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

Anonymous ACL submission

Abstract

Predicting when to initiate speech remains a fundamental challenge for conversational agents. We introduce EgoSpeak, a novel framework for real-time speech initiation prediction in egocentric streaming video. EgoSpeak uniquely advances existing turn-taking models by integrating a first-person point of view, RGB processing, online processing, and untrimmed video processing, bridging the gap between simplified experimental setups and complex natural conversations. We also present the YT-Conversation dataset, a diverse collection of conversational videos from YouTube including interviews and casual dialogues. Our experiments propose that EgoSpeak's effectiveness in realistic scenarios on EasyCom and Ego4D datasets, outperforming random and silence-based baselines in real-time. Moreover, we also show the importance of multimodal inputs and context length in utterance initiation prediction.

1 Introduction

Human-like conversational agents have long been a crucial objective in artificial intelligence. An essential aspect of human-like conversation is the ability to choose the appropriate timing for speech, involving turn-taking skills to accurately predict when to start and stop speaking (Duncan, 1972). Despite recent advances in turn-based dialogue systems (Achiam et al., 2023; Kim et al., 2022; Kang et al., 2024), this active speaking capability remains underexplored.

Most turn-taking studies assume simplified environments with a fixed number of interlocutors and clear, well-defined turn-taking cues (Li et al., 2022; Onishi et al., 2023; Kurata et al., 2023). However, these setups fail to fully capture the complexity and variability of real-world conversations, where factors such as varying number of speakers, and overlapping speech or interruptions can make turn-taking far more challenging (Skantze, 2017, 2021). Additionally, real-world conversations often involve multiple participants with dynamic roles, informal social cues, and contextual shifts, all of which complicate the turn-taking process.

Figure 1: We present EgoSpeak, a framework for conversational agents in the wild. EgoSpeak models speech initiation using an egocentric dataset across various real-world scenarios.

To bridge this gap, we propose EgoSpeak to predict utterance initiation in egocentric streaming videos. EgoSpeak incorporates four key innovations: (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. These innovations collectively enable EgoSpeak to handle the complexities of natural conversations more effectively than previous methods (see Table 1).

We evaluate 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

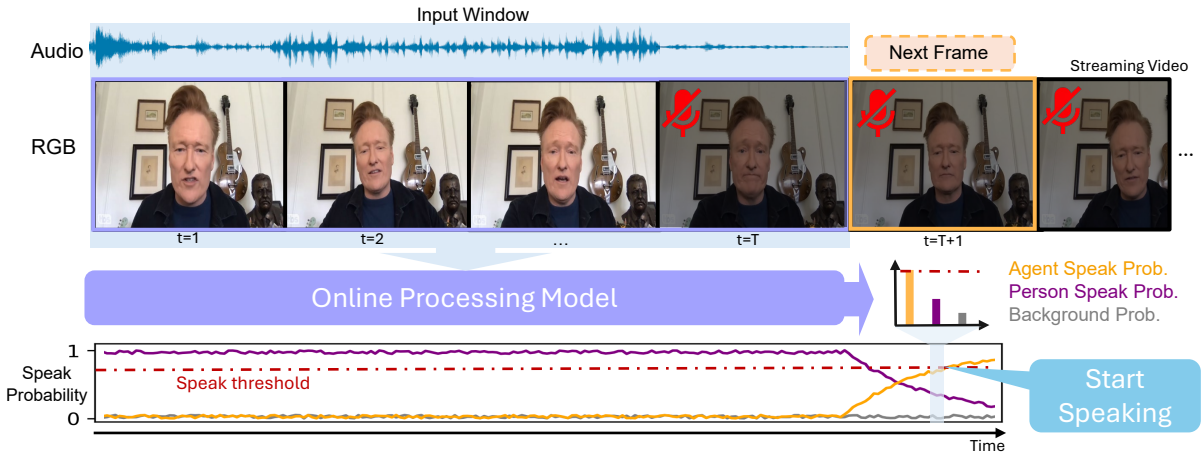

Figure 2: Overview of the EgoSpeak framework for real-time speech initiation prediction. The framework processes an untrimmed egocentric video stream (middle) and its corresponding audio (top), performing per-frame classification to predict probabilities for three classes: background (no speech), other person speaking, and target speaker (camera wearer) speaking (bottom). The model anticipates speech status for near-future frames, enabling proactive speech initiation for conversational agents.

advances the field of conversational AI, offering a robust solution for dynamic, intermittent conversations with varying numbers of speakers.

In summary, our contributions include:

1. The EgoSpeak framework for predicting utterance initiation in egocentric streaming video.
2. YT-Conversation, a pretraining corpus of in-the-wild conversation videos.
3. Empirical validation on two distinct datasets (EasyCom and Ego4D), demonstrating EgoSpeak's effectiveness in real-world scenarios.

2 Related Works

Turn-taking. Turn-taking research has evolved from simple audio-based models (Duncan, 1972; Khouzaimi et al., 2015) to sophisticated multimodal 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.

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 Overview

EgoSpeak is designed for in-the-wild conversational agents, where "in-the-wild" refers to real-world conditions outside controlled environments with unpredictable variables and numerous influ-

	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.

encing factors. Unlike previous simplified turn-taking or utterance initiation settings, our framework aims to predict the start of speech in natural human conversation using first-person perspective input. EgoSpeak is based on the intuition that the moments when the target speaker (camera wearer) speaks represent appropriate times for speech initiation. By predicting these moments, our framework learns natural turn-taking behavior, identifying when to speak even after long silences.

3.2 Task Definition

We aim to predict the target speaker's speech in an egocentric streaming video. In an egocentric video, the camera wearer is naturally identified as the target speaker. Given the real-time nature of streaming video, our framework is constrained to analyze only the information available up to the current moment.

We tackle this problem as a classification task aimed at predicting the speech status in the near future. This approach enables us to derive the start-of-speech label through thresholding. Formally, an online stream video up to timestep t is defined as $X_t = [x_1, \ldots, x_t]$, where x represents a data point that can manifest in multiple modalities, such as visual frame x^v or auditory signal x^a. We denote representation vectors obtained by processing each datapoint x_t through off-the-shelf feature extractors and then concatenated as z_t. We also define a temporal window of length L to create inputs with a consistent past view width, denoted as $Z_{t-L+1,t} = [z_{t-L+1}, \ldots, z_t]$. When the anticipation length is α, given the input stream up to timestep t ($Z_{t-L+1,t}$), the model performs a three-way classification for a future range from $t+1$ to $t+\alpha$. The model predicts the speech status for future timesteps, performing anticipation for timesteps up to $t+\alpha$. The final scores have $[\alpha, 3]$ shape, corresponding to the anticipation length and the three classes (background (no speech), target speaker speaking, other speaking).

Based on the confidence scores for each label, we can predict the start of speeches by identifying the moments when the confidence score for the label *target speaker speaking* crosses the threshold from below to above.

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.

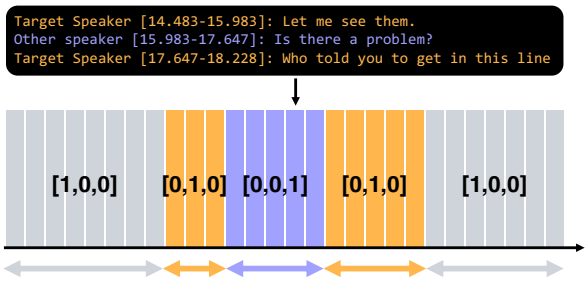

Figure 3: Converting Transcript to Per-Frame Labels. Colors indicate: gray - background, orange - target speaker speaking, purple - other speaker speaking. Labels are one-hot encoded for classification.

Frame-level Speech Labeling. Figure 3 illustrates this labeling process, showing how transcript timestamps are converted to per-frame, one-hot encoded labels. 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.

3.3 YT-Conversation: Dataset for Multimodal Conversation Pretraining

The development of robust conversational AI has been hindered by a lack of large-scale, diverse multimodal datasets. Previous approaches often relied on limited, controlled datasets from controlled settings like video calls or laboratory environments. These fail to capture the complexity of real-world interactions. To address this gap, we introduce YT-Conversation, a novel dataset derived from diverse YouTube content including interviews, podcasts, and casual dialogues. Although not primarily egocentric, YT-Conversation captures realistic scenarios that enhance the adaptability of conversational AI frameworks. By leveraging this rich dataset, we aim to develop more versatile systems capable of handling the nuances of real-world communication, even in first-person perspective interactions.

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.

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. This approach allows for easy expansion by simply including more channels or extending the sampling period.

Pseudo Per-frame Annotation for Collected Videos To create pseudo-labels for training without manual annotation, we employed the voice activity detection model from the Pyannote toolkit (Plaquet and Bredin, 2023; Bredin, 2023). This process generated per-frame pseudo-labels for speech activity based on detected timestamps. We applied a 0.2-second threshold for deleting small non-speech segments and short speech segments.

4 Experimental Setup

4.1 Dataset

We propose to use publicly available egocentric conversational video datasets for evaluation: EasyCom (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.

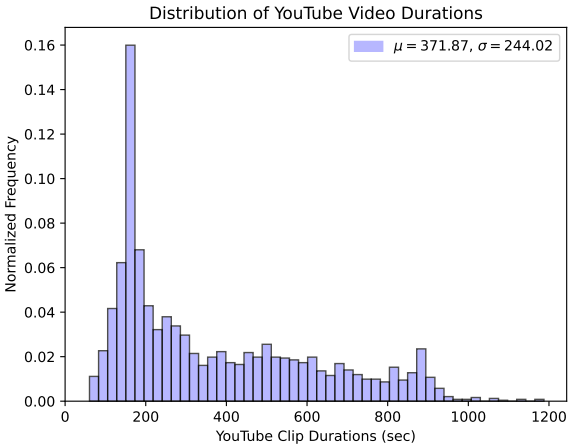

Figure 5: Video duration distribution for YT-Conversation. Our online formulation allows the use of long video clips, some even exceeding 900 seconds.

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 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 architectural specifications and hyperpa-

rameters for the neural models are provided in Appendix A.1.

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. This baseline simulates commercial spoken dialogue agents by initiating speech after detecting a 600ms silence period following utterances of other speakers. It uses ground truth speech activity labels of non-target speakers. For evaluation, we generously consider the entire subsequent speech segment as correct if the start is correctly identified, while incorrect classifications are limited to the specific misclassified timestep. This approach, while potentially overestimating performance compared to real-world implementations, provides a meaningful comparison point for our proposed models.

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 containing 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. Existing turn-taking evaluation protocols are not well-aligned with real-world conversational agents. Offline methods (Lee et al., 2023; Kurata et al., 2023) use the F1-score after processing an entire clip as a classification task. Even online methods (Ekstedt and Skantze, 2022; Onishi et al., 2023) still measure the F1-score on sampled events with thresholds around turn-takings. However, for real-world conversational agents, a decision should be made at each frame, even when there is no ongoing conversation or during overlapping speech. As Heldner and Edlund (2010) suggested, overlaps occur frequently in human conversation. To represent such natural conversational dynamics more effectively, the per-frame prediction should be measured.

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 allows us to evaluate the performance of anticipating the target speaker's speech up to 10 timesteps (2 seconds) into the future. To calculate mAP, we follow these steps: 1) Sort the confidence scores for each frame in descending order. 2) Set the highest score as the initial threshold. 3) Calculate precision and recall for all confidence scores based on this threshold. 4) Repeat steps 2-3 using each subsequent score as a new threshold. 5) Compute the average precision across all thresholds. 6) Repeat this process for each class and timestep, then average the results to obtain the final mAP.

5 Results and Analysis

5.1 Quantitative Results

Tables 2 and 3 present comprehensive comparisons of our models across different modalities, datasets, and baselines.

Model Performance Across Modalities As shown in Table 2, the multimodal (A+V) approach generally outperforms unimodal inputs. On EasyCom, the Transformer model achieves 59.0% mAP with pretrained A+V, compared to 57.0% for audio-only and 51.2% for visual-only. The GRU model performs best with A+V (61.2% mAP), while Mamba shows moderate improvements with multimodal input (57.6% mAP). On Ego4D, the Transformer and GRU models maintain strong performance with A+V (69.6% and 69.3% mAP respec-

Model	Modality	Pre.	mAP (%)										
			0.20s	0.40s	0.60s	0.80s	1.00s	1.20s	1.40s	1.60s	1.80s	2.00s	Avg
Transformer	A		72.3	65.0	60.4	56.9	54.5	53.0	52.5	51.9	51.7	51.7	57.0
	V		52.2	51.9	51.7	51.5	51.3	51.1	50.9	50.8	50.5	50.4	51.2
	A+V		**73.9**	67.0	**62.3**	58.7	56.5	55.1	54.3	53.8	53.4	52.8	58.8
	A+V	✓	**73.9**	**67.3**	**62.3**	**58.9**	**56.6**	**55.3**	**54.6**	**54.2**	**53.7**	**53.3**	**59.0**
GRU	A		71.9	65.7	60.8	57.4	55.5	54.1	52.9	51.9	50.8	49.8	57.1
	V		52.5	52.4	51.9	51.6	51.2	50.9	50.6	50.4	50.2	50.0	51.2
	A+V		**74.2**	**68.4**	**64.3**	**61.1**	**59.8**	**58.8**	**57.5**	**57.1**	**55.9**	**54.8**	**61.2**
	A+V	✓	69.2	63.9	59.5	56.4	54.7	53.5	53.1	52.6	52.1	51.2	56.6
Mamba	A		67.4	62.8	58.8	55.9	54.3	53.2	52.4	51.3	50.2	49.5	55.6
	V		52.2	52.0	51.8	51.7	51.4	51.3	51.2	50.9	50.4	50.1	51.3
	A+V		**71.8**	**65.4**	**61.0**	**57.4**	**55.4**	**54.4**	**53.5**	**52.9**	**52.2**	**51.9**	**57.6**
	A+V	✓	69.4	63.7	59.6	56.8	54.7	53.4	53.3	52.6	51.5	50.5	56.5

(a) Results on EasyCom

Model	Modality	Pre.	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	64.8	69.2
	V		58.3	58.2	58.0	57.8	57.7	57.5	57.4	57.3	57.3	57.2	57.7
	A+V		78.4	74.8	**72.1**	**70.0**	**68.6**	**67.6**	**66.9**	**66.3**	**65.9**	**65.5**	**69.6**
	A+V	✓	77.9	74.1	71.1	69.1	67.8	66.6	65.9	65.3	65.3	64.6	68.8
GRU	A		**77.8**	73.9	71.0	68.8	67.2	66.2	65.4	64.8	64.5	64.1	68.4
	V		57.1	56.9	56.9	56.7	56.5	56.6	56.6	56.5	56.3	56.2	56.6
	A+V		**77.8**	**74.1**	**71.6**	**69.6**	**68.2**	**67.3**	**66.7**	**66.4**	**65.9**	**65.5**	**69.3**
	A+V	✓	77.5	73.6	70.8	68.7	67.5	66.5	65.8	65.4	65.0	64.8	68.5
Mamba	A		**78.7**	**74.8**	**71.7**	**69.4**	**67.8**	**66.8**	**66.0**	**65.2**	**64.8**	**64.4**	**69.0**
	V		57.8	57.6	57.5	57.2	57.1	57.0	57.0	56.9	56.9	56.8	57.2
	A+V		76.5	73.0	70.3	68.3	67.0	66.0	65.2	64.6	64.2	63.9	67.9
	A+V	✓	73.5	70.1	67.5	65.3	63.7	62.9	62.5	62.0	62.0	61.6	65.1

(b) Results on Ego4D

Table 2: Performance comparison of models across different modalities and timesteps, discussed in Section 5.1. Mean average precision (mAP) scores at various timesteps on (a) EasyCom and (b) Ego4D (after 0.20s to 2.00s) for the transformer and the RNN (GRU) backbone using different modalities: Audio (A), Visual (V), and Audio+Visual (A+V). The "Pre." column indicates whether the model is initialized by pretraining on the YT-Conversation dataset. A+V generally provides the best performance.

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.

tively), slightly outperforming their audio-only counterparts. Interestingly, the Mamba model performs best with audio-only input (69.0% mAP) on this dataset.

Comparison with Baselines Table 3 compares our best-performing model (Transformer A+V) with random and silence-based baselines across both datasets. Our model significantly outperforms the baselines, achieving 52.7% AP on EasyCom and 66.8% AP on Ego4D, compared to baseline performances ranging from 26.1% to 27.7% AP. Notably, despite the inherent advantage of detection-based baselines over our predictive models that predict up to 2 seconds into the future, our approach still significantly outperforms them. Also, the silence-based baseline shows comparable results to the random baseline, indicating that systems that require a silence period cannot model real-world conversations effectively.

Pretraining Effects Our YT-Conversation pretraining experiments (Table 2) offer valuable insights into transfer learning for turn-taking prediction. The Transformer model demonstrates adaptability, maintaining or slightly improving performance across datasets, especially for longer-term

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.

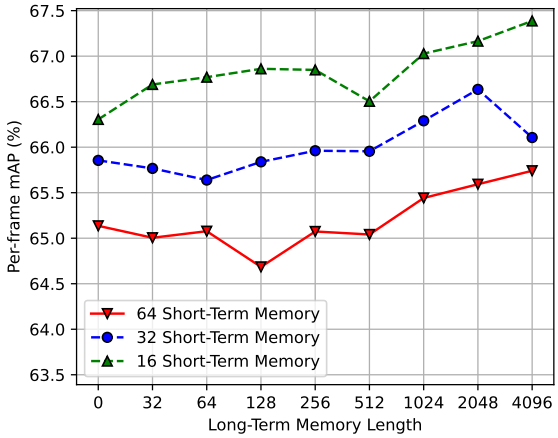

Figure 6: Utterance initiation prediction with varying transformer memory length. a shorter context window for short-term memory and a longer context window for long-term memory generally show better results.

predictions on EasyCom. Varied effects on GRU and Mamba models highlight the complexity of transfer learning in multimodal conversation analysis. The differing impacts between EasyCom and Ego4D suggest that domain characteristics may play a role in pretraining effectiveness. These findings open avenues for future research into optimizing pretraining strategies for specific turn-taking scenarios and architectures.

5.2 Motion Inputs Contributes 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 Dense-flow 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 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 adapts online processing models, which typically use a context length to determine how much historical information to consider. Given that turn-taking occurs within 600 to 1500 ms, the context length can significantly impact utterance initiation prediction. Figure 6 shows the performance of the transformer model depending on the long-term and short-term context lengths. While it is intuitive that having a longer context is beneficial for predicting utterance initiation, the results are counter-intuitive. Although having a longer long-term window is beneficial, having a longer short-term window decreases performance. This suggests that while a broader context can provide more comprehensive information for decision-making, an overly extended short-term context might introduce noise or irrelevant data, thereby impairing the model's accuracy. Therefore, these results emphasize the importance of considering such untrimmed videos and optimizing the balance between long-term and short-term context lengths to improve performance in predicting utterance initiation.

5.4 Runtime Analysis

We evaluated the computational efficiency of our models by measuring their frames per second (FPS) performance on a single RTX3090 GPU on the EasyCom dataset, as shown in Table 5. The RNN-based model demonstrated the highest throughput with an impressive 13,939.5 FPS, closely followed by the Mamba-based model at 12,009.3 FPS. These results indicate that both the RNN and Mamba architectures are highly efficient for the real-time processing of egocentric video streams. While the Transformer-based model operated at a significantly lower 99.8 FPS, it is still achieving real-time

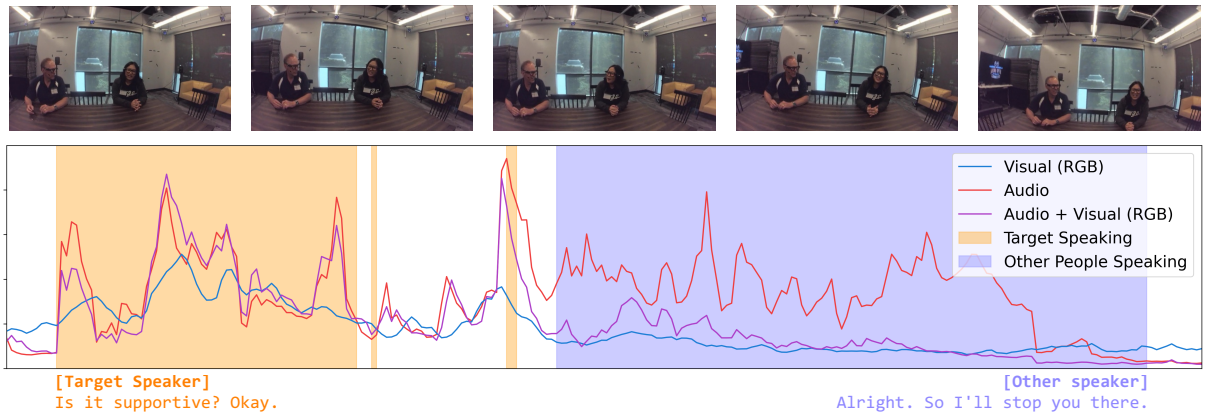

Figure 7: Qualitative results on EasyCom. The predicted scores are shown in lines and the ground-truth label is shown in regions. The blue line represents a model with RGB input, the red line represents a model with audio input, and the purple line represents a model with audio and visual input.

Model	FPS
Transformer	99.8
RNN	13939.5
Mamba	12009.3

Table 5: Frames per second (FPS) performance of different models.

performance. This result suggests that our models are capable of processing in real time.

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

6 Conclusion

We have introduced EgoSpeak, a novel framework for predicting utterance initiation in the wild for conversational agents. EgoSpeak integrates first-person perspective, RGB processing, online processing, and untrimmed video analysis to model real-world conversation. We also presented the YT-Conversation dataset, a diverse collection of YouTube videos for pretraining our models. Our experiments on EasyCom and Ego4D datasets demonstrated that EgoSpeak significantly outperforms random and silence-based baselines in real-time. Notably, our analysis revealed that the multimodal approach consistently outperformed unimodal inputs, incorporating optical flow significantly improves performance, and a counter-intuitive result that shorter context windows for short-term memory and longer windows for long-term memory generally showed better results. These findings underscore the potential of EgoSpeak to enhance real-time conversational dynamics in naturalistic settings, paving the way for more intuitive human-agent interactions.

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), developed for online action recognition, demonstrates improvements in both state-of-the-art performance and FPS using such an approach.

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

Lastly, our current approach does not explicitly model speaker-specific behaviors. Future re-

search could focus on incorporating these individual speaking patterns and tendencies, This could involve analyzing and clustering speaker behaviors from 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.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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 long-form 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.

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.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. *arXiv preprint arXiv:2402.13211*.

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.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, et al. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

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 multiparty 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*.

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.

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.

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 Models

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 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

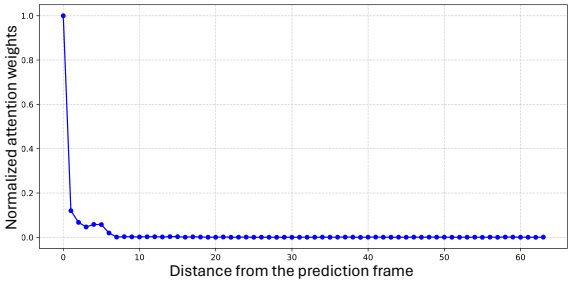

Figure 8: Attention weight of a transformer encoders. Transformer models focus on mostly local context for utterance initiation.

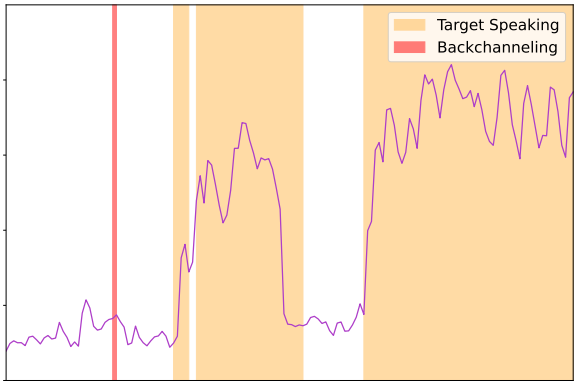

Figure 9: Failure case on EasyCom. The orange region represents the target speaker speaking, and the red region represents the target speaker's backchanneling.

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.

C Error Analysis

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 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.

Model	Modality	Pre.	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	✓	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	✓	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	✓	55.8 ± 0.97	65.8 ± 0.51

Table 6: 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 deviation.

D Descriptive Statistics of Experimental Results

Table 6 shows a detailed comparison of the models' performance on the EasyCom and Ego4D datasets. We conducted our experiments using five different seeds (0, 10, 20, 29, 42). The table reports the mean value of the mean average precision (mAP) across these five seeds, along with the corresponding standard deviation (± std) to reflect the variance in performance, ensuring transparency and robustness in the results.

E Use of AI Assistants

We used Claude 3.5 Sonnet for revising the paper and code and GitHub Copilot for writing the code.