Predicting Turn-Taking and Backchannel in Human-Machine Conversations Using Linguistic, Acoustic, and Visual Signals

Yuxin Lin[*], **Yinglin Zheng**[*], **Ming Zeng**[†], **Wangzheng Shi**

School of Informatics, Xiamen University

{linyx@stu., zhengyinglin@stu., zengming@, shiwangzheng@stu.}xmu.edu.cn

Abstract

This paper addresses the gap in predicting turn-taking and backchannel actions in human-machine conversations using multi-modal signals (linguistic, acoustic, and visual). To overcome the limitation of existing datasets, we propose an automatic data collection pipeline that allows us to collect and annotate over 210 hours of human conversation videos. From this, we construct a Multi-Modal Face-to-Face (MM-F2F) human conversation dataset, including over 1.5M words and corresponding turn-taking and backchannel annotations from approximately 20M frames. Additionally, we present an end-to-end framework that predicts the probability of turn-taking and backchannel actions from multi-modal signals. The proposed model emphasizes the interrelation between modalities and supports any combination of text, audio, and video inputs, making it adaptable to a variety of realistic scenarios. Our experiments show that our approach achieves state-of-the-art performance on turn-taking and backchannel prediction tasks, achieving a 10% increase in F1-score on turn-taking and a 33% increase on backchannel prediction. Our dataset and code are publicly available online to ease of subsequent research. The code and dataset are available at https://github.com/Linyx1125/MM-F2F.

1 Introduction

The ideal natural spoken dialogue system is inherently full-duplex. Human conversation is a typical example of full-duplex dialogue system, where listeners intuitively understand when to interject or engage in **turn-taking** (Sacks et al., 1974), based on the speaker's behavior. However, most existing spoken dialogue systems have yet to achieved truly full-duplex, instead, they require users to send an explicit completion signal to indicate the end of their problem description or use voice activity detection with a predefined threshold to determine

---

[*]Equal Contribution.
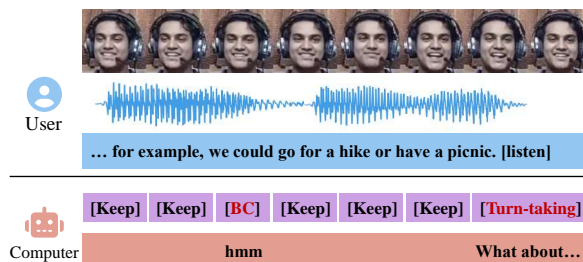[†]Corresponding Author.

Figure 1: In face-to-face conversation scenarios, the computer determines keeping, turn-taking and backchannel actions according to the user's linguistic, acoustic and visual signals (as shown above the line).

the appropriate response time. Besides turn-taking, human conversations often consist of short overlapping statements. For example, the listener might say "I see" to express understanding or respond with an "hmm" to indicate agreement. In linguistics domain, these responses are known as **backchannels** (Yngve, 1970), which primarily indicate the listener's understanding or agreement rather than conveying substantive information. To facilitate the development of natural full-duplex dialogue systems, we focus on the turn-taking and backchannel actions in human conversations. While existing dialogue systems (Wang et al., 2024b; Veluri et al., 2024) provide valuable insights into managing these actions, they are still limited to focusing on linguistic and acoustic information. In real human-machine conversation scenarios, some **linguistic (textual)**, **acoustic**, and **visual** signals from the speaker may affect the occurrence of these actions, as Fig. 1 illustrates.

In recent years, there has been a rising interest in predicting turn-taking and backchannel using neural networks. Initial research primarily concentrated on audio signal analysis, assuming that variations in pitch and rhythm could trigger turn-taking or backchannel actions (Skantze, 2017; Ward et al., 2018; Ekstedt and Skantze, 2022b). More recent studies (Wang et al., 2024b; Veluri et al., 2024; Défossez et al., 2024) have made pre-

liminary progress in achieving full-duplex audio-based dialogue. However, real human conversations also rely on linguistic (textual) and visual signals. Subsequent research illustrated that LLMs can perceive turn-taking actions using linguistic information (Ekstedt and Skantze, 2020), and that visual cues such as head movements, gaze direction, or facial expressions can also trigger turn-taking and backchannel (Kendon, 1967; Lee and Narayanan, 2010).

Despite these advancements, there is a gap in the research on using linguistic, acoustic, and visual signals to predict turn-taking and backchannel actions accurately. One significant reason is the lack of suitable datasets. The well-used datasets in this domain primarily capture only text and audio modalities. The recently proposed Egocom (Northcutt et al., 2020) dataset introduces the video modality but mainly focuses on multi-person communication. Face-to-face conversation data is more representative of real-world situations for turn-taking prediction in human-computer interactions. Moreover, most datasets often overlook backchannel actions.

To address these limitations in multi-modal turn-taking and backchannel prediction, we first proposed an automatic data collection pipeline to collect and annotate over 210 hours of human conversation videos. To protect personal privacy, we remove identity-related elements from these videos. Using these videos, we constructed a **M**ulti-**M**odal **F**ace-to-**F**ace (MM-F2F) human conversation dataset, which includes over 1.5 million words, for turn-taking and backchannel prediction. We present the text transcripts, raw audio, face video frames and corresponding turn-taking and backchannel annotations of these data.

Furthermore, we proposed an end-to-end framework for predicting the probability of turn-taking and backchannel actions from multi-modal signals. Unlike previous work that focused only on text and audio signals, we deeply emphasized the importance of each modality and their interrelations. Moreover, our framework supports uni-modal, bi-modal or tri-modal inputs of text, audio, and video signals, making it adaptable to a wide variety of realistic scenarios that may lack certain modalities. Through a set of experiments, our proposed approach achieves state-of-the-art performance on turn-taking and backchannel prediction tasks. In conclusion, the paper is featured in the following aspects:

- We proposed an automatic data collection pipeline to annotate turn-taking and backchannel actions from in-the-wild videos with minimal manual effort. Utilizing the pipeline, we presented Multi-Modal Face-to-Face Conversation Dataset. To our knowledge, this is the first face-to-face human conversation dataset encompassing text, audio, and video modalities, with word-level annotations for turn-taking and backchannel actions. Additionally, we de-identified the identity-related elements in our dataset for personal privacy protection.

- To our knowledge, we introduced the first tri-modal solution for turn-taking and backchannel prediction in natural conversation systems, which utilizes linguistic, acoustic, and visual information. Benefiting from our proposed flexible fusion module, our framework adapts to various backbone uni-models and supports any combination of text, audio, and video inputs.

- We comprehensively evaluated our framework on MM-F2F dataset and compared it with previous works. The experimental results indicated that our approach achieves state-of-the-art performance. We release our dataset and code for subsequent research.

2 Related Work

2.1 Turn-taking and Backchannel Prediction

2.1.1 Signals in Human Conversation

One of the earliest turn-taking systems (Sacks et al., 1978) considered that the turn-taking process can vary based on the context of the conversation. Later studies (De Ruiter et al., 2006) have found that humans can predict turn-taking simply using the context, highlighting the importance of linguistic signals. Some research indicated that acoustic signals (Ekstedt and Skantze, 2022a) and visual (Kendon, 1967) signals may lead to occurrences of turn-taking. Yang et al. identified that the speaker's pitch and expression are active during and after turn-taking occurs. These studies have demonstrated that multi-modal cues can affect turn-taking and backchannel actions in human face-to-face conversations.

2.1.2 Uni-modal Prediction

Several methods utilizing the aforementioned signals have been presented. The VAP model (Ekstedt and Skantze, 2022b) projected the acoustic signals

into a 256-dimensional vector to predict the probability of keep, turn-taking and backchannel actions, successfully achieving backchannel prediction. In linguistic domain, Ekstedt and Skantze fine-tuned the pre-trained GPT-2 (Radford et al., 2019) model with the addition of TURN tokens, achieving favorable results in turn-taking prediction. Recent work (Shukuri et al., 2023; Kim et al., 2025) has shown that LLMs may better understand context and extract linguistic signals from conversations than conventional methods. Additionally, studies (Lee and Narayanan, 2010) suggest that specific movements of the eyebrows and mouth in face-to-face conversation may also influence

2.1.3 Multi-modal Prediction

Unlike uni-modal approaches, fusing multi-modal signals leverages the strengths of each modality. Chang et al. classified turn-taking actions into six sub-categories and proposed an end-to-end ASR-based network for predicting turn-taking and backchannel actions with acoustic and linguistic signals. Yang et al. demonstrated that multi-modal fusion contributes to turn-taking prediction by combining acoustic and semantic signals with a gated fusion block. Their experiments showed that the incorporating text information boosts the accuracy of audio-only models. Kurata et al.; Wang et al. extracted and fused the multi-modal signals for turn-taking prediction, achieving impressive experimental results. Nevertheless, a unified approach integrating text, audio, and video for turn-taking prediction is still lacking, partly due to the absence of a comprehensive turn-taking and backchannel dataset encompassing all these modalities.

2.2 Human Conversation Dataset

Since turn-taking prediction is a relatively new field, open-source datasets are limited. Early research often used Automatic Speech Recognition (ASR) or Text-to-Speech (TTS) datasets, which included audio and text transcriptions of conversations. Researchers typically annotated the end of an utterance as turn-taking for prediction tasks based on these transcriptions. Acknowledging the role of backchannel actions in natural interactions, the FTAD (Chen et al., 2021) was introduced to annotate both turn-taking and backchannel actions in conversation texts. Concurrently, the EgoCom (Northcutt et al., 2020) dataset was developed to explore the impact of visual signals on turn-taking, recording multi-person conversations with specialized capture devices. However, EgoCom's third-

person perspective and partially obscured eyes limit its utility. Consequently, there is still a need for a dataset that accurately reflect realistic face-to-face conversation scenarios.

3 MM-F2F: Multi-Modal Face-to-Face Conversation Dataset

3.1 Automatic Data Collection and Annotation

To the best of our knowledge, there is still a lack of open-source multi-modal dataset that effectively captures natural face-to-face human conversations. To address this gap, we propose an automatic video collection pipeline with minimal manual annotation from in-the-wild videos. Utilizing this pipeline, we construct the first face-to-face human conversation dataset containing text, audio, and video tri-modal data annotated with turn-taking and backchannel actions. The key stages of our proposed pipeline are described in this section.
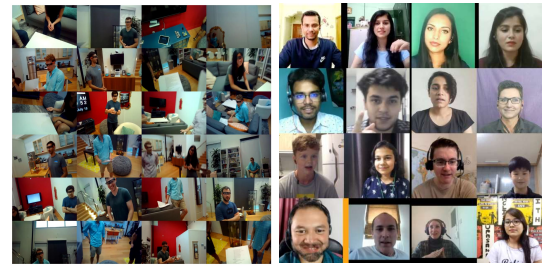
(a) EgoCom dataset   (b) Our MM-F2F dataset[1]

Figure 2: Samples of EgoCom vs. our MM-F2F dataset and visualized distribution of our dataset. (a) Samples of EgoCom dataset. (b) Samples of our MM-F2F dataset.

3.1.1 Stage 1. Video Collection and De-Identification

Video Collection. We first fetch nearly 1,000 in-the-wild online English conversation videos from Internet. Then we employ a face detection system to analyze these videos, selecting those that consist of two people positioned on the left and right sides of the screen, facing the camera.

De-Identification. To protect personal privacy, we remove identity-related elements such as faces, voiceprints, and surrounding backgrounds, while retaining other conversation-relevant information.

For facial identity, we first generate a face pool containing over 10,000 synthetic faces. Given a face f from the original video, we select the most similar face f' from the pool according to the face similarity (Deng et al., 2020). This guarantees the

removal of identifiable features while maintaining high video quality.

For voiceprints, we extract the latent code representing the voice ID (Li et al., 2023) and perturb each dimension of this code by 20% of its standard deviation. We then synthesize new audio (Li et al., 2023) based on the perturbed latent code, ensuring the generated audio remains synchronized with the original.

Additionally, we crop the video to retain only the facial regions, thus removing surrounding background information. In the experiment detailed in Sec. 5.1.1, we show that focusing solely on facial regions reduces the impact of redundant background information. To prevent private content from being included in conversations, we manually removed any conversations containing personal information. Besides, the experiments described in the appendix Sec. A.1 demonstrates that removing identity information does not hinder the understanding of conversational behavior.

3.1.2 Stage 2. Video Transcription

We divide each video into finer-grained word frames. Specifically, the audio tracks of all videos are processed through an Automatic Speech Recognition (ASR) system (Bain et al., 2023), which splits them into sentence-level clips and finer word-level frames, each with corresponding text transcriptions.

3.1.3 Stage 3. Speaker Verification

To determine which clips come from the same speaker, we first extract the embeddings of each clip through a ResNet-based encoder (Wang et al., 2023). These embeddings are then clustered into two classes using a clustering method to distinguish between the two speakers.

3.1.4 Stage 4. Active Speaker Detection

In order to correspond the clustered audio to the faces in each frame, we employ the Active Speaker Detection (ASD) method using the TalkNet (Tao et al., 2021) model to determine which person in the frames is speaking. The detection results are then cross-checked with the previous clustering results to confirm the speaker of each clip.

3.1.5 Stage 5. Annotation

Finally, the clips are categorized into three turn-taking actions: KEEP, TURN and BACKCHAN-

---

NEL. We first annotate the last word of each utterance as TURN. Next, following the approach of previous work (Ekstedt and Skantze, 2020), we match the BACKCHANNEL with a specific vocabulary. For the remaining words, we annotate them as KEEP. Through manual annotation and double-checking, we ensure that the automatic annotation results achieve a precision of better than 95%. More details of automated data collection and annotation pipeline, and manual double-checking are provided in the appendix Sec. A.2 and Sec. A.3.

3.2 Dataset Analysis

Utilizing our data collection and annotation pipeline, we selected 773 in-the-wild videos, totaling approximate 20 million frames, to construct the Multi-Modal Face-to-Face Conversation Dataset. Each video features two people on the left and right sides of the screen, engaged in daily English conversations, as illustrated in Fig. 2b. The videos are divided into sentence-level clips, each providing time-stamped text transcriptions, raw audio tracks, face annotations of the current speaker, and turn-taking and backchannel actions annotations of each word.

Our dataset includes approximately 169,029 utterances from about 955 speakers, featuring a diverse range of identities across different races and genders. We have annotated these utterances into over 1.5M word frames, which include about 51K instances of turn-taking and 22K instances of backchannel actions. Approximately 19K utterances contain backchannel actions. In these utterances, about 95% contain backchannel actions once or twice. Each utterances contains an average of 9.33 word frames and lasts approximately 4.18 seconds. In face-to-face human conversations, utterances typically range from about 2 to 12 words in length and last between 1 to 4 seconds. We observed a similar distribution in our dataset. Additionally, there are a small number of sentences consisting of one or two words used for short responses, as well as fewer long-length sentences of more than 20 words used to convey longer points.

Tab. 1 compares basic information of previous datasets with ours. Evidently, our MM-F2F dataset is the first tri-modal dataset with turn-taking and backchannel annotations. A notable innovation of our dataset is its focus on visual cues in face-to-face conversations. As the table demonstrates, while the EgoCom dataset also incorporates tri-modal signals, its third-person perspective limits its ability to capture face-to-face conversation sce-

---

[1]This figure serves solely to illustrate our dataset's diversity. In the publicly released version of the dataset, only the cropped face data will be included.

Name	Modality			Annotation		Utter.	Video Info		
	T	A	V	Turn	BC		Num.	Speaker	Dura.
MapTask (Anderson et al., 1991)	✓	✓	-	✓	-	27,084	-	-	-
Switchboard (Godfrey et al., 1992)	✓	✓	-	✓	✓	199,740	-	-	-
Librispeech (Panayotov et al., 2015)	✓	✓	-	✓	-	148,688	-	-	-
LJ-Speech (Ito and Johnson, 2017)	✓	✓	-	✓	-	13,100	-	-	-
VCTK (Veaux et al., 2017)	✓	✓	-	✓	-	44,583	-	-	-
Daily Dialog (Li et al., 2017)	✓	-	-	✓	-	102,980	-	-	-
FTAD (Chen et al., 2021)	✓	-	-	✓	✓	344,264	-	-	-
EgoCom (Northcutt et al., 2020)	✓	✓	✓	✓	-	20,436	28	34	38.5h
MM-F2F (Ours)	✓	✓	✓	✓	✓	169,029	773	955	210h

Table 1: Brief summary of previously presented datasets. **T**, **A**, **V** denote text, audio, and video modals respectively. **Turn** and **BC** denote turn-taking and backchannel while **Num.** and **Dura.** denote number and duration. Our dataset is the first tri-modal dataset with turn-taking and backchannel annotations. Compared to previous multi-modal datasets, our dataset exhibits a substantially greater scale.

narios, as shown in Fig. 2a. In contrast, our MM-F2F dataset (Fig. 2b) is derived from face-to-face human conversations to be more generalized for turn-taking and backchannel prediction. Additionally, our dataset includes a substantial increase in the number of utterances, videos, speaker identities, and video duration. Furthermore, it features annotations for backchannel actions within natural conversation scenario.

4 Multi-Modal Turn-Taking and Backchannel Prediction Framework

4.1 Overview

In this section, we present our turn-taking and backchannel prediction framework, as shown in Fig. 3. The text, audio and video data X_k are first fed into three uni-modal encoder models E_k to obtain the uni-modal features $\boldsymbol{z_k} \in \mathbb{R}^{256}, k \in \{T, A, V\}$ respectively. Then the features are fused with a flexible multi-modal fusion module F to predict the probability $\hat{y}$ of keep, turn-taking and backchannel actions.

$$\hat{y} = F(E_T(X_T), \ E_A(X_A), \ E_V(X_V)) \quad (1)$$

We train our framework in two stage of: (1) training the three uni-modal encoders E_k independently; (2) using the multi-modal data to train the end-to-end prediction framework jointly. With the flexibility of the fusion module, our framework supports arbitrary combination of input modalities. In the following subsections we will introduce each part of our proposed model.

4.1.1 Word-level Aligned Trimodal Input

To comprehensibly capture the contextual information embedded within conversations, we construct
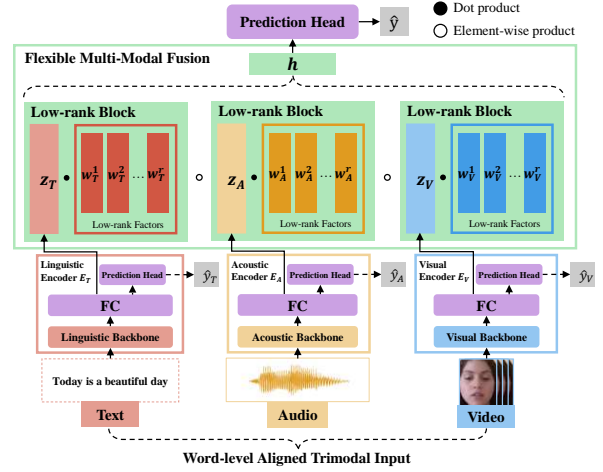

Figure 3: Architecture of our proposed multi-modal turn-taking and backchannel prediction framework. The text, audio and video inputs are first fed into uni-modal encoders E_T, E_A, E_V to extract the uni-modal features $\boldsymbol{z_T}, \boldsymbol{z_A}, \boldsymbol{z_V}$ respectively. The features are then fused in the flexible fusion model. The fused feature $\boldsymbol{h}$ will be input into a prediction head to predict the output probability $\hat{y}$. Our framework supports uni-modal, bi-modal and tri-modal inputs of text, audio and video signals.

an input sequence X for each word frame by incorporating preceding word frames from the same sentence. For the text input X_T, we use the raw text from the sequence. We extract the corresponding audio waveform data for the audio input X_A. To capture the continuous variations in the speaker's behavior leading up to turn-taking and backchannel actions, as input X_V, we fetch the last n frames of sequence which is timely-aligned with X_T.

4.2 Uni-Modal Encoder

Our uni-modal encoder model is adaptable to different backbone. For each uni-modal model, the raw data X_k is fed into a backbone model to obtain a 256-dimensional feature $\boldsymbol{z_k}$. During the uni-modal training phase, the feature is passed through a prediction head to estimate action probabilities $\hat{y}_k$, which are then supervised with cross-entropy loss.

4.2.1 Linguistic Encoder

To better exploit contextual information, we assert that the text backbone is capable of processing text sequence $X_T \in \mathbb{R}^{len\_text}$, generating a corresponding text embedding. This embedding is subsequently passed through a linear layer to produce a text feature $\boldsymbol{z_T} \in \mathbb{R}^{256}$.

4.2.2 Acoustic Encoder

Likewise, we employ the audio backbone to extract the audio embedding from the original audio waveform $X_A \in \mathbb{R}^{len\_audio}$. The audio embedding is then projected into $z_A \in \mathbb{R}^{256}$ and subsequently fed into a prediction head.

4.2.3 Visual Encoder

The visual backbone encodes the last few frames $X_V \in \mathbb{R}^{n \times c \times w \times h}$ of each sequence, where n, c, w, h denote the number of frames, channels, width and height of input frames. Following the linguistic and acoustic models, the embedding from the visual backbone is passed through a linear layer to obtain a feature vector $z_V \in \mathbb{R}^{256}$.

The selection of backbone models will be thoroughly explored in the experimental Sec. 5.1.

4.3 Flexible Multi-Modal Fusion

The key to multi-modal fusion is to fully exploit the correlations between multi-modal features. Zadeh et al. presented the fusion representation as a tensor $Z = \bigotimes_k^K z_k$, where $\bigotimes_k^K$ denotes the tensor outer product over multi-modal features of modality k, and $z_k \in \mathbb{R}^{d_k}$ denotes the feature of each modality. The tensor $Z \in \mathbb{R}^{d_1 \times d_2 \times \dots \times d_K}$ is then passed through a linear transformation to produce the output representation $h = W \cdot Z + b$, $h \in \mathbb{R}^{d_h}$, where $W \in \mathbb{R}^{d_h \times d_1 \times d_2 \times \dots \times d_K}$ and $b \in \mathbb{R}^{d_h}$ represent the weight and bias, respectively. To reduce computational effort and refine information, Liu et al. presented LMF block, decomposing W with r low-rank decomposition factors $w_k^{(i)} \in \mathbb{R}^{d_h \times d_k}$:

$$W = \sum_{i=1}^{r} \bigotimes_k^K w_k^{(i)} \tag{2}$$

Based on this low-rank decomposition, the multimodal fusion algorithm can be simplified as:

$$h = \Lambda_k^K [\sum_{i=1}^{r} w_k^{(i)} \cdot z_k] \tag{3}$$

where Λ_k^K is the element-wise product over the tensors. This simplification downgrades the multimodal features into a low-rank space, significantly improving computational efficiency.

4.3.1 Modality Selection Scheme

Considering the potential absence of modalities in real human-computer interaction scenarios, we introduce a modality selection scheme grounded in the idea of low-rank decomposition. This scheme enables the model to support any combination of modality inputs, adapting to various scenarios and enhancing system robustness even when some signal channels are accidentally lost:

$$h = I_T \left(\sum_{i=1}^{r} w_T^{(i)} \cdot z_T \right) \circ I_A \left(\sum_{i=1}^{r} w_A^{(i)} \cdot z_A \right) \circ I_V \left(\sum_{i=1}^{r} w_V^{(i)} \cdot z_V \right),$$

$$I_k(x) = \begin{cases} x & \text{if modality } k \text{ exists} \\ \mathbf{1} \in \mathbb{R}^{d_h} & \text{otherwise} \end{cases}$$

$$\tag{4}$$

where $I_k(x)$ is an indicator function, representing the existence of modality k. Built on the modality selection scheme, our fusion module can extend to accommodate additional modalities.

4.3.2 Random Modality Dropout Training

To fully leverage the capability of the modal selection scheme, we further design a Random Modality Dropout Training (RMDT) scheme. Specifically, we randomly drop one modality with a small probability during training, using the remaining two for fusion to improve robustness across different numbers of modalities. This allows our fusion module to seamlessly adapt to both bi-modal and tri-modal inputs.

The fused features are then fed into a linear prediction head that predicts the output probability $\hat{y}^{(i)}$ for each action. We train the entire framework end-to-end, optimizing it by minimizing the cross-entropy loss:

$$L = - \sum_i y^{(i)} \, log(\hat{y}^{(i)}), \ \ i \in \{Keep, Turn, BC\}$$

Benefiting from our flexible fusion module and training strategies, our framework accommodates any combination of input modalities. For single-modality data, the model defaults to the uni-modal encoder model, using its output probability $\hat{y}_k$ directly as the prediction results. For bi-modal or tri-modal inputs, we extract features from the uni-modal encoders and then fuse them according to Eq. 4. This approach enables the framework to be trained only once while supporting to infer on any combination of modalities. Detailed implementation and evaluation will be introduced in the Sec. 5.

5 Experiments

5.1 Instantiations

5.1.1 Choices of Backbone Models

Our proposed framework is generic and can be instantiated with various backbones. In this subsection, for the text backbone, we compare the performance of BERT (Devlin, 2018) and GPT-2

(Radford et al., 2019) models; for the audio backbone, we apply HuBERT (Hsu et al., 2021) and Wav2Vec2 (Baevski et al., 2020) models. For the visual backbone, we also employ two models: (1) using the ViT image extractor model (Dosovitskiy et al., 2020), which processes the last frame of each clip as input; (2) utilizing the VideoMAE (Tong et al., 2022) model to process the last 16-frame sequence of each clip. Given that non-facial parts of the video, such as body and gesture movements, may also affect conversational actions, we conducted another comparison experiment: (1) using only the face area from the video, and (2) using the entire frames.

We employ the backbone models from the Huggingface Transformer Library (Wolf et al., 2019) with default hyperparameters, keeping all parameters unfrozen during fine-tuning. To focus on capturing speaker actions at the end of the speech, we extract the hidden state of the last token as text embedding. For acoustic and visual backbones, we apply an additional average pooling layer to transform the hidden state from $[len\_seq, hidden\_size]$ to $[1, hidden\_size]$ along the temporal dimension.

Modality	Backbone	Accuracy	F1-Score		
			Keep	Turn	BC
Text	BERT	0.742	0.743	0.761	0.674
	GPT-2	**0.751**	**0.747**	**0.767**	**0.707**
Audio	Wav2Vec2	0.730	0.715	0.726	0.779
	HuBERT	**0.751**	**0.737**	**0.735**	**0.805**
Video	ViT (Single Frame)	0.473	0.535	0.470	0.271
	VideoMAE (Full)	0.533	0.516	0.523	0.482
	VideoMAE (Face)	**0.559**	**0.597**	**0.536**	**0.513**

Table 2: Comparison results of different backbones and exploitation of visual signals. We choose GPT-2, HuBERT and VideoMAE as backbones of the uni-modal models.

We evaluated the F1-score and accuracy of each uni-modal model for the prediction of keeping, turn-taking and backchannel actions. The experimental results are presented in Tab. 2. It can be observed that GPT-2 and HuBERT models outperform in text and audio feature extraction, respectively. Compared to the ViT model, VideoMAE demonstrates superior performance by capturing the continuous variation of the speaker before the turn-taking and backchannel actions occur. Besides, while the entire frame provides more information, it introduces redundant data. Focusing on the face area yields better results by emphasizing crucial facial information. According to the above results and analysis, for subsequent experiments,

we employ GPT-2, HuBERT, and VideoMAE as backbones of the uni-modal models and crop the face area from the video as input for the visual model. However, it should be pointed that our tri-modal framework flexibly support various backbone models, leaving interfaces for future exploration of other backbones.

5.1.2 Implementation Details

All experiments are implemented on a single NVIDIA RTX3090 GPU. During the uni-modal model training phase, we set the batch sizes for the text, audio, and video encoders to 4, 1, and 4, respectively. In the modal fusion training phase, the rank number in low-rank block is set to 16. The three unfrozen uni-modal encoders and the multi-modal fusion module are trained end-to-end with a batch size of 1. We used a three-layer MLP as the prediction head, with layer sizes of [256, 64, 3]. All training stages are optimized using Adam optimizer with a learning rate of 10^{-5}. The training process lasted for 20 epochs, taking approximately 30 hours for uni-modal training and 20 hours for end-to-end multi-modal training.

5.2 Ablation Study

To investigate the impact of different modal information on the prediction of turn-taking and backchannel actions, we conducted an ablation study on the modalities. Concretely, for the uni-modal model, we directly used the encoder from the first stage of our framework. For the bi-modal model, we retained the two-stage structure of our framework, utilizing only bi-modal data inputs for end-to-end training.

Training Modal	Accuracy	F1-Score		
		Keep	Turn	BC
Text	0.751	0.747	0.767	0.707
Audio	0.751	0.737	0.735	0.805
Video	0.559	0.597	0.536	0.513
Text+Audio	0.811	0.783	0.809	0.894
Text+Video	0.757	0.751	0.766	0.743
Audio+Video	0.742	0.742	0.770	0.829
Text+Audio+Video	**0.823**	**0.806**	**0.811**	**0.906**

Table 3: Turn-taking and backchannel prediction performance of different modal models and F1-score respectively.

Tab. 3 presents the evaluation results of uni-modal, bi-modal, and tri-modal fusion models on our test set. Generally, prediction accuracy increases with the number of modalities used. The tri-

modal model, utilizing all modalities, achieves optimal prediction accuracy and F1-scores for each action. The uni-modal models' results align with previous work, showing that predicting keep-speaking and turn-taking actions with text or audio alone is capable of achieving a high F1-score of nearly 0.75. Backchannel actions are well predicted using acoustic signals, likely due to pitch or tone discontinuities triggering these actions. While the video-only model performs satisfactorily, it is inferior to text and audio models, possibly due to individual differences in visual cues.

Multi-modal fusion significantly improves prediction accuracy for each action. The text-audio model leverages linguistic and acoustic signals effectively, achieving excellent results. Video information notably enhances backchannel prediction, as speakers may use expressions and motions to signal listeners about the occurrence of backchannel actions. Consequently, our tri-modal model achieves an F1-score of about 0.91 for backchannel prediction and about 0.82 for the other actions.

Figure 4: Qualitative results on visual modal ablation. Dynamics of visual signals transmit communication cues that text and audio signals cannot capture.

Based on the results of the ablation experiments, we observe that using only text and audio modalities sometimes fails to provide an accurate prediction of turn-taking and backchannel actions. Visual information can effectively compensate for the shortcomings. As demonstrated in the qualitative results in Fig. 4, subtle changes in head movements, eye gaze, and facial expressions serve as important visual cues. Consequently, we conducted an in-depth analysis to evaluate the effectiveness of visual signals on prediction.

5.3 Evaluation of Multi-modal Fusion Strategies

Previous work (Wang et al., 2024a; Kurata et al., 2023), uni-modal features were fused by concatenation and then passed through a linear projection layer for prediction. Additionally, the GMF (Yang et al., 2022) model has shown outstanding results in multi-modal fusion for turn-taking prediction tasks. We evaluated these fusion strategies on our model, as shown in Tab. 4. Concatenation offers high computational speed but fails to effectively fuse uni-modal information, yielding performance similar to uni-modal models. The GMF module, though less efficient, learns fusion weights in a supervised manner, improving feature integration. Our flexible fusion module combines the strengths of both approaches: it projects features into a low-rank space, learns fusion weights supervisedly, and maintains good computational efficiency.

Fusion Strategy	Accuracy	F1-Score		
		Keep	Turn	BC
Concatenate	0.771	0.764	0.774	0.784
GMF	0.807	0.791	0.795	0.889
Ours	**0.823**	**0.806**	**0.811**	**0.906**

Table 4: Turn-taking and backchannel prediction performance of different fusion strategies. Our flexible fusion module provides the best fusion results while maintaining efficient computational performance.

5.4 Comparison with State-of-the-Art

We compare our framework with: (1) TurnGPT (Ekstedt and Skantze, 2020), which supports only text inputs and predicts the probability of keeping and turn-taking actions; (2) Wang et al.'s model that utilizes text and audio signals for turn-taking and backchannel prediction; (3) Kurata et al.'s model, which incorporates text, audio and video information, fusing them for turn-taking prediction. Since Wang et al.'s and Kurata et al.'s models are not currently open source, we implement them based on the details and parameters provided in their original papers. Besides, we extend the Kurata et al.'s binary classifier to a triple classifier to enable backchannel prediction. All these models are trained and evaluated on our proposed MM-F2F for comparison.

The results of the comparison study are shown in Tab. 5. Our approach improves the F1-score for backchannel prediction by more than 0.2 compared

to previous work. Additionally, there is an approximate 0.1 improvement in predicting keep-speaking and turn-taking actions. Overall, our framework outperforms state-of-the-art approaches in all aspects.

Method	Modal	Acc.	F1-Score		
			Keep	Turn	BC
TurnGPT	T	0.645	0.745	0.420	-
Wang et al.'s	T+A	0.737	0.742	0.739	0.680
Kurata et al.'s	T+A+V	0.720	0.729	0.728	0.667
Ours	T+A+V	**0.823**	**0.806**	**0.811**	**0.906**

Table 5: The comparison study results on our proposed dataset. We implement TurnGPT and the framework of Wang et al. and Kurata et al. on our proposed dataset to compare with our model. The **T**, **A**, **V** denote input modal of text, audio and video.

5.5 Inference with Arbitrary Modality Combination

In a wide variety of real-world scenarios, human-computer interaction systems encounter various combinations of input modalities. Our framework is well-equipped to handle tri-modal inputs. When only bi-modal inputs are provided, benefiting from our Random Modality Dropout Training (RMDT) scheme, the Eq. 4 seamlessly degenerates to the element-wise product of the two modalities. For uni-modal input, we directly predict the turn-taking and backchannel probabilities using the uni-modal encoder model in the first stage. Thus, our framework is capable to adapt arbitrary modality combination input by training only once. We evaluated our method on bi-modal inputs, using the model trained on tri-modal data. For comparison, we also evaluated the performance of the model trained without the RMDT scheme, the results are illustrated in Tab. 6. It can be observed that the RMDT scheme enables the model to process bi-modal signals effectively, with minimal impact on the model's predictive performance for tri-modal inputs. Compared to the results of bi-modal training model in Tab. 3, the tri-modal based model can effectively capture information from all modalities during training, providing comparable results even on bi-modal evaluating.

6 Conclusion

In this paper, we introduce an automatic data collection and annotation pipeline, along with a novel MM-F2F dataset for turn-taking and backchannel prediction. Using this dataset, we explore a pioneering framework that leverages linguistic, acoustic

Inference Modalities	Accuracy	F1-Score		
		Keep	Turn	BC
Text+Audio (w/o RMDT)	0.552	0.619	0.597	0.017
Text+Audio (w/ RMDT)	0.816	0.803	0.803	0.896
Text+Video (w/o RMDT)	0.423	0.601	0.001	0.005
Text+Video (w/ RMDT)	0.760	0.757	0.765	0.747
Audio+Video (w/o RMDT)	0.433	0.058	0.640	0.041
Audio+Video (w/ RMDT)	0.765	0.748	0.752	0.845
Text+Audio+Video (w/o RMDT)	0.818	0.807	0.802	0.902
Text+Audio+Video (w/ RMDT)	0.823	0.806	0.811	0.906

Table 6: We train our model on tri-modal input data while evaluating on bi-modal data pairs. Benefit from our RMDT scheme, the model can still achieve good performance when the inputs consist of only two modalities.

and visual signals inputs to predict turn-taking and backchannel actions. We hope that our work could pave the way for future research on more natural human-computer interaction systems. Future work may lie in studying the fuller integration of more visual cues. Another promising direction could be to investigate the role of personalized information to enhance prediction capabilities.
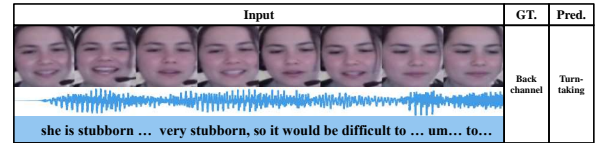

Figure 5: Failure case. When the speaker pauses to think while the talking context is semantically incomplete, our framework might sometimes mistakenly initiate turn-taking. In this case, in contrast, providing a backchannel would be more appropriate.

7 Limitations

There is still improvement space due to the subtle and complicated nature of multi-modal conversation signals. For example, when the speaker is semantically incomplete and pauses to think, a backchannel response might be expected. Instead, our framework may occasionally misinterpret this as an indication for turn-taking due to the speaker's prolonged silence and frozen expression, as shown in Fig. 5. To solve this problem, incorporating additional visual cues such as body movements or gestures could be a potential direction.

8 Acknowledgement

References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

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.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Shuo-yiin Chang, Bo Li, Tara N Sainath, Chao Zhang, Trevor Strohman, Qiao Liang, and Yanzhang He. 2022. Turn-taking prediction for natural conversational speech. *arXiv preprint arXiv:2208.13321*.

Kehan Chen, Zezhong Li, Suyang Dai, Wei Zhou, and Haiqing Chen. 2021. Human-to-human conversation dataset for learning fine-grained turn-taking action. In *Interspeech*, pages 3231–3235.

Jan-Peter De Ruiter, Holger Mitterer, and Nick J Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. 2020. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Proceedings of the IEEE Conference on European Conference on Computer Vision*.

Jiankang Deng, Jia Guo, Zhou Yuxiang, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019. Retinaface: Single-stage dense face localisation in the wild. In *arxiv*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Erik Ekstedt and Gabriel Skantze. 2020. Turngpt: a transformer-based language model for predicting turn-taking in spoken dialog. *arXiv preprint arXiv:2010.10874*.

Erik Ekstedt and Gabriel Skantze. 2022a. How much does prosody help turn-taking? investigations using voice activity projection models. *arXiv preprint arXiv:2209.05161*.

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

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63.

JiWoo Kim, Minsuk Chang, and JinYeong Bak. 2025. Beyond turn-taking: Introducing text-based overlap into human-llm interactions. *arXiv preprint arXiv:2501.18103*.

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.

Chi-Chun Lee and Shrikanth Narayanan. 2010. Predicting interruptions in dyadic spoken interactions. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5250–5253. IEEE.

Jingyi Li, Weiping Tu, and Li Xiao. 2023. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Curtis Northcutt, Shengxin Zha, Steven Lovegrove, and Richard Newcombe. 2020. Egocom: A multi-person multi-modal egocentric communications dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

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

Kotaro Shukuri, Ryoma Ishigaki, Jundai Suzuki, Tsubasa Naganuma, Takuma Fujimoto, Daisuke Kawakubo, Masaki Shuzo, and Eisaku Maeda. 2023. Meta-control of dialogue systems using large language models. *arXiv preprint arXiv:2312.13715*.

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.

Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. 2021. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 3927–3935.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093.

Christophe Veaux, Junichi Yamagishi, and Kirsten Mac-Donald. 2017. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.

Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. 2024. Beyond turn-based interfaces: Synchronous llms as full-duplex dialogue agents. *arXiv preprint arXiv:2409.15594*.

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. 2023. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pranav Dheram, Di He, Minhua Wu, Andreas Stolcke, and Venkatesh Ravichandran. 2024a. Turn-taking and backchannel prediction with acoustic and large language model fusion. *arXiv preprint arXiv:2401.14717*.

Peng Wang, Songshuo Lu, Yaohua Tang, Sijie Yan, Wei Xia, and Yuanjun Xiong. 2024b. A full-duplex speech dialogue scheme based on large language models. *arXiv preprint arXiv:2405.19487*.

Nigel G Ward, Diego Aguirre, Gerardo Cervantes, and Olac Fuentes. 2018. Turn-taking predictions across languages and genres using an lstm recurrent neural network. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 831–837. IEEE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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.

Liu Yang, Jieyeon Woo, Catherine Achard, and Catherine Pelachaud. 2023. Exchanging... watch out! *arXiv preprint arXiv:2311.04747*.

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.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

A Appendix

A.1 Effectiveness of the De-Identification Data

We also verify that our pre-processing of removing identification will not introduce domain gaps which hinder the understanding of talking behavior. We conducted comprehensive experiments on the de-identification process by testing the following four settings: 1. Train on de-identified data; test on de-identified data; 2. Train on de-identified data; test on original data; 3. Train on original data; test on original data; 4. Train on original data; test on de-identified data; As the results shown in Tab. 7, model performance is not significantly affected by de-identification. We believe that factors influencing turn-taking and backchannel actions mainly include cues such as pitch changes, facial expressions, emotions, and eye motions, which are not substantially impacted by de-identification. Our approach, which removes facial identity, voiceprint, and dialogue content, maximizes privacy protection while minimizing its impact on real-world model applications. Meanwhile, we invite volunteers to confirm that they are unable to recognize identities in the processed videos, even when provided with the original videos as references.

Setting	Accuracy	F1-Score		
		Keep	Turn	BC
Train De-id/Test De-id	0.823	0.806	0.811	0.906
Train De-id/Test Original	0.835	0.819	0.822	0.912
Train Original/Test Original	0.836	0.822	0.821	0.916
Train Original/Test De-id	0.835	0.812	0.805	0.896

Table 7: We conducted comprehensive experiments on the de-identification process by testing the following four settings: 1. Train on de-identified data; test on de-identified data; 2. Train on de-identified data; test on original data; 3. Train on original data; test on original data; 4. Train on original data; test on de-identified data; The results show that the model performance is not significantly affected by de-identification.

A.2 More Detail of Automatic Data Collection and Annotation Pipeline

Since our primary focus is on face-to-face human-computer interaction scenarios, where users typically engage with a computer while facing its camera, we collected a large corpus of online remote conversation videos from various video platforms. These videos feature two individuals conversing via video chat while directly facing their respective cameras, covering topics such as campus life, daily interactions, and cultural exchange. Compared to scripted or laboratory-recorded conversational datasets, these naturally occurring video conversations exhibit greater spontaneity and fluency, enabling a more faithful representation of fine-grained behaviors in real-world dialogue interactions. We propose an automatic annotation pipeline, as illustrated in Fig. 6. The detailed steps are as follows:

Face Detection. we applied a face detection model (Deng et al., 2019), with a confidence threshold of 0.95 to perform frame-by-frame analysis on approximately 1,000 collected videos. We retained only those videos in which, for at least 90% of the frames, exactly two individuals appeared symmetrically on the left and right sides of the frame while facing the camera.
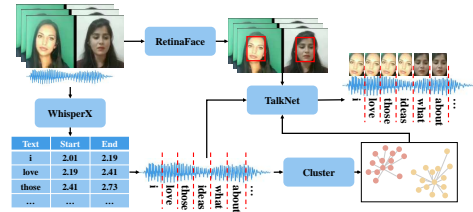

Figure 6: Our automatic data annotation pipeline.

Video Transcription. We employed WhisperX (Bain et al., 2023), built upon the whisper-large-v3 model, to generate automatic transcriptions of the videos. The output included word-level and sentence-level transcriptions, along with start and end timestamps for each word and utterance.

Speaker Verification. To associate transcribed speech with individual speakers, we first segmented the audio into clips corresponding to each utterance. We then extracted speaker embeddings using a speaker verification encoder (Wang et al., 2023) with default parameters. Since embeddings from the same speaker exhibit high cosine similarity, while those from different speakers demonstrate lower similarity, we applied KMeans-based clustering to partition the utterances into two distinct clusters, assuming that each cluster corresponds to a unique speaker.

Active Speaker Detection. Given that each video contains exactly two detected faces, we leveraged TalkNet (Tao et al., 2021), a deep learning-based active speaker detection model, with default parameters to establish correspondence between the audio clips and the visible speakers. TalkNet determines which individual in the video is actively speaking at the moment. By incorporating the previously detected face information, we assigned
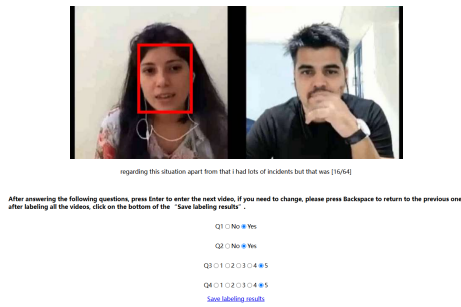
Figure 7: We developed a simple data validation system to assist validators in verifying the data more effectively. Validators were instructed to answer a series of straightforward questions based on the video clips and their corresponding transcriptions.

each utterance to a unique speaker's facial identity.

Ultimately, each video was segmented into multiple clips, with each clip annotated with its corresponding transcription, aligned audio, and the associated speaker's facial video. To ensure privacy protection, we exclusively release de-identified versions of the video, audio, and textual dialogue content, ensuring that personally identifiable information is effectively anonymized.

A.3 Manual Double-Check

We developed a simple data validation system, as illustrated in Fig. 7. This system enables data validators to systematically verify and rate each clip of every video. To ensure the reliability of our annotations, we recruited 100 volunteers to participate in a secondary review process. Each volunteer was assigned 50 videos for validation. The 773 videos processed by the automatic annotation pipeline were partitioned and distributed among the validators, ensuring that each video underwent review by approximately 6 to 7 individuals.

For each video clip, validators were instructed to assess the following criteria:

Q1: Frame composition and audio quality: Does the video contain only the frontal views of the two participants engaged in conversation, positioned on the left and right sides of the frame? Is the dialogue clear and entirely in English?

- The layout does not follow a left-right distribution; the number of participants is not two; the dialogue contains excessive noise; the conversation is not in English.

- Completely accurate.

Q2: Active speaker detection accuracy: Is the speaker highlighted by the red bounding box correctly identified (i.e., the individual within the red box is speaking at that moment)?

- Incorrect annotation.
- Completely accurate.

Q3: Subtitle alignment with spoken dialogue: Does the subtitle text accurately reflect the spoken content in the video?

- Subtitles do not correspond to the spoken dialogue. (1 point)

- Significant errors affecting sentence completeness or original meaning. (2 points)

- Some errors present, but they do not compromise sentence integrity or meaning. (3 points)

- Minor errors that do not impact semantic understanding. (4 points)

- Completely accurate. (5 points)

Q4: Naturalness of facial expressions and speech: Do the facial expressions and lip movements appear natural and expressive? Are the vocal tone and speech rate natural?

- Noticeable face-swapping artifacts; synthetic voice with evident electronic distortion; audio-visual misalignment. (1 point)

- Missing facial expressions; poor lip-sync accuracy; unnatural vocal tone. (2 points)

- Slight electronic distortion in voice; rigid facial expressions; suboptimal lip-sync. (3 points)

- Minor unnaturalness, discernible only upon close inspection. (4 points)

- Highly natural, indistinguishable from real human expressions and speech. (5 points)

Following the collection of validation responses, we applied a secondary filtering process to the video clips. Based on the first two criteria, we discarded videos exhibiting visual or audio inconsistencies, as well as those with erroneous speaker annotations. The latter two criteria assessed the accuracy of textual transcription and the naturalness of de-identification. Any segment receiving a score of 3 or lower on either of these criteria was removed from the dataset.

Ultimately, approximately 4% of the clips were discarded. The remaining clips achieved average scores of 4.72 and 4.74 on Q3 and Q4, respectively, indicating that our automatic annotation pipeline exhibits a high degree of accuracy.