Modeling Turn-Taking with Semantically Informed Gestures

Varsha Suresh[1]*, **M. Hamza Mughal**[2]*, **Christian Theobalt**[1,2], **Vera Demberg**[1,2]

[1]Saarland University, [2]Max Planck Institute for Informatics, Saarland Informatics Campus

Correspondence: {vsuresh,vera}@lst.uni-saarland.de, {mmughal,theobalt}@mpi-inf.mpg.de

Abstract

In conversation, humans use multimodal cues, such as speech, gestures, and gaze, to manage turn-taking. While linguistic and acoustic features are informative, gestures provide complementary cues for modeling these transitions. To study this, we introduce DnD Gesture++, an extension of the multi-party DnD Gesture corpus enriched with 2,663 semantic gesture annotations spanning iconic, metaphoric, deictic, and discourse types. Using this dataset, we model turn-taking prediction through a MoE framework integrating text, audio, and gestures. Experiments show that incorporating semantically guided gestures yields consistent performance gains over baselines, demonstrating their complementary role in multimodal turn-taking.

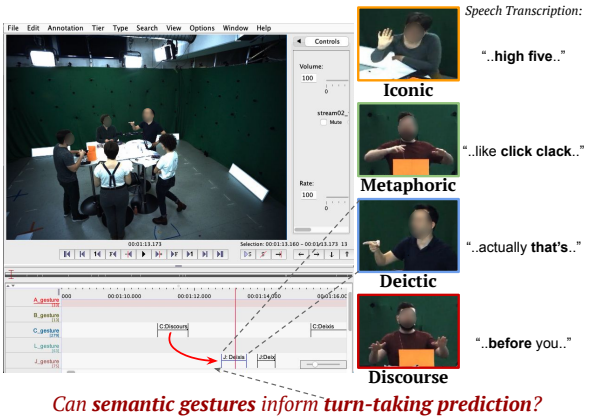

Figure 1: **Gesture Type Annotations.** Time-aligned labels contain semantic gesture types, that are determined by speech context. These labels can help learn gesture representations which improve turn-taking prediction in a multi-party conversation.

1 Introduction

Multi-party conversations are rich, dynamic interactions in which participants coordinate through both verbal and non-verbal channels. A fundamental mechanism that structures these interactions is turn-taking, the implicit management of who speaks next, when a speaker should continue, and when they should yield the floor (Sacks et al., 1974). Effective turn-taking relies on anticipating when a speaker will finish and when another should begin, a problem that becomes increasingly challenging as the number of participants grows.

Existing computational models of turn-taking have predominantly relied on verbal cues, such as lexical content (Ekstedt and Skantze, 2020) and prosody (Ekstedt and Skantze, 2022). However, human communication extends far beyond speech. Social interaction is inherently multimodal, with non-verbal behavior such as gesture and gaze, playing a crucial role in signaling intentions to hold or yield turns (Duncan et al., 1979; Skantze, 2021). Relying solely on text and audio therefore overlooks important communicative signals. Among non-verbal behaviors, semantic gestures (iconic, deictic, metaphoric, and discourse gestures) are especially informative because they are shaped by conversational meaning rather than prosodic rhythms (McNeill, 1992). These gestures help structure dialogue, convey contextual information, and regulate the flow of interaction (Bavelas et al., 1995; Holler et al., 2018). For this reason, we focus on studying how semantic gestures contribute to turn-taking dynamics in multi-party settings and how incorporating them can improve predictive models.

To investigate the role of semantic gestures in turn-taking prediction, we build upon the DnD Gesture dataset (Mughal et al., 2024), a multi-party conversational corpus containing synchronized 3D motion, audio, and transcripts from participants in a tabletop game. We manually annotated the six-hour corpus with gesture-type labels following McNeill (1992). The resulting DnD Gesture++ corpus includes 2,663 gesture instances across six hours (444 labels/hour), forming the most densely annotated English dataset of its kind. For com-

*These authors contributed equally to this work.

Findings of the Association for Computational Linguistics: EACL 2026, pages 2034–2041
March 24-29, 2026 ©2026 Association for Computational Linguistics

parison, BEAT (Liu et al., 2024), one of the few large-scale datasets with semantic gesture-type annotations contains approximately 288 labels per hour. The higher annotation density in DnD Gesture++ provides richer coverage of gesture behavior and more fine-grained supervision for downstream multimodal modeling tasks. We further reformat this data for turn-taking prediction task, labeling 12k turns as either *hold* or *yield*.

In this work, we model turn-taking using a Mixture-of-Experts (MoE) framework that fuses text, audio, and gesture modalities through a gating network. Gesture embeddings are semantically enriched using our annotations to better align motion cues with linguistic and acoustic information. Our experiments demonstrate consistent gains from incorporating gestures, particularly when gesture embeddings are semantically supervised. We further analyze the latent space and modality weights from the MoE framework to better understand the contribution of semantic gesture representations. Beyond turn-taking, the dense annotations in DnD Gesture++ provide a valuable resource for tasks such as co-speech gesture generation, enabling more semantically grounded gesture synthesis (Kucherenko et al., 2021; Mughal et al., 2025).

2 Related Work

2.1 Modeling Turn-Taking for Spoken Dialog

The coordination of speaking turns in human dialogue is inherently multimodal (Skantze, 2021). Speakers use non-verbal signals such as prosody, gaze, and hand gestures to project turn boundaries and manage conversational flow (Duncan et al., 1979; Kendrick et al., 2023). Gestures serve key semantic functions: pragmatic or discourse-related gestures often signal a *yield*, while representational gestures like iconic forms indicate to *hold* the floor (Bavelas et al., 1995; Holler et al., 2018). Moreover, they serve as reliable cues for turn-taking and help maintain conversational smoothness (Holler et al., 2018; Ter Bekke et al., 2024; Kendrick et al., 2023; Holler and Levinson, 2019; Hofstetter, 2021).

Early dialogue systems detected turn boundaries using fixed silence thresholds, while recent data-driven approaches use syntactic and pragmatic cues from dialogue transcripts (Ekstedt and Skantze, 2020). Multimodal approaches further integrate prosody (Ekstedt and Skantze, 2022), face features (Russell and Harte, 2025; Lin et al., 2025; Kurata et al., 2023) to enhance accuracy. However, the effect of semantic gestures remains underexplored in turn-taking models. Our work examines whether semantic gesture types enhance turn-taking prediction in language models.

2.2 Role of Semantic Gestures in Spoken Dialog

Co-speech gestures are typically classified as rhythmic (beat) gestures, aligned with prosody, or semantic gestures, aligned with meaning (McNeill, 1992; Nyatsanga et al., 2023). Semantic gestures play a key role in communication by enhancing comprehension (Holler et al., 2018) and clarifying speaker intent (Goldin-Meadow, 2014). They also provide valuable signals for tasks like multimodal reference resolution (Ghaleb et al., 2025), discourse marker disambiguation (Suresh et al., 2025), and co-speech gesture synthesis (Mughal et al., 2025; Kucherenko et al., 2021; Ram et al., 2025).

Semantic gestures are often classified by McNeill's taxonomy (McNeill, 1992), which reflects their communicative intent. The main types include iconic, metaphoric and deictic gestures. Discourse gestures that structure dialogue, such as signaling topic shifts, are also considered semantic (Bavelas et al., 1995). Refer to Sec. 3.1 for details. Datasets like SAGA (Lücking et al., 2013; Kucherenko et al., 2021) and BEAT (Liu et al., 2022) include semantic type labels but are limited to scripted or monadic data (Lücking et al., 2013; Liu et al., 2022). The DnD Group Gesture Dataset (Mughal et al., 2024) captures natural multiparty interaction but lacks such labels. Our work extends it with semantic type annotations, yielding the first multiparty dataset for studying gesture-informed turn-taking modeling.

3 DnD Group Gesture++

The DnD Group Gesture Dataset (Mughal et al., 2024) captures co-speech gestures in multi-party conversations during a Dungeons & Dragons (DnD) roleplaying game. It includes full-body motion and fine-grained finger articulation from five English-speaking participants over four sessions (6 hours total).

3.1 Gesture type annotation

We extend the DnD Gesture (Mughal et al., 2024) with gesture type annotations based on McNeill's framework (McNeill, 1992, 2005), classifying gestures as iconic, metaphoric, deictic, or discourse (Goldin-Meadow et al., 1993; Kendon,

Overall	Deixis	Metaphoric	Iconic	Discourse
0.522	0.576	0.458	0.674	0.081

Table 1: Interrater agreement scores

Label	% of gesture type				turns w sem gesture (total # of turns)
	De	Di	Ic	Me	
hold	42.3	23.1	28.6	5.9	1550 (7722)
yield	44.4	24.2	25.7	5.7	990 (5087)

Table 2: Distribution of semantic gesture types across turn-taking labels. For each turn type (hold or yield), the table shows the relative frequency (%) of each gesture type, deictic (de), discourse (di), iconic (ic), and metaphoric (me), and the number of turns containing at least one semantic gesture, along with the total number of turns in that label.

1995; Khosrobeigi et al., 2022). The annotations include (# of samples), **iconic** (724) gestures, which depict concrete objects; **metaphoric** (151) gestures, which represent mental images of abstract concepts; **deictic** (1155) gestures, which involve referential gestures like pointing; and **discourse** (633) gestures, which are elicited by the structure of the spoken discourse. We focus on these semantic gesture types as they capture meaning-related aspects of communication that extend beyond prosodic or lexical information and may therefore provide complementary signals for turn-taking. Beat gestures are not included, as their close association with prosody overlaps with cues already available in the audio modality. Table 1 reports interrater agreement, with an average Cohen's κ of 0.52. It also presents per-class κ scores, computed by treating each gesture category as a binary classification task, where the target class is considered positive and all others negative. Agreement is notably lower for discourse gestures, which are more challenging to annotate, as they are defined primarily by communicative function rather than by gesture form. Additional details on the annotation process are provided in Appendix A.1.

3.2 Converting to turn-taking data

To model turn-taking, we convert continuous multiparty conversations in DnD Gesture++ into a structured dataset of discrete prediction instances, following previous works (Kelterer and Schuppler, 2025; Jokinen et al., 2013; Ekstedt et al., 2023). We segment audio into Inter-Pausal Units (IPUs), continuous speech segments bounded by silence, with each IPU ending at a Transition Relevance Place (TRP) where a speaker change may occur. Turns are labeled *yield* if another participant speaks next, or *hold* if the same speaker continues, using a 200 ms silence threshold (Lee et al., 2023; Hara et al., 2019). Short IPUs are merged with adjacent utterances. This results in 12k turns (7k hold, 5k yield). We split the data into 70% train, 10% validation, and 20% test. Refer to Table 2 for more details.

4 Multimodal Modeling of Turn-Taking Prediction

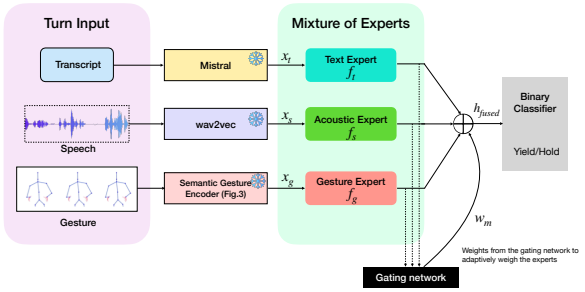

Figure 2: MoE modeling of turn taking

We model turn-taking using text, audio, and gestures captured via motion data. Each modality is processed by a dedicated expert within a MoE framework, where the outputs of all experts are fused via a gating network that dynamically weighs each modality based on context. Let x_m denote the input features for modality m and $f_m(x_m)$ its expert output. The gating network computes modality weights w_m using softmax over concatenated weights. The fused representation is a weighted sum of expert outputs $h_{\text{fused}} = \sum_{m=1}^{M} w_m \cdot f_m(x_m)$.

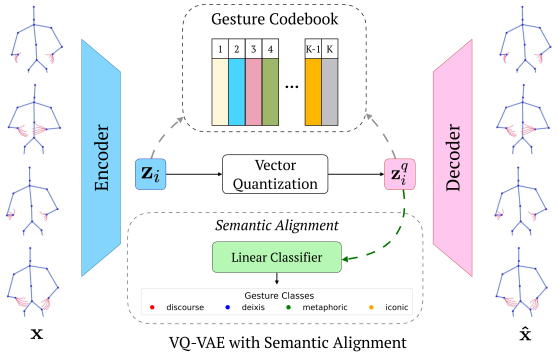

Figure 3: Learning semantically-aligned gesture representations.

Finally, h_{fused} is passed through a linear classifier to predict *hold* or *yield*. Text and audio features are obtained from pretrained embeddings, sentence

embeddings for text[1] and Wav2Vec2 for audio[2] and projected via MLP layers. Gestures are encoded following recent VQ-VAE based gesture tokenization approaches (Liu et al., 2024; Suresh et al., 2025). Given a sequence of 3D upper body motion $\mathbf{x}$, the encoder network compresses it into a sequence of tokens. Each token $\mathbf{z}_i$ is then quantized via a codebook (Van Den Oord et al., 2017) to produce $\mathbf{z}_i^q$. The sequence of quantized embeddings is then used to reconstruct the input motion $\mathbf{x}$. Consequently, the VQ-VAE is trained through an MSE loss on reconstructed motion $\hat{\mathbf{x}}$ and input motion $\mathbf{x}$. Since the base model lacks semantics, we use gesture type annotations to inject semantic information (see Figure 3). For each turn, we obtain the gestures token sequences of codebook embeddings from the trained VQ-VAE which are then processed by a transformer encoder (Vaswani et al., 2017) which forms the gesture expert. The output of the Transformer is mean pooled over the sequence length to produce a fixed-length representation.

5 Experiments

Our work addresses two key research questions: (i) Does the inclusion of body movement features such as gestures improve turn-taking prediction beyond text and audio? (ii) Do semantically informed gesture representations offer advantages over raw motion features?

5.1 Model Comparisons

We systematically compare our Text+Audio+Gesture model against single- and dual-modality baselines. We then evaluate the impact of semantic gesture embeddings from a VQ-VAE versus raw motion features. We also benchmark our MoE-based fusion against existing multimodal turn-taking approaches using the same modality experts for fair comparison. Importantly, our main aim is to assess the effect of gesture representations on turn-taking beyond text and audio, and the MoE setup allows us to quantify the contribution of each modality in this task. Further implementation details can be found in the Appendix A.2.

[1] https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1

[2] https://huggingface.co/facebook/wav2vec2-large-960h

	Overall		Per-class	
	Acc	F1	*hold*	*yield*
Majority Class	60.4	37.6	75.3	0.0
Text	68.8 $\pm$0.5	66.6 $\pm$0.4	75.1	58.2
Audio	67.1 $\pm$0.5	63.2 $\pm$1.6	75.1	51.3
Gesture	66.2 $\pm$0.2	61.2 $\pm$1.6	75.1	47.4
Gesture (w/o sem)	66.1 $\pm$0.3	61.8 $\pm$0.5	74.6	48.9
Text+Audio	69.7 $\pm$2.1	67.9 $\pm$1.2	74.9	61.1
Text+Gesture	70.0 $\pm$0.2	67.8 $\pm$0.2	76.2	59.4
Audio+Gesture	67.7 $\pm$0.4	64.2 $\pm$0.2	75.4	53.0
Text+Audio+Gesture	**71.5** $\pm$0.3	**69.9** $\pm$0.1	**76.7**	**63.1**
Text+Audio+Gesture (w/o sem)	70.4$\pm$0.4	68.7 $\pm$0.5	75.9	61.5

Table 3: Performance of multimodal variations for turn-taking prediction using text, audio, and gesture modalities with MoE based fusion modeling.

Gesture Type	*hold*	*yield*
Discourse	78.3	74.1
Deixis	78.2	65.4
Iconic	72.5	58.3
Metaphoric	89.5	66.7
Overall	76.7	63.1

Table 4: Per-class F1 contributions for *hold* and *yield* predictions across each semantic gesture type

5.2 Turn-taking Prediction

From Table 3, text performs best among single-modality experts, followed by audio. Gesture only models show lower performance, with little difference between non-semantic and semantic variants. This result reflects the nature of turn-taking cues—gestures complement rather than replace lexical and prosodic structure. Importantly, gestures consistently improve performance in fusion settings, indicating that they provide important information, especially for subtle transitions where speech cues are weak or ambiguous. Multimodal fusion outperforms single-modality models with Text+Audio+Gesture model achieving the highest overall macro-F1 compared to the Text+Audio variant ($p < 0.05$). Notably, the semantically aligned gesture representation outperforms the non-semantic one ($p < 0.05$), showing that semantic supervision enhances multimodal fusion.

When ablating with fusion techniques from prior multimodal turn-taking studies (Table 5), we observe per-class F1 imbalances, with the minority class particularly affected. In contrast, MoE-based fusion with semantically aligned gestures yields more balanced predictions across classes. Future work can explore fusion strategies to further im-

	Overall	*hold*	*yield*
ConcatFusion (Kurata et al., 2023)	68.2	**78.5**	58.0
LMF (Lin et al., 2025)	67.9	77.2	58.7
MoE (Ours)	**69.9**	76.7	**63.1**

Table 5: Ablation study on fusion techniques: F1 scores of the Text+Audio+Gestures model evaluated with different existing multimodal fusion methods for turn-taking prediction.

prove alignment between text, audio, and gestures. Table 4 reports the per-class F1 scores for *hold* and *yield* predictions across semantic gesture categories. The results further demonstrate that different gesture types provide distinct cues for turn-taking: discourse gestures strongly contribute to *yield* transitions, whereas metaphoric gestures are more indicative of *hold* behavior. All results are averaged over three random seeds.

5.3 Analysing gesture representations

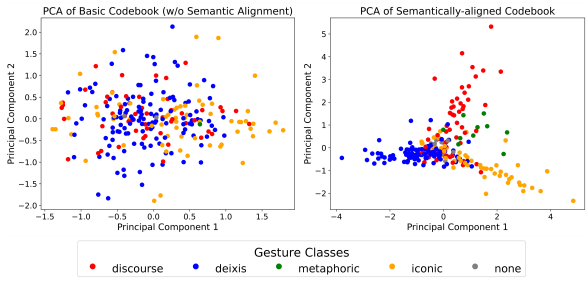

Figure 4: Visualization of Gesture Representations.

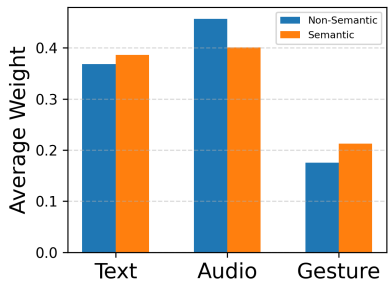

Figure 5: MoE modality weight contributions for semantic vs w/o semantic gesture representations

Figure 4, we visualize the gesture embeddings with and without semantic supervision. The semantically aligned embeddings form more distinct clusters corresponding to different gesture types, which facilitates better alignment with text and audio modalities. Figure 5 further shows the average weight contribution for each modality in the MoE framework for the test samples. Using semantically aligned gestures increases the weights for both gestures and text, likely due to improved cross-modal alignment, which in turn supports better overall task performance.

6 Conclusion

In this work, we investigate the role of semantic gestures in multimodal turn-taking prediction. We extended the multi-party DnD Gesture corpus with semantic annotations to create DnD Gesture++. Our results show that semantically guided gestures improve turn-taking modeling, especially when fused with text and audio. Beyond turn-taking, the corpus supports a range of multiparty interaction tasks, including discourse-level gesture analysis, co-speech gesture generation, and listener backchannel prediction, while addressing a key gap in semantic gesture resources for multi-party settings.

Limitations

While this work is based on a naturalistic, game-based interaction setting, the domain specificity may limit generalisability. Future work should assess the role of semantic gestures in other interaction contexts, including task-oriented dialogues, meetings, remote communication, and cross-cultural settings. Although our semantic gesture annotations are carefully curated, they show only moderate inter-annotator agreement, reflecting the inherent subjectivity of gesture interpretation, particularly for discourse gestures. This annotation noise may cap achievable performance and could be mitigated through additional annotators. Our study focuses on whether semantic gestures provide complementary information for IPU-based turn-taking when combined with text and audio. While we employ a MoE fusion framework, future work could investigate more expressive fusion and temporal modeling approaches that jointly capture modality interactions and turn structure.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft, Funder Id: http://dx.doi.org/10.13039/501100001659, SFB 1102: "Information Density and Linguistic Encoding", project number 232722074, and funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2853/1 "Neuroexplicit Models of Language, Vision, and Action" - project number 471607914.

References

Janet Beavin Bavelas, Nicole Chovil, Linda Coates, and Lori Roe. 1995. Gestures specialized for dialogue. *Personality and social psychology bulletin*, 21(4):394–405.

Starkey Duncan, Lawrence J Brunner, and Donald W Fiske. 1979. Strategy signals in face-to-face interaction. *Journal of Personality and Social Psychology*, 37(2):301.

Erik Ekstedt and Gabriel Skantze. 2020. Turngpt: a transformer-based language model for predicting turn-taking in spoken dialog. *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Erik Ekstedt and Gabriel Skantze. 2022. How much does prosody help turn-taking? investigations using voice activity projection models. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 541–551. Edinburgh UK: Association for Computational Linguistics.

Erik Ekstedt, Siyang Wang, Éva Székely, Joakim Gustafson, and Gabriel Skantze. 2023. Automatic evaluation of turn-taking cues in conversational speech synthesis. In *Proc. Interspeech 2023*, pages 5481–5485.

Esam Ghaleb, Bulat Khaertdinov, Asli Özyürek, and Raquel Fernández. 2025. I see what you mean: Co-speech gestures for reference resolution in multimodal dialogue. In *The 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*.

Susan Goldin-Meadow. 2014. How gesture works to change our minds. *Trends in neuroscience and education*, 3(1):4–6.

Susan Goldin-Meadow, Martha Wagner Alibali, and R Breckinridge Church. 1993. Transitions in concept acquisition: using the hand to read the mind. *Psychological review*, 100(2):279.

Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2019. Turn-taking prediction based on detection of transition relevance place. In *Interspeech*, pages 4170–4174.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Emily Hofstetter. 2021. Achieving preallocation: Turn transition practices in board games. *Discourse processes*, 58(2):113–133.

Judith Holler, Kobin H Kendrick, and Stephen C Levinson. 2018. Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic bulletin & review*, 25(5):1900–1908.

Judith Holler and Stephen C Levinson. 2019. Multimodal language processing in human communication. *Trends in cognitive sciences*, 23(8):639–652.

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):1–30.

Anneliese Kelterer and Barbara Schuppler. 2025. Turn-taking annotation for quantitative and qualitative analyses of conversation. *arXiv preprint arXiv:2504.09980*.

Adam Kendon. 1995. Gestures as illocutionary and discourse structure markers in southern italian conversation. *Journal of pragmatics*, 23(3):247–279.

Kobin H Kendrick, Judith Holler, and Stephen C Levinson. 2023. Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical transactions of the royal society B*, 378(1875):20210473.

Zohreh Khosrobeigi, Maria Koutsombogera, and Carl Vogel. 2022. Gesture and part-of-speech alignment in dialogues. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue*, pages 172–182.

Taras Kucherenko, Rajmund Nagy, Patrik Jonell, Michael Neff, Hedvig Kjellström, and Gustav Eje Henter. 2021. Speech2Properties2Gestures: Gesture-property prediction as a tool for generating representational gestures from speech. In *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents*.

Fuma Kurata, Mao Saeki, Shinya Fujie, and Yoichi Matsuyama. 2023. Multimodal turn-taking model using visual cues for end-of-utterance prediction in spoken dialogue systems. In *Proc. Interspeech*, pages 2658–2662.

Meng-Chen Lee, Mai Trinh, and Zhigang Deng. 2023. Multimodal turn analysis and prediction for multiparty conversations. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 436–444.

Yuxin Lin, Yinglin Zheng, Ming Zeng, and Wangzheng Shi. 2025. Predicting turn-taking and backchannel in human-machine conversations using linguistic, acoustic, and visual signals. *arXiv preprint arXiv:2505.12654*.

Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. 2024. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling.

Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. *European Conference on Computer Vision*.

Andy Lücking, Kirsten Bergman, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2013. Data-based analysis of speech and gesture: The bielefeld speech and gesture alignment corpus (saga) and its applications. *Journal on Multimodal User Interfaces*, 7:5–18.

David McNeill. 1992. Hand and mind: What gestures reveal about thought.

David McNeill. 2005. *Gesture and Thought*. University of Chicago Press.

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

Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. 2024. Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis. In *Computer Vision and Pattern Recognition (CVPR)*.

Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A comprehensive review of data-driven co-speech gesture generation. In *Computer Graphics Forum*, volume 42, pages 569–596. Wiley Online Library.

Ashwin Ram, Varsha Suresh, Artin Saberpour Abadian, Vera Demberg, and Jürgen Steimle. 2025. Gesturecoach: Rehearsing for engaging talks with llm-driven gesture recommendations. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, UIST '25, New York, NY, USA. Association for Computing Machinery.

Sam O'Connor Russell and Naomi Harte. 2025. Visual cues enhance predictive turn-taking for two-party human interaction. *arXiv preprint arXiv:2505.21043*.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735.

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

Varsha Suresh, M Hamza Mughal, Christian Theobalt, and Vera Demberg. 2025. Enhancing spoken discourse modeling in language models using gestural cues. *arXiv preprint arXiv:2503.03474*.

Marlijn Ter Bekke, Linda Drijvers, and Judith Holler. 2024. Gestures speed up responses to questions. *Language, Cognition and Neuroscience*, 39(4):423–430.

Aaron Van Den Oord, Oriol Vinyals, and 1 others. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

A Appendix

A.1 Additional Annotation Details

The dataset contains 4 recording sessions. To label the gesture type in each recording, the annotators utilize the speech from the group conversation and multi-view videos to identify who made the gesture and mark it with label for the duration of said gesture. Therefore, the annotators considered both linguistic and visual modalities to identify the gesture types. Annotation was performed in ELAN [3] where each recording participant was assigned a separate track for labeling. For annotation, two English-speaking young students were recruited from the university. After that, authors randomly sampled the annotations and verified the annotations manually and cleaned conflicting or wrong annotations. To acquire the transcriptions we use, WhisperX [4].

A.2 Implementation Details

For obtaining gesture representations: In order to train the VQ-VAE, we pass 3D upper body motion encoded as joint positions. The input skeleton is normalized relative to pelvis (root) joint and translation is fixed to zero, such that the network only models hand and arm movements. We utilize convolutional encoder and decoder networks based on ResNet backbone (He et al., 2016). To apply semantic alignment loss, we utilize the annotations by training using 4 annotated semantic classes i.e. iconic, deictic, metaphoric, discourse, and a "none" class consisting of beat and other non-semantic gestures. During training, cross entropy loss is applied on the linear classifier to identify gesture categories. This loss ignores the "none" class. The whole framework is trained for 120 epochs with following hyperparameters: number of residual blocks = 2 in both encoder and decoder, embedding dim = 256, codebook size = 256, learning rate = 3e-4, optimizer = AdamW, reconstruction loss

[3] https://archive.mpi.nl/tla/elan

[4] https://github.com/m-bain/whisperX

weight = 1, semantic loss weight = 0.1, and batch size = 512.

For gesture transformer encoder, we use a 1-layer transformer with a feedforward dimension of 128 and 4 attention heads. The final hidden states are mean-pooled to obtain the gesture expert representation. MoE training is conducted on a single NVIDIA V100 GPU with a batch size of 32 for 20 epochs. The learning rate is selected via hyperparameter tuning from 1e-4, 5e-5, 1e-5, with 5e-5 yielding the best performance.