EgoSpeak: Learning When to Speak
for Egocentric Conversational Agents in the Wild

Junhyeok Kim♣   **Min Soo Kim**♣   **Jiwan Chung**♣   **Jungbin Cho**♣
Jisoo Kim♣   **Sungwoong Kim**♣   **Gyeongbo Sim**◇   **Youngjae Yu**♣

♣ Yonsei University   ◇ Multimodal AI Lab., NC Research, NCSOFT Corporation
junhyeok@yonsei.ac.kr

Abstract

Predicting when to initiate speech in real-world environments remains a fundamental challenge for conversational agents. We introduce EgoSpeak, a novel framework for real-time speech initiation prediction in egocentric streaming video. By modeling the conversation from the speaker's first-person viewpoint, EgoSpeak is tailored for human-like interactions in which a conversational agent must continuously observe its environment and dynamically decide when to talk.

Our approach bridges the gap between simplified experimental setups and complex natural conversations by integrating four key capabilities: (1) first-person perspective, (2) RGB processing, (3) online processing, and (4) untrimmed video processing. We also present YT-Conversation, a diverse collection of in-the-wild conversational videos from YouTube, as a resource for large-scale pretraining. Experiments on EasyCom and Ego4D demonstrate that EgoSpeak outperforms random and silence-based baselines in real time. Our results also highlight the importance of multimodal input and context length in effectively deciding when to speak. Code and data are available at website.

1 Introduction

Human-like conversational agents have long been a key objective in artificial intelligence. A critical aspect of human conversation is not only understanding what to say but also when to say it—often framed as the study of turn-taking (Duncan, 1972). While most are designed under simplified assumptions where turn boundaries are well-defined or where only audio-based cues are available, real-world conversations can be highly fluid, with overlapping speech, unclear speaker roles, and frequent interruptions (Skantze, 2017, 2021).

To address these complexities, we introduce EgoSpeak, a framework that predicts when an agent

Figure 1: EgoSpeak models speech initiation in real time from the camera wearer's (camera icon) egocentric video stream, mirroring how a real-world agent would perceive and engage in dynamic, multi-speaker environments.

should begin speaking based on egocentric streaming video. Concretely, EgoSpeak models speech initiation from the first-person perspective of the camera wearer, capturing exactly what the agent sees at each moment in real time. Unlike a third-person or fixed camera view, the egocentric perspective is especially relevant for real-world conversational agents such as social robots that must decide on the fly whether to speak or remain silent. By leveraging the camera wearer's immediate visual context (e.g., facing another person, noticing body language or gaze direction), EgoSpeak can more naturally detect subtle cues that signal an appropriate moment to start speaking. This is particularly crucial for a real-world agent that must not only process inputs in real time, but also respond autonomously in dynamic, multi-speaker environments to appear natural and engaging.

EgoSpeak incorporates four key capabilities: (1) first-person perspective: aligns closely with real-world interactions for conversational agents, (2) RGB feature processing: handles scenarios where audio or non-verbal cues may be unreliable, (3) dynamic real-time turn-taking: enables more natural and fluid conversations, and (4) continuous untrimmed video stream processing: captures periods of silence and sporadic interactions.
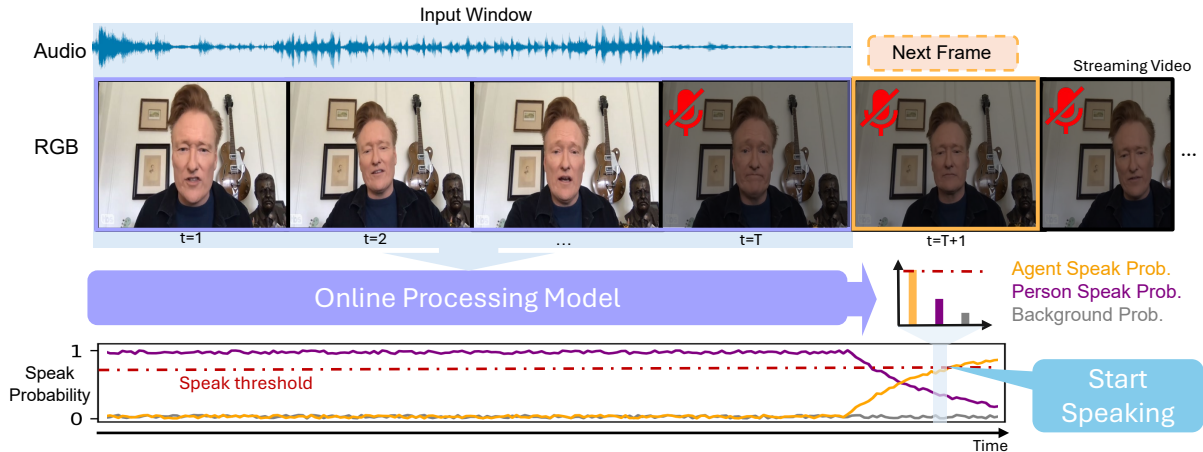
Figure 2: Overview of the EgoSpeak framework. At each time step, the model processes an untrimmed egocentric video and audio stream, classifying them in real time into three categories: background (no speech), other person speaking, and target speaker (camera wearer) speaking. These probabilities are visualized at the bottom, where the model anticipates near-future frames and enables proactive speech initiation for conversational agents.

These four collectively enable EgoSpeak to handle the complexities of real-world conversations more effectively than previous methods (see Table 1). Figure 2 provides an overview of our real-time pipeline, illustrating how EgoSpeak processes continuous video streams to decide when to speak. EgoSpeak outputs a continuous speak-probability that a conversational agent can leverage in real time (e.g., by triggering speech once the probability surpasses a threshold).

We validate EgoSpeak on two distinct datasets: EasyCom and Ego4D, demonstrating its effectiveness across various conversational contexts. Additionally, we introduce the YT-Conversation dataset, a collection of in-the-wild conversation videos including interviews and casual conversations from YouTube, designed for scalable pretraining.

By addressing the critical challenge of when to speak in a natural, human-like manner, EgoSpeak advances the field of conversational AI, offering a robust solution for dynamic, intermittent conversations with varying numbers of speakers.

In summary, our key contributions are:

1. EgoSpeak, a novel framework for speech initiation prediction from egocentric streaming video in real time.
2. YT-Conversation, a large-scale corpus of in-the-wild conversational videos, suitable for pretraining multimodal turn-taking models.
3. Experimental results on EasyCom and Ego4D demonstrate effectiveness in real-world scenarios and provide a comprehensive analysis of the role of multimodal inputs and context length.

2 Related Works

Turn-taking. Turn-taking research has evolved from simple audio-based models (Duncan, 1972; Khouzaimi et al., 2015) to sophisticated multi-modal approaches (Maier et al., 2017; Lee et al., 2023; Mizuno et al., 2023; Kurata et al., 2023). Early offline methods, which process entire clips, often result in unnatural pauses. This prompted the development of continuous (online) methods (Skantze, 2017; Ekstedt and Skantze, 2022; Li et al., 2022), including recent multimodal models incorporating non-verbal cues (Onishi et al., 2023). However, these approaches typically rely on controlled dyadic conversations, limiting real-world applicability. EgoSpeak addresses these limitations by adopting a first-person perspective, processing both RGB and audio features, and handling untrimmed video streams, aiming to better align turn-taking models with the complexities of natural conversations.

	Egocentric	RGB	Online	Untrimmed
Skantze (2017)	✗	✗	✓	✗
Ekstedt and Skantze (2022)	✗	✗	✓	✗
Li et al. (2022)	✗	✗	✓	✗
Yang et al. (2022)	✗	✗	✗	✗
Kurata et al. (2023)	✗	✓	✗	✗
Lee et al. (2023)	✗	✓	✗	✗
Mizuno et al. (2023)	✗	✗	✗	✗
Onishi et al. (2023)	✗	✗	✓	✗
Fatan et al. (2024)	✓	✓	✗	✗
EgoSpeak (Ours)	✓	✓	✓	✓

Table 1: Comparison of EgoSpeak with existing utterance initiation and turn-taking methods. EgoSpeak uniquely addresses all four key aspects of real-world conversational dynamics.

Online Processing. Online systems, which predict based only on past and present information continuously, have gained popularity in real-time applications across various fields from computer vision to speech (Fan et al., 2018; De Geest et al., 2016; Kang et al., 2021; Bewley et al., 2016; Rettig et al., 2019; Miao et al., 2020). EgoSpeak applies this approach to turn-taking, enabling real-time prediction of speech initiation points in natural conversations without relying on future information. This capability allows EgoSpeak to adapt to dynamic conversational scenarios, making it more suitable for real-world interactions.

3 EgoSpeak Framework

3.1 Framework Overview

EgoSpeak is designed for in-the-wild conversational agents, building on the challenges discussed in Section 1, where "in-the-wild" refers to real-world conditions outside controlled environments with unpredictable variables and numerous influencing factors.

EgoSpeak is grounded in the intuition that, in the egocentric video, the camera wearer's speaking moments naturally serve as cues for speech initiation. By predicting these moments from the agent's perspective, our framework learns natural turn-taking behavior, identifying when to speak even after long silences. Moreover, by anticipating these moments in advance, EgoSpeak effectively mirrors human turn-taking, deciding when to begin speaking as a real-world agent would. To achieve this, we train the model with a cross-entropy objective, akin to next-token prediction in language modeling, since it must anticipate speaking before the camera wearer actually speaks.

3.2 Task Definition

Guided by this intuition, we formulate the problem of predicting the target speaker's speech in an egocentric streaming video, where the camera wearer is naturally identified as the target speaker. Given the real-time nature of the stream, EgoSpeak only analyzes information available up to the current moment. This design allows our system to capture the continuously unfolding context and prepare a speech onset before a turn-shift occurs in complex, dynamic conversations.

Formally, let $X_t = [x_1, \ldots, x_t]$ be an online stream up to timestep t, where each x_i can include multiple modalities x_i^m including visual frames x_i^v
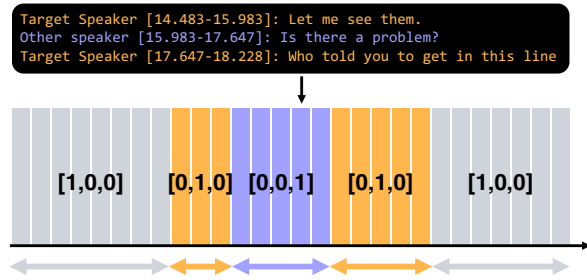

Figure 3: Converting Transcript to Per-Frame Labels. Colors indicate: gray - background, orange - target speaker speaking, purple - other speaker speaking. Labels are one-hot encoded for classification.

or auditory signals x_i^a. We transform each x_i^m into a representation z_i^m via off-the-shelf feature extractors, concatenating them into z_i. Next, we define a temporal window $Z_{t-L+1,t} = [z_{t-L+1}, \ldots, z_t]$ of length L. Given an anticipation length α, the model performs a three-way classification (background / target speaker speaking / other speaking) for the future range $t+1$ to $t+\alpha$. This anticipatory modeling gives the system extra time to prepare responses, rather than reacting only after a silence threshold. The final model output is a probability tensor of shape $[\alpha, 3]$, where the dimensions correspond to the anticipated future timesteps and the three classes, respectively.

Prediction vs Detection. A naive approach for determining when to speak is detection which occurs based on silence threshold. However, detection offers an inadequate response time of only 200ms for listeners. A psycholinguistic study (Levinson and Torreira, 2015) estimates that actual response time ranges from 600 to 1500ms, as humans begin preparing their responses while the other person is still speaking. Additionally, turn-shifts often occur as overlapping without any gaps (Skantze, 2021). The prediction will give conversational systems more time to generate reactions and enable human-like conversation.

Frame-level Speech Labeling. Figure 3 illustrates how transcript timestamps convert into per-frame, one-hot encoded labels. As our framework requires per-frame speech labels which are expensive to annotate, we developed a method to convert transcript annotations from egocentric videos into per-frame speech classification labels. At each timestep t, we label the datapoint x_t as *target speaker speaking* if the camera wearer is speaking, *other person speaking* if others are speaking, and *no speech* otherwise.

Figure 4: Sample frames from YT-Conversation dataset. The dataset includes a diverse range of conversational scenarios from YouTube, such as podcasts, interviews, and informal dialogues, representing various real-world conversation formats.

3.3 YT-Conversation: Dataset for Multimodal Conversation Pretraining

Existing turn-taking resources often stem from controlled laboratory setups or video calls, which are expensive to annotate and capture only a fraction of the complexity found in real-world interactions, limiting scalability. To address this gap, we introduce YT-Conversation, a novel dataset derived from diverse YouTube content including interviews, podcasts, and casual dialogues.

While YT-Conversation is not fully egocentric, it offers realistic face-to-face and multi-person interactions that can effectively transfer to first-person scenarios in egocentric video understanding (Zhang et al., 2022; Lin et al., 2022). By leveraging content from real-world YouTube videos through an automatic pipeline, YT-Conversation aims to provide a more scalable resource for turn-taking pretraining.

Collecting Conversational Videos. We curated our dataset from four manually selected YouTube channels, covering diverse conversational formats including podcasts, interviews, and face-to-face dialogues. Videos were randomly sampled without further filtering, ensuring scalability. We preprocessed the videos by downsampling to 20 FPS for video and 16 kHz for audio, and trimmed opening segments. Our final dataset comprises 414 videos totaling 41 hours, with durations ranging from 1 to 60 minutes.

Pseudo Per-frame Annotation for Collected Videos Since manual annotation of each video frame is labor-intensive, we employ voice activity detection (VAD) from Pyannote (Plaquet and Bredin, 2023; Bredin, 2023) to generate pseudo-labels for speech activity. Specifically, we remove
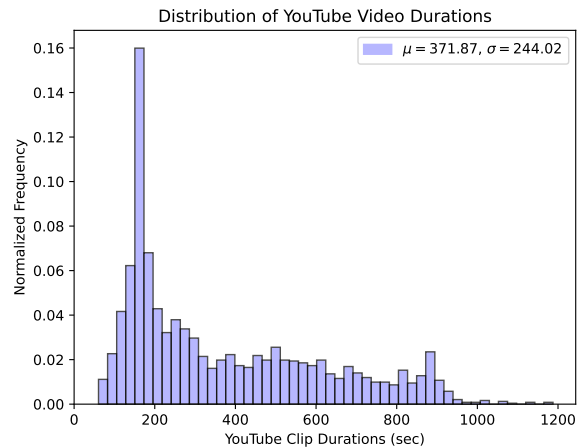

Figure 5: Video duration distribution for YT-Conversation. Our online formulation allows the use of long video clips, some even exceeding 900 seconds.

any speech segments under 200 ms (or non-speech gaps under 200 ms) to match our 200 ms resolution. This approach yields a speech vs. no-speech label per frame, effectively approximating the ground truth for large-scale pretraining. Figure 5 shows a distribution of video durations and illustrates the diversity of conversation styles in YT-Conversation. For the validation of pseudo-annotation quality, see Appendix E.

4 Experimental Setup

4.1 Dataset

We propose to use publicly available egocentric conversational video datasets for evaluation: Easy-Com (Donley et al., 2021) and Ego4D (Grauman et al., 2022).

EasyCom. The EasyCom dataset contains egocentric videos of 3-5 participants conversing around a table in a room for about 30 minutes per session. It comprises 12 sessions totaling approximately 5 hours and 18 minutes. We use sessions 1-3 for testing and 4-12 for training. The dataset features human-annotated transcripts with precise timestamps and mono-channel audio.

Ego4D. We use the Audio-Visual Diarization benchmark from Ego4D (Grauman et al., 2022), a large-scale, in-the-wild egocentric video dataset. This subset contains 5-minute clips from diverse scenarios, including both indoor and outdoor settings. However, the original test split is mostly limited to indoor settings. To ensure robust evaluation, we randomly split the combined original train and test sets into 346 training clips and 87 test

2993

clips.

The EasyCom and Ego4D datasets offer complementary scenarios for evaluating our framework. EasyCom provides a controlled setting with continuous conversations among fixed participants, while Ego4D presents diverse, real-world scenarios with varying numbers of speakers, environments, and intermittent speech patterns. This combination allows us to assess EgoSpeak's performance in both relatively structured and unstructured environments, testing its ability to predict utterance initiation across a range of conversational dynamics.

4.2 Baselines & Models

We evaluate our framework using three trained models with different architectural backbones: RNN (An et al., 2023), Transformer (Xu et al., 2021), and State-Space-Model (Gu and Dao, 2023). Additionally, we implement two static baselines: a random baseline and a rule-based algorithm using silence as decision threshold (Bell et al., 2001). Detailed implementation details including architectural specifications, hyperparameters, training objective and feature extraction for the neural models are provided in Appendix A.

Random Baseline. This baseline randomly assigns one of the three possible labels (background, target speaker speaking, or other speaking) to each frame with uniform probability.

Silence-based Algorithm. Simulating commercial spoken dialogue agents, this approach triggers speech only after a 600 ms silence interval following other speakers. Our evaluation likely overestimates its real-world performance, since we use ground-truth labels to detect non-target speech and count the entire subsequent speech segment as correct once the start is identified (only a single timestep is penalized when incorrect).

Transformer-based Model. We adopt Long Short-term TRansformer (LSTR) (Xu et al., 2021) for temporal modeling. LSTR uses long-term and short-term memory mechanisms to handle sequence data, with an encoder-decoder structure. The encoder leverages long context windows by compressing inputs, while the decoder processes shorter context windows, allowing for flexible temporal modeling.

RNN-based Model. Inspired by An et al. (2023), we used a simple and effective RNN model contain-

ing one GRU layer. This model was chosen for its computational efficiency and strong performance.

Mamba-based Model. We implement a Mamba-based model (Gu and Dao, 2023) similar to the RNN architecture. Given the recent success of Mamba across various tasks, we include this model to explore its potential to predict speech initiation in egocentric videos while maintaining computational efficiency.

4.3 Settings

Feature Extraction. We process features at 5 FPS, predicting every 0.2 seconds to align with typical human response times (Skantze, 2021). For RGB features, we use a ResNet-50 (He et al., 2016) model pretrained on Kinetics-400 (Kay et al., 2017). Audio features are extracted using wav2vec2 (Baevski et al., 2020). These features are concatenated to create our multimodal input. Further details on feature extraction are provided in Appendix A.3.

Evaluation Protocol. Most existing turn-taking evaluations rely on offline F1-scores after processing the entire clip (Lee et al., 2023; Kurata et al., 2023), or on sample-based F1-scores around turn-taking events using threshold-based detection (Ekstedt and Skantze, 2022; Onishi et al., 2023). However, both approaches fail to capture the continuous, overlapping nature of real-world conversations, where a decision must be made at every frame. As Heldner and Edlund (2010) suggested, overlaps occur frequently in human conversation. Consequently, we measure performance per frame to better reflect these natural conversational dynamics.

To address this need, we propose using per-frame mean average precision (mAP), inspired by prior work on online tasks (De Geest et al., 2016). This metric evaluates how well the model anticipates the target speaker's speech up to 10 timesteps (2 s) into the future. We compute mAP by 1) sorting all frame-level confidence scores in descending order, 2) iteratively using each score as a threshold, 3) calculating precision and recall at each threshold, and 4) averaging all precision values. This procedure is repeated for each class and timestep, then averaged to yield the final mAP.

Model	Modality	mAP (%)										
		0.20s	0.40s	0.60s	0.80s	1.00s	1.20s	1.40s	1.60s	1.80s	2.00s	Avg
Transformer	A	72.2	65.2	60.3	56.8	54.4	53.1	52.4	52.0	51.6	51.4	56.9 ± 0.05
	V	52.0	51.7	51.6	51.3	51.1	50.9	50.8	50.5	50.3	50.1	51.0 ± 0.08
	A+V	73.8	66.9	62.1	58.5	56.3	55.0	54.1	53.7	53.3	53.0	58.7 ± 0.13
	A+V[P]	73.4	66.8	61.8	58.3	56.1	54.8	54.1	53.5	53.2	52.7	58.5 ± 0.26
GRU	A	71.5	65.0	60.1	57.0	55.0	53.8	52.9	52.2	51.5	50.9	57.0 ± 0.30
	V	53.0	52.7	52.4	52.0	51.7	51.6	51.2	51.1	50.8	50.6	51.7 ± 0.29
	A+V	73.5	68.1	63.7	60.7	59.1	58.1	57.2	56.3	55.4	54.4	60.6 ± 0.17
	A+V[P]	70.8	64.9	60.1	56.9	55.0	53.8	53.0	52.4	51.8	51.4	57.0 ± 0.29
Mamba	A	67.5	62.2	58.4	55.7	54.0	52.9	52.0	51.1	50.2	49.6	55.4 ± 0.62
	V	52.2	51.8	51.5	51.1	50.9	50.7	50.5	50.4	50.0	49.7	50.9 ± 0.21
	A+V	71.8	65.4	60.5	57.1	55.0	53.9	53.5	53.1	52.3	51.8	57.4 ± 0.26
	A+V[P]	68.9	63.2	59.1	56.0	54.0	52.7	51.8	51.4	50.7	50.1	55.8 ± 0.43

(a) Results on EasyCom

Model	Modality	mAP (%)										
		0.20s	0.40s	0.60s	0.80s	1.00s	1.20s	1.40s	1.60s	1.80s	2.00s	Avg
Transformer	A	78.8	74.9	71.8	69.7	68.1	67.0	66.3	65.7	65.1	64.7	69.2 ± 0.03
	V	58.7	58.5	58.4	58.2	58.1	58.0	57.9	57.8	57.7	57.7	58.0 ± 0.27
	A+V	78.1	74.3	71.5	69.4	68.0	67.0	66.3	65.7	65.3	64.9	69.0 ± 0.24
	A+V[P]	78.4	74.5	71.5	69.4	67.9	66.7	65.9	65.4	65.0	64.5	68.9 ± 0.18
GRU	A	78.6	74.8	71.8	69.6	68.1	66.9	66.2	65.6	65.2	64.8	69.2 ± 0.25
	V	58.6	58.3	58.1	57.9	57.8	57.8	57.7	57.6	57.5	57.5	57.9 ± 0.61
	A+V	76.4	73.0	70.4	68.5	67.1	66.3	65.6	65.2	64.7	64.4	68.2 ± 0.42
	A+V[P]	76.9	73.4	70.6	68.6	67.3	66.3	65.6	65.1	64.7	64.4	68.3 ± 0.18
Mamba	A	77.4	73.6	70.5	68.5	66.9	65.8	65.0	64.3	63.9	63.5	67.9 ± 0.37
	V	58.2	58.1	57.9	57.8	57.6	57.5	57.5	57.4	57.4	57.3	57.7 ± 0.28
	A+V	76.0	72.5	69.8	67.9	66.6	65.6	64.8	64.2	63.8	63.5	67.5 ± 0.18
	A+V[P]	74.1	70.8	68.1	66.2	64.8	63.9	63.2	62.7	62.3	62.0	65.8 ± 0.23

(b) Results on Ego4D

Table 2: Mean average precision (mAP) scores on (a) EasyCom and (b) Ego4D at time steps 0.20 s to 2.00 s for Transformer, GRU, and Mamba architectures under Audio (A), Visual (V), and Audio+Visual (A+V) modalities. Models with [P] are pretrained on YT-Conversation. Per-timestep values come from a single random seed, while "Avg" shows mean ± SE over five seeds (see Appendix D for full multi-seed results). Using both A and V yields the best performance overall.

Dataset	Model	Target Speaker AP
EasyCom	Transformer (A+V)	**52.7**
	Random	27.2
	Silence-based	26.6
Ego4D	Transformer (A+V)	**66.8**
	Random	26.1
	Silence-based	27.7

Table 3: Performance comparison between our predictive Transformer model (A+V) and detection-based baselines on EasyCom and Ego4D datasets.

5 Results and Analysis

5.1 Quantitative Results

Tables 2 and 3 present comprehensive comparisons of our models across different modalities, datasets, and baselines. To ensure the robustness of our findings, we report performance over five random seeds with error bars in Appendix D.

Model Performance Across Modalities As shown in Table 2, the multimodal (A+V) approach generally outperforms unimodal inputs on Easy-Com, with the Transformer model achieving 58.7% mAP (compared to 56.9% for audio-only and 51.0% for visual-only). The GRU model performs best with A+V (60.6% mAP), while Mamba sees moderate improvements (57.4% mAP). On Ego4D, the Transformer with A+V attains 69.0% mAP, which is roughly on par with its audio-only counterpart. Interestingly, GRU and Mamba actually do better with audio alone, at 69.0% and 67.9% mAP respectively.

Pretraining Effects Our YT-Conversation pre-training results show that overall gains are modest. However, we do observe a small but consistent improvement in detecting *other person speaking*

Model	Modality	Avg. mAP
Transformer	A (+F)	57.0 (+8.0)
	V (+F)	51.2 (+6.4)
	A+V (+F)	**58.8 (+9.6)**
GRU	A (+ F)	57.1 (+3.9)
	V (+ F)	51.2 (+4.2)
	A+V (+ F)	61.2 (+0.9)
Mamba	A (+ F)	55.6 (+2.6)
	V (+ F)	51.3 (+3.9)
	A+V (+ F)	57.6 (+2.7)

Table 4: Impact of optical flow on utterance initiation prediction for EasyCom dataset. Values show average mAP, with performance gains from optical flow in parentheses. A: Audio, V: Visual, F: Flow. Bold indicates the best overall performance.

class, which is especially valuable in egocentric scenarios. In contrast, GRU and Mamba show little or no net gain, aligning with prior work that certain recurrent/state-space models often struggle with large-scale pretraining (Wang et al., 2023). We attribute these results to domain mismatch and the inherently noisier nature of real-world conversational videos. For a per-class breakdown and further discussion, refer to Appendix C.

Comparison with Baselines Table 3 compares our best Transformer (A+V) model with both baselines on EasyCom and Ego4D. Even though the silence-based approach benefits from an evaluation bias, our predictive model still achieves significantly higher AP (52.7% vs. 26.6% on EasyCom, and 66.8% vs. 27.7% on Ego4D). Moreover, the silence-based method performs similarly to random, indicating that requiring a fixed silence interval fails to accommodate the fluid, overlapping speech found in real-world conversations.

5.2 Motion Inputs Contribute to Turn-Taking Prediction

Since many non-verbal cues involve motion, we hypothesized that incorporating optical flow could improve utterance initiation prediction. To test this, we extracted optical flow using the Denseflow toolkit (Wang et al., 2020) with the TV-L1 algorithm (Zach et al., 2007), following a similar process to our RGB feature extraction. Optical flow is a computer vision technique that estimates object motion between consecutive video frames by calculating the apparent motion of brightness patterns. This is useful for tracking movement and analyzing dynamic scenes.
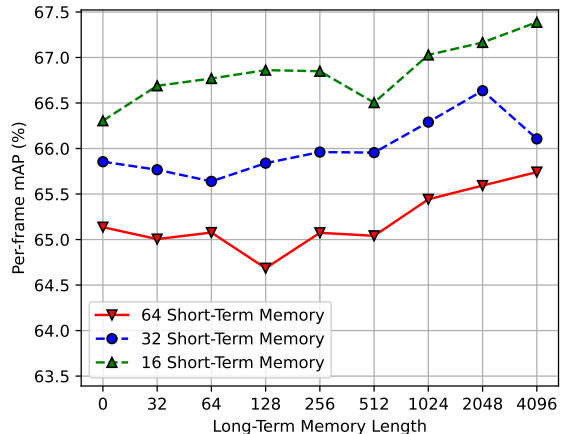
Table 4 presents the results of our experiment on

Figure 6: Utterance initiation prediction with varying transformer memory length. a shorter context window for short-term memory and a longer context window for long-term memory generally show better results.

EasyCom. Incorporating optical flow consistently improved performance across all model types and input combinations. These results suggest that motion information provides valuable cues for predicting utterance initiation, complementing static visual and audio features to enable more accurate predictions of speech onset.

5.3 Models Do Not Exploit Short-Term Information Well

Our framework uses online processing models that rely on context length to capture historical information. Because dialogue context is crucial in turn-taking (Skantze, 2021), the choice of context can strongly affect utterance initiation. Figure 6 shows how the Transformer model's performance varies with different long-term and short-term window sizes. While extending the long-term window helps, increasing the short-term window unexpectedly degrades performance. This suggests that although a broader context provides valuable cues, an overly large short-term window may introduce noise or irrelevant data, reducing accuracy. These findings highlight the importance of balancing long-term and short-term context in untrimmed videos to optimize turn-taking predictions.

5.4 Runtime Analysis

We evaluated the computational efficiency of our models by measuring their frames per second (FPS), parameter counts, and floating-point operations (GFLOPs) on a single RTX3090 GPU using the EasyCom dataset, as shown in Table 5. The
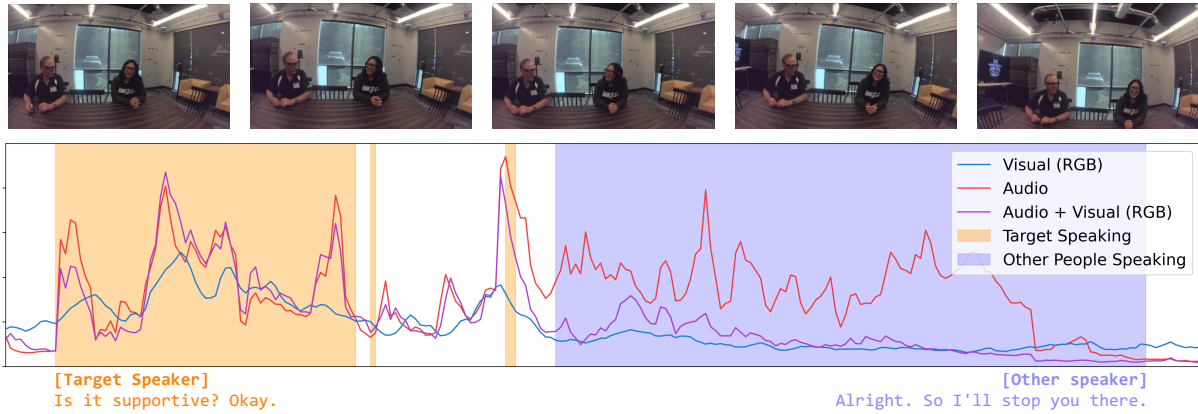
Figure 7: Qualitative results on EasyCom. The predicted scores are shown in lines and the ground-truth label is shown in regions. The blue line represents a model with RGB input, the red line represents a model with audio input, and the purple line represents a model with audio and visual input.

Model	FPS	Parameters	GFLOPs
Transformer	99.8	67.21M	129.48
RNN	13939.5	34.6M	206.52
Mamba	12009.3	83.1M	610.93

Table 5: Runtime analysis of different models on a single RTX3090 GPU.

RNN-based model achieved the highest throughput with 13,939.5 FPS, while maintaining a relatively low parameter count of 34.6M and requiring 206.52 GFLOPs. The Mamba-based model followed with 12,009.3 FPS, though it has the highest parameter count (83.1M) and the largest computational requirement (610.93 GFLOPs). The Transformer-based model ran at 99.8 FPS with 67.21M parameters and 129.48 GFLOPs, achieving real-time processing but at a lower frame rate than RNN and Mamba. Overall, these results indicate that all three architectures are capable of real-time processing.

5.5 Qualitative Results

Figure 7 shows the qualitative results based on Transformer. Our observations indicate that the model using only RGB features struggles to effectively distinguish between speaking and non-speaking segments, leading to frequent misclassifications. In contrast, the model utilizing audio input shows notable improvement in predicting the target speaker's speech. However, the audio-only model often assigns high probabilities to the speech of other individuals, resulting in less accurate turn-taking. Notably, the model that integrates both audio and visual inputs demonstrates superior performance. This multimodal model effectively distin-

guishes the target speaker from others, accurately identifying speaking segments while minimizing false positives from other speakers.

6 Conclusion

We introduced EgoSpeak, a novel framework for real-time speech initiation prediction from an in-the-wild, first-person viewpoint. EgoSpeak integrates four key capabilities to better handle complex, dynamic real-world conversations. We also presented YT-Conversation, a large-scale dataset of in-the-wild YouTube videos for pretraining.

Empirical results on two egocentric conversational video datasets demonstrate that EgoSpeak significantly outperforms random and silence-based baselines in real time. Notably, our analysis revealed that incorporating optical flow significantly improves performance and highlighted a counterintuitive finding on context length. We encourage further exploration of large-scale conversational datasets, improved multimodal modeling techniques, and integration with large language models to generate responses, advancing turn-taking in real-world settings.

7 Limitations

Our proposed method relies on pre-encoded features, which can limit both performance and frames per second (FPS). A potential solution is to adopt an end-to-end framework that learns to extract relevant features directly from raw input. For example, E2E-LOAD (Cao et al., 2023) demonstrates improvements in both state-of-the-art performance and FPS in online action recognition task through such an approach.

While our YT-Conversation dataset provides diverse pretraining data, it may not fully capture the nuances of first-person interactions. Future work could explore methods to augment the dataset with more egocentric conversational data, potentially improving model performance in first-person scenarios.

Lastly, our current approach does not explicitly model speaker-specific behaviors. Future research could incorporate individual speaking patterns and tendencies, by analyzing larger egocentric conversational datasets. By capturing these speaker-specific nuances, future models may better anticipate utterance initiation points, particularly in prolonged conversations with familiar participants.

8 Ethical Statement

Data Collection and Privacy Considerations Although the YT-Conversation, derived from publicly shared YouTube videos, provides natural conversations for training AI models, there is a possibility it may capture the facial features of the participants. However, the YT-Conversation dataset was collected under the principles of informed consent and data anonymization, adhering to the ACM Code of Ethics 1.6 (Respect privacy).

Informed consent We selected videos that participants are likely aware of and have consented to be recorded and publicly shared, such as podcasts, interviews, and face-to-face conversations.

Data Anonymization We only released the YouTube IDs rather than the raw YouTube videos so that content creators can remove their videos from YouTube anytime, which will automatically exclude them from our dataset. Moreover, our transcripts only include the time ranges for the start and end of the speech, along with the corresponding video frames without personal information.

9 Acknowledgements

References

Joungbin An, Hyolim Kang, Su Ho Han, Ming-Hsuan Yang, and Seon Joo Kim. 2023. Miniroad: Minimal rnn framework for online action detection. In *International Conference on Computer Vision (ICCV)*.

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

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

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Shuqiang Cao, Weixin Luo, Bairui Wang, Wei Zhang, and Lin Ma. 2023. E2e-load: end-to-end longform online action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10422–10432.

MMAction2 Contributors. 2020. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2.

Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. 2016. Online action detection. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 269–284. Springer.

Jacob Donley, Vladimir Tourbabin, Jung-Suk Lee, Mark Broyles, Hao Jiang, Jie Shen, Maja Pantic, Vamsi Krishna Ithapu, and Ravish Mehra. 2021. Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. *arXiv preprint arXiv:2107.04174*.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283.

Erik Ekstedt and Gabriel Skantze. 2022. Voice activity projection: Self-supervised learning of turn-taking events. *arXiv preprint arXiv:2205.09812*.

Ruchao Fan, Pan Zhou, Wei Chen, Jia Jia, and Gang Liu. 2018. An online attention-based model for speech recognition. *arXiv preprint arXiv:1811.05247*.

Mehdi Fatan, Emanuele Mincato, Dimitra Pintzou, and Mariella Dimiccoli. 2024. 3m-transformer: A multi-stage multi-stream multimodal transformer for embodied turn-taking prediction. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8050–8054. IEEE.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

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

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.

Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Hyolim Kang, Kyungmin Kim, Yumin Ko, and Seon Joo Kim. 2021. Cag-qil: Context-aware actionness grouping via q imitation learning for online temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13729–13738.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Hatim Khouzaimi, Romain Laroche, and Fabrice Lefevre. 2015. Optimising turn-taking strategies with reinforcement learning. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 315–324.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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. *Proc. Interspeech 2023*, pages 2658–2662.

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*, pages 436–444.

Stephen C Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:136034.

Siyan Li, Ashwin Paranjape, and Christopher D Manning. 2022. When can i speak? predicting initiation points for spoken dialogue agents. *arXiv preprint arXiv:2208.03812*.

Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. 2022. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586.

Angelika Maier, Julian Hough, David Schlangen, et al. 2017. Towards deep end-of-turn prediction for situated spoken dialogue systems.

Haoran Miao, Gaofeng Cheng, Changfeng Gao, Pengyuan Zhang, and Yonghong Yan. 2020. Transformer-based online ctc/attention end-to-end speech recognition architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6084–6088. IEEE.

Saki Mizuno, Nobukatsu Hojo, Satoshi Kobashikawa, and Ryo Masumura. 2023. Next-speaker prediction based on non-verbal information in multi-party video conversation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura. 2023. Multimodal voice activity prediction: Turn-taking events detection in expert-novice conversation. In *Proceedings of the 11th International Conference on Human-Agent Interaction*, pages 13–21.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Laura Rettig, Mourad Khayati, Philippe Cudré-Mauroux, and Michał Piórkowski. 2019. Online anomaly detection over big data streams. *Applied Data Science: Lessons Learned for the Data-Driven Business*, pages 289–312.

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*, pages 220–230.

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

Junxiong Wang, Jing Nathan Yan, Albert Gu, and Alexander Rush. 2023. Pretraining without attention. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 58–69, Singapore. Association for Computational Linguistics.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer.

Shiguang* Wang, Zhizhong* Li, Yue Zhao, Yuanjun Xiong, Limin Wang, and Dahua Lin. 2020. denseflow. https://github.com/open-mmlab/denseflow.

Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, and Nong Sang. 2021. Oadtr: Online action detection with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7565–7575.

Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. 2021. Long short-term transformer for online action detection. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Jiudong Yang, Peiying Wang, Yi Zhu, Mingchao Feng, Meng Chen, and Xiaodong He. 2022. Gated multimodal fusion with contrastive learning for turn-taking prediction in human-robot dialogue. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7747–7751. IEEE.

Victor H Yngve. 1970. On getting a word in edgewise. In *Papers from the sixth regional meeting Chicago Linguistic Society, April 16-18, 1970, Chicago Linguistic Society, Chicago*, pages 567–578.

Christopher Zach, Thomas Pock, and Horst Bischof. 2007. A duality based approach for realtime tv-l 1 optical flow. In *Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29*, pages 214–223. Springer.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. 2022. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer.

Mario Zusag, Laurin Wagner, and Bernhad Thallinger. 2024. Crisperwhisper: Accurate timestamps on verbatim speech transcriptions. In *Proc. Interspeech 2024*, pages 1265–1269.

A Implementation Details

A.1 Architecture & Hyperparameters

This section provides detailed information on the architectures and hyperparameters used for each model in our experiments. We set the anticipation length to 10 timesteps for all models, predicting up to 2 seconds into the future. All experiments were done with a single RTX3090 GPU within one day.

Transformer-based Model (LSTR) For the transformer model, we configured 16 attention heads and 1024-dimensional hidden units in the transformer blocks. The LSTR encoder processes long context windows up to 2048 frames, while the decoder handles shorter context windows up to 32 frames. We trained this model using the Adam optimizer (Kingma and Ba, 2014) with a weight decay of 5×10^{-5}. The learning rate was scheduled to increase linearly from zero to 7×10^{-5} during the first 40% of training iterations, then decrease to zero following a cosine function. We trained the transformer model for 50 epochs with a batch size of 16.

RNN-based Model For the RNN model, we used 2048-dimensional embeddings and 1024-dimensional hidden units. This model was trained for 30 epochs with a batch size of 64. We used the same optimizer and learning rate schedule as the transformer model.

Mamba-based Model The Mamba-based model builds upon the RNN architecture, replacing the GRU layer with a Mamba block. We set the SSM state factor to 16, local convolution width to 4, and block expansion factor to 2. The training settings were kept consistent with the RNN model.

A.2 Training Objective

For training, we use cross-entropy loss between predicted confidence scores s_T at time T and the ground-truth label $y_T \in \{0, 1, \ldots, K\}$. K is the number of classes and s_T^k is the k-th element of the probability vector s_T. For Transformer-based model, α_T is always 1. For RNN-based and Mamba-based models, α_T is used to modulate the contribution of intermediate time steps during the computation of the loss. Specifically, α_T takes the value 1 only at a designated step $t = L$ and 0 otherwise.

We also define a temporal window of length L, which determines the final step contributing to the
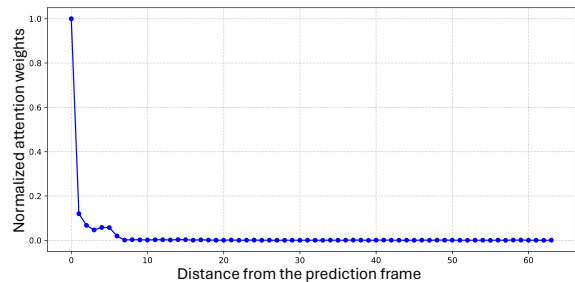

Figure 8: Attention weight of a transformer encoders. Transformer models focus on mostly local context for utterance initiation.

objective function:

$$J(y_T, s_T; T) = -\sum_{k=0}^{K} \alpha_T \delta(k - y_T) \log s_T^k,$$

A.3 Feature Extraction

RGB Features As mentioned in Section 4.3, videos are downsampled to 20 FPS and processed in 4-frame chunks, resulting in a 5 FPS prediction rate. We use ResNet-50 (He et al., 2016) initialized with weights from a video action recognition model (Wang et al., 2016), implemented via MMAction2 (Contributors, 2020). The center frame of each chunk is sampled for feature extraction. For the EasyCom dataset, we cropped all clips in each session to remain only video frames and merged them to make one video per session.

Audio Features We use wav2vec2's (Baevski et al., 2020) multi-layer convolutional feature encoder, as noted in Section 4.3. Every 10 encoded audio features are concatenated temporally to match the 5 FPS RGB features.

B Importance of Recent Frames

Figure 8 shows the distribution of attention weights across the encoder layers of a transformer model in the context of predicting utterance initiation in real-world conversations (Wang et al., 2021). The attention weights of the test set were averaged with respect to the layers, multi-heads, and batch and then normalized. These weights reveal the significance assigned to each frame in the sequence during prediction. Our analysis shows that the model focuses predominantly on the recent frames, with attention weights diminishing notably as the distance from the current frame increases. This pattern indicates that recent frames have a greater impact on the model's predictions for utterance initiation.

Dataset	Model	Avg mAP	Per-Class AP (%)		
			Background	Target Speaker	Other Speaker
EasyCom	Transformer	58.79	43.17	52.74	80.46
	Transformer[P]	59.01	42.94	52.91	81.17
Ego4D	Transformer	69.61	73.50	66.78	68.56
	Transformer[P]	68.79	71.80	64.46	70.11

Table 6: Per-class average precision (AP) for the Transformer with and without ([P]) YT-Conversation pretraining on EasyCom and Ego4D. Although overall gains are modest, we observe a notable improvement for *Other Speaker* detection.
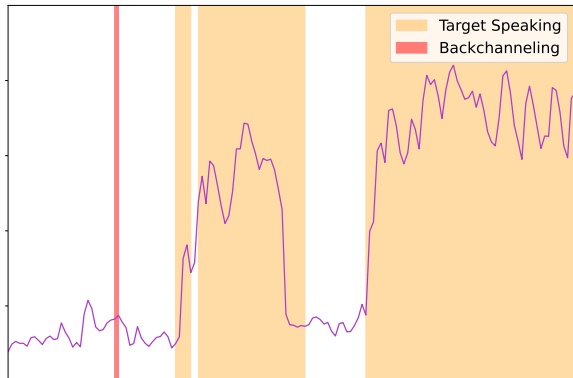

Figure 9: Failure case on EasyCom. The orange region represents the target speaker speaking, and the red region represents the target speaker's backchanneling.

C   Error Analysis

Pretraining on YT-Conversation   We observed that YT-Conversation pretraining yields modest overall gains, but a notable improvement for *other person speaking* class (+0.7% on EasyCom, +1.5% on Ego4D). Table 6 lists the per-class average precision (AP) for the Transformer model with and without pretraining. Although this benefit can be crucial in egocentric scenarios—where identifying others' speech fosters smoother turn-taking—gains for *Background* and *Target Speaker* remain unchanged or slightly negative. We attribute this to domain mismatch (YouTube interviews vs. dynamic ego footage) and noisy data from real-world conversational videos such as visual effects or subtitles. Future work might address these limitations by bridging domain gaps—e.g., with domain adaptation—or introducing video filtering to obtain higher-quality conversational clips.

Backchannels   While our method aims to predict any utterance initiation point, there is a short and brief response that occurs when one participant is speaking and the listener reacts to signify the listener's attention, understanding, or emotion rather than take turns and speak. This behavior is referred

Model	Modality	EasyCom	Ego4D
Transformer	A	56.9 ± 0.12	69.2 ± 0.08
	V	51.0 ± 0.18	58.1 ± 0.60
	A+V	58.7 ± 0.29	69.6 ± 0.54
	A+V[P]	58.5 ± 0.59	68.8 ± 0.41
GRU	A	57.0 ± 0.68	69.2 ± 0.55
	V	51.7 ± 0.65	57.9 ± 1.36
	A+V	60.6 ± 0.38	68.2 ± 0.95
	A+V[P]	57.0 ± 0.65	68.3 ± 0.41
Mamba	A	55.4 ± 1.39	67.9 ± 0.83
	V	50.9 ± 0.48	57.7 ± 0.64
	A+V	57.4 ± 0.58	67.5 ± 0.41
	A+V[P]	55.8 ± 0.97	65.8 ± 0.51

Table 7: Performance comparison of models across five different seeds on the EasyCom and Ego4D datasets. Each value represents the average mAP across the seeds, along with the standard error.

to as "backchannels" (Yngve, 1970; Skantze, 2021). We observed that prediction scores usually do not increase before backchanneling. Figure 9 illustrates this phenomenon, showing how the model's prediction scores do not significantly increase before a backchanneling event, in contrast to regular speaking turns.

D   Descriptive Statistics of Experimental Results

We evaluated each model with five random seeds {0, 10, 20, 29, 42} to measure performance variance. Table 7 shows the multi-seed mean average precision (mAP) on EasyCom and Ego4D, while Tables 8 and 9 provide per-timestep results (mean ± standard error). These tables complement the main text figures (Tables 2 and 3) by offering a full breakdown of multi-seed performance at each time step, ensuring transparency and robustness in our results.

E   YT-Conversation Pseudo Annotation Quality Validation

To validate the quality of pseudo-annotations in our YT-Conversation dataset, we conducted a human evaluation study on 100 segments randomly sampled from 10 videos, excluding the first five segments of each (typically non-conversational teasers). Each segment received a label alignment score on a 5-point scale: (1) completely misaligned, with timestamps far off from actual speech; (2) poor alignment, missing large portions, or labeling silence as speech; (3) adequate but potentially off by 0.5–1 second; (4) good alignment, within

about 0.5 second of true boundaries; and (5) excellent alignment, nearly matching human labels. Across all evaluated segments, the average alignment score was 2.147. We want to note that as ASR models continue to advance (Zusag et al., 2024), the pseudo-label will be precise as well. We also use these pseudo-labels only for pretraining, ensuring the evaluations remain robust with human-annotated labels.

F   Use of AI Assistants

We used Claude 3.5 Sonnet to revise the paper and code, and GitHub Copilot to write the code.

Model	Modality	mAP (%)				
		0.20s	0.40s	0.60s	0.80s	1.00s
Transformer	A	72.2 ± 0.05	65.2 ± 0.08	60.3 ± 0.09	56.8 ± 0.09	54.4 ± 0.08
	V	52.0 ± 0.06	51.7 ± 0.08	51.6 ± 0.06	51.3 ± 0.09	51.1 ± 0.06
	A+V	73.8 ± 0.12	66.9 ± 0.16	62.1 ± 0.14	58.5 ± 0.16	56.3 ± 0.15
	A+V^P	73.4 ± 0.17	66.8 ± 0.18	61.8 ± 0.22	58.3 ± 0.31	56.1 ± 0.30
GRU	A	71.5 ± 0.30	65.0 ± 0.38	60.1 ± 0.29	57.0 ± 0.30	55.0 ± 0.31
	V	53.0 ± 0.28	52.7 ± 0.34	52.4 ± 0.25	52.0 ± 0.34	51.7 ± 0.33
	A+V	73.5 ± 0.30	68.1 ± 0.20	63.7 ± 0.30	60.7 ± 0.21	59.1 ± 0.25
	A+V^P	70.8 ± 0.41	64.9 ± 0.26	60.1 ± 0.29	56.9 ± 0.31	55.0 ± 0.31
Mamba	A	67.5 ± 0.88	62.2 ± 0.98	58.4 ± 0.76	55.7 ± 0.71	54.0 ± 0.61
	V	52.2 ± 0.18	51.8 ± 0.18	51.5 ± 0.18	51.1 ± 0.19	50.9 ± 0.17
	A+V	71.8 ± 0.22	65.4 ± 0.15	60.5 ± 0.17	57.1 ± 0.14	55.0 ± 0.13
	A+V^P	68.9 ± 0.51	63.2 ± 0.47	59.1 ± 0.43	56.0 ± 0.44	54.0 ± 0.41

(a) 5 Different Seeds Results on EasyCom - Timesteps from 0.20s to 1.00s

Model	Modality	mAP (%)				
		1.20s	1.40s	1.60s	1.80s	2.00s
Transformer	A	53.1 ± 0.06	52.4 ± 0.09	52.0 ± 0.08	51.6 ± 0.07	51.4 ± 0.10
	V	50.9 ± 0.06	50.8 ± 0.07	50.5 ± 0.11	50.3 ± 0.12	50.1 ± 0.11
	A+V	55.0 ± 0.18	54.1 ± 0.14	53.7 ± 0.21	53.3 ± 0.15	53.0 ± 0.15
	A+V^P	54.8 ± 0.31	54.1 ± 0.31	53.5 ± 0.36	53.2 ± 0.28	52.7 ± 0.31
GRU	A	53.8 ± 0.43	52.9 ± 0.35	52.2 ± 0.43	51.5 ± 0.46	50.9 ± 0.42
	V	51.6 ± 0.29	51.2 ± 0.31	51.1 ± 0.31	50.8 ± 0.28	50.6 ± 0.28
	A+V	58.1 ± 0.24	57.2 ± 0.15	56.3 ± 0.23	55.4 ± 0.17	54.4 ± 0.15
	A+V^P	53.8 ± 0.37	53.0 ± 0.40	52.4 ± 0.44	51.8 ± 0.34	51.4 ± 0.40
Mamba	A	52.9 ± 0.50	52.0 ± 0.47	51.1 ± 0.43	50.2 ± 0.50	49.6 ± 0.47
	V	50.7 ± 0.23	50.5 ± 0.26	50.4 ± 0.34	50.0 ± 0.31	49.7 ± 0.30
	A+V	53.9 ± 0.36	53.5 ± 0.43	53.1 ± 0.46	52.3 ± 0.52	51.8 ± 0.43
	A+V^P	52.7 ± 0.41	51.8 ± 0.51	51.4 ± 0.46	50.7 ± 0.44	50.1 ± 0.50

(b) 5 Different Seeds Results on EasyCom - Timesteps from 1.20s to 2.00s

Table 8: Per-frame performance over 5 different random seeds on EasyCom.

Model	Modality	mAP (%)				
		0.20s	0.40s	0.60s	0.80s	1.00s
Transformer	A	78.8 ± 0.06	74.9 ± 0.05	71.8 ± 0.04	69.7 ± 0.05	68.1 ± 0.02
	V	58.7 ± 0.25	58.5 ± 0.25	58.4 ± 0.25	58.2 ± 0.24	58.1 ± 0.24
	A+V	78.1 ± 0.20	74.3 ± 0.23	71.5 ± 0.24	69.4 ± 0.26	68.0 ± 0.27
	A+V^P	78.4 ± 0.21	74.5 ± 0.17	71.5 ± 0.18	69.4 ± 0.20	67.9 ± 0.19
GRU	A	78.6 ± 0.27	74.8 ± 0.30	71.8 ± 0.27	69.6 ± 0.27	68.1 ± 0.28
	V	58.6 ± 0.59	58.3 ± 0.58	58.1 ± 0.58	57.9 ± 0.59	57.8 ± 0.61
	A+V	76.4 ± 0.41	73.0 ± 0.42	70.4 ± 0.40	68.5 ± 0.39	67.1 ± 0.40
	A+V^P	76.9 ± 0.18	73.4 ± 0.15	70.6 ± 0.15	68.6 ± 0.17	67.3 ± 0.19
Mamba	A	77.4 ± 0.53	73.6 ± 0.44	70.5 ± 0.35	68.5 ± 0.35	66.9 ± 0.35
	V	58.2 ± 0.29	58.1 ± 0.28	57.9 ± 0.29	57.8 ± 0.30	57.6 ± 0.29
	A+V	76.0 ± 0.23	72.5 ± 0.22	69.8 ± 0.20	67.9 ± 0.18	66.6 ± 0.16
	A+V^P	74.1 ± 0.38	70.8 ± 0.34	68.1 ± 0.34	66.2 ± 0.30	64.8 ± 0.33

(a) 5 Different Seeds Results on Ego4D - Time Steps from 0.20s to 1.00s

Model	Modality	mAP (%)				
		1.20s	1.40s	1.60s	1.80s	2.00s
Transformer	A	67.0 ± 0.03	66.3 ± 0.03	65.7 ± 0.04	65.1 ± 0.04	64.7 ± 0.04
	V	58.0 ± 0.26	57.9 ± 0.25	57.8 ± 0.26	57.7 ± 0.24	57.7 ± 0.25
	A+V	67.0 ± 0.26	66.3 ± 0.25	65.7 ± 0.25	65.3 ± 0.26	64.9 ± 0.27
	A+V^P	66.7 ± 0.20	65.9 ± 0.20	65.4 ± 0.21	65.0 ± 0.21	64.5 ± 0.18
GRU	A	66.9 ± 0.24	66.2 ± 0.25	65.6 ± 0.25	65.2 ± 0.23	64.8 ± 0.26
	V	57.8 ± 0.60	57.7 ± 0.61	57.6 ± 0.62	57.5 ± 0.64	57.5 ± 0.65
	A+V	66.3 ± 0.39	65.6 ± 0.45	65.2 ± 0.50	64.7 ± 0.47	64.4 ± 0.44
	A+V^P	66.3 ± 0.21	65.6 ± 0.21	65.1 ± 0.24	64.7 ± 0.22	64.4 ± 0.28
Mamba	A	65.8 ± 0.35	65.0 ± 0.35	64.3 ± 0.33	63.9 ± 0.36	63.5 ± 0.37
	V	57.5 ± 0.27	57.5 ± 0.28	57.4 ± 0.29	57.4 ± 0.29	57.3 ± 0.28
	A+V	65.6 ± 0.17	64.8 ± 0.17	64.2 ± 0.19	63.8 ± 0.22	63.5 ± 0.22
	A+V^P	63.9 ± 0.27	63.2 ± 0.19	62.7 ± 0.18	62.3 ± 0.09	62.0 ± 0.11

(b) 5 Different Seeds Results on Ego4D - Time Steps from 1.20s to 2.00s

Table 9: Per-frame performance over 5 different random seeds on Ego4D.