This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DyaDiT: A Multi-Modal Diffusion Transformer for Socially Favorable Dyadic Gesture Generation

Yichen Peng[1,2] Jyun-Ting Song[2] Siyeol Jung[2,3] Ruofan Liu[1] Haiyang Liu[4,5]
Xuangeng Chu[4,5] Ruicong Liu[4,5] Erwin Wu[1,4] Hideki Koike[1] Kris Kitani[2]

[1]Institute of Science Tokyo [2]Carnegie Mellon Unversity
[3]UNIST [4]Shanda AI Research Tokyo [5]The University of Tokyo

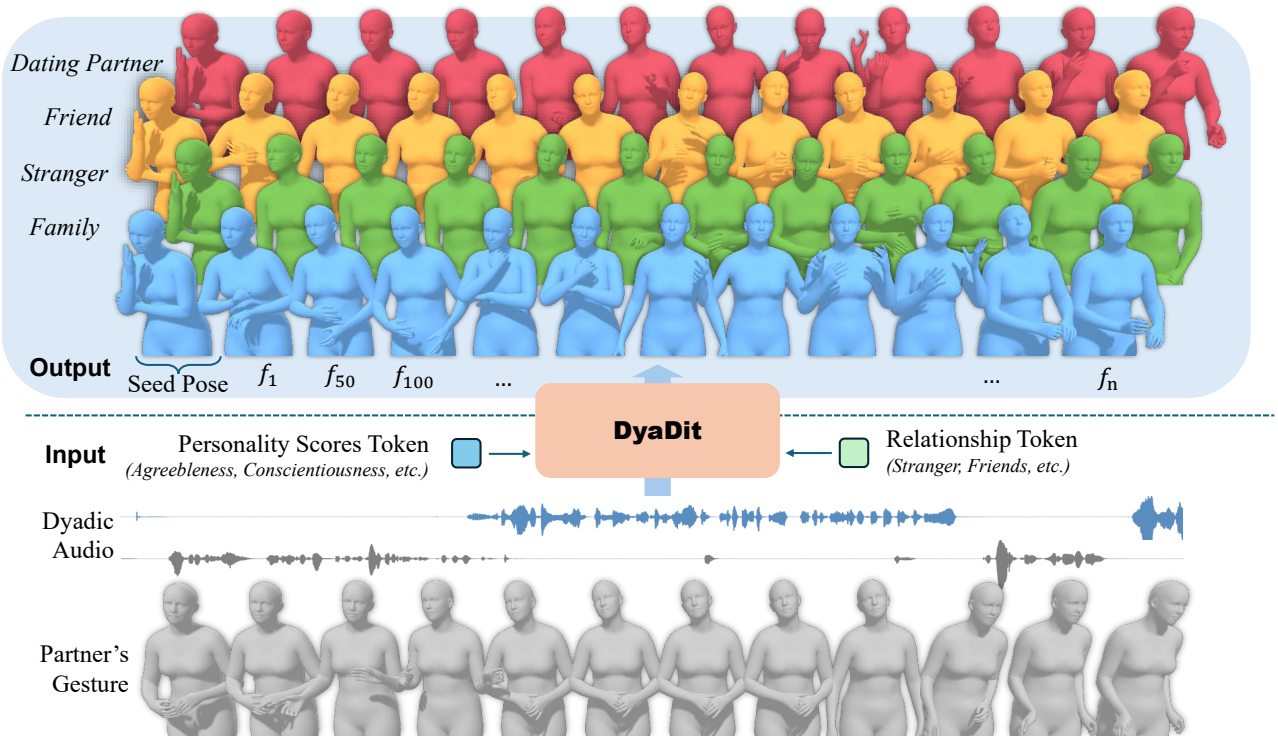

Figure 1. **DyaDiT** generates socially aware conversational gestures from dyadic audio, conditioned on social factors such as relationship and personality traits, achieving natural and contextually appropriate reactions that outperform prior methods in both quantitative and user evaluations.

Abstract

Generating realistic conversational gestures are essential for achieving natural, socially engaging interactions with digital humans. However, existing methods typically map a single audio stream to a single speaker's motion, without considering social context or modeling the mutual dynamics between two people engaging in conversation. We present DyaDiT, a multi-modal diffusion transformer that generates contextually appropriate human motion from dyadic audio signals. Trained on Seamless Interaction Dataset [1], DyaDiT takes dyadic audio with optional social-context tokens to produce context-appropriate motion. It fuses information from both speakers to capture interaction dynamics, uses a motion dictionary to encode motion priors, and can optionally utilize the conversational

partner's gestures to produce more responsive motion. We evaluate DyaDiT on standard motion generation metrics and conduct quantitative user studies, demonstrating that it not only surpasses existing methods on objective metrics but is also strongly preferred by users, highlighting its robustness and socially favorable motion generation. Code and models can be found at https://puckikk1202.github.io/dyadit_hp/.

1. Introduction

Building synthetic agents (also known as digital humans, AI agents, avatars or androids) that can interact naturally with people, is essential for the future of human–computer interfaces. Recent language models such as GPT-4.5 [34] and LLaMA-3.1 [13] already demonstrate impressive conversational ability, and many users even feel as if they are speaking with another person [16]. However, currently this illusion still remains confined to a text window. In reality, human interaction involves more than just spoken words. People gesture, respond to each other, and express subtle social cues through body motion. In order for humans to truly feel that a synthetic agent is interactive, the agent must accompany its speech with gestures that evolve naturally with the conversation.

However, generating such gestures is challenging because they are shaped not only by speech content, but also by rich social context and the dynamics of interaction. Human gestures depend on factors such as personality, the relationship between speakers, and their conversational roles, all of which influence how people move, respond, and coordinate with one another. Yet most existing gesture generation models do not explicitly model these social factors [27, 28, 44], causing the generated motions to appear generic or unnatural. The problem becomes even harder in dyadic conversations, where interaction unfolds dynamically. Two people may speak simultaneously, interrupt one another, or rapidly alternate between speaking and listening [8, 30]. Capturing such moment-to-moment coordination requires the model to distinguish speaker-specific and listener-specific cues, but most existing dyadic gesture generation methods do not explicitly disentangle the two input audio streams [31, 33], making it difficult for them to generate gestures that faithfully reflect the underlying interaction.

To tackle these limitations, we introduce DyaDiT, a diffusion based transformer that generates socially aware gestures from dyadic audio. Unlike previous work [10, 23, 24, 37] that focuses only on the alignment between audio and generated motion, DyaDiT conditions its generation on explicit social cues such as relationship and personality. Furthermore, in order to address the strong entanglement of two overlapping audio streams during conversation, we propose Orthogonalization Cross Attention (ORCA), a simple yet effective module that disambiguates the two audio streams, resulting in a cleaner audio representation for better gesture generation. Lastly, since human motion in a dyadic setting is often affected by the partner's gestures, DyaDiT can optionally take the partner's movements as an additional input, allowing the model to generate gestures that are more coordinated, responsive, and natural.

Our experiments show that DyaDiT consistently outperforms existing dyadic gesture generation methods across standard quantitative metrics, achieving clear improvements in gesture quality and distribution alignment. In addition, we conduct extensive user studies to assess human perceptual preference. The results show that participants strongly favor gestures generated by DyaDiT, indicating that our generated motion appears more socially aware and better suited for real conversational settings.

In summary, our main contributions are as follows:

- We propose DyaDiT, a diffusion transformer (DiT) that generates social context aware gestures in dyadic conversations.
- We introduce an Orthogonalization Cross Attention (ORCA) module that reduces interference between two individual's audio streams, enabling a cleaner audio representation for better gesture generation in dyadic conversation.
- Through extensive quantitative evaluations and user studies, we show that DyaDiT consistently outperforms existing methods on standard metrics and is preferred by users in terms of perceived realism and social consistency.

2. Related Work

2.1. Co-Speech Gesture Generation

Co-speech gesture generation focuses on synthesizing body movements aligned with a single speaker's speech. Early approaches formulate gesture generation as a translation problem from multimodal speech cues, combining text, audio, and speaker identity through recurrent or adversarial models, such as the trimodal framework of Yoon et al. [50], which treated gesture generation as a translation problem, combining speech transcripts, audio, and speaker identity through an adversarial recurrent framework. Liu et al. [23] introduced the BEAT dataset and CaMN, a cascaded multimodal adversarial network capable of generating synchronized body and hand gestures. Other works, such as EMAGE [24], TalkSHOW [49], MECo [5], etc. [12, 22, 25, 35, 48] enhanced realism by disentangling rhythmic and semantic motion features or integrating discrete latent spaces via VQ-VAEs.

Although these advances are notable, most of them focuses on single-speaker co-speech gestures and overlooks the interactive nature of human communication. More complex settings such as dyadic gesture generation, which re-

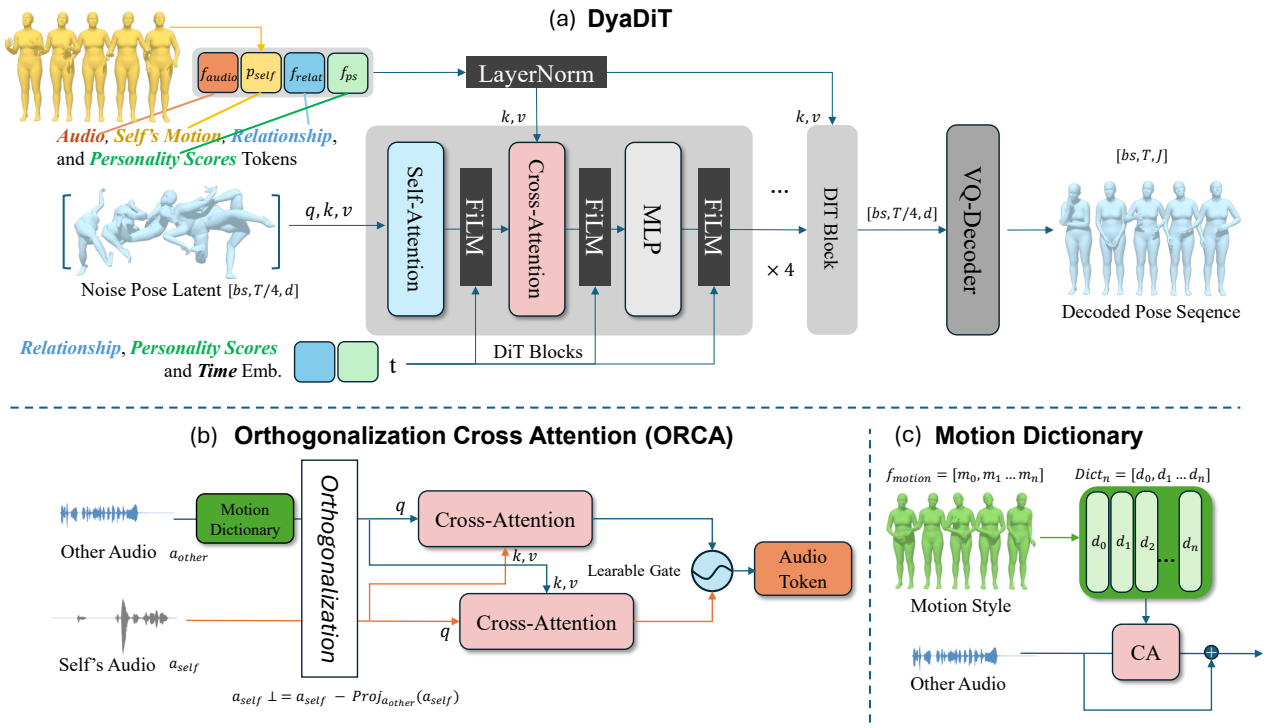

Figure 2. **Overview of DyaDiT.** DyaDiT conditions on multiple input modalities, including audio, partner motion, relationship type, and personality scores. It employs an Audio Orthogonalization Cross Attention (ORCA) module to obtain cleaner audio representations and a motion dictionary to guide style aware gesture generation.

quire modeling both participants' behaviors and their interpersonal dynamics, remain largely unexplored.

2.2. Dyadic Gesture and Reaction Generation

Unlike co-speech gesture synthesis, dyadic gesture generation must model the coordinated behavior between two participants, including interpersonal timing, mutual attention, and responsiveness. Most prior works in this domain stem from facial reaction generation (FRG), where the goal is to predict non-deterministic listener facial responses to a speaker's behavior [17, 26–28, 32, 41, 51]. Beyond FRG, related efforts have also been made in broader interactive motion domains, such as couple dance modeling [29] and two-person motion dataset & synthesis [11, 18, 21], both of which emphasize the role of interpersonal synchrony and response modeling. Although FRG provides a basis for modeling two-person interactions, it primarily focuses only on facial or head movements rather than full-body gestures. Due to this limitation, several recent works have begun exploring body gesture generation in dyadic settings. A few efforts such as Audio2Photoreal [33], ConvoFusion [31], and TAG2G [9] extend single-speaker frameworks to two-party settings. Yet, these models often either (1) overlook the social context between the two individuals or (2) treat dyadic audio as a single blended signal without explicitly modeling cross-person dynamics, which often leads to ambiguity in the roles and interaction patterns presented to the model. These limitations highlight the need for explicit feature disentanglement and better support for socially contextual reasoning.

2.3. Diffusion-based Gesture Generation

Diffusion models have emerged as powerful generative tools for human motion due to their ability to model multimodal and many-to-many distributions. Several works have adapted diffusion frameworks originally designed for text-conditioned motion generation, such as MotionDiffuse [53], FineMoGen [52], and MDM [43], etc. [3, 6, 38, 42], to the speech–gesture domain. Alexanderson et al. [2] reformulated DiffWave for co-speech gestures, while Zhu et al. [56] proposed DiffGesture, integrating noisy gesture sequences with contextual embeddings for temporal modeling. DiffuseStyleGesture+ [47] further introduced conditioning on audio, text, style, and seed gestures, leveraging transformer-based denoising with attention control. Other works, such as UnifiedGesture [46], LivelySpeaker [54], and AMUSE [7], emphasized semantic and rhythmic consistency, disentangling emotional and stylistic latent factors. Considering the success of diffusion-based approaches in gesture generation, and the inherently nondeterministic na-

ture of dyadic conversation, we build our method upon a diffusion model to better capture the variability and dynamics present in two-person interactions.

3. DyaDiT

We introduce **DyaDiT**, a diffusion transformer (DiT) with multi-modal input for dyadic gesture generation. Given the conversational audio streams from two individuals (denoted as *self* and *other*), our objective is to generate plausible upper-body gestures for the *other* speaker in accordance with the conversational context. To further enhance the realism of the generated gestures, DyaDiT optionally incorporates the relationship, personality traits, and gesture sequence of *self* as auxiliary conditioning signals during inference. We train our model on a subset of the Seamless Interaction dataset [1], with dataset specifications and preprocessing procedures detailed in Section 3.1. An overview of the full architecture is shown in Fig. 2. Details of the Audio Orthogonalization Cross Attention (ORCA) module are provided in Section 3.3, while the motion dictionary and motion tokenizer are presented in Sections 3.4 and 3.5, respectively.

3.1. Seamless Interaction Dataset

The Seamless Interaction dataset [1] is a large-scale corpus of dyadic conversation containing synchronized audio, full-body motion, and facial expressions. In this work, we use a curated subset of 3,000 clips (approximately 182 hours) from the dataset's *naturalistic* scenario collection, which contains rich and spontaneous dyadic conversations suited for modeling conversational gestures. We focus on generating upper-body gestures, represented as $g \in \mathbb{R}^{T \times J \times 6}$ using the 6D rotation representation [55], where J denotes the number of upper-body joints. In addition to motion data, we incorporate two high level social annotations provided in the dataset: a relationship type label $f_{rs} \in \{0,1\}^4$ indicating whether the speakers are friends, strangers, family members, or dating partners; and a personality score vector $f_{ps} \in \mathbb{R}^5$ that quantifies five major personality traits, which are extraversion, agreeableness, conscientiousness, neuroticism, and openness. For audio inputs, we extract the dyadic speech signals from both individuals and process them with a pretrained Wav2Vec2 encoder [4] to obtain audio feature embeddings for conditioning in our model.

3.2. DiT Backbone

Our model adopts a diffusion transformer (DiT) backbone that follows the Denoising Diffusion Probabilistic Model (DDPM) [15] framework. The network takes a noisy latent pose $\mathbf{x}_t$ and predicts the added Gaussian noise according to time step t, $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})$ conditioned on multiple contextual inputs $\mathbf{c}$, where $\mathbf{c} = (ORCA(a_{\text{self}}, a_{\text{other}}), p_{\text{self}}, f_{\text{relat}}, f_{\text{ps}})$ including audio, partner's motion, relationship type, and personality scores. The training objective follows the standard ϵ-prediction loss:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{x}_0, t, \boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})\|_2^2 \right], \tag{1}$$

where $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$.

As shown in Figure 2(a), each DiT block consists of a self-attention layer for modeling temporal dependencies within the latent pose sequence and a cross-attention layer for integrating contextual information from multimodal cues. In addition, the relationship and personality embeddings are injected through both FiLM-based [36] modulation and cross-attention, enabling DyaDiT to jointly capture social attributes and individual expressive styles in gesture generation.

3.3. Audio Orthogonalization Cross Attention (ORCA)

To effectively capture conversational dynamics between two speakers, we introduce a **orthogonalization cross attention** module for audio fusion (Figure 2(b)). Given the audio features from both speakers a_{self} and a_{other}, encoded by Wav2Vec2, our goal is to separate redundant factors while aligning complementary information into a joint representation.

We first apply an *orthogonalization* process to filter out redundant components between the two audio streams:

$$a_{\text{self}}^{\perp} = a_{\text{self}} - \text{Proj}_{a_{\text{other}}}(a_{\text{self}}), \tag{2}$$

where $\text{Proj}_{a_{\text{other}}}(\cdot)$ denotes the projection of the self audio feature a_{self} onto the subspace of the other's audio feature, implemented as a lightweight MLP:

$$\phi(\mathbf{x}) = W_2\sigma(W_1\mathbf{x} + b_1) + b_2. \tag{3}$$

This operation enforces complementary conditioning and reduces correlated information across dyadic audio.

Subsequently, we employ two symmetric cross-attention modules to achieve bidirectional information exchange. The first module uses a_{other} as query and attends to $a_{\text{self}}^{\perp}$, capturing the speaker's response to the partner's utterance. Conversely, the second module takes $a_{\text{self}}^{\perp}$ as query and attends to a_{other}, modeling the listener's reactive cues. The outputs of the two cross-attention streams are then adaptively fused through a learnable gating mechanism:

$$f_{\text{audio}} = \sigma(\mathbf{W}_g) \cdot h_{\text{self}\to\text{other}} + (1 - \sigma(\mathbf{W}_g)) \cdot h_{\text{other}\to\text{self}}, \tag{4}$$

where $\sigma(\cdot)$ is a sigmoid function and $\mathbf{W}_g$ is a learnable gate parameter.

The resulting fused token f_{audio} is used as the final audio conditioning input to the DyaDiT, providing a plausible acoustic embedding that reflects both interlocutors' vocal behaviors. As shown in Figure 3, with ORCA, DyaDiT have the capability to generate either speaker or listener gestures.

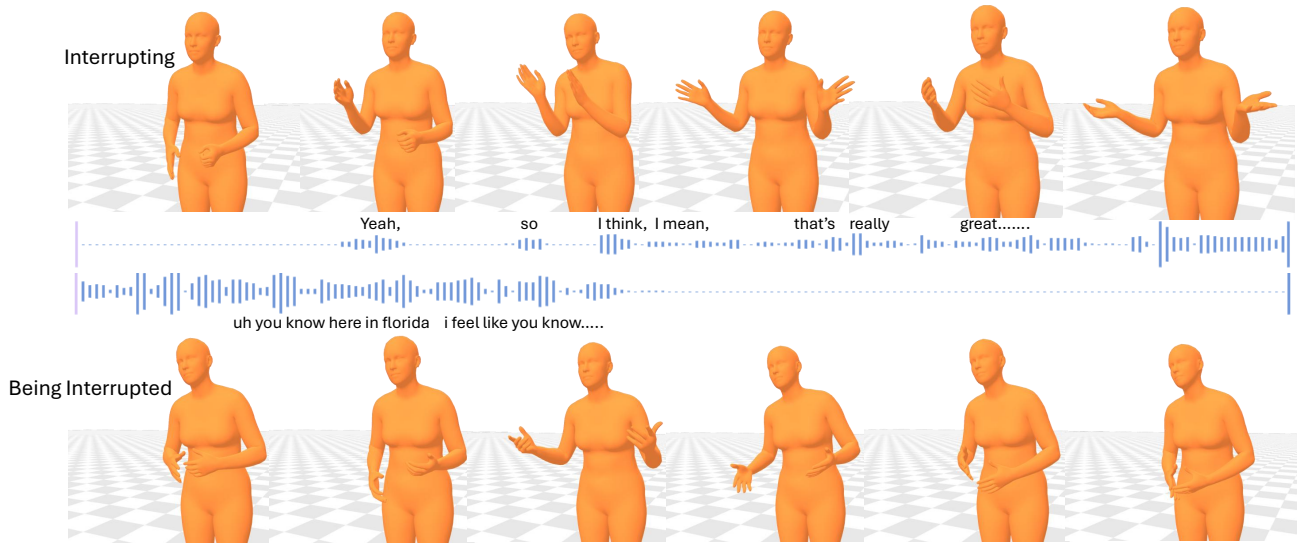

Figure 3. **ORCA** reduces ambiguity between the two audio streams, allowing **DyaDiT** to generate realistic motion even when one person interrupts the other during the conversation. The example demonstrates the generated motions adjusts naturally as the conversation shifts.

3.4. Motion Dictionary (MD)

Inspired by prior work LIA [45] which learns a set of orthogonal motion-directions by enforcing orthonormality at initialization, we similarly introduce a learnable orthogonal motion dictionary to modulate the partner's audio feature a_{other} according to the current style motion. The motion dictionary is a module designed to incorporate motion style conditioning when user requires motion style control. As shown in Figure 2(c), it consists of a set of learnable motion bases $\{d_0, d_1, \ldots, d_n\}$ that encode representative gesture primitives. During training, we include ground truth motion style features $f_{\text{motion}} = [m_0, m_1, \ldots, m_n]$ to guide the model in learning style-aware correspondences between audio cues and motion patterns.

Given the other-speaker audio feature a_{other}, the dictionary is integrated through a cross-attention (CA) operation followed by a weighted combination with the motion bases:

$$a'_{\text{other}} = \text{CA}(a_{\text{other}}, \sum_{k=0}^{n} m_k d_k) + a_{\text{other}}, \tag{5}$$

where $\text{Dict} = [d_0, d_1, \ldots, d_n]$ and m_k denotes the style-dependent modulation weight derived from f_{motion}.

The Motion Dictionary is jointly trained with DyaDiT without orthogonalization, as strict phase alignment is unnecessary for gestures and degrades learning. At inference, the motion dictionary can be optionally activated. With classifier-free guidance (CFG) [14], the style conditioning can be strengthened to generate gestures with distinct stylistic patterns, or completely dropped to produce an unconditional, style-agnostic motion driven purely by the learned audio–motion prior.

3.5. Motion Tokenizer

Instead of directly training our DiT model using the raw gesture sequences $x'_0 \in \mathbb{R}^{T \times J}$, we first train a VQ-VAE to discretize the continuous motion space into compact latent representations. Our implementation follows the residual vector quantization framework introduced in prior works [19], where multiple codebooks are cascaded to progressively refine quantization errors. Specifically, we adopt a residual VQ-VAE with a residual length of 4, utilizing four hierarchical codebooks to capture progressively finer-grained motion details. In addition, to reduce temporal redundancy, we downsample the input gesture sequence by a factor of 4 using 1D convolution layers in the encoder, resulting in one latent token for every four frames. During training of DyaDiT, the VQ-VAE encoder compresses gesture sequences into a sequence of quantized tokens, $x'_t \in \mathbb{R}^{T/4 \times d}$, where $d = 64$ in our experiment. At inference time, the VQ-VAE decoder then reconstructs continuous motions from the generated quantized embeddings. Note that since our diffusion model operates directly in the latent space, the encoder is not used during inference, and the decoder is not used during training. This hierarchical and temporally compact quantization allows the diffusion model to capture long-range dependencies while maintaining high-fidelity motion reconstruction.

Method	BC	FD		Diversity	
	($\times 10^{-1}$)	Sta. ↓	Kine. ↓	Sta. ↑	Kine. ↑
GT	-	-	-	28.42	1.97
GT Random	-	14.94	3.74	33.85	2.05
ConvoFusion	-	9.22	1.74	18.33	1.10
Audio2PhotoReal	-	8.77	1.84	19.35	1.05
DyaDiT (w/o ORCA)	7.25	7.32	1.79	23.57	1.24
DyaDiT (CrossAttn)	7.10	7.82	1.91	18.87	1.13
DyaDiT (w/o MD)	7.37	6.88	1.75	18.34	1.29
DyaDit (MD contin)	7.39	6.69	1.72	21.47	1.33
DyaDiT (w/o self)	7.61	6.64	1.48	26.43	1.25
DyaDiT (Uncond)	7.63	7.40	1.63	21.65	1.16
DyaDiT (Random)	7.64	8.24	1.53	21.94	**1.43**
DyaDiT	**7.71**	**6.40**	**1.37**	**27.46**	1.38

Table 1. Quantitative comparison of DyaDiT and baselines in terms of Fréchet Distance (FD) and Diversity. Lower FD indicates higher realism, and higher diversity values indicate more varied motion generation.

4. Experiments

We evaluate the effectiveness of DyaDiT for dyadic gesture generation through both quantitative and qualitative experiments. For the quantitative evaluation, we compute the Fréchet Distance (FD) and Diversity metrics to compare DyaDiT with two existing dyadic gesture generation models, ConvoFusion [31] and Audio2PhotoReal [33]. In addition, we report scores with respect to the ground truth gestures, which serve as a reference for the underlying motion distribution in the dataset. For the qualitative evaluation, we conduct a user preference study to evaluate both the overall perceived quality of the generated gestures and their alignment with the social relationship and personality characteristics.

4.1. Evaluation Metrics

Following prior works on gesture generation [2, 3, 33], we adopt both distribution and diversity metrics for quantitative evaluation. Specifically, we report the *Fréchet Distance* (FD) to measure the realism of generated motions, *Diversity* [20] to quantify motion variability across samples For reference, we also compute all metrics on ground truth samples from the Seamless Interaction dataset[1] to provide a baseline for comparison. Below, we provide the definitions of the quantitative metrics used in our evaluation.

- **BC** measures beat consistency between generated gestures and audio, reflecting how well motion accents are synchronized with speech beats.
- **FD (Static)** measures the pose realism for individual frames in the pose space $\mathbb{R}^{3\times j}$, where j=43 (including finger joints).
- **FD (Kinetic)** evaluates the realism of temporal dynamics by computing the FD over gesture velocities in $\mathbb{R}^{T\times 3\times j}$, where T=300 frames in our experiments.
- **Diversity (Static)** quantifies the variation of single-frame poses by computing the average mean squared error (MSE) between clips.
- **Diversity (Kinetic)** measures temporal motion diversity by averaging the pairwise velocity differences across generated gesture sequences.

4.2. Baseline Methods

We compare DyaDiT with two representative dyadic gesture generation baselines: **ConvoFusion** [31], a diffusion-based multimodal fusion model with stacked cross-attention layers, and **Audio2PhotoReal** [33], which predicts gesture keyframes from audio and refines them with a diffusion transformer. For a fair comparison, both models are adapted to our gesture representation, retrained on the same training subset of the Seamless Interaction Dataset [1], and conditioned on the same seed pose. During inference, both methods use a DDIM scheduler with 50 denoising steps (CFG = 2) to generate 300-frame (10 s) gesture sequences.

4.3. Quantitative Results

The quantitative comparison between **DyaDiT** and baseline models is shown in Table 1. For reference, **GT** diversity reflects the upper bound of motion variation in the dataset, while **GT Random** *FD* provides a lower-bound reference for cross-sample mismatch. Overall, **DyaDiT** achieves the best overall performance while maintaining strong motion diversity. The ablation results further confirm the contributions of ORCA, the discrete Motion Dictionary, and self-conditioning, each of which improves generation quality and interaction modeling. These results demonstrate the effectiveness of the proposed social context-aware DiT framework over prior baselines such as ConvoFusion [31] and Audio2PhotoReal [20].

4.4. Ablation Studies

We further conduct several ablation studies to analyze how different components and conditioning signals contribute to DyaDiT's performance. Below, we explicitly explain the difference between each variant.

- **DyaDiT (w/o ORCA)** removes the ORCA module and directly concatenates the raw dyadic audio features.
- **DyaDiT (CrossAttn)** replaces the ORCA module with a standard cross-attention for combining *self* and *other* audio features.
- **DyaDiT (w/o MD)** removes the motion dictionary.
- **DyaDiT (MD contin)** replaces the discrete motion dictionary with a continuous motion representation.
- **DyaDiT (w/o self)** removes the *self* motion conditioning branch.
- **DyaDiT (Uncond)** generates gestures without any conditioning.
- **DyaDiT (Random)** performs inference with randomly assigned relationship and personality labels.

As expected, removing **ORCA** degrades performance, especially in *FD*, confirming its role in improving motion

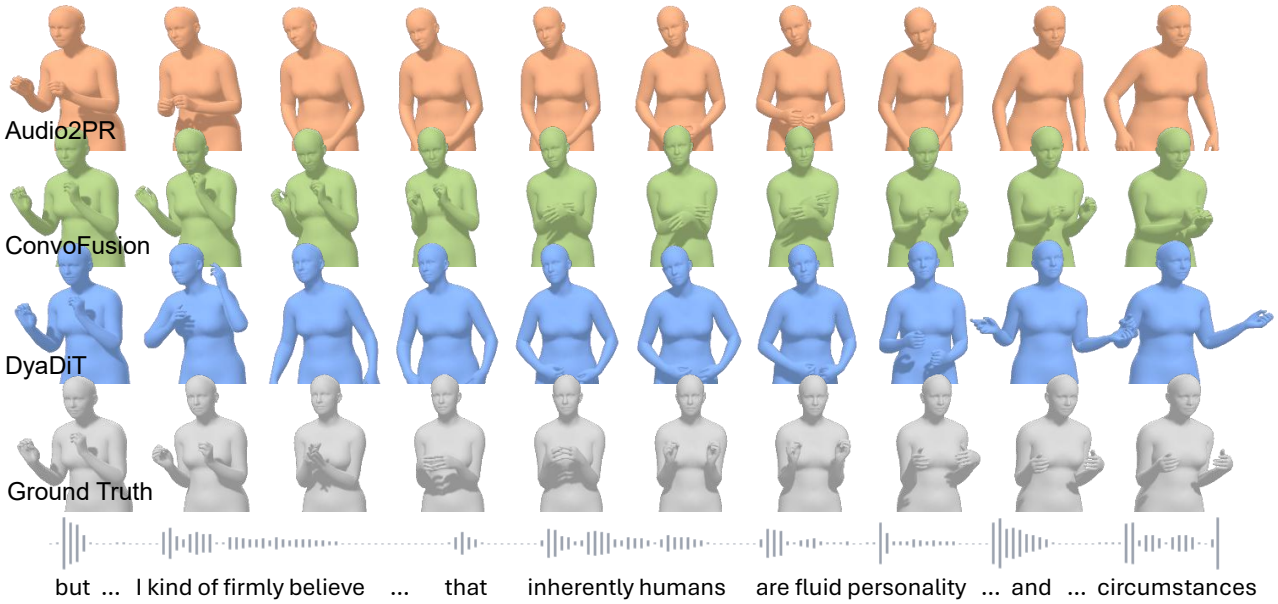

Figure 4. **Qualitative Results.** Comparison of visualization results between **DyaDiT**, **ConvoFusion** [31], and **Audio2PhotoReal** [3]. The gestures generated by DyaDiT exhibit higher diversity and greater realism compared to the other methods.

realism. Replacing the discrete **Motion Dictionary** with a continuous variant also hurts performance, particularly in *Diversity (Static)*, indicating that discrete motion bases better capture diverse interaction styles. We also find that the full **DyaDiT** outperforms both **Uncond** and **Random** social conditioning in overall generation quality, showing the importance of social context for realistic gesture generation. In particular, removing or mismatching social context noticeably reduces *Diversity (Static)*, suggesting that proper social cues help the model produce a broader motion distribution. Finally, **DyaDiT (Uncond)** does not achieve higher diversity than the fully conditioned model. We attribute this to the dyadic setting, where many segments correspond to listener behavior with inherently limited motion variation; additional social conditioning helps disambiguate such cases and encourages richer gestures.

5. User Study

We conduct an A/B preference study with sixteen participants to assess user preference on human motion in dyadic conversation settings. We compare our method with ConvoFusion [31] as well as with ground truth motion. The user study evaluates three aspects of the generated gestures: overall quality, relationship consistency, and personality consistency. Participants are shown paired clips and select their preferred gesture sequence, with the video pairs presented in randomized order to avoid ordering bias. An example video pair is shown in Figure 6.

We randomly selected 56 ten-second sequences from the validation set of the Seamless Interaction dataset to conduct the experiment. Below we show sample questions presented to the participants in each section.

- **overall quality**: *"Which gesture of the orange character looks more human-like?"*
- **relationship consistency**: *"Which pair seems more likely to be friends (or strangers, or family members, or dating partners)?"*
- **personality consistency**: *"Which gesture better reflects the personality trait agreeable (or conscientious, or extraverted, or neurotic)?"*

For each video pair, participants were asked to provide both their preference and their confidence level (strongly or slightly). For the full questionnaire and interface design used in the experiment, please refer to the supplementary material. As shown in Figure 7, DyaDiT achieves notably higher user preference scores than *ConvoFusion* [31], and interestingly, even slightly outperforms the ground truth gestures from the Seamless Interaction dataset.

Specifically, DyaDiT is preferred by 73.9%, 69.8%, and 66.7% of participants in overall gesture quality, relationship consistency, and personality consistency, respectively, showing that it generates both higher-quality and more socially appropriate gestures. Each participant evaluated 56 video pairs (10 s each), requiring about 20-25 minutes in

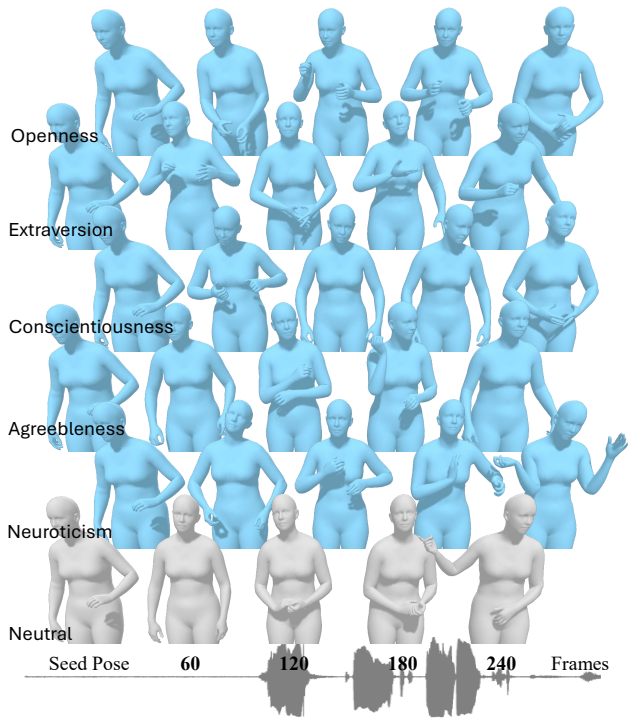

Figure 5. Visualization results under different personality score conditionings. All samples are generated using classifier-free guidance with CFG = 2.5.

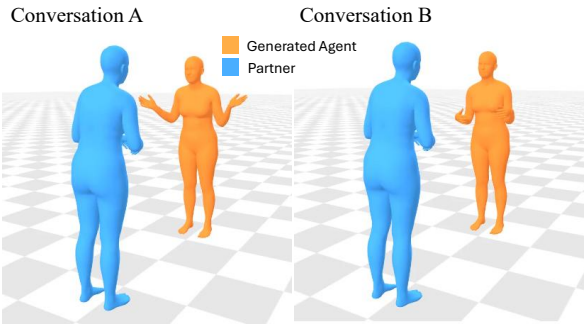

Figure 6. Example video pairs used in the user study for evaluating participant preference in conversational gesture generation.

total, and each pair was rated by 16 participants (aged 25-35, with CS backgrounds). Both quality and personality show strong deviations from chance against the baseline (χ^2, $p < 10^{-8}$), while personality also remains significantly preferred over the ground truth ($\chi^2, p = 1.5 \times 10^{-4}$). Our generated gestures are further preferred over the ground truth by 1.0% and 1.7% in two settings, likely because diffusion produces smoother motions and social conditioning encourages slightly more expressive gestures. We do not compare against the ground truth in personality consistency,

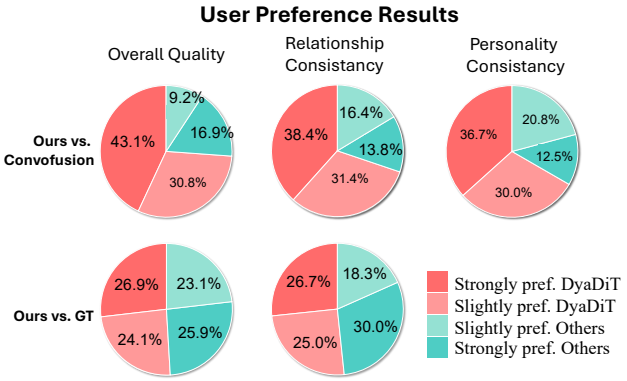

Figure 7. A/B subjective evaluation percentages comparing our method with ConvoFusion[31] and with ground truth. Participants preferred our generated motion due to its more natural and socially aware conversational behavior.

since personality annotations are continuous and difficult for participants to judge reliably.

6. Conclusion

In this work, we present **DyaDiT**, a multi-modal dyadic gesture generation model designed to produce socially consistent and natural conversational behaviors between two speakers. By jointly conditioning on dyadic audio, social attributes, and the partner's motion, DyaDiT effectively captures interpersonal dynamics in dyadic conversation. Our orthogonalization cross-attention (ORCA) module helps clarify the contribution of each speaker's audio, and the motion dictionary enhances expressive richness by providing style-aware motion priors. Both quantitative and subjective evaluations demonstrate that DyaDiT achieves superior performance compared to existing methods in terms of realism, diversity, and social coherence. We believe this work takes an important step toward socially favorable synthetic agents and opens up future research directions, such as socially aware dual-agent gesture generation.

7. Limitation & Future Work

Currently, our model is limited by the dataset and may still suffer from conditioning conflicts, as personality cues can be implicitly encoded in the audio, reducing controllability and diversity. In future work, we plan to explore audio neutralization to better disentangle social cues, collect extensive dyadic gesture datasets using our existing human reconstruction pipeline [39, 40] for full-body motion generation.

8. Acknowledgment

This work was supported by JST ASPIRE JPMJAP2404, and JST CRONOS JPMJCS24N8.

References

[1] Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, and et al. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. 2025. 1, 4, 6

[2] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Trans. Graph.*, 42 (4), 2023. 3, 6

[3] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Trans. Graph.* 3, 6, 7

[4] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 4

[5] Bohong Chen, Yumeng Li, Youyi Zheng, Yao-Xiang Ding, and Kun Zhou. Motion-example-controlled co-speech gesture generation leveraging large language models. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, New York, NY, USA, 2025. Association for Computing Machinery. 2

[6] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 3

[7] Kiran Chhatre, Radek Daněček, Nikos Athanasiou, Giorgio Becherini, Christopher Peters, Michael J. Black, and Timo Bolkart. AMUSE: Emotional speech-driven 3D body animation via disentangled latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1942–1953, 2024. 3

[8] Michael Dunne and Sik Hung Ng. Simultaneous speech in small group conversation: All-together-now and one-at-a-time? *Journal of Language and Social Psychology*, 13(1): 45–71, 1994. 2

[9] Filippo Favali, Viktor Schmuck, Valeria Villani, and Oya Celiktutan. Tag2g: A diffusion-based approach to interlocutor-aware co-speech gesture generation. *Electronics*, 13(17), 2024. 3

[10] Chencan Fu, Yabiao Wang, Jiangning Zhang, Zhengkai Jiang, Xiaofeng Mao, Jiafu Wu, Weijian Cao, Chengjie Wang, Yanhao Ge, and Yong Liu. Mambagesture: Enhancing co-speech gesture generation with mamba and disentangled multi-modality fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 10794–10803, New York, NY, USA, 2024. Association for Computing Machinery. 2

[11] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXVI*, page 418–437, Berlin, Heidelberg, 2024. Springer-Verlag. 3

[12] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3492–3501, 2019. 2

[13] Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri et al. The llama 3 herd of models, 2024. 2

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 5

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 4

[16] Cameron R. Jones and Benjamin K. Bergen. Large language models pass the turing test, 2025. 2

[17] Siyeol Jung and Taehwan Kim. Difflistener: Discrete diffusion model for listener generation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. 3

[18] Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris Kitani. Harmony4d: A video dataset for in-the-wild close human interactions. *Advances in Neural Information Processing Systems*, 37:107270–107285, 2024. 3

[19] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization, 2022. 5

[20] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Linchao Bao, and Zhenyu He. Audio2gestures: Generating diverse gestures from audio. *IEEE Transactions on Visualization and Computer Graphics*, 30(8):4752–4766, 2024. 6

[21] Siyao Li, Tianpe Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. In *ICLR*, 2024. 3

[22] Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 3764–3773, New York, NY, USA, 2022. Association for Computing Machinery. 2

[23] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, page 612–630, Berlin, Heidelberg, 2022. Springer-Verlag. 2

[24] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J. Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1144–1154, 2024. 2

[25] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei

Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10462–10472, 2022. 2

[26] Cheng Luo, Siyang Song, Weicheng Xie, Micol Spitale, Zongyuan Ge, Linlin Shen, and Hatice Gunes. Reactface: Online multiple appropriate facial reaction generation in dyadic interactions. *IEEE Transactions on Visualization and Computer Graphics*, 31(9):6190–6207, 2025. 3

[27] Cheng Luo, Siyang Song, Siyuan Yan, Zhen Yu, and Zongyuan Ge. Reactdiff: Fundamental multiple appropriate facial reaction diffusion model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, page 5607–5616, New York, NY, USA, 2025. Association for Computing Machinery. 2

[28] Cheng Luo, Jianghui Wang, Bing Li, Siyang Song, and Bernard Ghanem. Omniresponse: Online multimodal conversational response generation in dyadic interactions. *arXiv preprint arXiv:2505.21724*, 2025. 2, 3

[29] Vongani Maluleke, Lea Müller, Jathushan Rajasegaran, Georgios Pavlakos, Shiry Ginosar, Angjoo Kanazawa, and Jitendra Malik. Synergy and synchrony in couple dances, 2024. 3

[30] Antje S Meyer. Timing in conversation. *Journal of Cognition*, 6(1):20, 2023. 2

[31] Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. ConvoFusion: Multi-Modal Conversational Diffusion for Co-Speech Gesture Synthesis . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1388–1398, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 2, 3, 6, 7, 8

[32] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20395–20405, 2022. 3

[33] Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1001–1010, 2024. 2, 3, 6

[34] OpenAI, Josh Achiam, and Steven Adler et al. Gpt-4 technical report, 2024. 2

[35] Kunkun Pang, Dafei Qin, Yingruo Fan, Julian Habekost, Takaaki Shiratori, Junichi Yamagishi, and Taku Komura. Bodyformer: Semantics-guided 3d body gesture synthesis with transformer. *ACM Trans. Graph.*, 42(4), 2023. 2

[36] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2018. 4

[37] Xingqun Qi, Chen Liu, Lincheng Li, Jie Hou, Haoran Xin, and Xin Yu. Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation. *Trans. Multi.*, 26: 10420–10430, 2024. 2

[38] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[39] Soyong Shin, Chaeeun Lee, Holly Chen, Jyun-Ting Song, Eni Halilaj, and Kris Kitani. Bodycontact4d: A multi-view video dataset for understanding human and environment interactions. In *Thirteenth International Conference on 3D Vision*. 8

[40] Jyun-Ting Song, JungEun Kim, Jinkun Cao, Yu Lei, Takuma Yagi, and Kris Kitani. Contact4d: A video dataset for whole-body human motion and finger contact in dexterous operations. In *Thirteenth International Conference on 3D Vision*. 8

[41] Siyang Song, Micol Spitale, Xiangyu Kong, Hengde Zhu, Cheng Luo, Cristina Palmero, German Barquero, Sergio Escalera, Michel Valstar, Mohamed Daoudi, et al. React 2025: the third multiple appropriate facial reaction generation challenge. In *Proceedings of the ACM International Conference on Multimedia*, 2025. 3

[42] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022. 3

[43] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[44] Minh Tran, Di Chang, Maksim Siniukov, and Mohammad Soleymani. Dyadic interaction modeling for social behavior generation. *arXiv preprint arXiv:2403.09069*, 2024. 2

[45] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Lia: Latent image animator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2024. 5

[46] Sicheng Yang, Zilin Wang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Qiaochu Huang, Lei Hao, Songcen Xu, Xiaofei Wu, Changpeng Yang, and Zonghong Dai. Unifiedgesture: A unified gesture synthesis model for multiple skeletons. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 1033–1044, New York, NY, USA, 2023. Association for Computing Machinery. 3

[47] Sicheng Yang, Haiwei Xue, Zhensong Zhang, Minglei Li, Zhiyong Wu, Xiaofei Wu, Songcen Xu, and Zonghong Dai. The diffusestylegesture+ entry to the genea challenge 2023. In *Proceedings of the 25th International Conference on Multimodal Interaction*, page 779–785, New York, NY, USA, 2023. Association for Computing Machinery. 3

[48] Sicheng Yang, Zunnan Xu, Haiwei Xue, Yongkang Cheng, Shaoli Huang, Mingming Gong, and Zhiyong Wu. Freetalker: Controllable speech and text-driven gesture gen-

eration based on diffusion models for enhanced speaker naturalness. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. 2

[49] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 469–480, 2023. 2

[50] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Trans. Graph.*, 39(6), 2020. 2

[51] Jiangning Zhang, Liu Liang, Zhucun Xue, and Yong Liu. Apb2face: Audio-guided face reenactment with auxiliary pose and blink signals. pages 4402–4406, 2020. 3

[52] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *NeurIPS*, 2023. 3

[53] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(6): 4115–4128, 2024. 3

[54] Yihao Zhi, Xiaodong Cun, Xuelin Chen, Xi Shen, Wen Guo, Shaoli Huang, and Shenghua Gao. Livelyspeaker: Towards semantic-aware co-speech gesture generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20807–20817, 2023. 3

[55] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. *CoRR*, abs/1812.07035, 2018. 4

[56] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10553, 2023. 3