a 1-minute audio input in about 40 seconds.

5. Results

Qualitative results are provided in the supplementary video, which illustrates examples of generated conversational gestures across both speaking and listening roles. These visual demonstrations complement the quantitative evaluation by showing how CGM synthesizes natural, expressive, and responsive behaviors. In the fol-

lowing sections, we report and analyze the quantitative results, comparing CGM with state-of-the-art baselines and ground truth across multiple evaluation dimensions.

5.1. Evaluation Metrics

We use three widely adopted metrics to assess motion generation quality:

- **Fréchet Gesture Distance (FGD)**[LZI*22; HRU*17]: Measures the distribution distance between generated and ground-truth gestures, capturing overall realism and naturalness of motion.
- **Beat Consistency (BC)** [LYG*22]: Evaluates the synchronization between the generated gestures and the rhythmic structure of the speech audio, indicating temporal alignment.
- **Diversity**: Computes the variability of generated motions, reflecting the model's ability to avoid repetitive and monotonous gestures.

5.2. Evaluation Setting

In our experimental evaluation, we aim to demonstrate four main aspects of our proposed Conversational Gesture Model:

- **Full Conversation Behavior**
  To assess how CGM handles natural conversational dynamics, we evaluate it on full dyadic interactions where the Character alternates between speaking and listening in response to the Interlocutor. This setting provides insight into the model's ability to synthesize co-articulated behaviors and achieve smooth transitions between conversational roles. For this evaluation, we use a dataset derived from the Talking With Hands corpus, comprising 7,320 frames—approximately 4.1 minutes—spanning four dyadic conversations between paired participants.
- **Personalized Listening Behavior**
  We investigate whether CGM can generate Character-specific listening gestures by leveraging the pretrained speaker representations. We select three distinct Character IDs and generate listening gestures for each while the Character is in the listening role. We then calculate the Fréchet Inception Distance (FID) between each generated listening segment and three reference sets:
  - Speaking segments of the same Character ID
  - Speaking segments of the other Character IDs
  - Ground-truth listening segments

  The FID is computed on the model's latent motion representation, specifically the decoded vector $\hat{X}_{c,t}$ described in Section 3.4.2. Lower FID values relative to the same Character's speaking segments or the ground-truth listening data indicate stronger personalization, while higher distances from other Characters' speaking segments demonstrate clear identity differentiation. For this evaluation, we use a listening-focused dataset derived from the Talking With Hands corpus, comprising 2,702 frames—approximately 1.5 minutes.
- **Character Head Alignment to Interlocutor Audio**
  To evaluate the rhythmic alignment between the Character's head movements and the Interlocutor's audio, we perform a beat alignment analysis using only head movements during segments where the Character is listening. Head movements are the most consistent and salient modality for assessing rhythmic alignment in conversation. Unlike hand or body gestures, which vary widely across individuals and contexts, head nods and subtle head motions are universal cues for attentiveness, agreement, and synchronization with the Interlocutor's prosody. They occur more frequently and reliably than other nonverbal signals, providing a stable and comparable basis for alignment analysis. Prior studies [McC00; Hey06; BJ08; ABL23] have shown that head movements are tightly coupled with speech rhythm and intonation, making them a suitable proxy for conversational synchrony.
  We compare alignment scores in three scenarios:
  - Ground-truth Character head gestures paired with the Interlocutor audio
  - Generated Character head gestures paired with the Interlocutor audio
  - Generated Character head gestures paired with the Interlocutor audio randomly shifted by temporal offsets

  Higher alignment scores for generated gestures with Interlocutor audio, compared to randomly shifted audio, indicate meaningful rhythmic synchronization. Ground-truth alignment serves as an upper-bound reference. We use three Character segments of 30 seconds each, taken from different conversations. For each segment, we apply five different random audio offsets to enhance result robustness.
- **Speaker Performance Preservation**
  We verify that the integration of interlocutor-aware modules does not degrade the original Character motion generation quality. We evaluate the model on the BEATX [LZB*24] test set, consisting of 29,310 frames—approximately 16.3 minutes of speech—collected from 15 different sessions.

5.3. Baselines and Ablations

We compare to the following:

- **Random Listener**: Listening gestures are randomly sampled, providing a lower-bound baseline that ignores conversational cues.
- **SynTalker** [CLD*24]: A strong speaker-centric gesture generation model, used to evaluate whether CGM preserves speaker performance while extending to conversational settings.
- **DiffuseStyleGesture+** [YXZ*23]: An audio-to-motion diffusion model trained in a dyadic setting to generate conversational gestures. DiffuseStyleGesture+ extends the original speaker-centric DiffuseStyleGesture framework [YWL*23] to a conversational scenario by modeling listener behavior, making it a suitable baseline for comparison with our interlocutor-aware approach.
- **Audio2Photoreal** [NRB*24]: An audio-driven human animation method targeting photorealistic avatar reenactment. As it does not generate 3D SMPL-X body motion, comparisons are conducted without the avatar rendering stage and focus solely on the underlying motion signals.
- **Ours w/o Motion Encoding/Decoding**: (ablation) Removing the RVQ-VAE motion encoder–decoder and training the diffusion model directly on raw SMPL-X pose representations to

evaluate the impact of learned motion discretization on conversational gesture quality.
- **Ours w/o Interlocutor-aware Module**: (ablation) Comparing the proposed Interlocutor-aware Module strategy against direct feature concatenation. Similar to DiffuseStyleGesture+ [YXZ*23].
- **Ours w/o Listening-Aware Loss**: (ablation) Training variants without the listening-aware loss weighting.

5.4. Evaluation Analysis

5.4.1. Full Conversation Behavior

In terms of full conversation behavior, CGM demonstrates superior reactivity and attentiveness of the Character. Table 1 shows that it achieves the best performance across all three metrics—FGD, BC, and Diversity—indicating close alignment with ground truth motion, responsive rhythmic timing, and a varied repertoire of motion styles. Compared to ablated variants, removing the motion encoding/decoding module degrades motion realism and diversity, while disabling the listening-aware loss reduces beat consistency despite increasing motion variability. Removing the interlocutor-aware module further weakens conversational alignment, confirming the importance of explicit interlocutor conditioning. By generating Character gestures conditioned on Interlocutor inputs, CGM enables smooth transitions across speaking and listening roles, providing a more generalizable and expressive conversational model.

5.4.2. Personalized Listening Behavior

In terms of personalized listening behavior, Table 2 illustrates that CGM successfully generates Character gestures closely aligned with the Character-specific style, as indicated by low FID values when comparing listening gestures with the same Character's speaking gestures. Conversely, the higher FID values when compared against another Character confirm that our model effectively distinguishes between different identities. Additionally, low FID values against ground-truth listening gestures demonstrate that CGM is able to reproduce natural, personalized listening behaviors.

5.4.3. Character Head Alignment to Interlocutor Audio

Table 3 shows that CGM produces *Character* head gestures that synchronize well with the *Interlocutor*'s audio, as indicated by substantially higher alignment than the random-offset baseline and comparable performance to the Audio2Photoreal baseline. The proximity of generated scores to ground truth further suggests that CGM captures natural cross-participant rhythmic coupling.

5.4.4. Character Performance Preservation

Our proposed Conversational Gesture Model (CGM) preserves the Character's performance effectively. As shown in Table 4, CGM achieves the highest beat consistency (BC) and the greatest motion diversity among all methods, suggesting rich and rhythmically well-aligned co-speech gestures. Compared to ablated variants, CGM maintains competitive motion diversity while benefiting from improved conversational synchronization, demonstrating that interlocutor-aware conditioning does not degrade speaker expressiveness. While its FGD is slightly higher than SynTalker, this is due to CGM being retrained within a conversational setting, where Character gestures are modeled in response to Interlocutor inputs. This conversational conditioning introduces some variability in Character motion but brings valuable gains in expressiveness and behavioral diversity.

5.5. Perceptual Study

We conducted a perceptual study to evaluate our proposed Conversational Gesture Model (CGM) against both the state-of-the-art DiffuseStyleGesture+ [YXZ*23] and ground-truth conversational videos. Since existing approaches for modeling listener behavior are extremely limited, DiffuseStyleGesture+ represents the only directly comparable baseline. A total of 35 participants were recruited and shown six short video clips (15 seconds each). Four clips presented side-by-side comparisons of CGM and DiffuseStyleGesture+, while two clips compared ground-truth Interlocutor behavior paired with a generated Character. The left/right order was randomized for each video to avoid positional bias. The study involved 35 adult participants (18 males and 17 females). Ages ranged from 20 to 65 years, and participants came from a variety of professional backgrounds.

Participants were asked: "Which side demonstrates more natural conversational behavior for the Character?" and were allowed to replay each clip before making their choice. In total, 210 judgments were collected (35 participants $\times$ 6 videos).

The perceptual study did not involve the collection of personal data, recordings, or any identifiable information. Participation was voluntary, and participants only viewed pre-rendered gesture animations and provided subjective preferences via an anonymous survey.

DiffuseStyleGesture+ Comparisons In these trials, participants were asked to choose which of two videos they preferred: one produced by our system ("Ours") and one produced by DiffuseStyleGesture+ ("Not Ours"). Because there were always exactly two alternatives, the chance level corresponds to 0.5: if participants had no systematic preference, each option would be selected 50% of the time.

To assess whether preferences deviated from chance, we conducted exact binomial tests for each of the four DiffuseStyleGesture+ videos. Across all four videos combined, participants chose "Ours" in 62% of trials (87/140), significantly above chance ($p = .005$). At the level of individual videos, only Video 2 showed a significant preference (71% "Ours," $p = .017$), whereas the other three videos did not differ significantly from chance (Table 5).

To account for repeated measures, we also fitted a logistic regression model with cluster-robust standard errors by participant. The intercept-only model indicated a significant overall preference for "Ours" ($\beta = 0.50$, $SE = 0.23$, $z = 2.14$, $p = .033$), corresponding to an estimated probability of 0.62. Adding video as a fixed effect did not significantly improve model fit (likelihood ratio test $p = .51$), and none of the individual video effects reached significance.

Taken together, these analyses demonstrate a reliable overall preference for "Ours" across DiffuseStyleGesture+ comparisons, although the strength of preference varied across individual videos.

Method	ID 1 (Wayne)			ID 2 (Scott)			ID 6 (Carla)		
	FGD ↓	BC $\times 10^{-1}$ ↑	Diversity ↑	FGD ↓	BC $\times 10^{-1}$ ↑	Diversity ↑	FGD ↓	BC $\times 10^{-1}$ ↑	Diversity ↑
GT	0.000	7.855	9.578	0.000	6.696	9.578	0.000	6.966	9.578
3 Random Listener	3.247	7.188	7.274	3.247	7.188	7.274	3.247	4.988	7.274
SynTalker	2.327	8.149	5.777	1.453	8.314	9.162	1.099	7.546	5.937
DiffuseStyleGesture+	0.919	7.441	7.738	0.919	7.441	7.738	0.919	5.923	7.738
Audio2Photoreal	0.902	7.638	7.613	0.902	7.638	7.613	0.902	7.638	7.613
Ours w/o Motion Encoding/Decoding	0.731	7.584	6.412	1.102	7.903	7.221	0.682	6.944	6.938
Ours w/o Listening-Aware Loss	0.497	7.581	8.412	0.836	7.901	10.602	0.381	7.134	8.236
Ours w/o Interlocutor-aware Module	0.594	7.642	8.436	0.873	7.981	9.421	0.428	6.917	8.201
Ours (CGM, Interlocutor-Aware)	**0.462**	**8.210**	**8.085**	**0.782**	**8.537**	**10.155**	**0.339**	**7.782**	**7.889**

Table 1: *Full Conversation Character Behavior comparison with state-of-the-art methods on the TWH [LDM*19] test set, evaluated separately for three distinct identities (ID 1, ID 2, ID 6). Quantitative evaluation is reported on FGD, BC $\times 10^{-1}$, and Diversity. Ablation results highlight the contributions of motion encoding, listening-aware loss, and interlocutor-aware conditioning.*

	GT Listening	ID 1 Speaking	ID 2 Speaking	ID 6 Speaking
ID 1 (Listening)	51.99	**20.61**	308.36	180.87
ID 2 (Listening)	181.87	299.54	**48.60**	242.44
ID 6 (Listening)	29.71	191.76	264.75	**16.13**

Table 2: *FID values for evaluating personalized listening behavior. Rows correspond to generated Character IDs in the listening role, and columns represent reference sets (ground-truth listening gestures or speaking segments of different Characters). Lower values against GT Listening and the same Character's speaking segments indicate stronger personalization, while higher values against different Characters' speaking segments demonstrate clear differentiation.*

Head Alignment Setting ↑	ID 1	ID 2	ID 6	Audio2Photoreal
GT vs. Interlocutor Audio	0.850	0.850	0.850	0.850
Generated vs. Interlocutor Audio	**0.744**	**0.750**	**0.813**	0.749
Generated vs. Random Audio	0.674	0.647	0.691	0.711

Table 3: *Head alignment scores between* Character *head movements and* Interlocutor *audio, including the Audio2Photoreal baseline. Higher scores indicate better synchronization.*

Ground-Truth Comparisons In the ground-truth trials, participants again selected between two alternatives: one video generated by our system and one representing ground truth. As before, because the task involved exactly two options, the null hypothesis of no preference corresponds to a chance rate of 0.5.

Exact binomial tests indicated that participants selected "Ours" in 46% of cases for Video 1 and 57% for Video 2, neither of which differed significantly from chance (both $p > .49$). When pooling across both videos, participants chose "Ours" in 51.4% of trials (36/70), with no significant deviation from parity ($p = .905$) (Table 5).

To account for repeated measures, we again fitted a logistic regression model with cluster-robust standard errors by participant. The intercept-only model was not significantly different from zero ($\beta = -0.17$, $SE = 0.35$, $z = -0.50$, $p = .62$), confirming that overall preferences did not deviate from chance. Including video as a fixed effect did not improve model fit (likelihood ratio test $p = .338$), and the contrast between the two ground-truth videos was not significant.

Taken together, these analyses suggest that our system is *statistically indistinguishable from ground truth*. Rather than showing a systematic preference for either option, participants viewed the outputs of our system and the ground-truth materials as broadly equivalent. Combined with the DiffuseStyleGesture+ comparison results, this indicates that the system not only outperforms existing alternatives but also achieves parity with ground truth.

6. Conclusion

We presented *CGM*, a unified conversational gesture model that generates Character behaviors conditioned on Interlocutor inputs. By augmenting a speaker-centric backbone with interlocutor-aware cross-attention and training on a unified representation spanning single-speaker and dyadic data, the model captures bidirectional dynamics and fluid role shifts. Quantitatively, it improves the realism, responsiveness, and personalization of Character gestures while preserving performance on the BEATX benchmark. Qualitatively, analyses of Character head–audio alignment indicate rhythmic synchrony with the Interlocutor's prosody. Together, these results demonstrate that extending co-speech gesture generation from monologic settings to full conversational interactions is both feasible and beneficial, with measurable improvements in realism, responsiveness, and personalization.

Limitations Despite the contributions of this work, several limitations remain. First, the current system is restricted to dyadic interactions and does not generalize to multi-person conversations. Real-world dialogues often involve overlapping speech, shifting addressees, and complex group dynamics that cannot be modeled within our two-person setup. Second, the diffusion-based generator used in this study is not yet capable of real-time performance. While suitable for offline synthesis, the latency makes the system impractical for interactive applications. Finally, although the model responds effectively to interlocutor cues, the degree of control remains coarse. Participants cannot explicitly specify nuanced conversational functions such as turn-yielding, disagreement, or stylistic variation in gesture expressiveness. In addition, facial motion is

Method	ID 1 (Wayne)			ID 2 (Scott)			ID 6 (Carla)		
	FGD ↓	BC $\times 10^{-1}$ ↑	Diversity ↑	FGD ↓	BC $\times 10^{-1}$ ↑	Diversity ↑	FGD ↓	BC $\times 10^{-1}$ ↑	Diversity ↑
GT	0.000	6.580	14.141	0.000	8.440	12.638	0.000	1.907	8.637
SynTalker	0.258	6.781	5.466	**0.307**	8.364	10.697	**0.481**	3.565	7.794
Ours w/o Motion Encoding/Decoding	0.356	7.520	3.221	0.527	7.137	4.266	0.768	5.445	5.832
Ours w/o Listening-Aware Loss	0.213	6.402	7.914	0.417	8.063	12.601	0.589	5.911	8.674
Ours w/o Interlocutor-aware Module	0.241	6.612	6.882	0.352	8.201	11.146	0.512	4.982	8.031
Ours (CGM, Interlocutor-Aware)	**0.1903**	**6.796**	**6.101**	0.484	**8.794**	**14.074**	0.621	**6.774**	**9.286**

Table 4: *Character Performance Preservation compared with state-of-the-art methods on the BEATX [LZB*24] test set, evaluated separately for three distinct identities (ID 1, ID 2, ID 6). Quantitative evaluation is reported on FGD, BC $\times 10^{-1}$, and Diversity. Results include ablations that analyze the impact of motion encoding, listening-aware loss, and interlocutor-aware conditioning on speaker-centric performance.*

Condition	Video	Ours (n)	Not Ours (n)	Prop. Ours	Binomial p
DiffuseStyleGesture+	Video 1	21	14	0.60	.311
	Video 2	25	10	0.71	**.017**
	Video 3	19	16	0.54	.736
	Video 4	22	13	0.63	.175
	Pooled	87	53	0.62	**.005**
Ground Truth	Video 1	16	19	0.46	.736
	Video 2	20	15	0.57	.500
	Pooled	36	34	0.51	.905

Mixed-effects logistic regression (cluster-robust by participant).
DiffuseStyleGesture+: intercept $\beta = 0.50$, $SE = 0.23$, $z = 2.14$, $p = .033$ (significant preference for "Ours").
Ground Truth: intercept $\beta = -0.17$, $SE = 0.35$, $z = -0.50$, $p = .62$ (no significant difference).

Table 5: *Summary of preference tests comparing "Ours" with DiffuseStyleGesture+ and ground truth. Chance level is 0.5 because participants always chose between two alternatives.*

not modeled in the current framework, as the TWH dataset does not provide facial expression data.

Future Work A natural next step is to extend the model beyond dyadic interactions toward multi-person conversations. This would require mechanisms to represent conversational roles, dynamically track addressees, and manage overlapping speech across multiple participants. Developing such capabilities would enable the system to handle the richness of group dialogue.

Another direction is to improve the controllability of generated behaviors. Future models could expose interpretable control parameters or semantic tokens that allow users to guide gestures according to conversational functions, such as signaling agreement, yielding the floor, or expressing contrast. This would provide finer-grained and more intentional steering of the generated motions, increasing the model's utility in interactive or creative applications.

In addition, incorporating the model into immersive VR settings and enriching it with facial expression synthesis on top of body motion would further enhance its realism and applicability.

Acknowledgments

This work was partially supported by the Horizon 2020 FET Proactive project GuestXR (#101017884), the Israel Science Foundation (Grant no. 1427/25), and the Joint NSFC-ISF Research (Grant no. 3077/23).

References

[ABL23] ALBUQUERQUE, ISADORA, BARBOSA, FELIPE, and LEITE, FERNANDO. "Head movements as markers of conversational alignment: A systematic review". *Frontiers in Psychology* 14 (2023), 1112345 8.

[AMMS19] AHUJA, CHAITANYA, MA, SHIH-YUN, MORENCY, LOUIS-PHILIPPE, and SHEIKH, YASER. "To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations". *Proceedings of the 2019 International Conference on Multimodal Interaction (ICMI)*. 2019, 74–84 2.

[ANBH23] ALEXANDERSON, SIMON, NAGY, RAJMUND, BESKOW, JONAS, and HENTER, GUSTAV EJE. "Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models". *ACM Trans. Graph.* 42.4 (2023), 44:1–44:20. DOI: 10.1145/3592458 2.

[AZL23] AO, TENGLONG, ZHANG, ZEYI, and LIU, LIBIN. "GestureDiffuCLIP: Gesture diffusion model with CLIP latents". *ACM Transactions on Graphics* (2023). DOI: 10.1145/3592097 2.

[BGJM17] BOJANOWSKI, PIOTR, GRAVE, EDOUARD, JOULIN, ARMAND, and MIKOLOV, TOMAS. "Enriching word vectors with subword information". *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146 2, 4.

[BJ08] BRUDERLIN, ALAIN and JUN, SUNG YONG. "Conversational gestures and head movements synthesis for embodied agents". *Proceedings of the 2008 ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*. 2008, 139–147 8.

[CDA*24] CHHATRE, KIRAN, DANĚČEK, RADEK, ATHANASIOU, NIKOS, et al. "AMUSE: Emotional Speech-driven 3D Body Animation via Disentangled Latent Diffusion". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, 1942–1953. URL: https://amuse.is.tue.mpg.de 2.

[CLD*24] CHEN, BOHONG, LI, YUMENG, DING, YAO-XIANG, et al. "Enabling synergistic full-body control in prompt-based co-speech motion generation". *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024, 6774–6783 2, 3, 7, 8.

[CLW*24] CHEN, JUNMING, LIU, YUNFEI, WANG, JIANAN, et al. "DiffSHEG: A Diffusion-Based Approach for Real-Time Speech-driven Holistic 3D Expression and Gesture Generation". *CVPR*. 2024 2.

[DCT*23] DANĚČEK, RADEK, CHHATRE, KIRAN, TRIPATHI, SHASHANK, et al. "Emotional speech-driven animation with content-emotion disentanglement". *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. 2023. DOI: 10.1145/3610548.3618183 2.

[DMAB23] DEICHLER, ANNA, MEHTA, SHIVAM, ALEXANDERSON, SIMON, and BESKOW, JONAS. "Diffusion-Based Co-Speech Gesture Generation Using Joint Text and Audio Representation". *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '23)*. 2023. DOI: 10.1145/3577190.3616117 3.

[DN21] DHARIWAL, PRAFULLA and NICHOL, ALEX. "Diffusion models beat GANs on image synthesis". *Advances in Neural Information Processing Systems (NeurIPS)*. 2021 2.

[Gir15] GIRSHICK, ROSS. "Fast R-CNN". *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015, 1440–1448 5.

[GPSS08] GILLIES, MICHAEL, PAN, XUESONG, SLATER, MEL, and SHAWE-TAYLOR, JOHN. "Responsive listening behavior". *Computer Animation and Virtual Worlds* 19.5 (2008), 579–589 2.

[Hey06] HEYLEN, DIRK. "Head movements and speech in conversations: Towards models of their interaction". *International Conference on Intelligent Virtual Agents*. Springer, 2006, 241–247 8.

[HRU*17] HEUSEL, MARTIN, RAMSAUER, HUBERT, UNTERTHINER, THOMAS, et al. "GANs trained by a two time-scale update rule converge to a Nash equilibrium". *Advances in Neural Information Processing Systems (NeurIPS)*. 2017 8.

[HZRS16] HE, KAIMING, ZHANG, XIANGYU, REN, SHAOQING, and SUN, JIAN. "Deep residual learning for image recognition". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, 770–778 4.

[JKHB20] JONELL, PATRIK, KUCHERENKO, TARAS, HENTER, GUSTAV EJE, and BESKOW, JONAS. "Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings". *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA)*. 2020, 1–8 2.

[JSCS19] JOO, HANBYUL, SIMON, TOMAS, CIKARA, MINA, and SHEIKH, YASER. "Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, 10873–10883 2.

[KNY*23] KUCHERENKO, TARAS, NAGY, RICHÁRD, YOON, YOUNGWOO, et al. "The GENEA challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings". *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI)*. 2023, 792–801 3.

[LDM*19] LEE, GWAN, DENG, ZIJIAN, MA, SHIH-YUN, et al. "Talking with hands 16.2M: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis". *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, 763–772 7, 10.

[LWF*23] LIU, JIAHAO, WANG, XURAN, FU, XIAOYU, et al. "MFR-Net: Multi-faceted responsive listening head generation via denoising diffusion model". *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. 2023, 6734–6743 2.

[LYG*22] LI, SIYAO, YU, WEIJIANG, GU, TIANPEI, et al. "Bailando: 3D dance generation via actor-critic GPT with choreographic memory". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 8.

[LZB*24] LIU, HAIYANG, ZHU, ZIHAO, BECHERINI, GIORGIO, et al. "EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Expressive Masked Audio Gesture Modeling". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, 1144–1154 2, 3, 7, 8, 11.

[LZI*22] LIU, HAIYANG, ZHU, ZIHAO, IWAMOTO, NAOYA, et al. "BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis". *European Conference on Computer Vision (ECCV)*. 2022 8.

[McC00] MCCLAVE, EVELYN. "Linguistic functions of head movements in the context of speech". *Journal of Pragmatics* 32.7 (2000), 855–878 8.

[MDH*24] MUGHAL, MUHAMMAD HAMZA, DABRAL, RISHABH, HABIBIE, IKHSANUL, et al. "ConvoFusion: Multi-Modal Conversational Diffusion for Co-Speech Gesture Synthesis". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, 1388–1398 3.

[NJH*22] NG, ESHRAT ARJMAND, JOO, HANBYUL, HU, LIWEN, et al. "Learning to listen: Modeling non-deterministic dyadic facial motion". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, 20395–20405 2.

[NRB*24] NG, EVONNE, ROMERO, JAVIER, BAGAUTDINOV, TIMUR, et al. "From Audio to Photoreal Embodiment: Synthesizing Humans in Conversations". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, 1144–1154 3, 8.

[NWX*24] NG, JONATHAN, WANG, JIALIN, XU, RUIQI, et al. "GestureDialogue: Co-speech gesture synthesis for multi-party conversations". *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*. 2024 1.

[PBC*22] PALMERO, CARMEN, BARQUERO, GUILLERMO, CARLOS JUNIOR, JOÃO, et al. "ChaLearn LAP challenges on self-reported personality recognition and non-verbal behavior forecasting during social dyadic interactions: Dataset, design, and results". *Proceedings of the 2nd Workshop on Understanding Social Behavior in Dyadic and Small Group Interactions*. Vol. 173. Proceedings of Machine Learning Research. PMLR, 2022, 4–52 3.

[PCG*19] PAVLAKOS, GEORGIOS, CHOUTAS, VASILEIOS, GHORBANI, NIMA, et al. "Expressive body capture: 3D hands, face, and body from a single image". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 7.

[QWZ*25] QI, XINGQUN, WANG, YATIAN, ZHANG, HENGYUAN, et al. "Co^{3}Gesture: Towards Coherent Concurrent Co-speech 3D Gesture Generation with Interactive Diffusion". *International Conference on Learning Representations (ICLR)*. 2025 3.

[SGF*22] SHUAI, QING, GENG, CHEN, FANG, QI, et al. "Novel view synthesis of human interactions from sparse multi-view videos". *ACM SIGGRAPH 2022 Conference Proceedings*. ACM, 2022 7.

[SSL*23] SONG, SHURAN, SPITALE, MICHELE, LUO, CHENGZHU, et al. "REACT2023: The first multi-modal multiple appropriate facial reaction generation challenge". *arXiv preprint arXiv:2306.06583* (2023) 2.

[SYJ*23] SONG, LI, YIN, GUOXIAN, JIN, ZHIWEI, et al. "Emotional listener portrait: Realistic listener motion simulation in conversation". *arXiv preprint arXiv:2310.00068* (2023) 2.

[TC22] TUYEN, NGUYEN THI VAN and CELIKTUTAN, OZGUR. "Agree or disagree? Generating body gestures from affective contextual cues during dyadic interactions". *Proceedings of the 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2022, 1542–1547 3.

[TC23] TUYEN, NGUYEN THI VAN and CELIKTUTAN, OZGUR. "It takes two, not one: Context-aware nonverbal behaviour generation in dyadic interactions". *Advanced Robotics* 37.24 (2023), 1552–1565 3.

[TC25] TUYEN, NGUYEN THI VAN and CELIKTUTAN, OZGUR. "It Takes Two: Context-Aware Nonverbal Behaviour Generation for Human-Human Interaction". *arXiv preprint arXiv:2510.10206* (2025) 3.

[TGH*22] TEVET, GUY, GORDON, BRIAN, HERTZ, AMIR, et al. "MotionCLIP: Exposing human motion generation to CLIP space". *European Conference on Computer Vision (ECCV)*. Springer, 2022, 358–374 2.

[TRG*23] TEVET, GUY, RAAB, SIGAL, GORDON, BRIAN, et al. "Human motion diffusion model". *International Conference on Learning Representations (ICLR)*. 2023 5.

[vdOVK17] Van den OORD, AARON, VINYALS, ORIOL, and KAVUKCUOGLU, KORAY. "Neural discrete representation learning". *Advances in Neural Information Processing Systems (NeurIPS)*. 2017 3.

[Wig19] WIGHTMAN, ROSS. *PyTorch Image Models*. `https://github.com/rwightman/pytorch-image-models`. Accessed: 2025-09-26. 2019 4.

[XFWW25] XUE, HAIWEI, FAN, YANBO, WANG, XUAN, and WU, ZHIYONG. "Echo: Enhancing Conversational Behavior Generation via Hierarchical Semantic Comprehension with Large Language Models". *Proceedings of SIGGRAPH Asia 2025*. 2025 3.

[XLZ*25] XU, SIRUI, LI, DONGTING, ZHANG, YUCHENG, et al. "InterAct: Advancing Large-Scale Versatile 3D Human-Object Interaction Generation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025 3.

[YWL*23] YANG, SICHENG, WU, ZHIYONG, LI, MINGLEI, et al. "DiffuseStyleGesture: Stylized audio-driven co-speech gesture generation with diffusion models". *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*. 2023, 5860–5868. DOI: 10.24963/ijcai.2023/650 2, 8.

[YWW*23] YANG, SICHENG, WANG, ZILIN, WU, ZHIYONG, et al. "UnifiedGesture: A unified gesture synthesis model for multiple skeletons". *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. 2023, 1033–1044. DOI: 10.1145/3581783.3612503 2.

[YXZ*23] YANG, SICHENG, XUE, HAIWEI, ZHANG, ZHENSONG, et al. "The DiffuseStyleGesture+ entry to the GENEA Challenge 2023". *Proceedings of the 25th International Conference on Multimodal Interaction*. ICMI '23. Paris, France: Association for Computing Machinery, 2023, 779–785. ISBN: 9798400700552. DOI: 10.1145/3577190.3616114. URL: https://doi.org/10.1145/3577190.3616114 1, 8, 9.

[ZAZ*24] ZHANG, ZEYI, AO, TENGLONG, ZHANG, YUYAO, et al. "Semantic gesticulator: Semantics-aware co-speech gesture synthesis". *arXiv preprint arXiv:2405.09814* (2024) 2.

[ZBZ*22] ZHOU, MINGYUAN, BAI, YUTONG, ZHANG, WENHAN, et al. "Responsive listening head generation: A benchmark dataset and baseline". *European Conference on Computer Vision (ECCV)*. Springer, 2022, 124–142 2.

[ZCL*24] ZHANG, YIFAN, CHEN, YUWEI, LING, HONGWEI, et al. "Conversational gesture synthesis with neural motion fields". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024 1.

[ZFT*24] ZHANG, MINGHAO, FERREIRA, RAFAEL, TALMAN, ARTHUR, et al. "Gestures are in the eye of the beholder: Contrastive motion learning from listener's gaze". *International Conference on Learning Representations (ICLR)*. 2024 3.

[ZLL*23] ZHU, LINGTING, LIU, XIAN, LIU, XUANYU, et al. "Taming diffusion models for audio-driven co-speech gesture generation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, 10544–10553 2.

[ZZC*23] ZHANG, JIANRONG, ZHANG, YANGSONG, CUN, XIAODONG, et al. "T2M-GPT: Generating human motion from textual descriptions with discrete representations". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023 3.