Learning Multimodal Motion Cues for Online End-of-Turn Prediction in Multi-Party Dialogue

Meng-Chen Lee
Department of Computer Science
University of Houston
Houston, Texas, USA
mlee45@uh.edu

Zhigang Deng
Department of Computer Science
University of Houston
Houston, Texas, USA
zdeng4@central.uh.edu

ABSTRACT

Accurate end-of-turn prediction in multiparty conversations is essential for enabling dialogue systems to participate in fluid and socially responsive interactions. While prior work has explored linguistic, prosodic, and gaze-based cues, the role of continuous bodily motion in modeling turn transitions remains relatively underexplored. This study investigates the contribution of head, hand, and full-body movements by learning symbolic motion representations through Vector Quantized Variational Autoencoders (VQ-VAE). Each motion modality is encoded independently into a discrete latent space, allowing us to assess their individual and combined predictive value for end-of-turn classification. Using a triadic conversation dataset with synchronized audio, gaze, and motion streams, we evaluate the independent and combined contributions of each motion modality. Our results show that hand motion, particularly when combined with gestural backchannel features, significantly improves performance. In contrast, head motion encoded via VQ-VAE provides only marginal gains and may overlap with discrete gesture labels. Compared to prior work, our model achieves higher precision, recall, and F1-score while maintaining real-time inference speed. These findings highlight the potential of structured, data-driven motion embeddings in developing socially aware dialogue systems.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Collaborative interaction**.

KEYWORDS

multi-party conversations; multimodal interaction; non-verbal gesture; human-human interaction; turn prediction

ACM Reference Format:
Meng-Chen Lee and Zhigang Deng. 2025. Learning Multimodal Motion Cues for Online End-of-Turn Prediction in Multi-Party Dialogue. In *Proceedings of the 27th International Conference on Multimodal Interaction (ICMI '25), October 13–17, 2025, Canberra, ACT, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3716553.3750756

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICMI '25, October 13–17, 2025, Canberra, ACT, Australia
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1499-3/2025/10
https://doi.org/10.1145/3716553.3750756

1 INTRODUCTION

In human communication, turn taking is a fundamental mechanism governing the natural flow of conversation. In multiparty settings, where interactions involve more than two individuals, this process becomes particularly complex due to the dynamic interplay of verbal and nonverbal signals exchanged among participants. These interactions are shaped not only by what is said but also by how and when it is expressed through voice, gaze, body posture, and gestures. Understanding and modeling these multimodal behaviors has become a key area of interest across diverse research fields including conversational AI, human-computer interaction, and social signal processing [16, 22, 31, 46]. In particular, the task of predicting end-of-turn (EoT) moments, which determines when a speaker is likely to yield the floor, is critical for developing conversational systems capable of seamless and socially aware interactions [15, 48].

As dialogue systems increasingly shift toward real-time and multiuser applications, accurate EoT prediction in multiparty conversations plays a central role in managing speech coordination and conversational floor control [5, 45]. This task extends beyond merely detecting silences or pauses; it requires interpreting complex conversational cues to assess whether a speaker is concluding their turn, pausing to think, addressing a specific interlocutor, or expecting an interruption. The challenge is heightened in multiparty scenarios where decisions must account for speaker roles, listener dynamics, and potential overlaps or interruptions [10, 12]. Incorrect predictions in turn transitions can result in unnatural interactions, such as prematurely interrupting speakers or delayed responses, ultimately degrading the user experience [36, 37].

Recent research has demonstrated the advantages of incorporating multimodal signals, such as prosodic features, gaze behavior, and discrete gestural annotations like nods and shakes, to enhance EoT prediction [30]. However, quantifying the contribution of continuous bodily motion, especially full-body and hand gestures beyond simple binary labels, remains largely unexplored. Continuous motion data provides rich temporal and spatial details signaling engagement, hesitation, or readiness to speak [2, 43], nuances frequently lost in discrete annotations.

To address these limitations, this study introduces a comprehensive framework for online EoT prediction in multiparty conversations, emphasizing the *integration of learned motion representations into a multimodal architecture*. Specifically, we investigate the predictive value of head, hand, and full-body motion encoded using Vector Quantized Variational Autoencoders (VQ-VAE) [13], which transform high-dimensional motion sequences into discrete symbolic spaces. By modeling each motion modality separately, we examine their individual and combined contributions to predicting

turn transitions. Additionally, we compare these learned representations against traditional binary gesture annotations to assess whether VQ-VAE effectively captures more subtle conversational signals.

Recognizing that lightweight architectures such as distilled pretrained language models like DistilBERT combined with Gated Recurrent Units (GRU) efficiently handle EoT prediction tasks with minimal inference latency [30], we maintain this fundamental design for real-time responsiveness. Simultaneously, we extend the architecture by incorporating structured motion embeddings and cross-modal fusion strategies to enhance predictive accuracy, without sacrificing system efficiency.

Our approach is evaluated on a triadic conversation dataset featuring synchronized multimodal streams, including audio, text, gaze, and motion. We focus specifically on predicting various EoT events such as regular turn-taking, interruptions, and speech overlaps. Through extensive experiments, we demonstrate that full-body and hand motions provide complementary information that enhances predictive accuracy, and that VQ-VAE-encoded head motion slightly increase the performance than heuristic gesture annotations. These findings highlight the potential of structured motion representations in real-time multimodal dialogue systems, contributing to more adaptive and socially intelligent human-computer interactions.

Contributions. The main contributions of this work include:

- Proposing a real-time end-of-turn prediction framework for multi-party conversations that achieves a low latency by leveraging a distilled pretrained language model (DistilBERT) and GRU, while enhancing performance with symbolic motion representations.
- Introducing VQ-VAE to independently encode head, hand, and full-body motion into discrete latent spaces, enabling structured modeling of continuous gestural behaviors.
- Evaluating our model on triadic conversational data and demonstrating that hand motions provide extensive and valuable predictive cues.

2 RELATED WORK

IPU-based Turn Prediction Models. Early work on EoT prediction often adopted an Inter-Pausal Unit (IPU) framework, where utterances are segmented based on silence thresholds, typically 200ms or longer, to delineate potential turn boundaries [28]. Rule-based approaches such as semantic parsers [1] and early data-driven models like decision trees [41] were used to classify long pauses using syntactic, semantic, and prosodic cues. Subsequent studies refined these methods by shortening pause durations and integrating additional behavioral features [35, 42].

In a series of influential works, Ishii et al. [19–21] extended IPU-based approaches by incorporating multimodal cues such as gaze transitions, respiratory rhythms, and head movements to improve EoT and speaker-change prediction. These studies demonstrated that modeling mutual gaze, breathing, and speaker/listener head motion toward the end of an utterance can significantly enhance performance. Recently, Lee et al. [32] introduced the concept of Relative Engagement Level (REL) to predict turn-taking behavior in multiparty conversations, again using an IPU-based formulation.

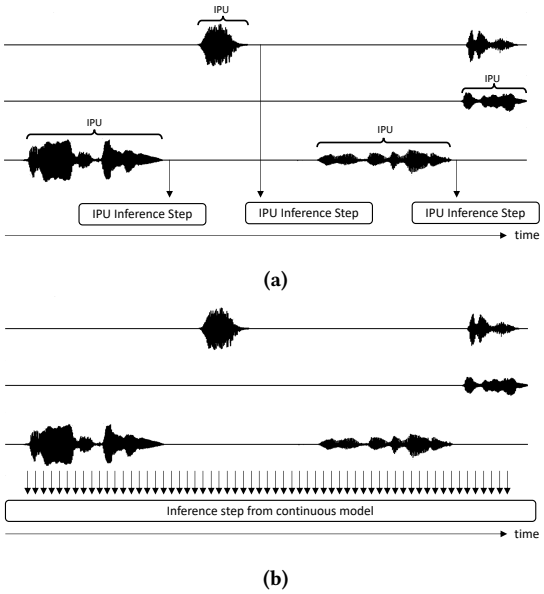


Figure 1: Illustration of the inference process for (a) IPU-based models and (b) continuous models. The IPU-based model makes predictions at the end of each Inter-Pausal Unit (IPU), while the continuous model operates over the entire speech stream without explicit segmentation.

Despite their conceptual clarity, IPU-based models are limited in interactive settings. Because predictions are triggered only after detecting a pause, these methods introduce a latency that undermines real-time responsiveness. Moreover, they often fail to capture nuanced timing cues necessary for handling interruptions and overlaps [29]. Figure 1(a) illustrates the contrast between IPU-based and continuous prediction paradigms.

Continuous Turn Prediction Models. To overcome the latency of IPU-based methods, continuous prediction models were proposed to forecast EoT events at regular intervals (e.g., every 500ms) [17, 29, 38, 44]. These models treat turn-taking as a real-time sequence modeling task, processing multimodal features frame by frame to anticipate future speech activity. Skantze [44] introduced a continuous turn-taking model based on prosodic cues using Long Short-Term Memory (LSTM) models to forecast speech activity for both speakers over a 3-second horizon. Roddy et al. [38] expanded this framework with a multiscale LSTM to fuse modalities at different temporal resolutions, demonstrating improved performance on dyadic data.

However, these models are predominantly designed for dyadic conversations and often struggle to generalize to multiparty scenarios, where speaker roles, addressee dynamics, and floor control are inherently more complex. Prior attempts to extend continuous multimodal modeling to multiparty contexts (e.g., [8]) have encountered scalability and performance challenges, highlighting the need for more robust modeling of diverse social signals.

Multimodal Cues for End-of-Turn. Research on EoT cues has emphasized the additive role of multimodal features in signaling conversational intent. Verbal indicators such as syntactic

completion and semantic closure have long been established as key turn-yielding signals [14, 39]. Prosodic features including intonation, loudness, energy, and pitch contours have also been shown to reliably distinguish turn-holding from turn-yielding behavior [15, 28, 48]. Visual signals, especially gaze, play a central role: speakers often look toward interlocutors to yield the floor and avert their gaze to hold it [9, 18].

While previous work has explored discrete gestures such as head nods and shakes to improve EoT prediction [30], the potential of continuous motion, particularly full-body and hand gestures, remains unexplored. These signals can reflect subtle behavioral cues such as hesitation, emphasis, or readiness to speak, which may not be captured through binary labels or static pose features.

Learned Motion Representations. Recent advances in generative modeling have led to increased interest in learning symbolic representations of human motion by discretizing continuous sequences into tokenized forms. These approaches offer an appealing alternative to traditional coordinate-based models by enabling compact, interpretable, and reusable embeddings for downstream tasks. A prominent technique in this domain is VQ-VAE [13, 47], which encode high-dimensional input sequences into discrete latent codes from a learned codebook.

In gesture and behavior synthesis, VQ-VAE has emerged as a popular technique for discretizing continuous motion into symbolic representations that support controllable and diverse generation. Bhattacharya et al. [3] proposed a text-to-gesture pipeline that first learns a codebook of motion primitives using VQ-VAE and then maps discrete gesture tokens to continuous gestures using a transformer-based decoder. Similarly, Yi et al. [50] developed a speech-driven gesture generation model that utilizes quantized embeddings to produce stylized and semantically aligned gestures. Yazdian et al. [49] introduced a discrete gesture embedding space that enables retrieval and synthesis, while Liu et al. [33] leveraged VQ-VAE to disentangle gesture style and content from video data for flexible gesture generation. These approaches illustrate the value of symbolic motion representations in modeling expressive human behavior and enabling high-level control over gesture synthesis.

Despite these advances, the application of discrete motion representations has largely focused on generative tasks, such as gesture synthesis or action generation. Their use in predictive settings, such as turn-taking or engagement modeling in multiparty dialogue, remains limited. Our work addresses this gap by applying VQ-VAE to encode head, hand, and full-body motion into discrete latent spaces and integrating these symbolic representations into a real-time multimodal EoT prediction model. Our approach allows for richer, data-driven modeling of non-verbal behaviors that are not easily captured by predefined gestures or raw feature aggregation. To the best of our knowledge, this is the first application of VQ-VAE based motion representations for EoT prediction in multiparty conversations, and it significantly expands the scope of motion-informed interaction modeling.

3 DATA ACQUISITION

We constructed an in-house multimodal dataset capturing face-to-face three-party conversations, originally collected for our earlier study presented at ICMI 2023 [32]. In this work, we reuse the raw recordings, comprising motion, gaze, and audio streams from the previous study, while introducing new annotations, symbolic motion representations, and downstream modeling components tailored for end-of-turn prediction. The data acquisition setup is illustrated in Figure 2.

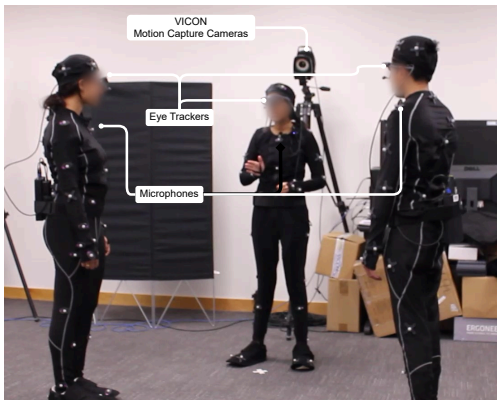


Figure 2: A snapshot from our three-party conversational data acquisition setup.

Motion Capture. Body movements were recorded using a Vicon optical motion capture system with eight high-speed infrared cameras—four mounted above and four below—to ensure comprehensive spatial coverage. Participants wore motion capture suits fitted with reflective markers to track full-body motion, including head, hand, torso, and leg movements. The captured 3D trajectories were processed via inverse kinematics to extract joint angles for downstream analysis (see Figure 2).

Gaze and Audio. Gaze behavior was captured using Ergoneers Dikablis Glass 3 eye trackers (Ergoneers Group, Germany), selected for their lightweight, wireless design and compatibility with D-LAB software. Their eyeglass-like appearance minimized behavioral interference and participant discomfort. Participants reported that the devices had negligible impact on their natural conversational behavior. Each participant also wore a wireless microphone clipped to the neckline, allowing for clear and synchronized audio recording. Gaze and audio data streams were recorded and synchronized using D-LAB.

Capture Design. We recruited 21 participants (12 male, 9 female; aged 20–30) from a university campus and randomly assigned them to seven triads. All participants were native English speakers with no prior familiarity, ensuring spontaneous and unbiased conversational dynamics.

Each group engaged in 3 to 5 unscripted sessions, freely discussing topics of their choice. Participants were positioned in an equilateral triangle with approximately one meter of spacing, consistent with established practices in multiparty conversation research [6, 11, 23–25]. Although they were asked to maintain their positions, natural body movement within the capture space was permitted to preserve interaction authenticity.

Session durations across the seven groups were 46'18", 44'47", 46'17", 42'49", 46'31", 44'27", and 48'19", totaling 319 minutes and 28

Figure 3: Visualization of the GTV computation process. The resulting Direction-of-Focus (DFoc) is depicted as a blue line.

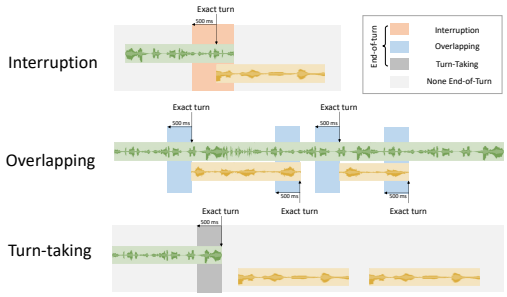


Figure 4: Illustration of three categories for end-of-turn: interruption, overlapping, and turn-taking.

seconds of recorded conversation. All modalities—motion capture, eye tracking, and audio—were synchronized using a clapperboard at the start of each session and uniformly downsampled to 60 Hz to ensure precise temporal alignment.

4 FEATURE EXTRACTION

We extend the multimodal feature extraction pipeline in [30] for this study, which includes frame-level representations of speaker identity, prosody, gaze, and gestural backchanneling. In this section, we briefly describe these components and then elaborate our motion representation strategy.

Interlocutor States. Each frame is assigned a 1×3 *Interlocutor State Vector* (ISV) that represents the communicative state of each interlocutor. Speech activity was extracted using manually corrected transcripts generated by OpenAI Whisper. Following the criteria in [26], we identified active speech and verbal backchanneling using keyword-based rules (e.g., "Yeah," "Um," "Cool," "Oh," "Okay," "Right," and "Uh"). Each ISV entry is labeled as 2 for active speech, 1 for backchanneling, and 0 for silence. All annotations were manually verified by human annotators to ensure accuracy.

Prosody Features. Prosodic cues were extracted using the Praat toolkit [4], focusing on pitch and intensity at 16 ms resolution. Each frame is represented by a 1×6 vector (pitch and intensity for all the three interlocutors), capturing acoustic features that are known to signal turn-yielding or turn-holding behavior.

Gaze Targets. Visual attention was modeled using the Direction-of-Focus (DFoc) metric, computed from the orientation and position of the hips, spine, head, and eyes [24]. Smoothed gaze vectors were projected forward, and their intersections with the bounding spheres of the heads of other interlocutors were used to estimate the interlocutor's focus of attention (illustrated in Figure 3). The resulting 1×3 *Gaze Target Vector* (GTV) specifies which interlocutor each individual is looking at during each frame.

Gestural Backchanneling (GB). Head gestures — including nods and shakes — were detected using yaw and pitch dynamics, through the adaptation of the algorithm in [27]. Frames were classified into stable, transient, or extreme categories based on angular thresholds. We then generated a 1×6 binary *Gestural Backchannel Vector* (GBV), consisting of two 1×3 sub-vectors that indicate head nods and shakes respectively, for each non-speaking interlocutor.

Motion Representations. To effectively model gestural behavior in multiparty conversations, we adopt a symbolic motion representation framework based on VQ-VAE, drawing inspiration from the architecture of T2M-GPT [51]. While traditional approaches typically rely on raw joint trajectories or hand-crafted kinematic features, these representations tend to be noisy, redundant, and difficult to generalize across various interlocutors and conversational contexts. In contrast, our objective is to obtain compact and expressive motion embeddings that capture high-level gestural patterns while enabling efficient integration into downstream EoT prediction models.

To support modality-specific abstraction and analysis, we train separate VQ-VAE models for head, hand, head-and-hand, and full-body motion. Each model encodes its respective motion stream into discrete latent tokens, which are then used as symbolic inputs to our EoT prediction framework, alongside other modalities. This design facilitates a unified and scalable representation of gesture while allowing for fine-grained evaluation of each motion modality. Further details on the architecture and training of the VQ-VAE models are provided in Section 5.1.

5 METHODOLOGY

Label Annotations. Inspired by the prior work on EoT modeling [30], we also categorize EoT events into three distinct subtypes: *interruption*, *overlapping*, and *turn taking*. Unlike traditional IPU based annotation schemes, which typically label turn transitions at coarse boundaries, our method provides high-resolution annotations at 100 millisecond intervals. This design enables our approach to track conversational dynamics at a much finer temporal granularity, offering greater responsiveness and predictive capability.

To reduce the imbalance between positive and negative examples in the training data, we adopt a forward-looking labeling strategy consistent with previous literature [29], where a time frame is labeled as a positive EoT instance if a speaker change is expected to occur within the next 500 milliseconds. This anticipatory framing allows the model to forecast upcoming transitions rather than react to completed ones.

Figure 4 illustrates the three EoT subtypes. (i) *Interruption* occurs when the next interlocutor begins speaking before the current interlocutor has completed the utterance. Common in fast-paced conversations, such instances often reflect high engagement or competitive floor dynamics. In our annotations, we label both the

500 milliseconds leading up to the transition and the overlapping region itself as positive instances. (ii) *Overlapping* describes scenarios in which multiple interlocutors start speaking simultaneously, typically following a silence. This is a frequent phenomenon in multiparty interactions. We explicitly exclude brief listener responses such as backchannels, since they generally do not lead to speaker change. The "overlapping" label captures not only the lead-up to the concurrent speech but also the expected resolution of the overlap. (iii) *Turn-taking* represents smooth (regular) speaker transitions, usually marked by a silence exceeding 200 milliseconds between utterances. This category aligns closely with the traditional understanding of orderly floor exchange.

To accurately reflect the diversity of turn-transition behavior, we formulate the prediction task as a four-class classification problem: *interruption*, *overlapping*, *turn-taking*, and *not end-of-turn*. This multi-class structure enables the model to distinguish between nuanced speaker dynamics rather than collapsing all transitions into a single positive label. By departing from IPU-based timing and adopting a fine-grained and semantically meaningful label design, our framework offers a robust foundation for real-time EoT prediction in natural dialogue. The statistics of these sub-categories in our dataset are presented in Table 1.

Table 1: Statistics of end-of-turn sub-categories in our dataset

Categories	Not End-of-Turn	End-of-Turn		
		Interruption	Overlapping	Turn-Taking
Instances	134,394	18,495	6,982	6,826
Ratio	80.6%	11.1%	4.2%	4.1%

5.1 Motion VQ-VAE

In this work, a gesture motion sequence is denoted as $\mathbf{X} = [\mathbf{x}_f]_{f=1}^{F}$, where each frame $\mathbf{x}_f \in \mathbb{R}^d$ captures both the global translation of the character and the Euler rotations of its joints. Although this representation preserves high-resolution motion details, it is computationally intensive and introduces redundancy, which can affect both model training and inference efficiency.

To obtain a more compact and symbolic representation, we employ VQ-VAE to discretize the motion sequence into a latent space. The goal is to reconstruct the input sequence from a set of learned codebook vectors $\mathbf{C} = [\mathbf{c}_k]_{k=1}^{K}$, where each code vector $\mathbf{c}_k \in \mathbb{R}^{d_c}$ resides in the low-dimensional latent space. The encoder E_{VQ} transforms the input $\mathbf{X}$ into a downsampled latent sequence $\mathbf{Z} = [\mathbf{z}_\ell]_{\ell=1}^{L}$, with $\mathbf{z}_\ell \in \mathbb{R}^{d_c}$ and a downsampling ratio of F/L.

Each latent vector $\mathbf{z}_\ell$ is then quantized by locating the nearest codebook entry,

$$\hat{z_\ell} = \underset{c_k \in C}{\arg\min} \|z_\ell - c_k\|_2, \tag{1}$$

to produce a discrete index sequence $\mathbf{S} = [\mathbf{s}_\ell]_{\ell=1}^{L}$, where each $\mathbf{s}_\ell$ denotes the selected codebook entry. This sequence of symbolic tokens serves as a compact, discrete representation of the original motion, suitable for downstream applications (see Section 5.2).

To reconstruct the motion, each index s_ℓ is mapped back to its corresponding code vector $\mathbf{c}_{s_\ell}$, forming the quantized sequence $\hat{\mathbf{Z}} = [\hat{\mathbf{z}}_\ell]_{\ell=1}^{L}$, which is then passed to the decoder D_{VQ} to recover the motion sequence $\hat{\mathbf{X}}$. The full encoding-decoding pipeline is expressed as:

$$\mathbf{Z} = E_{\text{VQ}}(\mathbf{X}), \quad \hat{\mathbf{X}} = D_{\text{VQ}}(\hat{\mathbf{Z}}). \tag{2}$$

Inspired by the T2M-GPT architecture [51], our VQ-VAE design employs a 1D convolutional encoder composed of residual blocks and ReLU activations, paired with a transpose convolutional decoder. The encoder reduces the temporal resolution of the motion input through stride-based convolutions, producing latent embeddings that are quantized via the learned codebook. The decoder then reconstructs the full-resolution motion sequence from these discrete embeddings. Training is guided by the standard VQ-VAE objective [47], which combines a reconstruction loss with a commitment loss to ensure stability and consistency in the latent space.

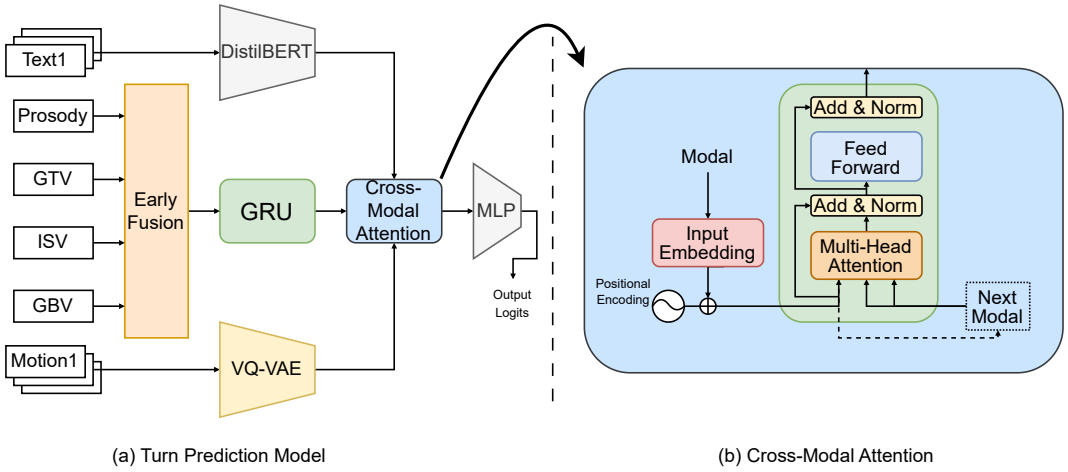
5.2 Turn Prediction Model

Our primary objective is to design an efficient, online-capable (potentially real-time) framework for predicting end-of-turn events in multiparty conversations. Inspired by the recent work [30], we leverage DistilBERT [40], a lightweight transformer-based language model known for its compact architecture and rapid inference capabilities, thus offering an efficient alternative to larger transformer models. Additionally, instead of employing the commonly used LSTM networks for modeling temporal features, we utilize a GRU model. The GRU architecture, characterized by fewer gates and parameters, offers faster computation speeds while maintaining performance comparable to LSTM models [7].

We empirically set our temporal analysis window to 500 milliseconds (ms) and select a step size of 100 ms, implying that our model predicts EoT occurrences every 100 ms based on features captured within the previous 500 ms. To effectively capture linguistic context, we extend the text feature window significantly, incorporating textual content from up to 100 seconds prior to each prediction point. This extended time frame was selected based on empirical observations indicating that shorter windows frequently omit critical conversational elements or encounter silences that hinder accurate interpretation of speaker contributions. For motion data specifically, we synchronize the window length to match the VQ-VAE window size K.

Figure 5 illustrates an architecture overview of our model. For text features, we employ DistilBERT to extract semantic embeddings from input tokens, while motion features are derived from modality-specific VQ-VAE models, as detailed in Section 5.1. Both DistilBERT and the VQ-VAE encoders are frozen during the training of our EoT prediction model. We intentionally avoid fine-tuning these encoders due to the inherent complexity and multimodal nature of our task, thereby mitigating potential overfitting risks associated with the richness and redundancy of multimodal features.

The textual modality is first tokenized, with each token transformed into embeddings that encode semantic and contextual information through DistilBERT's transformer layers. Similarly, the VQ-VAE encoders convert continuous motion sequences into compact, quantized latent embeddings. These contextual embeddings from both modalities are subsequently integrated with additional numeric and time-series features that have undergone Min-Max normalization to maintain consistent feature scales.



(a) Turn Prediction Model (b) Cross-Modal Attention

Figure 5: The architecture of our window-based prediction model

To fuse other inputs effectively, we employ an early fusion strategy, concatenating all feature vectors into a single unified representation. This combined representation is then fed into a single-layer GRU network with a hidden dimension of h units. The GRU model, distinguished by its reset and update gates, efficiently captures temporal dependencies and patterns critical to accurately predicting conversational dynamics.

Following the GRU-based temporal modeling stage, the GRU's final hidden states are combined with the embeddings from the DistilBERT and VQ-VAE encoders. This integrated feature representation undergoes further refinement through a cross-modal attention layer, as illustrated in Figure 5(b), ultimately yielding probabilities for each EoT class.

6 EVALUATION

6.1 System Setup

For motion quantization, we adopted a VQ-VAE model with a codebook size of 512×512. The temporal downsampling ratio was $F/L = 4$, and each motion sequence was segmented to a fixed length of $F = 120$ frames for training. Each frame contained motion features with dimensionality $d = 46$. The model was trained using the AdamW optimizer [34] with $\beta_1 = 0.9$, $\beta_2 = 0.99$, a batch size of 128, and an exponential moving average coefficient of $\lambda = 0.99$. The learning rate was scheduled in stages: 2×10^{-4} for the first 50K iterations, followed by 1×10^{-5} until 350K iterations, and finally decayed to 5×10^{-7} for iterations beyond 400K.

For the end-of-turn prediction model, we used a GRU-based classifier with a hidden state dimensionality of $h = 256$. The input size was either 12 or 18 depending on the inclusion of gestural backchanneling (GB) features. To improve robustness and prevent overfitting, we applied a dropout rate of 0.1 and employed early stopping based on validation performance. The model was trained using the Adam optimizer with an initial learning rate of 0.001 and optimized via the cross-entropy loss function for multiclass classification.

The input feature windows were temporally aligned as follows: text input was processed using a window of 100 seconds, motion features used a 2-second window aligned with the VQ-VAE segment length (F), and all other features used a 500-millisecond window. To mitigate class imbalance, the training data was downsampled, while evaluation was conducted on the original unbalanced test set. We trained all models for 80 epochs to ensure a stable convergence.

All experiments were performed on an off-the-shelf computer equipped with an Intel i9-13900K CPU @ 3.00GHz, 80 GB of RAM, and a single NVIDIA GeForce RTX 4090 GPU.

6.2 Results

To rigorously assess the performance of our model, we implemented a comprehensive and systematic evaluation methodology designed to ensure reliability across diverse conversational scenarios. We employed a 10-fold cross-validation approach, widely recognized for robustly evaluating predictive models. Specifically, our dataset was randomly partitioned into ten equally sized subsets, iteratively using nine subsets for training and the remaining subset for testing. This process was repeated ten times, with each subset serving once as the test set, thereby thoroughly assessing the model's generalization capabilities. To measure the performance comprehensively, we calculated precision, recall, and F1 scores using a macro-average approach, ensuring that each class or conversational dynamic was weighted equally in our evaluation.

Table 2 provides a detailed overview of our model's performance metrics across each EoT category. We observed notable variations in precision, recall, and F1-score across these categories. For example, our model achieved a higher precision (0.805) and recall (0.836) for the interruption category, indicating its strong capability in accurately identifying interruption events. However, the model exhibited comparatively lower precision and recall for overlapping and interruption events, underscoring the inherent difficulty in distinguishing these nuanced conversational behaviors. Although

Table 2: Performance metrics for the four-class categorization, highlighting individual category performance.

Class	Metrics			
	Accuracy	Precision	Recall	F-Measure
Macro-Average	0.933	0.802	0.807	0.804
Not End-of-Turn	-	0.975	0.970	0.972
Interruption	-	0.805	0.836	0.821
Overlapping	-	0.647	0.647	0.647
Turn-Taking	-	0.781	0.770	0.775

Table 3: Quantitative comparison of turn prediction between our model and the state of the art work [30].

Model	Inference time (ms)	Precision	Recall	F-Measure
Lee et al. [30]	2 ± 0.06	0.758	0.755	0.756
Ours	5.53 ± 0.12	0.802	0.807	0.804

overlapping and turn-taking occur with similar frequency, overlapping speech often features rapid and uncoordinated onsets, making it more ambiguous and harder to model. This ambiguity also contributes to reduced precision for the interruption class, as interruptions are frequently misclassified as overlapping events. In contrast, turn-taking typically presents clearer temporal boundaries, resulting in better model performance. Overall, the macro-averaged precision (0.802), recall (0.807), and F1-score (0.804) offer a balanced evaluation of the model's effectiveness across diverse EoT subcategories.

Furthermore, in Table 3, we present a detailed comparative analysis between our model and the previous state of the art method [30] for predicting EoT events in multiparty conversations. This comparison serves as a crucial benchmark for highlighting the improvements achieved by our approach. Our model demonstrates superior performance across all evaluated metrics compared to the state of the art method [30]. Specifically, our approach attains a precision of 0.802, underscoring its accuracy in correctly identifying true end-of-turn events. Additionally, a recall of 0.807 highlights the model's effectiveness in capturing a significant proportion of actual turn transitions. The resulting F1-score of 0.804 further emphasizes the balanced and robust predictive capability of our model, clearly outperforming previous approaches.

6.3 Ablation Study

To investigate the role of motions and gestures in end-of-turn classification, we conducted an ablation study that compare explicit GB features with learned motion representations derived from vector quantized (VQ) embeddings. Specifically, we aimed to evaluate whether removing GB improves performance, whether it can be effectively replaced by latent motion features from head VQ, and how performance varies when using whole-body, hand-only, or head-and-hand VQ representations.

The evaluation results (Figure 4) reveal several important insights regarding the contribution of motion and GB to this task. Overall, incorporating motion data leads to consistent improvements in both macro-averaged and per-class performance metrics. Among all configurations, the combination of hand motion and GB yields the best results, achieving a macro F1-score of 0.8041 and an overall accuracy of 0.9328. This setup provides balanced precision and recall across all four classes, suggesting that hand motion enriched with GB offers strong and discriminative cues for distinguishing between different turn-taking behaviors.

In contrast, combining GB with head-only motion results in a noticeably lower performance, with the macro F1 decreasing to 0.6980. This outcome may reflect an overreliance on head movements alone, which may be insufficient to capture more complex interaction dynamics. Interestingly, when GB is removed, the performance of head-only motion improves significantly, with the macro F1 rising to 0.7787. This indicates a potential redundancy or interference between the explicit GB features and the latent head motion representations extracted by the VQ-VAE model. In fact, most configurations that include gestural backchanneling show a reduction in performance unless paired specifically with hand-only motion. This suggests that the combination of GB and head VQ provides overlapping signals that may not be complementary. Notably, the latent representation of head motion, although capable of capturing a broader range of semantic motion features, does not outperform the discrete GB features. This observation implies that head nodding and shaking alone likely represent the most salient and informative components of head motion for this task.

Furthermore, we observe that when both head and hand VQ or whole-body VQ representations are included, performance often declines relative to using hand-only motion. This pattern indicates that hand motion serves as the most critical nonverbal cue for end-of-turn detection, while the inclusion of additional body regions may introduce redundancy rather than improve the model. Taken together, these findings highlight the importance of selecting complementary and non-overlapping motion signals and suggest that discrete GB combined with hand VQ offers the most effective configuration for multimodal end-of-turn prediction.

Nevertheless, as shown in Table 5, all model configurations achieve inference speeds compatible with real-time applications, demonstrating their suitability for deployment in interactive systems. The reported inference time includes motion vector quantization, text embedding extraction, and multimodal fusion. However, due to the offline nature of our data collection, we do not account for the additional latency introduced by upstream processes such as speech-to-text transcription, eye gaze tracking, or motion capture. In practice, researchers may choose different sensing and processing pipelines based on their specific requirements, which can result in varying latency.

7 DISCUSSION AND CONCLUSION

We present a real-time multimodal framework for end-of-turn (EoT) prediction in multiparty conversations, with a particular emphasis on learning symbolic gesture representations through VQ-VAE. By systematically evaluating head, hand, and full-body motion representations, both independently and in combination with traditional gestural backchannel features, we uncovered several key insights into the role of motion in conversational dynamics.

Table 4: Full evaluation results for all configurations, showing accuracy (Acc), macro-averaged scores, and per-class precision (P), recall (R), and F1.

GB	Motion	Overall				Not End-of-Turn			Interruption			Overlapping			Turn-Taking		
		Acc	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
False	-	0.895	0.6901	0.8216	0.7379	0.9855	0.916	0.9495	0.802	0.8438	0.8223	0.5056	0.6618	0.5732	0.4672	0.8649	0.6066
True		0.9052	0.7132	0.8329	0.7606	0.9817	0.9265	0.9533	0.8191	0.849	0.8338	0.47	0.6912	0.5595	0.5818	0.8649	0.6957
False	whole	0.9172	0.7736	0.8161	0.7898	0.976	0.9467	0.9612	0.7714	**0.8438**	0.806	0.4792	0.6765	0.561	0.8676	0.7973	0.831
True		0.9016	0.755	0.8075	0.7643	0.9763	0.9287	0.9519	0.8256	0.8385	0.832	0.3429	0.7059	0.4615	0.875	0.7568	0.8116
False	head	0.9076	0.7848	0.8117	0.7787	0.969	0.9385	0.9535	0.8478	0.8125	**0.8298**	0.3926	0.7794	0.5222	**0.9298**	0.7162	0.8092
True		0.856	0.655	0.8271	0.698	0.9906	0.8657	0.9239	0.8351	0.8177	0.8263	0.4752	0.7059	0.568	0.3192	0.9189	0.4739
False	hand	0.9202	0.7747	0.8225	0.7824	0.9785	0.958	0.9682	**0.9119**	0.7552	0.8262	0.4109	**0.7794**	0.5381	0.7973	0.7973	0.7973
True		**0.9328**	**0.802**	0.8065	**0.8041**	0.9751	**0.97**	**0.9725**	0.805	0.8385	0.8214	**0.6471**	0.6471	**0.6471**	0.7808	0.7703	0.7755
False	head and hand	0.898	0.7575	0.8229	0.7621	0.9826	0.9295	0.9553	0.9051	0.7448	0.8171	0.3046	0.7794	0.438	0.8378	0.8378	**0.8378**
True		0.8326	0.6459	**0.8292**	0.6786	**0.992**	0.8395	0.9094	0.875	0.7656	0.8167	0.4417	0.7794	0.5638	0.2749	**0.9324**	0.4246

Table 5: Runtime statistics of our approach for different configurations.

GB	Motion	Inference time (ms)
True	whole	5.54 ± 0.18
	head	5.57 ± 0.21
	hand	5.53 ± 0.12
	head & hand	5.54 ± 0.13
False	whole	5.49 ± 0.15
	head	5.52 ± 0.15
	hand	5.53 ± 0.18
	head & hand	5.47 ± 0.11

Our findings reveal that hand motion provides the most salient cues for predicting turn transitions, consistently outperforming head and full-body motion when used alone or in combination. The VQ-VAE-based hand motion representation, especially when fused with discrete gestural features such as head nods and shakes, captures rich temporal dynamics and generalizes well across conversation types and speaker behaviors. In contrast, incorporating head motion, particularly when combined with backchannel gestures, often introduces redundant or overlapping signals. Our ablation study shows that this redundancy can lead to performance degradation, likely due to VQ-VAE's tendency to overuse certain codebook entries, diminishing its ability to preserve fine-grained variations.

While the proposed model outperforms existing methods in predictive accuracy and meets the latency requirements for real-time applications, several limitations remain. Most notably, the current VQ-VAE framework may struggle to retain subtle gestural nuances, such as micro head movements, due to its reliance on discrete quantization and potential codebook imbalance. To address this, future work will explore techniques such as soft or entropy-regularized quantization to improve symbolic diversity and better capture nuanced motion patterns.

Looking ahead, our research offers several promising directions. We aim to extend our framework beyond triadic conversations to more complex multiparty interactions, including four- and five-party scenarios, enabling a deeper exploration of real-world dialogue dynamics. In parallel, we plan to develop more sophisticated models capable of distinguishing interruptions and overlaps from standard end-of-turn events, a distinction that holds significant relevance for virtual agents and conversational interfaces.

Another important avenue is the deployment of our framework in embodied agents and robotic systems. Applying our end-of-turn prediction model to physical robots or avatars in shared virtual environments, such as multi-agent interaction in online platforms or the metaverse, would allow us to evaluate the framework in more interactive, dynamic, and socially situated contexts. Such testing can reveal new challenges related to embodiment, responsiveness, and coordination, and would help bridge the gap between offline modeling and real-world deployment.

Although our current system supports real-time inference on pre-recorded data, we have not yet realized a fully integrated, live-processing system. Toward this goal, we plan to leverage emerging platforms such as Mixed Reality (MR) and Extended Reality (XR) devices, which offer built-in multimodal sensing capabilities. These technologies can facilitate more naturalistic deployments and enable responsive turn-taking in immersive environments.

8 SAFE AND RESPONSIBLE INNOVATION STATEMENT

Our research upholds ethical and responsible innovation by exclusively utilizing anonymized IRB approved data collected from consenting adult participants in naturalistic multiparty conversations. We maintain strict privacy standards through secure data handling and storage protocols, and our models are designed to operate on symbolic representations that do not infer or expose personally identifiable information. To address potential bias, our dataset encompasses a diverse set of speakers, conversation dynamics, and multimodal behaviors, ensuring inclusivity and reducing representational harm. Although our models aim to support real time conversational systems, we recognize the risk of misuse in surveillance or manipulative applications. We strongly advocate for their deployment in transparent user consensual settings, including but not limited to assistive communication, social robotics, and education such as assistive technology or educational tools.

ACKNOWLEDGMENTS

This work is supported in part by US NSF IIS grant 2005430.

REFERENCES

[1] Linda Bell, Johan Boye, and Joakim Gustafson. 2001. Real-time handling of fragmented utterances. In *Proc. NAACL workshop on adaptation in dialogue systems*. 2–8.

[2] Atef Ben-Youssef, Giovanna Varni, Slim Essid, and Chloé Clavel. 2019. On-the-fly detection of user engagement decrease in spontaneous human–robot interaction using recurrent and deep neural networks. *International Journal of Social Robotics* 11, 5 (2019), 815–828.

[3] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. 2021. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*. IEEE, 1–10.

[4] Paul Boersma and Vincent Van Heuven. 2001. Speak and unSpeak with PRAAT. *Glot International* 5, 9/10 (2001), 341–347.

[5] Dan Bohus and Eric Horvitz. 2011. Decisions about turns in multiparty conversation: from perception to action. In *Proceedings of the 13th international conference on multimodal interfaces*. 153–160.

[6] Geert Brône, Bert Oben, Annelies Jehoul, Jelena Vranjes, and Kurt Feyaerts. 2017. Eye gaze and viewpoint in multimodal interaction management. *Cognitive Linguistics* 28, 3 (2017), 449–483.

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[8] Iwan De Kok and Dirk Heylen. 2009. Multimodal end-of-turn prediction in multi-party meetings. In *Proceedings of the 2009 international conference on Multimodal interfaces*. 91–98.

[9] Ziedune Degutyte and Arlene Astell. 2021. The role of eye gaze in regulating turn taking in conversations: a systematized review of methods and findings. *Frontiers in Psychology* 12 (2021), 616471.

[10] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review* 54 (2021), 755–810.

[11] Yu Ding, Yuting Zhang, Meihua Xiao, and Zhigang Deng. 2017. A multifaceted study on eye contact based speaker identification in three-party conversations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3011–3021.

[12] Nia MM Dowell, Tristan M Nixon, and Arthur C Graesser. 2019. Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. *Behavior research methods* 51 (2019), 1007–1041.

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12873–12883.

[14] Cecilia E Ford and Sandra A Thompson. 1996. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. *Studies in interactional sociolinguistics* 13 (1996), 134–184.

[15] Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language* 25, 3 (2011), 601–634.

[16] Dilek Hakkani-Tür, Malcolm Slaney, Asli Celikyilmaz, and Larry Heck. 2014. Eye gaze for spoken language understanding in multi-modal conversational interactions. In *Proceedings of the 16th International Conference on Multimodal Interaction*. 263–266.

[17] Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Prediction of turn-taking using multitask learning with prediction of backchannels and fillers. *Listener* 162 (2018), 364.

[18] Simon Ho, Tom Foulsham, and Alan Kingstone. 2015. Speaking and listening with the eyes: Gaze signaling during dyadic interactions. *PloS one* 10, 8 (2015), e0136905.

[19] Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2015. Multimodal fusion using respiration and gaze for predicting next speaker in multi-party meetings. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 99–106.

[20] Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2015. Predicting next speaker based on head movement in multi-party meetings. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2319–2323.

[21] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Masafumi Matsuda, and Junji Yamato. 2013. Predicting next speaker and timing from gaze transition patterns in multi-party meetings. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. 79–86.

[22] Aobo Jin, Qixin Deng, and Zhigang Deng. 2020. A Live Speech-Driven Avatar-Mediated Three-Party Telepresence System: Design and Evaluation. *PRESENCE: Virtual and Augmented Reality* 29 (2020), 113–139.

[23] Aobo Jin, Qixin Deng, and Zhigang Deng. 2022. S2M-Net: Speech Driven Three-party Conversational Motion Synthesis Networks. In *Proceedings of the 15th ACM SIGGRAPH Conference on Motion, Interaction and Games*. 2:1–2:10.

[24] Aobo Jin, Qixin Deng, Yuting Zhang, and Zhigang Deng. 2019. A deep learning-based model for head and eye motion generation in three-party conversations. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 2, 2 (2019), 1–19.

[25] Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. 2013. Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3, 2 (2013), 1–30.

[26] Tatsuya Kawahara, Takuma Iwatate, and Katsuya Takanashi. 2012. Prediction of Turn-Taking by Combining Prosodic and Eye-Gaze Information in Poster Conversations.. In *Interspeech*. 727–730.

[27] S. Kawato and J. Ohya. 2000. Real-time detection of nodding and head-shaking by directly detecting and tracking the "between-eyes". In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. 40–45.

[28] Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and speech* 41, 3-4 (1998), 295–321.

[29] Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2019. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In *2019 International Conference on Multimodal Interaction*. 226–234.

[30] Meng-Chen Lee and Zhigang Deng. 2024. Online Multimodal End-of-Turn Prediction for Three-party Conversations. In *Proceedings of the 26th International Conference on Multimodal Interaction*. 57–65.

[31] Meng-Chen Lee, Wu Angela Li, and Zhigang Deng. 2024. A Computational Study on Sentence-based Next Speaker Prediction in Multiparty Conversations. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. 1–4.

[32] Meng-Chen Lee, Mai Trinh, and Zhigang Deng. 2023. Multimodal Turn Analysis and Prediction for Multi-party Conversations. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 436–444.

[33] Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. 2022. Audio-driven co-speech gesture video generation. *Advances in Neural Information Processing Systems* 35, 21386–21399.

[34] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[35] Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. 2014. Data-driven models for timing feedback responses in a Map Task dialogue system. *Computer Speech & Language* 28, 4 (2014), 903–922.

[36] Nicole Mirnig, Manuel Giuliani, Gerald Stollnberger, Susanne Stadler, Roland Buchner, and Manfred Tscheligi. 2015. Impact of robot actions on social signals and reaction times in HRI error situations. In *International conference on social robotics*. Springer, 461–471.

[37] Hannah Pelikan and Emily Hofstetter. 2023. Managing delays in human-robot interaction. *ACM Transactions on Computer-Human Interaction* 30, 4 (2023), 1–42.

[38] Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Multimodal continuous turn-taking prediction using multiscale rnns. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 186–190.

[39] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*. Elsevier, 7–55.

[40] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[41] Ryo Sato, Ryuichiro Higashinaka, Masafumi Tamoto, Mikio Nakano, and Kiyoaki Aikawa. 2002. Learning decision trees to determine turn-taking by spoken dialogue systems.. In *Interspeech*. 861–864.

[42] David Schlangen. 2006. From reaction to prediction: Experiments with computational models of turn-taking. *Proceedings of Interspeech* (2006).

[43] Candace L Sidner, Christopher Lee, Cory D Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence* 166, 1-2 (2005), 140–164.

[44] Gabriel Skantze. 2017. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 220–230.

[45] Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language* 67 (2021), 101178.

[46] Gabriel Skantze, Martin Johansson, and Jonas Beskow. 2015. Exploring turn-taking cues in multi-party human-robot discussions about objects. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 67–74.

[47] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).

[48] Nigel G Ward. 2019. *Prosodic patterns in English conversation*. Cambridge University Press.

[49] Payam Jome Yazdian, Mo Chen, and Angelica Lim. 2022. Gesture2Vec: Clustering Gestures using Representation Learning Methods for Co-speech Gesture Generation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 3100–3107.

[50] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. 2023. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*

Pattern Recognition. 469–480.
[51] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. 2023. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14730–14740.