

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

Supplementary Material

This supplementary material contains sections below:

- 1. Relationships & Personality Clustering of Generated Results
- 2. Implementation Details of DyaDiT
- 3. Details of A/B Test Questionnaire

In addition to this `supplementary.pdf`, we also include a `narration_video.mp4`, which provides a brief overview of the paper along with several qualitative gesture generation examples. We further provide our implementation in `dyadit_code.zip`; please refer to the included `README.md` for instructions on running the code. The trained models will be released upon acceptance.

1. Clustering of Generated Gestures

In the main paper, we conduct an A/B test to evaluate the relationship and personality consistency of the generated gestures. To further assess the controllability of the conditional inputs in DyaDiT, we perform a t-SNE clustering analysis on the generated motion embeddings.

Figure 1 visualizes the t-SNE embeddings of the generated gestures under different conditioning signals. On the left, we generate gestures using various *relationship* types while fixing the personality scores. On the right, we discretize the continuous *personality score* features into five “one-hot” vectors and generate gestures for each vector to examine personality controllability.

We observe that the personality clusters form clearly separable groups, indicating that DyaDiT effectively captures the global behavioral tendencies associated with different personality traits. In contrast, the relationship clusters are less clearly separated. We consider this to be consistent with the nature of dyadic conversational gestures: the styles between *Friend*, *Family*, and *Dating* share some overlap. As a result, the generated gestures also show a more continuous manifold across these categories rather than sharp cluster boundaries.

2. Implementation Details

Diffusion Transformer The input pose sequence is encoded into a latent space aligned with the VQ-VAE representation, resulting in a latent embedding in \mathbb{R}^{64} , instead of the original $\mathbb{R}^{43 \times 6}$ 6D rotation matrix [7] joint representation. A linear layer projects the noisy latent input from \mathbb{R}^{64} to a hidden space in \mathbb{R}^{512} , followed by a symmetric projection back to \mathbb{R}^{64} at the output.

The model contains 4 Transformer blocks, each

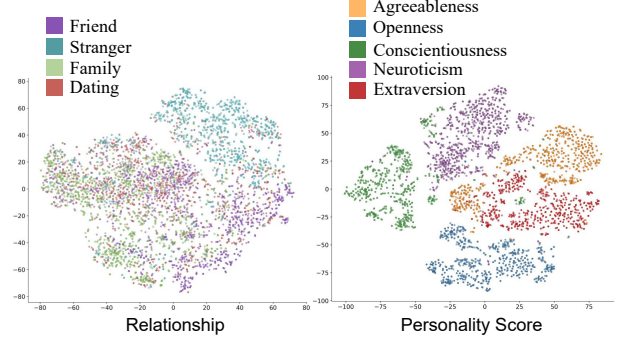


Figure 1. t-SNE clustering results of Relationships (left), Personality Scores (right).

equipped with 4-head multi-head attention (\mathbb{R}^{128}) and a \mathbb{R}^{2048} feed-forward network. We employ a \mathbb{R}^{512} sinusoidal time embedding, which is injected into each block through FiLM [5] modulation.

Two independent Wav2Vec2 [2] processors extract high-level audio features from the conversational speech of both speakers. These yield feature sequences denoted as $a_{\text{self}}, a_{\text{other}} \in \mathbb{R}^{T \times 768}$, where T denotes the number of audio frames. Each of these is projected to $\mathbb{R}^{T \times 512}$ via a linear transformation, followed by LayerNorm and a gated fusion mechanism to combine self and other speaker cues.

A learnable motion bank contains 1000 prototype vectors in \mathbb{R}^{512} , providing contextual priors via cross-attention. similar to the time embedding, relationship and personality embeddings are projected into \mathbb{R}^{512} . These vectors are injected into the DiT blocks via FiLM-style adaptive scaling. All contextual cues are concatenated into a unified sequence in \mathbb{R}^{512} and injected into each DiT block via cross-attention.

Motion Tokenizer (VQ-VAE). We implement a temporal VQ-VAE [4] to discretize pose sequences before diffusion. Given an input sequence of joint features $X \in \mathbb{R}^{T \times 6 \times 43}$, the encoder is a 1D CNN consisting of three Conv1d layers, with LeakyReLU applied after the first two layers, and an overall temporal downsampling factor of 4, producing a latent sequence in $\mathbb{R}^{(T/4) \times 64}$. This continuous latent is quantized by a residual vector quantizer with depth 4, each equipped with a 512-entry codebook, which maps each time step to a stack of discrete code indices. The decoder is a 1D CNN consisting of an initial Conv1d-LeakyReLU layer, two upsampling blocks (linear interpolation followed by Conv1d and LeakyReLU) interleaved with additional

Conv1d-LeakyReLU refinement layers, and a final Conv1d projection, achieving an overall temporal upsampling factor of 4 and recovering the original temporal resolution to reconstruct poses in $\mathbb{R}^{T \times 6 \times 43}$. The final latent representation used by the DiT denoiser is obtained from the quantized codes as a compact 64-dimensional embedding per $4 \times T$ frames in \mathbb{R}^{64} .

Seamless Interactive Dataset. We conduct our experiments on a part of the Seamless Interaction dataset [1]. In particular, we adopt the *naturalistic* split of the dataset. For training, we utilize the first 10 official training archives (provided as zip files), which contain approximately 182 hours of naturalistic interactions and 3000 paired motion-audio samples. For testing, we select the first archive from the official test split to ensure a consistent evaluation setting.

We observe that the SMPL-H parameters provided in the dataset exhibit noticeable inaccuracies in lower body estimation, likely due to limited camera views and body occlusions during data capture. To avoid introducing artifacts into our motion modeling, we discard lower-body joints and only retain the upper body comprising 43 joints, including fingers. For visualization, all unused joints, along with global orientation and root translation, are set to zero.

In addition to pose data, the dataset includes high level annotations such as *relationship* and *personality scores*. While the dataset provides *Interpersonal Communicative Dynamic* (IPC) tags for social dynamics, we found the annotations to be too noisy and ambiguous where it is unclear which speaker they apply to. Consequently, we do not employ IPC tag supervision in our current study and instead focus on the cleaner relationship and personality cues. We note that once the IPC annotations are refined in future dataset releases, we plan to extend our framework with an IPC-aware conditioning module to further capture communicative intent in dyadic gestures.

In the future, we plan to re-annotate the video data using advanced human pose estimation tools such as SMPLest-x [6], Harmony4D [3] or recent state-of-the-art models, with the aim of obtaining more reliable full body motion supervision.

3. Questionnaire

We provide a reconstructed version of the A/B test questionnaire used in our user study. To view the questionnaire, please first extract the `questionnaire_video.zip` file inside the `questionnaire` folder. After extraction, open `questionnaire.html` in any modern web browser.

The original questionnaire was conducted through Google Forms (see Figure 2). It consists of 28×2 questions

in total, including 10 questions on overall gesture quality, 8 questions on relationship consistency, and 10 questions on personality consistency. Each question presents paired gesture videos for comparison under two settings: *DyaDiT* vs. *ConvoFusion* and *DyaDiT* vs. *Ground Truth*.

For an accurate viewing experience, please wear headphones. The left audio channel corresponds to the partner’s speech, while the right audio channel corresponds to the target speaker’s speech.

The reconstructed interface allows reviewers to browse all questions and play the corresponding videos to experience the same evaluation procedure as our participants.

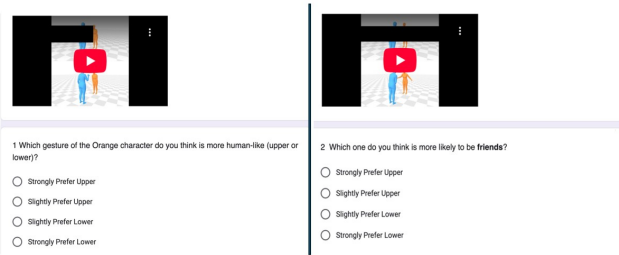


Figure 2. Example of Questionnaires in GoogleForm

References

- [1] Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, and et al. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. 2025. 2
- [2] 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. 1
- [3] Rawal Khrodgar, 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. 2
- [4] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization, 2022. 1
- [5] 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. 1
- [6] Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang,

Mingyuan Zhang, Lei Zhang, Chen Change Loy, Atsushi Yamashita, Lei Yang, and Ziwei Liu. Smplest-x: Ultimate scaling for expressive human pose and shape estimation. *arXiv preprint arXiv:2501.09782*, 2025. [2](#)

- [7] 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. [1](#)