MIBURI: Towards Expressive Interactive Gesture Synthesis

Supplementary Material

1. Online Generation Demo

To demonstrate the real-time capabilities of our gesture generation framework, we build an interactive demo in which a user can converse naturally with an Embodied Conversational Agent (ECA). We urge readers to watch the supplementary video, which highlights how our system supports online, continuous, and responsive gesture generation during live interaction.

1.1. Architecture

The primary goal of the demo is to showcase MIBURI's ability to generate gestures and speech in real time during a fully interactive conversation with the user. Unlike traditional turn-based systems, our setup supports full-duplex interaction, allowing both the user and the ECA to speak, interrupt, and respond fluidly—mirroring natural human dyadic communication. Achieving such responsiveness requires maintaining low latency while processing both the user's input and MIBURI's output continuously.

To this end, we implement the demo using three parallel processes, executed on a workstation equipped with an NVIDIA RTX 3090 GPU. These processes run concurrently and communicate through lightweight websocket channels to ensure synchronized, low-overhead data exchange. The three processes operate as follows:

- **Inference Process (Main Process).** This process runs the core inference loop for both Moshi [2] and MIBURI. It handles real-time speech–text token streaming and generates gesture tokens frame by frame.
- **Speech/Text Visualization Process.** At every inference step, the raw audio waveform and the decoded text tokens are sent via websocket to this process. It visualizes the user's speech and the agent's responses, allowing real-time inspection of the conversational flow.
- **Motion Visualization Process.** In parallel, the gesture generation module sends a time-aligned SMPL-X [15] mesh for each frame to a dedicated visualization process. This process renders the full-body motion—including hands and facial expressions—on the user's screen in real time.

Together, these components enable seamless, continuous interaction with the embodied agent, as illustrated in Fig. 1. The system maintains low latency at each stage, enabling a fluid and immersive demonstration of real-time embodied dialogue.

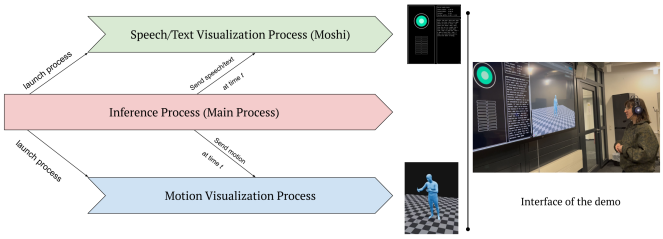

Figure 1. **System architecture of our real-time demo.** The main inference process runs Moshi and MIBURI in a continuous loop, while two parallel processes handle speech/text visualization and motion rendering. Data is streamed between processes at each timestep via websockets to support low-latency, full-duplex interaction. Right: the user-facing interface of the demo.

Table 1. **Quantitative evaluation on the Embody3D dataset.**

	FGD↓	BeatAlign→	L1-Div→
GT	–	0.453	5.97
EMAGE [8]	3.786	0.022	1.79
GestureLSM [9]	3.744	0.776	13.75
MIBURI	1.642	0.605	10.18

2. Additional Results on Embody3D [10]

Embody3D is a recently released dataset containing 59 hours of dyadic interaction recordings. In this setup, two interlocutors face each other and communicate naturally, mirroring human–human conversational dynamics. We evaluate on this dataset because our long-term goal is to develop fully interactive embodied agents capable of behaving like humans in real conversational settings.

We finetune our multi-speaker BEAT2 models on Embody3D and report performance using FGD, BeatAlign, and L1 Divergence. Across all metrics, MIBURI achieves the best quantitative results, showing lower gesture distribution divergence, improved alignment with speech prosody, and motion diversity closer to GT. This performance trend mirrors our findings on the BEAT2 multi-speaker evaluation, further demonstrating that our causal token-based framework generalizes well to new conversational settings. For fair comparison, we retrain the FGD network following EMAGE [8] and recompute the mean velocity used in the BeatAlign metric. All evaluations are conducted using only the upper body and hands

3. Analyzing Autoregressive Dependency in Kinematic Transformer.

Since we model body-part level details through an autoregressive transformer, this leads to a dependency in which

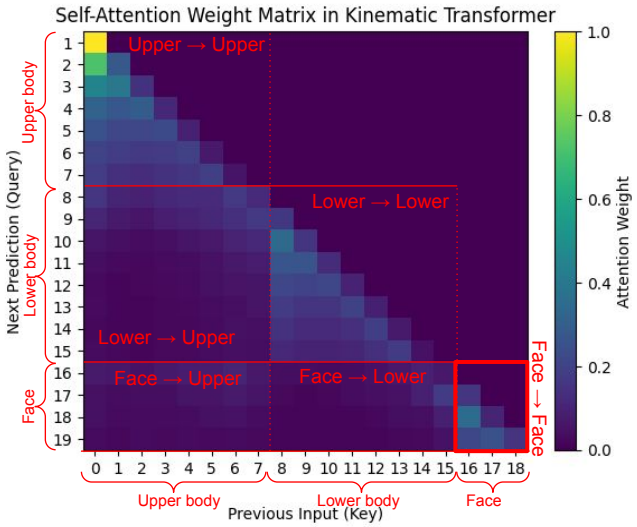

Figure 2. **Kinematic Dependency Analysis.** Here, "→" means "attends to".

body parts predicted later (lower body and face) depend on body parts predicted earlier (upper body). Therefore, we analyze the effect of this ordering to examine whether it imposes a specific dependency chain. We plot the causal self-attention between the face, upper-body, and lower-body tokens in Fig. 2. Even though face tokens are predicted after lower-body tokens, the attention weights show that the model implicitly learns to ignore lower-body tokens when predicting face tokens. We observe that face self-attention is concentrated in the "Face → Face" block, as the face does not depend on other parts. Lower-body tokens exhibit small attention to the upper body, since both are linked in terms of motion dynamics.

4. On Causality-Quality Trade-off

Recall that MIBURI is designed to be an interactive, embodied conversational agent (ECA). This necessitates that the model is not only causal, but also real-time and our design choices have been profoundly dictated by these considerations. Naturally, causality comes at the cost of quality. However, this trade-off is not merely a consequence of having a limited context. In this section, we highlight the underlying nuances associated with the design of causal synthesis for human gestures.

Where do Gestures originate? We submit that the premise of causal co-speech gesture synthesis is rather ill-posed, as it tacitly assumes that the agent (or humans) gesture on the basis of speech uttered within the *past* context. This assumption, however, is not true. In reality, human speech and gestures are driven in parallel through a shared *intent* [1, 3–5]. This is reflected in the observation that gestures can often be stroked even before the speech has been uttered [11, 13]. This is also observed in turn-taking between multiple interlocutors [16]. Likewise, it is also possible that gestures occur with a delay (for example, to reinforce or qualify the argument in the speech) [12]. While the latter case can be, in principle, modeled by a causal model, the former is inherently challenging to achieve. Consequently, we observe that causal modeling incentivizes more prominent beat gestures, as the temporal correlation between the speech and the gestures is easier to discover while training. On the other hand, non-causal, full-context models, thrive in the luxury of future context availability and are able to model more nuanced and semantically meaningful gestures [14, 21].

Should we Model the Intent Behind the Gestures? We believe that a common modeling of the intent before generating the speech and gestures is a goal worth pursuing. This could be achieved by first inferring the intent behind an LLM's output, and then generating the speech and gestures jointly based on the inferred intent. However, this would be a suboptimal approach that breaks the constraints of causality, while also being slow. For truly interactive ECAs, we either need a real-time LLM that is trained to generate the intent before the final output, or, we need an approach to disentangle the LLM's intent from the intermediate features of the model. While fascinating directions for future research, both the potential solutions remain out of the scope of our current submission.

5. Implementation Details of Gesture Codecs.

We build streaming codecs for each body part using Residual VQ-VAE [19]. These codecs consist of an encoder-decoder architecture, where the encoder downsamples the input motion sequence by a factor of 2 and the decoder upsamples it back for reconstruction. Given an input of 250 frames during training, the encoder outputs 125 tokens for a 10-second sequence. During training, the input sequence length is randomly sampled between 2 and 250 at every iteration.

Upper and lower body codecs consist of 2 1d-convolutional layers and 8 transformer layers with 4 attention heads. Face codec contains 2 1d-convolutional layers and 4 transformer layers with 2 attention heads. Every codec is trained with a set of reconstruction and geometric losses along with commitment losses for each codebook. We apply Geodesic Loss on rotation matrices and standard MSE losses on 6D, axis-angle and joint position representation of the motion. Moreover, we also apply additional MSE losses to optimize velocity/acceleration of motion [14]. Lastly, we apply loss on foot contact predictions during codec training to reduce foot sliding [14, 20].

6. Evaluation Metrics

FGD. We adopt the Fréchet Gesture Distance (FGD), following Yoon *et al*. [18]. For evaluation, we use the gesture encoder released with BEAT2 [8] to extract gesture embeddings and compute FGD, without retraining the encoder.

Beat Alignment Score. Originally proposed to assess synchronization between music beats and dance motion [6], the Beat Alignment Score has been adapted for gesture synthesis to measure how well gesture beat events align with audio beat events. It captures temporal correlation between gesture dynamics and speech prosody.

L1 Divergence. Also referred to as L1 variance, this metric computes the average L1 distance between each generated pose and the mean pose of the sequence. Lower values indicate motion collapse toward static poses, making it useful for detecting unexpressive or frozen gesture generation.

Facial-MSE. This metric, introduced by EMAGE [8], measures mean squared error between the ground truth facial expressions and predicted facial expressions. FLAME [7] is used as the representation to calculate this loss.

Mean Per Joint Position Error (MPJPE). This is a standard metric for evaluating motion reconstruction and pose estimation. It measures the average Euclidean distance between predicted and ground-truth joint positions across all joints and frames. Formally, it is computed as the mean L2 distance in 3D space, providing a direct measure of pose accuracy.

7. Details on User Study

To evaluate perceptual quality, we conducted a user study with 53 participants. Each participant was presented with 15 forced-choice questions randomly sampled from 45 questions. These questions display a side-by-side animation, comparing our method against state-of-the-art baselines and ground truth. Each question displayed a side-by-side animation of our model and one of EMAGE [8], GestureLSM [9], or the ground truth. For every pairwise comparison, participants answered two questions:

- "Which gesture sequence looks more natural?"
- "Which appears better aligned with the spoken content?"

Across all comparisons, results were statistically significant with p-values < 0.001, except for the appropriateness comparison against GestureLSM [9], which remained significant at $p < 0.05$.

8. Baseline Implementations

For the single-speaker evaluation, we use the publicly released checkpoints provided by each baseline method. For the multi-speaker evaluation, many baselines do not release multi-speaker models. To ensure a fair comparison, we retrain EMAGE [8] and MambaTalk [17] on the 23-speaker subset of BEAT2 (excluding *carla* and *itoi*), following the training configurations described in their respective papers.

Beyond comparing to non-causal baselines, we also create causal variants of GestureLSM and MambaTalk to evaluate them under the same online, real-time constraints as our method. In both cases, causality is enforced by applying a causal attention mask to all transformer layers during training. This allows us to report quantitative comparisons against models operating under equivalent causal conditions.

References

[1] Jan P. de Ruiter. The production of gesture and speech. In *Language and Gesture*, pages 248–311. Cambridge University Press, 2000. 2

[2] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024. 1

[3] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004. 2

[4] Sotaro Kita and Aslı Özyürek. What does cross-linguistic variation in semantic coordination of speech and gesture reveal? evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48(1):16–32, 2003.

[5] Willem J. M. Levelt. *Speaking: From Intention to Articulation*. MIT Press, 1989. 2

[6] Ruilong Li, Sha Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021. 3

[7] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 3

[8] 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 *CVPR*, 2024. 1, 3

[9] Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. Gesturelsm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. *arXiv preprint arXiv:2501.18898*, 2025. 1, 3

[10] Claire McLean, Makenzie Meendering, Tristan Swartz, Orri Gabbay, Alexandra Olsen, Rachel Jacobs, Nicholas Rosen, Philippe de Bree, Tony Garcia, Gadsden Merrill, Jake Sandakly, Julia Buffalini, Neham Jain, Steven Krenn, Moneish Kumar, Dejan Markovic, Evonne Ng, Fabian Prada, Andrew

Saba, Siwei Zhang, Vasu Agrawal, Tim Godisart, Alexander Richard, and Michael Zollhoefer. Embody 3d: A large-scale multimodal motion and behavior dataset, 2025. 1

[11] David McNeill. *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, 1992. 2

[12] David McNeill. *Gesture and thought*. University of Chicago Press, 2005. 2

[13] Palmer Morrel-Samuels and Robert M Krauss. Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(3):615, 1992. 2

[14] M. Hamza Mughal, Rishabh Dabral, Merel C. J. Scholman, Vera Demberg, and Christian Theobalt. Retrieving semantics from the deep: an rag solution for gesture synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[15] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1

[16] Varsha Suresh, M. Hamza Mughal, Christian Theobalt, and Vera Demberg. Modeling turn-taking with semantically informed gestures. In *Findings of the Association for Computational Linguistics: EACL 2026*, 2026. 2

[17] Zunnan Xu, Yukang Lin, Haonan Han, Sicheng Yang, Ronghui Li, Yachao Zhang, and Xiu Li. Mambatalk: Efficient holistic gesture synthesis with selective state space models. *Advances in Neural Information Processing Systems*, 37:20055–20080, 2024. 3

[18] 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 TOG*, page 1–16, 2020. 3

[19] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021. 2

[20] Wanyue Zhang, Rishabh Dabral, Thomas Leimkühler, Vladislav Golyanik, Marc Habermann, and Christian Theobalt. Roam: Robust and object-aware motion generation using neural pose descriptors. In *2024 International Conference on 3D Vision (3DV)*, pages 1392–1402, 2024. 2

[21] Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. Semantic gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Trans. Graph.*, 2024. 2