MIBURI: Towards Expressive Interactive Gesture Synthesis

M. Hamza Mughal[1] Rishabh Dabral[1] Vera Demberg[1,2] Christian Theobalt[1,2]
[1]Max Planck Institute for Informatics, SIC [2]Saarland University
vcai.mpi-inf.mpg.de/projects/MIBURI

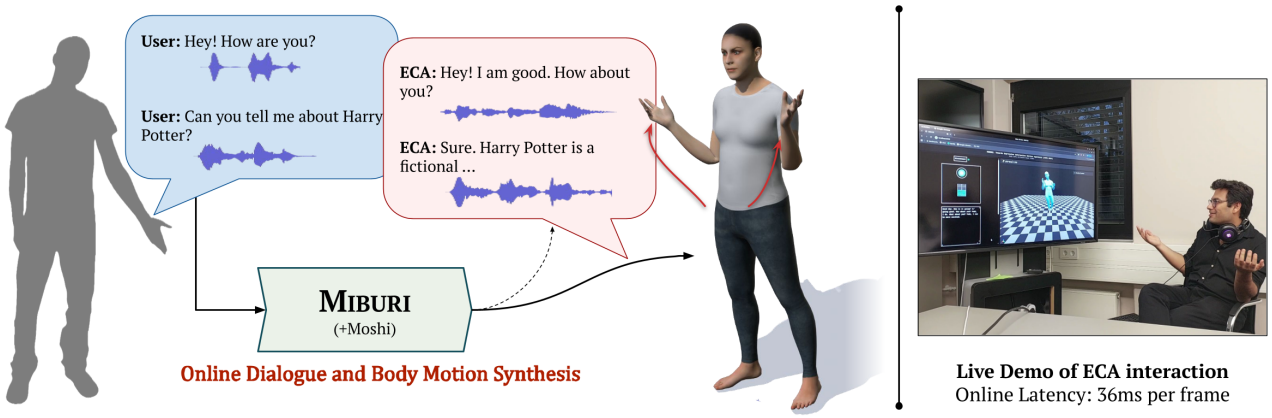

Figure 1. **MIBURI: An online, causal framework for real-time dialogue and gesture generation.** Given live speech, the system produces full-duplex responses with synchronized full-body gestures. Right: Interactive demo using our approach.

Abstract

Embodied Conversational Agents (ECAs) aim to emulate human face-to-face interaction through speech, gestures, and facial expressions. Current large language model (LLM)-based conversational agents lack embodiment and the expressive gestures essential for natural interaction. Existing solutions for ECAs often produce rigid, low-diversity motions, that are unsuitable for human-like interaction. Alternatively, generative methods for co-speech gesture synthesis yield natural body gestures but depend on future speech context and require long run-times. To bridge this gap, we present MIBURI, *the first online, causal framework for generating expressive full-body gestures and facial expressions synchronized with real-time spoken dialogue. We employ body-part aware gesture codecs that encode hierarchical motion details into multi-level discrete tokens. These tokens are then autoregressively generated by a two-dimensional causal framework conditioned on LLM-based speech-text embeddings, modeling both temporal dynamics and part-level motion hierarchy in real time. Further, we introduce auxiliary objectives to encourage expressive and diverse gestures while preventing convergence to static poses. Comparative evaluations demonstrate that our causal and real-time approach produces natural and contextually aligned gestures against recent baselines. We urge the reader to explore demo videos on our project page.*

1. Introduction

Human Computer Interaction has evolved from punch card based interfaces to LLM-driven conversational agents. Throughout this journey, these interfaces have progressed to emulate a more "human" way of interaction. Current textual chat assistants are the latest iteration in this evolution, which feature a strong understanding of linguistically encoded world knowledge. We interact with these digital assistants naturally through our voice or text, without the need to navigate a Graphical User Interface. However, human communication is not limited to verbal interaction but also involves non-verbal elements, such as body gestures and facial expressions, which are non-existent in these assistants. Full-body gestures not only convey meaningful contextual information in a conversation but also structure human interactions, serving as another important means of communication.

Introducing this new communication channel in digital assistants paves the way for *Embodied Conversational Agents* [7]: interfaces that are more interactive and natural for human communication [40], marking a step toward a deeper understanding of the physical world knowledge beyond language. To achieve this goal of interactive agents, the seminal work of Cassell *et al.* [7] outlines architectural requirements specifying that the system should produce *expressive* body gestures alongside spoken dialogue in *real-time*. Building on this foundation, both early rule-

based [4, 7] and recent data-driven [33] approaches have attempted real-time gesture generation synchronized with speech. However, they often yield less expressive, low-diversity motion and exhibit artificial turn-taking interaction patterns with distinct speaking and listening phases.

In contrast, recent generative approaches [2, 31, 32, 50] produce more natural and expressive co-speech gestures, leveraging neural architectures through diffusion [35, 51] or masked modeling in transformers [25]. However, these models typically operate in an *offline*, non-causal manner, requiring access to both past and future speech context to synthesize motion for a given time step, and thus cannot run in parallel with live speech generation. It is important to note that *causal* and *real-time* processing are related but distinct requirements: causal models, such as autoregressive transformers, rely only on past inputs, with no regard to any latency requirement, whereas real-time interactive systems must additionally meet strict time constraints to maintain conversational fluidity along with expressive gestures. Consequently, existing generative gesture approaches, while expressive, cannot be used as plug-and-play solution to build the embodied agents outlined by Cassell *et al*. [7].

To address this gap, we introduce MIBURI– an online, fully causal generative framework that generates expressive co-speech body gestures and facial expressions along with spoken dialogue in real-time. We build this framework upon Moshi [11], a speech-text foundation model that generates full-duplex spoken dialogue, and leverage its rich contextual speech-text embeddings to generate synchronized body motion. While several LLM-based gesture synthesis approaches exist [9, 32], they typically involve a bulky pipeline in which the LLM outputs are converted to speech, which is then tokenized to condition the gesture synthesis model (Fig. 2 top). We propose an alternative paradigm. In order to be causal and real-time, we exploit the speech-text aligned token stream of Moshi, and build our gesture synthesis architecture by directly tapping-on to the internal Moshi tokens. As illustrated in Fig. 2 (bottom), this allows us to avoid the latency-inducing steps of the conventional pipelines while benefiting from the rich semantic and acoustic contexts provided by the token embeddings.

Architecturally, our causal generative network leverages these internal tokens to generate gestures through two transformers: one incorporating the *temporal* context and the other generating per-frame, skeleton-aware *kinematic* features. To facilitate such decomposition, we propose a two-dimensional gesture encoding through Residual VQ-VAE, which is trained to perform causal decoding of the generated gesture tokens. Noteworthy is that our tokens encode a short temporal window (2 frames) in order to keep the latency low. We encode gestures by dividing the body into three groups (face, upper and lower body) and tokenize them seperately through individual codecs.

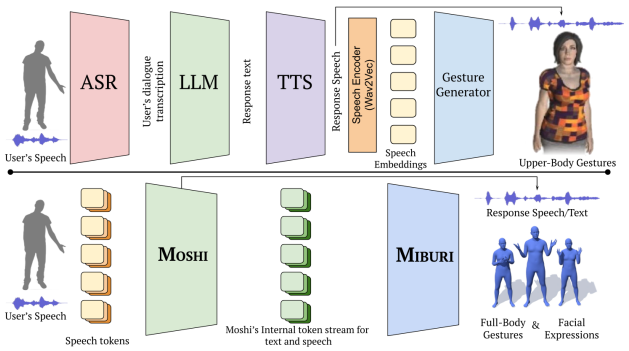

Figure 2. **Overview.** Existing solutions [9, 33] to animate ECAs involve a complex pipeline (above) of multiple components to generate gestures with speech. MIBURI (below) generates full body co-speech gestures directly by utilizing internal semantic/acoustic tokens of speech-text foundation model [11].

In summary, our contributions are threefold:

- We contribute a new paradigm for online, real-time and causal gesture generation, which leverages the internal token-stream of a speech-based Large Language Model to perform interactive gesture synthesis.
- We carefully design the network architecture and tokenization strategy, which facilitate causal gesture synthesis without compromising the expressiveness of the generated gestures.
- We present a comprehensive analysis of the several design choices involved in our method. Through perceptual and numerical experiments, we demonstrate that MIBURI advances the state-of-the-art of Embodied Conversational Agents (ECAs).

2. Related Work

We first review works on co-speech gesture synthesis, followed by methods for building Embodied Conversational Agents. Although both aim to generate co-speech gestures, their differing requirements make bridging the two fields together non-trivial.

2.1. Co-Speech Gesture Synthesis

Co-speech gestures are speech-synchronized body and hand movements that convey semantically aligned meaning [30]. Existing works on gesture generation range from early rule-based systems [8, 43, 45] to modern learning-based systems [12, 13, 15, 19, 48]. Learning based methods [2, 25, 31, 35] are typically data-driven and employ deep networks to convert speech input into synchronized natural-looking motion. CaMN [24] and EMAGE [25] introduce large-scale speech-aligned motion datasets and transformer-based gesture synthesis methods. GestureDiffuCLIP [2] utilizes diffusion transformers with causal attention over past and future speech frames. ConvoFusion [31]

uses diffusion to generalize generation across single- and two-person interactions, while Audio2Photoreal [35] also generates dyadic interactions with photorealistic avatars. RAG-Gesture [32] and SemanticGesticulator [50] develop retrieval-based paradigms to improve the semantic alignment in generated gestures. These methods cannot run in real time due to heavy computation, making them unsuitable for online gesture synthesis.

To address long runtimes, MambaTalk [47] uses selective state-space models with non-causal cross-attention for low-latency generation. GestureLSM [26] tackles this with a real-time flow-matching framework and shortcut sampling. Both methods also require seed gesture sequences during inference. However, these methods remain offline and non-causal, relying on past and future speech, and therefore cannot support online ECAs. This highlights the need for a real-time, causal framework that generates expressive gestures directly from speech without future context, seed gestures, or long runtimes.

2.2. Embodied Conversational Agents (ECAs)

In language generation, LLMs [38, 44] have shown strong capability to generate and understand text. Similarly, recent spoken dialogue systems [11, 37] aim to perform conversations in real-time while maintaining knowledge and reasoning abilities exhibited by LLMs. However, these natural language interfaces lack the full-body dynamics for an embodied avatar. In the avatar space, recent approaches have tried to enhance LLM-driven conversations with virtual characters through articulated body movements. Digital Life Project [5] uses an LLM backbone to synthesize instruction-driven motion for virtual characters. TaoAvatar [9] focuses on producing a full-body photorealistic avatar in real-time, given gesture input from motion library.

Full-fledged solutions for ECAs mainly include rule-based systems, while there are only a few recent data-driven systems. Rule-based systems [4, 6, 7] usually utilize pre-recorded animations to synthesize body motion in real time. Hybrid systems [27, 46] use neural methods for lip animation synthesis and a rule-based approach for body gestures. Recently, Abel *et al*. [1] propose a GRU-based pipeline to generate real-time co-speech gestures. Nagy *et al*. present Gesturebot [33] that utilizes data-driven methods like [20] to create an embodied avatar for body gestures. However, Gesturebot is limited to manual turn-based interactions and animates gestures only during speech, using a non-causal model. In contrast, our framework operates causally at both speech and gesture token levels, enabling real-time, continuous interaction.

3. Approach

The goal of our approach is to generate full-body gestures and facial expressions synchronized with speech for ECAs.

Table 1. Approaches for offline Gesture Synthesis and ECAs

Method	Approach	Expressive	Causal	Real-time
Cassell et. al. [6, 7]	Rule-based	✗	✓	✓
DigitalEinstein [46]	Rule-based	✗	✓	✓
Gesturebot [33]	Autoregressive	✗	✗	✓
EMAGE [25]	Masked Gesture Modelling	✓	✗	✗
ConvoFusion [31]	Diffusion	✓	✗	✗
Audio2Photoreal [35]	VQ+Diffusion	✓	✗	✗
RAG-Gesture [32]	Retrieval+Diffusion	✓	✗	✗
GestureLSM [26]	Flow-Matching	✓	✗	✓
MambaTalk [47]	SSM	✓	✗	✓
MIBURI (Ours)	RVQ+Autoregressive	✓	✓	✓

To enable such interactive agents, an ideal framework must produce spoken dialogue and then leverage the underlying verbal and prosodic context to synthesize *expressive* and diverse body gestures. According to the Cassell et al. [7], there are two key requirements for interactive gesture synthesis: (1) it must be *causal*, $i.e.$ one cannot assume the availability of future utterance, (2) and it must be *realtime* with low latency. Here, it is important to emphasize that simply having a low amortized rate of generation, as is typical for diffusion-based methods, is not enough.

Our approach builds upon a speech-text foundation model [11] to generate full-duplex spoken dialogue and extract its internal speech-text token stream to provide rich contextual input for gesture synthesis (Sec. 3.1). Our gesture generator then autoregressively produces body-region aware motion tokens (Sec. 3.2) using two-dimensional temporal and kinematic transformers(Sec. 3.3).

However, this base generation framework is insufficient to achieve diverse and expressive body gestures, which are crucial for natural communication. Therefore, we propose additional objectives for our autoregressive framework to achieve human-like gesture quality (Sec. 3.4). Finally, we ensure causal and real-time inference by carefully designing attention contexts and efficient cache mechanisms (Sec. 3.5). Fig. 3 illustrates our proposed architecture.

3.1. Preliminaries: Moshi

To produce real-time conversational speech and language, we utilize an open-source spoken dialogue system, i.e. Moshi [11]. Built on a textual LLM backbone, this framework autoregressively generates text and speech tokens. Crucially, it enables full-duplex conversations by jointly modeling its own speech and the user's speech in parallel token streams. At output, it generates speech and text tokens i.e. $\mathbf{f}^{\text{speech}} \in \mathbb{R}^{T \times K^{\text{speech}} \times d}$ and $\mathbf{f}^{\text{text}} \in \mathbb{R}^{T \times K^{\text{text}} \times d}$, where T represents the number of tokens along the time axis and d denotes the embedding dimension. Moshi also utilizes Residual Vector Quantization [49] to encode speech into multiple levels of semantic and acoustic tokens, and K^{speech} and K^{text} represent those quantization levels for speech and text respectively. MIBURI aims to leverage these semantic and prosodic details from Moshi for generating its own skeleton-aware token stream of gestures.

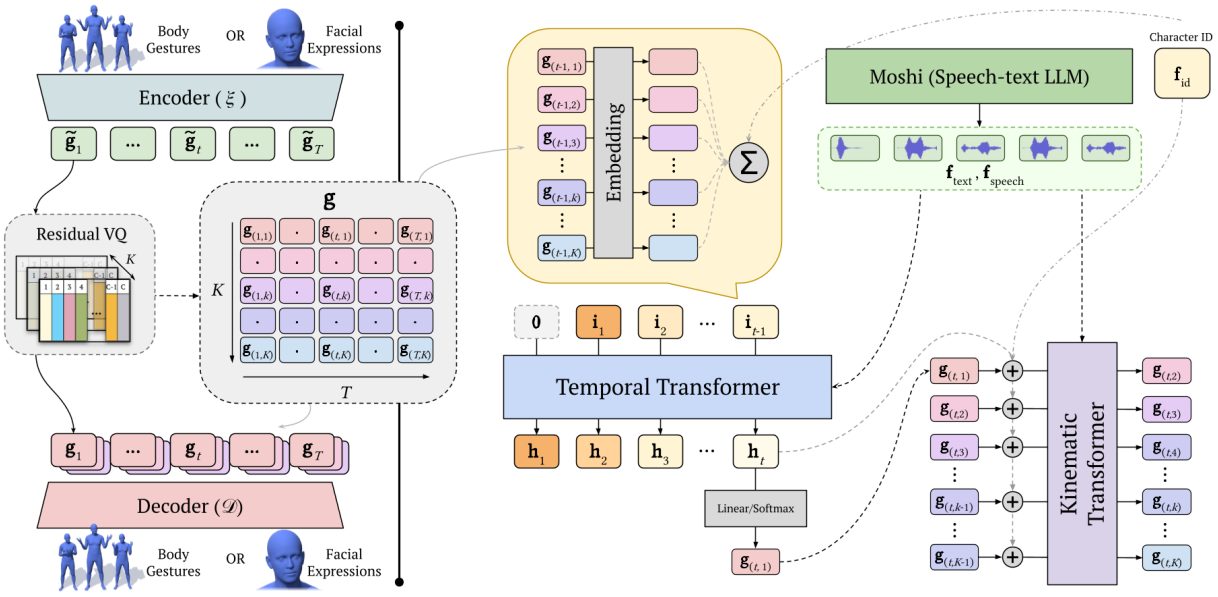

Figure 3. **MIBURI Architecture.** Given Moshi's speech/text tokens(Sec. 3.1), our approach generates a sequence of gesture tokens, which are obtained through Body-part aware Gesture Codecs(Sec. 3.2). This online framework takes in Moshi's text/speech token as input and predict gesture tokens through autoregressive *temporal* and *kinematic* transformers(Sec. 3.3).

3.2. Body-part wise Gesture Codecs

As the first step in our framework, we build a robust motion prior that encodes gesture frames into discrete tokens, which can then be used for downstream gesture generation using an autoregressive transformer. Since co-speech articulation in each body region happens at different scales [31] and different body regions relate to speech seperately [25, 32], we decouple each pose in the gesture sequence $\mathbf{x}$ into three body regions: upper body with hands $\mathbf{x}^u \in \mathbb{R}^{N\times 6J^u}$, lower body with global translation and foot contacts $\mathbf{x}^l \in \mathbb{R}^{N\times(6J^l+3+4)}$, and facial expressions using FLAME parameters $\mathbf{x}^f \in \mathbb{R}^{N\times(100+6J^f)}$ [23]. Here N is the number of frames of human motion, while J^u, J^l and J^f represent upper body, lower body and jaw joints respectively [39, 52]. Each region-specific gesture sequence is encoded through a separate Gesture Codec, which utilizes Residual VQ-VAE for motion tokenization.

Residual VQ-VAE for Gestures. Co-speech body articulation contains multiple aspects of detail, ranging from large arm jerks to subtle finger-level gestures. Naïvely tokenizing gestures through VQ-VAE quantization schemes [25, 41, 42] can result in coarse and choppy motion due to the loss of finer kinematic details (see Sec. 4.4). To encapsulate these subtle motion details, we train gesture codecs for each body region using Residual VQ-VAE [49]. Each region-wise codec consists of an encoder-decoder architecture (Fig. 3) with encoder $\mathcal{E}^b$ containing downsampling 1-d convolution layers and a transformer encoder with causal self-attention. It encodes motion for a given body region as $\tilde{\mathbf{g}}^b = \mathcal{E}^b(\mathbf{x}^b)$, whose output is quantized into tokens $\mathbf{g}^b \in \mathbb{R}^{T\times K^b}$ with K^b levels of motion details via Residual Vector Quantization (RVQ). Here T is the temporal length of the token sequence, downsampled from N. Each residual level learns a codebook $\mathbf{C}_k \in \mathbb{R}^{V\times d}$ that is used for vector quantization of the corresponding residual. Consequently, motion can be reconstructed through the decoder: $\hat{\mathbf{x}}^b = \mathcal{D}^b(\mathbf{g}^b)$, which consists of a similar causal transformer encoder and upsampling transpose 1D-convolution layers.

These gesture tokenizing codecs are trained with a set of reconstruction/geometric losses and a latent embedding loss at each quantization level (detailed in the supplemental). Finally, the resulting gesture sequence $\mathbf{g} \in \mathbb{R}^{T\times K}$ can be defined as a concatenation :

$$\mathbf{g} = \text{Concat}(\mathbf{g}^u, \mathbf{g}^l, \mathbf{g}^f)$$

along the K level axis, with $K = K^u + K^l + K^f$. This tokenized sequence $\mathbf{g} = \{\mathbf{g}_{(t,k)} \mid t = 1\dots T, , k = 1\dots K\}$ represents motion along *temporal* and *kinematic* dimensions, where former encompasses kinematic details across time and the latter contains part-level details across body regions at each t.

3.3. Autoregressive & Causal Transformers

Recall that our objective is to design a *causal* gesture synthesis framework, which needs to generate gesture tokens $\mathbf{g}$, given speech $\mathbf{f}^{\text{speech}}$ and text $\mathbf{f}^{\text{text}}$ tokens from Moshi and a character identity embedding $\mathbf{f}^{\text{id}}$ as input. Autoregressive transformers are commonly used in causal next-token

prediction tasks, where attention layers attend to the previous T tokens. However, in our case, each token frame in T also contains K token levels representing hierarchical motion details. A naïve implementation would require us to model $T.K$ tokens autoregressively, where attention layers would need the context of at least $> K$ tokens to learn temporal dynamics across motion frames. This automatically increases the size of context window in attention layers, while being harder to train and computationally expensive at inference (see Sec. 4.4). Therefore, we disentangle the prediction of both temporal and kinematic dimensions of gesture codecs $\mathbf{g}$ with two transformers inspired by RQ-Transformer [11, 21, 53].

Temporal Transformer. First, we build our base *temporal* transformer $\mathcal{T}_{\text{temporal}}$ to focus on the temporal dynamics across time. This causal transformer is trained to autoregressively predict the first level token $\mathbf{g}_{(t,1)}$ (among the K kinematic levels), conditioned on the tokens from previous timesteps.

$$\mathbf{h}_t = \mathcal{T}_{\text{temporal}}\big(\mathbf{g}_{(<t)}, \mathbf{f}^{\text{speech}}_{(\leq t)}, \mathbf{f}^{\text{text}}_{(\leq t)}, \mathbf{f}^{\text{id}}\big) \tag{1}$$

$$\mathbf{g}_{(t,1)} = \text{Softmax}\big(\text{Linear}(\mathbf{h}_t)\big) \tag{2}$$

Internally, the embeddings for $\mathbf{g}_{(<t)}$ along the kinematic dimension K are summed up to form a single input $\mathbf{i}_{t-1}$ for each t (see Fig. 3). The output of transformer $\mathbf{h}_t$ is converted to logits $\mathbf{o}_{(t,1)} \in \mathbb{R}^V$ through a simple classification layer and then we obtain $\mathbf{g}_{(t,1)}$ through Softmax. This module is implemented as a transformer decoder with a causal self-attention over past gesture tokens and dual causal cross-attention layers attending to preceding and current speech and text tokens. Note that we also learn per-identity feature embeddings that are added at each timestep.

Kinematic Transformer. Next, we model the *kinematic* dimension of gesture tokens through a transformer $\mathcal{T}_{\text{kinematic}}$, which autoregressively predicts the next body-part level at each timestep t. In addition to the previously generated levels, the kinematic transformer conditions on the temporal context $\mathbf{h}_t$, as well as speech, text, and identity embeddings:

$$\mathbf{g}_{(t,k)} = \mathcal{T}_{\text{kinematic}}\big(\mathbf{h}_t, \mathbf{g}_{(t,<k)}, \mathbf{f}^{\text{speech}}_t, \mathbf{f}^{\text{text}}_t, \mathbf{f}^{\text{id}}\big) \tag{3}$$

Here, the timestep t remains fixed for each level prediction. Therefore, the speech and text inputs correspond only to embeddings at time t. This transformer is also implemented as a decoder with a causal self-attention layer and cross-attention layers for speech and text. The identity embedding and temporal context $\mathbf{h}_t$ are added to the input of each level-step. Finally, the output of each step k is projected through classification layers to predict $\mathbf{g}_{(t,k)}$. Note, that at the first level-step, the $\mathcal{T}_{\text{kinematic}}$ receives $\mathbf{g}_{(t,1)}$ from the temporal transformer as input and predicts the next level $\mathbf{g}_{(t,2)}$ and further.

We jointly train both transformers using cross-entropy loss $\mathcal{L}_{\text{CE}}$ over the ground-truth tokens and employ teacher-forcing during the training process.

3.4. Improving Expressiveness

Autoregressive architectures for motion synthesis excel at generating coherent motion sequences, especially in causal scenarios. However, they tend to converge to mean-poses and accumulate drifts along the temporal dimension [10, 20]. To obtain expressive gestures, we introduce auxiliary objectives that explicitly encourage motion diversity and prevent collapse into static or repetitive gestures.

During training, we apply a contrastive InfoNCE loss [36] over the predicted tokens to improve gesture expressiveness. However, sampling from a discrete distribution is non-differentiable and will not allow this loss to contribute during training. Hence, we resort to the Gumbel-Softmax reparameterization trick [18] to approximate the discrete sampling process. This allows us to obtain probabilities from the logit outputs $\tilde{\mathbf{o}} \in \mathbb{R}^{T\times K\times V}$ of temporal and kinematic transformers, which are then converted to the latent output of RVQ step (from Sec. 3.2):

$$\mathbf{z} = \sum_{k=1}^{K} \text{GumbelSoftmax}(\tilde{\mathbf{o}}_k)\mathbf{C}_k \in \mathbb{R}^{T\times d} \tag{4}$$

We use GumbelSoftmax with a temperature of 0.4 and sample one-hot vectors at its output using differentiable straight-through estimator. The latent output $\mathbf{z}$ is calculated separately for each body region by using its corresponding RVQ codebooks. Given ground-truth latents $\mathbf{z}^{\text{GT}}$ and generated latents $\mathbf{z}^{\text{pred}}$, we compute a similarity matrix across all real-fake pairs and apply an Info-NCE loss:

$$\mathcal{L}_{\text{con}} = -\mathbb{E}_i\left[\log \frac{\exp(\text{sim}(\mathbf{z}^{\text{GT}}_i, \mathbf{z}^{\text{pred}}_i)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(\mathbf{z}^{\text{GT}}_i, \mathbf{z}^{\text{pred}}_j)/\tau)}\right]. \tag{5}$$

where $\text{sim}(\cdot)$ denotes cosine similarity and τ is the temperature parameter. This loss enforces high similarity between matching GT–predicted latents while pushing apart mismatched pairs across the batch B, resulting in more expressive and speech-aligned motion generation.

In practice, we apply this loss over temporal segments of $\mathbf{z}$ instead of the complete temporal length T, in order to encourage similarity in motion trajectories across gesture phases.

Voice Activation Loss. Our framework generalizes to both listening and speaking states of body gestures. Since, humans gesticulate differently while listening or speaking, we explicitly enforce our network to learn the distinction

between the two states. This is achieved by projecting the transformer output $\mathbf{h}_t$ onto a binary-classification head that classifies $\mathbf{h}_t$ into listening (0) or speaking (1) states. Trained with a Binary Cross-Entropy loss $\mathcal{L}_{\text{va}}$, this auxiliary task head prevents *phantom* gestures during the listening state and forces speech-aligned expressive gestures during the speaking stage.

Finally, the complete network is optimized through joint loss $\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{con}} + \beta\mathcal{L}_{\text{va}}$, with α, β being loss weights.

3.5. Implementation

To enable gesture generation in real-time that is time-aligned with Moshi, we implement efficient techniques to achieve faster synthesis times. Moshi's latency is 200ms at a rate of 12.5 tokens per second, where each token represents 0.08 seconds of audio, and hence, MIBURI also generates 0.08 seconds of gestures and facial expressions at each timestep. Our training data contains 25 FPS motion, which means our framework generates 2 frames at each step. Gesture codecs contain $K_u = K_l = 8$ and $K_f = 4$ residual levels and we use $T = 125$ during training which amounts to a 10-second motion sequence. Temporal transformer consists of 4 layers with 2 attention heads and the kinematic transformer consists of 2 layers and 1 attention head. Training optimization is done using AdamW [28] with starting learning rate of 1e-4, which is annealed across epochs.

For efficient attention inference, we store key and values for previous timesteps in a KV-Cache to retain the context required during attention. We limit the attention context of self-attention layers to 25 tokens and keep a longer context of 50 tokens for cross-attention layers with speech and text. The temporal transformer starts inference with a zero initial token to predict $\mathbf{g}_{(1,1)}$ (refer to Fig. 3). In practice, due to the small relationship between the lower body and speech/text, we mask out cross-attention for lower body tokens to save runtime. During training, we set α and β to 0.1 and 0.01, respectively. At inference time, we generate tokens using top-p (nucleus) sampling [17], instead of greedy sampling, to maintain diversity. We set top-p for the temporal transformer to 0.8 and for the kinematic transformer to 0.95, with the softmax temperature of 0.9 for both. Moreover, we apply classifier-free guidance (CFG) [16] during sampling to improve gesture alignment with Moshi's rich semantic and acoustic information.

4. Experiments

We evaluate our approach against state-of-the-art baselines for co-speech gesture synthesis. We perform perceptual (Sec. 4.1) and quantitative (Sec. 4.2) evaluations to measure gesture quality, motion naturalness and speech appropriateness. Moreover, we also analyze generation times for each baseline to measure real-time capability. Lastly, we validate our design choices through ablative analysis (Sec. 4.4).

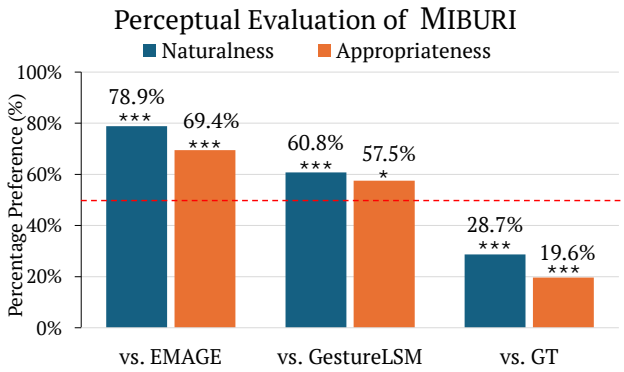

Figure 4. **User Study for Perceptual Evaluation.** Here, the red line indicates chance level (50%), * stands for $p < 0.05$ and *** for $p < 0.001$.

Baseline methods include two types of approaches: (1) Non-causal and non real-time approaches like RAG-Gesture [32], EMAGE [25] and CaMN [24], which aim to synthesize meaningful expressive motion, and (2) Real-time approaches like GestureLSM [26] and MambaTalk [47], which have fast sampling times during generation. Since there are no causal neural baselines, we also implement causal versions of real-time methods [26, 47] to compare our approach with naïve implementations of causal gesture synthesis (details in supplemental). It is important to note that all baselines (except [32]) require a seed sequence and leverage its context to generate motion, whereas our framework does not.

Dataset. We train our approach on the BEAT2 dataset [25] and evaluate its performance on standard train/val/test split from the dataset. The dataset originally contains 25 speakers, but we remove 2 speakers (*carla* & *itoi*) to ensure good quality motion-tracking for our training/evaluation data. Following [32] and unlike other baseline BEAT2 methods, we evaluate on both single-speaker (*scott*) and multi-speaker test sets to assess performance on large-scale multi-speaker setting. Our test set contains 15 and 249 full-length utterances for 1-speaker and 23-speaker setting respectively. We retrain baselines for the multi-speaker setting if their multi-speaker variant is unreleased. Lastly, we also provide an evaluation on the recently released Embody3D dataset [29] in the supplemental material.

4.1. Perceptual Evaluation

Quantitative metrics focus on singular aspects of the gesture generation problem, and have yet to represent correlation with human perception of gestures [34]. Therefore, we perform a perceptual evaluation on BEAT2 test set to holistically evaluate aspects of gesture synthesis like *naturalness* of motion and *appropriateness* to given speech (Fig. 4). Par-

ticipants perform pair-wise comparison between MIBURI's gesture outputs and generations from baseline methods. Results demonstrate MIBURI's ability to generate expressive and natural motion over standard non-causal baselines like EMAGE [25] and GestureLSM [26]. However, we observe that our framework has yet to achieve similar quality and speech appropriateness against ground truth data. Further details are given in the supplementary material.

4.2. Quantitative Evaluation

Evaluation metrics for gesture synthesis include Beat-Alignment [22], Frechet Gesture Distance [48], L1 Divergence and Diversity. Each metric aims to measure a specific aspect of gesture quality, with FGD measuring distribution alignment to ground-truth data and BeatAlign gauging prosodic alignment of motion with speech. To be consistent with BEAT2 baseline methods, we first evaluate our approach on single speaker setting on BEAT2 (Tab. 3) and then perform multi-speaker evaluation across 23 speakers (Tab. 2). For single-speaker training data, we observe comparable performance against non-causal baselines in terms of BeatAlign. Methods which generate gestures from ground-truth seed sequences understandably perform better in single-speaker setting, achieving lower FID. We set MIBURI's CFG scale to 1.5 in single speaker setting.

More importantly, our framework achieves state-of-the-art metric performance in FGD and BeatAlign, when trained with a larger number of speakers. Firstly, this entails that our causal approach benefits from larger and more diverse motion data and scales well across multiple identities, without the need for seed sequences and future context. Secondly, when comparing causal versions of existing methods, we find that naïvely converting baselines to be trained in a causal fashion, leads to worse performance even if the method is real-time. This also shows current architectures' dependance on future speech context in order to achieve good quality. Lastly, we also trained larger versions of MIBURI in this setting to gauge the effect of model sizes without being limited from real-time constraints. However, we find that leaner versions of MIBURI are equivalent or better. We set the CFG scale to 2.3 in multi-speaker setting.

4.3. Latency Analysis

Recall that having low latency is critical for enabling seamless interactions with the end-user. Consequently, keeping the latency low has been one of the key design considerations in our method. Our online demo system achieves a latency of **36ms per frame** on RTX3090. This includes model's runtime and rendering on a web dashboard (see Suppl. Mat.). Moreover, we present a comparative analysis of MIBURI's latency with respect to existing state-of-the-art methods in Tab. 4. Having a low token context (2 frames) helps our autoregressive design and we achieve the lowest latency. In contrast, non-autoregressive diffusion-based methods need to wait for all the context-frames to be generated in order to render the output, thereby leading to high latency. Interestingly, while MambaTalk [47] is based on the inherently causal Mamba [14] architecture, their decision to inject speech conditioning through a cross-attention layer becomes counter-productive for generating low-latency outputs. Our proposed MIBURI architecture strikes a balance between gesture quality and generation latency.

Table 2. **Multi-speaker evaluation.** Facial-MSE scaled by 10^{-8}. * refers to retrained methods.

	Multiple speakers (23)			
	FGD↓	BeatAlign→	L1-Div→	Facial-MSE↓
GT		0.446	8.45	
CaMN	0.736	0.176	6.73	–
EMAGE*	0.850	0.236	6.58	4.6
RAG-Gesture	0.515	0.648	10.09	–
GestureLSM	0.537	0.481	8.41	–
GestureLSM (Causal*)	2.792	0.684	9.11	–
MambaTalk*	1.375	0.080	3.73	4.12
MambaTalk (Causal*)	1.222	0.102	4.61	4.17
MIBURI-L	0.555	0.431	9.45	–
MIBURI-L (+Face)	0.582	0.434	9.31	7.63
MIBURI	0.585	0.415	9.75	–
MIBURI(+Face)	0.480	0.461	10.44	7.77

Table 3. **Single-speaker evaluation.** Facial-MSE scaled by 10^{-8}.

	Single-speaker (*Scott*)			
	FGD↓	BeatAlign→	L1-Div→	Facial-MSE↓
GT		0.749	13.22	–
CaMN	0.969	0.698	10.61	–
EMAGE	0.552	0.795	13.06	7.68
RAG-Gesture	0.879	0.730	12.62	–
GestureLSM	0.410	0.719	13.42	–
GestureLSM (+Face)	0.424	0.729	13.76	10.20
MambaTalk	0.530	0.779	12.99	6.25
MIBURI	0.806	0.790	17.5	–
MIBURI(+Face)	0.753	0.790	15.85	8.85

Table 4. **Latency and Causality Comparison.** Wall-clock time is measured from the beginning of the forward pass to the conversion of outputs into SMPL-X parameters. Render times are excluded here. #Frames / Step indicates the number of frames generated per forward pass.

	Causal	Latency$_{\text{A100}}$↓	#Frames / Step
GestureLSM (8 steps)	✗	0.1447 ± 0.0034	124
EMAGE	✗	0.0374 ± 0.0004	60
MambaTalk	✗	0.0529 ± 0.0039	60
MIBURI (ours)	✓	0.0349 ± 0.0017	2

4.4. Ablation Studies

We perform ablative analysis over different aspects of our framework, ranging from choice of speech encodings, architecture/loss design and motion tokenization strategy.

Table 5. **Wav2vec ablation against moshi features.**

	FGD↓	BeatAlign→	L1-Div→
GT	–	0.446	8.45
MIBURI-L (+Face) w/ wav2vec	0.595	0.404	7.92
MIBURI(+Face) w/ wav2vec	0.665	0.363	7.07
MIBURI-L (+Face)	0.582	0.434	9.31
MIBURI(+Face)	0.480	0.461	10.44

Table 6. **Comparison of Model Variants on Gesture Generation and Runtime.**

	FGD↓	BeatAlign→	L1-Div→	Step Time (s)↓
GT		0.446	8.45	
Single Transformer	1.256	0.731	5.48	0.096
Ours	0.480	0.461	10.44	0.035

Comparison of Speech/Text Encodings. Since existing systems utilize a multi-step pipeline to generate body gestures in ECAs (Fig. 2), we analyze the most important part of that pipeline for gesture synthesis i.e. speech input encoding, and compare it with our approach of leveraging Moshi's internal token stream. We compare the performance of our gesture synthesis model MIBURI, by training it with internal embeddings of Moshi tokens and also, by using standard wav2vec [3] based encoding, which is common in gesture synthesis frameworks [32]. Tab. 5 shows higher FGD and worse BeatAlign scores when using wav2vec, which also incurs an additional computation cost of computing audio embeddings. In contrast, using Moshi [11]'s internal text and speech token stream gives us better quantitative metrics and saves time for encoding and decoding speech.

Two-dimensional Transformer Design. We ablate our design choice of using a two-tier arrangement of temporal and kinematic transformers. As discussed in Sec. 3, the disadvantage of using a single stream for both dimensions T and K is the scale-up in context-length of attention layers. This manifests itself during training in terms of bad convergence, leading to overall worse performance in metrics. Tab. 6 demonstrates that using a single transformer results in higher FID, worse BeatAlign scores and lower diversity. Not to mention, the step times are almost doubled due to increased attention context.

Effect of additional losses. We ablate the contribution of auxiliary losses to our training by evaluating final models on the evaluation sets. Our base losses consist of $\mathcal{L}_{CE}$ and $\mathcal{L}_{\text{va}}$. We evaluate two different losses that are applied on estimated latents and ground-truth: (1) contrastive loss $\mathcal{L}_{\text{con}}$ and (2) MSE-loss. Tab. 7 shows that contrastive loss improves FGD from the base setup of cross-entropy loss, while applying direct MSE on estimated latents increases FGD.

Table 7. **Quantitative Effect of Losses on Generation.**

	FGD↓	BeatAlign→	L1-Div→
GT		0.446	8.45
$\mathcal{L}_{CE} + \mathcal{L}_{\text{va}}$	0.499	0.450	10.25
with MSE-loss	0.577	0.438	9.79
with $\mathcal{L}_{\text{con}}$	0.480	0.461	10.44

Table 8. **Effect of Number of Codebooks K on Motion Reconstruction.** MPJPE is represented in meters.

	FGD↓	MPJPE (m)↓
K=1	0.55	0.043
K=2	0.42	0.032
K=4	0.135	0.022
K=8	0.059	0.016

Evaluation of Gesture Codec across K levels. Since we divide our gesture token structure in K levels to represent finer kinematic details, we evaluate how many of these levels are necessary for gesture tokenization. Tab. 8 demonstrates the relation between the increasing number of levels and reconstruction quality. We report Mean Per-Joint Position Error as a metric to evaluate the reconstruction quality for varying number of K levels. Lastly, we observe that generative FGD also follows a similar pattern as MPJPE.

5. Limitations & Future Work

Our current framework models only the agent's motion and does not incorporate the user's body dynamics or full dyadic context, limiting its ability to handle interactive, multi-party gestures. Extending MIBURI to perceive and respond to a partner's gestures is an important direction for future work.

6. Conclusion

In this work, we present MIBURI – an online, causal framework for generating expressive co-speech gestures and facial expressions synchronized with real-time dialogue. Through body-part–aware gesture codecs and a two-dimensional causal generator, our method models both temporal and kinematic motion structure at low latency. Contrastive objectives further enhance gesture diversity and expressiveness. Experiments across single- and multi-speaker settings show that MIBURI produces natural, contextually aligned gestures and outperforms recent baselines. Our approach moves ECAs closer to truly interactive, human-like embodied communication.

Acknowledgments. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2853/1 “Neuroexplicit Models of Language, Vision, and Action” - project number 471607914. We also thank Anton Zubekhin & Andrea Boscolo Camiletto for their help with the demo.

References

[1] Louis Abel, Vincent Colotte, and Slim Ouni. Towards realtime co-speech gestures synthesis using stargate. In *25th Interspeech Conference (INTERSPEECH 2024)*, 2024. 3

[2] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM TOG*, 42(4): 1–18, 2023. 2

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 8

[4] Timothy Bickmore and Justine Cassell. Social dialogue with embodied conversational agents. *Advances in natural multimodal dialogue systems*, 30:23–54, 2005. 2, 3

[5] Zhongang Cai, Jianping Jiang, Zhongfei Qing, Xinying Guo, Mingyuan Zhang, Zhengyu Lin, Haiyi Mei, Chen Wei, Ruisi Wang, Wanqi Yin, Liang Pan, Xiangyu Fan, Han Du, Peng Gao, Zhitao Yang, Yang Gao, Jiaqi Li, Tianxiang Ren, Yukun Wei, Xiaogang Wang, Chen Change Loy, Lei Yang, and Ziwei Liu. Digital life project: Autonomous 3d characters with social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 582–592, 2024. 3

[6] Justine Cassell. Embodied conversational interface agents. *Commun. ACM*, 2000. 3

[7] Justine Cassell, Timothy Bickmore, Mark Billinghurst, Lee Campbell, Kenny Chang, and Hao Yan. An architecture for embodied conversational characters. In *Proceedings of the First Workshop on Embodied Conversational Characters*, 1998. 1, 2, 3

[8] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. Beat: The behavior expression animation toolkit. In *SIGGRAPH Conference Proceedings*, 2001. 2

[9] Jianchuan Chen, Jingchuan Hu, Gaige Wang, Zhonghua Jiang, Tiansong Zhou, Zhiwen Chen, and Chengfei Lv. Taoavatar: Real-time lifelike full-body talking avatars for augmented reality via 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10723–10734, 2025. 2, 3

[10] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *CVPR*, 2023. 5

[11] 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. 2, 3, 5, 8

[12] Ylva Ferstl, Michael Neff, and Rachel McDonnell. Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics*, 89:117–130, 2020. 2

[13] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. *Computer Graphics Forum*, 42(1):206–216, 2023. 2

[14] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 7

[15] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the International Conference on Intelligent Virtual Agents*, 2021. 2

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[17] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019. 6

[18] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 5

[19] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019. 2

[20] Taras Kucherenko, Patrik Jonell, Sanne Van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020. 3, 5

[21] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11523–11532, 2022. 5

[22] Ruilong Li, Sha Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021. 7

[23] 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. 4

[24] 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 *ECCV*, 2022. 2, 6

[25] 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. 2, 3, 4, 6, 7

[26] 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. 3, 6, 7

[27] Jose Llanes-Jurado, Lucía Gómez-Zaragozá, Maria Eleonora Minissi, Mariano Alcañiz, and Javier Marín-Morales. Developing conversational virtual humans for social emotion elicitation based on large language models. *Expert Systems with Applications*, 246:123261, 2024. 3

[28] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[29] 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. 6

[30] David McNeill. *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, 1992. 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 *CVPR*, 2024. 2, 3, 4

[32] 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, 3, 4, 6, 8

[33] Rajmund Nagy, Taras Kucherenko, Birger Moell, André Pereira, Hedvig Kjellström, and Ulysses Bernardet. A framework for integrating gesture generation models into interactive conversational agents. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2021. 2, 3

[34] Rajmund Nagy, Hendric Voss, Thanh Hoang-Minh, Mihail Tsakov, Teodor Nikolov, Zeyi Zhang, Tenglong Ao, Sicheng Yang, Shaoli Huang, Yongkang Cheng, et al. Towards reliable human evaluations in gesture generation: Insights from a community-driven state-of-the-art benchmark. *arXiv preprint arXiv:2511.01233*, 2025. 6

[35] 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 *CVPR*, 2024. 2, 3

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[37] OpenAI. Gpt-4o system card, 2024. 3

[38] OpenAI, Achiam J, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report, 2024. *arXiv preprint arXiv:2303.08774*, 2024. 3

[39] 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. 4

[40] Maha Salem, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In *2011 ro-man*, pages 247–252, 2011. 1

[41] Varsha Suresh, M. Hamza Mughal, Christian Theobalt, and Vera Demberg. Enhancing spoken discourse modeling in language models using gestural cues. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18109–18123, 2025. 4

[42] 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. 4

[43] Marcus Thiebaux, Stacy Marsella, Andrew N Marshall, and Marcelo Kallmann. Smartbody: Behavior realization for embodied conversational agents. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, 2008. 2

[44] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 3

[45] Petra Wagner, Zofia Malisz, and Stefan Kopp. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232, 2014. 2

[46] Rafael Wampfler, Chen Yang, Dillon Elste, Nikola Kovacevic, Philine Witzig, and Markus Gross. A platform for interactive ai character experiences. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025. 3

[47] 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, 6, 7

[48] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, 2019. 2, 7

[49] 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. 3, 4

[50] 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, 3

[51] Weiyu Zhao, Liangxiao Hu, and Shengping Zhang. Diffugesture: Generating human gesture from two-person

dialogue with diffusion models. In *International Conference on Multimodal Interaction*, 2023. 2

[52] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 4

[53] Yongxin Zhu, Dan Su, Liqiang He, Linli Xu, and Dong Yu. Generative pre-trained speech language model with efficient hierarchical transformer. *arXiv preprint arXiv:2406.00976*, 2024. 5