Friends-MMSI: A Speaker Identification Dataset for Multi-modal Multi-party Dialogue Understanding

Anonymous ACL submission

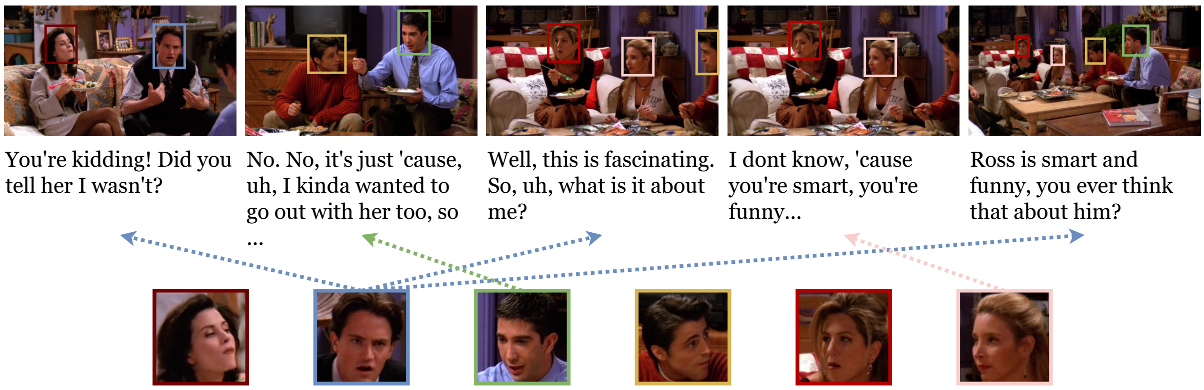


Figure 1: **An overview of Friends-MMSI, a multi-modal multi-party speaker identification dataset.** The goal is to find out the speaker of every utterance from characters that appear in the visual context (i.e., dotted arrows), by considering the entire dialogue as a whole and leveraging multi-modal information. Best viewed in color.

Abstract

Multi-modal multi-party dialogue understanding is a less studied yet important topic of research due to that it well fits real-world scenarios and thus potentially has more widely-used applications. In this paper, we pay attention to an important prerequisite of knowing whom is speaking for better understanding multi-modal multi-party dialogues, and thus propose this new format of task: Multi-modal Multi-party Speaker Identification (MMSI), where the system is required to identify the speaker of each utterance given the dialogue contents and corresponding visual context within a session. We construct Friends-MMSI, the first dataset of MMSI, which contains 24,000+ unique utterances annotated with speakers and faces in corresponding frames collected from TV Series *Friends*. We also propose a simple yet effective baseline method for MMSI, with results indicating that our proposed task and benchmark are still challenging, and we provide insightful knowledge to well understand this task. The code and dataset will be publicly available.

1 Introduction

Multi-modal dialogue systems have attracted extensive attention in recent studies (Zang et al., 2021; Zheng et al., 2022; Feng et al., 2023; Zhu et al., 2023; Liu et al., 2023; Li et al., 2023). However, there are two main deficiencies of existing work: (1) As most multi-modal datasets are collected from human annotations or chat history on social media, these dialogues are designed between human and system, instead of among several human interlocutors; (2) human interlocutors are bystanders (Das et al., 2016) and discuss the given visual content such as an image, instead of really being situated into the visual context. In addition, those dialogue datasets are mostly presented in Question-Answer format (AlAmri et al., 2019).

However, in real-world conversations, the interlocutors are often situated into the visual contexts, which means conversations can change the visual content. And real conversations can be much more diverse than merely responding to human-annotated questions, *i.e.,* QA. Therefore, we emphasize that multi-modal multi-party dialogue, especially when interlocutors are really situated in the visual context, is a more important for real application yet less studied topic.

To better study this topic, we first focus on an important prerequisite of it: **M**ulti-modal **M**ulti-party **S**peaker **I**dentification (**MMSI**). Apparently, for multi-party dialogue sessions that include many interlocutors, identifying the speaker of each utterance is crucial for dialogue understanding. In particular, for multi-modal dialogue sessions, it is also important to connect the utterance and its speaker to the person from the visual context. However, such annotations are often expensive and require tedious manual efforts, which indicates the necessity to study how to automatically perform speaker identification, and thus better supports the understanding of multi-party dialogues.

Currently, there are two tasks related to speaker identification: 1) **multi-party speaker identification for text-only dialogues**. Given a dialogue of m utterances and the speakers of the first $m-1$ utterances, this speaker identification aims to choose one from the previously appeared speakers for the last utterance. However, it is very limited since it requires previous utterances of a dialogue to be labelled with speakers, while only the last utterance is not. This largely hinders its application in most real-world scenarios where all speakers are unknown. Besides, this task does not take multi-modal contexts into account. 2) **Active speaker detection** for videos, on another hand, is to judge whether each track of face is speaking or not, given a video clip of a high frame rate. Recent works (Tao et al., 2021; Wuerkaixi et al., 2022; Datta et al., 2022) of this task focuses on facial movements, neglecting other information such as dialogue content and history. Moreover, this task setting relies much on high-quality videos. The model performance is largely affected if the frame rate of video is low (e.g., only very few or even one frame is available), or when the speaker does not even appear in the current visual frame.

To address existing issues and better meet real-world needs, we propose the new task **MMSI** (**M**ulti-modal **M**ulti-party **S**peaker **I**dentification): identifying the speaker of each utterance in a dialogue given the dialogue content and visual context of each utterance. Formally, a dataset of multi-modal multi-party speaker identification $\mathbf{D}$ consists of n sessions: $\mathbf{D} = \{e_1, \cdots, e_n\}$, and each session e_i consists of m consecutive utterances u, and each utterance is paired with a frame v: $e_i = \{(u_{i1}, v_{i1}), \cdots, (u_{im}, v_{im})\}$. Each utterance u_{ij} contains a dialogue content x_{ij} and speaker y_{ij}, each frame v_{ij} contains an image img_{ij}, and is labeled with f faces ($f = 0$ if there are no faces in the frame), where each face contains a bounding box b and a character name c: $v_{ij} = img_{ij}, \{(b_{ij1}, c_{ij1}), \cdots, (b_{ijf}, c_{ijf})\}$. To foster this newly proposed task, we build Friends-MMSI, a multi-modal multi-party speaker identification dataset collected from the famous TV series *Friends*. An overview of Friends-MMSI is shown in 1. Compared to the two described tasks, our task of MMSI, and our proposed dataset Friends-MMSI, have some traits worth emphasizing:

a) Modalities of available data are more diverse, including but not limited to: texts in utterance content, visual contexts in frames, face feature including the appearances, bounding boxes and character names, etc. Utilizing all of these modalities can be challenging for existing multi-modal models; **b) Reasoning** can be very complex. In our scenarios, a speaker can not appear in the given frame. Therefore, the preceding or succeeding textual and visual contexts, as well as their temporal relations, should be taken into account to infer the speaker, which is quite difficult to solve even for humans in our experiments. **c) Conversations** are taken from daily life such as TV series, which are more natural and diverse compared to existing multi-party datasets (Ouchi and Tsuboi, 2016; Hu et al., 2019) that are collected from chats only about computers.

We present a baseline method, which consists of a CNN-based for speaking face recognition, a transformer-encoder-based for modelling multi-turn speaker relation, and a speaker identification problem solver to assign speakers to utterances by optimizing outputs of these two models. We verify its performance on Friends-MMSI, and find that though basically effective, our method is still far from satisfactory and this new MMSI task is indeed challenging. In summary, our contributions are three-fold: (1) We propose **MMSI**, a new task of identifying speakers of each utterance given multi-modal contexts; (2) We build **Friends-MMSI**, a benchmark of multi-modal multi-party speaker identification; (3) We design a baseline for the MMSI task, validate its performance on Friends-MMSI, and provide insightful results to well understand this task.

2 Related Work

2.1 Multi-party Conversations

Multi-party conversations (MPC), as opposed to two-party conversations, is a more practical and challenging scenario of conversation that involves more than two interlocutors. Research on MPC understanding consists of three sub-topics: speaker prediction, utterance prediction, and addressee prediction. Ouchi and Tsuboi (2016) construct an MPC dataset from Ubuntu IRC Logs, and propose an RNN-based dual encoder model for addressee and utterance selection. Hu et al. (2019) also construct an MPC dataset from Ubuntu Dialogue Corpus, and use a graph-based model to understand the structure of dialogue history and generate response. Recently, studies on MPC usually train and evaluate models jointly on those three objectives. Gu et al. (2021) propose MPC-BERT, which fine-tunes BERT (Devlin et al., 2019) on several self-supervised tasks , and achieve state-of-the-art results on the above MPC tasks. GIFT (Gu et al., 2023) revises the model structure of transformer encoders to make the self-attention layer be aware of the information flow of MPC. Details regarding MPC can be found in works (Le et al., 2019; Gu et al., 2020) and this survey (Gu et al., 2022).

2.2 Active Speaker Detection

Active speaker detection (ASD) aims to detect which face is speaking in a video consisting of multiple speakers. The most widely-used dataset of ASD is AVAActiveSpeaker (Roth et al., 2019), where many video clips from movies are provided, and candidate models are required to label whether each face in each frame is speaking or not. ASC (Alcazar et al., 2020) and MAAS (Le'on-Alc'azar et al., 2021) first exploit the temporal and relational information from multiple speakers in consecutive frames, and more methods (Köpüklü et al., 2021; Min et al., 2022) further improve its performance. TalkNet (Tao et al., 2021) proposes to use cross-attention to aggregate video and audio features and achieve good performance. SyncTalkNet (Wuerkaixi et al., 2022), ADE-Net (Xiong et al., 2022) and ASD-Transformer (Datta et al., 2022) further improves this idea of video-audio aggregation by introducing novel structures like attention module, layer normalization, and position encoding. SPELL (Min et al., 2022) introduces graph structure to model spatial and temporal relations of speaker faces from a video, and then formalize ASD as a node classification task.

However, to the best of our knowledge, none of the existing works attempt to use semantic information in visual or textual dialogue contexts. More importantly, our motivation is not to replicate the traditional ASD task in a more tricky setting. We can simply improve ASD by leveraging more modalities, such as the high-rate frames and voice of each speaker. We aim to propose a new format of task that reflects how our existing multi-modal models can really understand aspects of multi-modal multi-party conversations, and we believe one of the most important aspects should be speaker identification.

2.3 Multi-Modal Dialogue Datasets

There have been a number of works on constructing multi-modal dialogue datasets. Das et al. (2016) introduce Visual Dialog, in which task an agent is asked to hold a natural and meaningful dialog with humans about a given image. Similar datasets include IGC (Mostafazadeh et al., 2017) and ImageChat (Shuster et al., 2020). AlAmri et al. (2019) propose AVSD, a dialogue dataset using videos as visual context. However, as these datasets are collected by asking crowd-sourced workers to discuss a given image/video, utterances are usually strongly grounded by the visual context, which is inconsistent with daily conversations. To address this issue, MMChat (Zheng et al., 2022) is a multi-modal dialogue corpus collected from Chinese social media, where dialogues are more in line with real-world scenarios, and each dialogue may correspond to one or multiple images. In PhotoChat (Zang et al., 2021) and MMDialog (Feng et al., 2023), images are not provided initially as visual context, but sent during the conversation. Despite the diversity in the position of images and videos, the above datasets are limited as the interlocutors are outside the visual contexts rather than "situated" inside them.

Dialogue in movie/TV series is a typical data source with "situated" visual context. Recently proposed large-scale movie dialogue datasets include OpenViDial (Meng et al., 2020; Wang et al., 2021) and VSTAR (Wang et al., 2023). However, these datasets do not consider modeling speaker information, which hinders a deeper-level understanding and utilization of the dialogue content. Perhaps the data most similar to ours is MELD (Poria et al., 2018), which is also a speaker-aware multi-modal multi-party dialogue dataset collected from *Friends* but focuses on emotion recognition, and does not annotate faces in the visual context.

3 The Friends-MMSI Dataset

In this section, we describe the dataset collection and annotation procedure we followed for constructing the Friends-MMSI dataset, which covers all the 220 episodes from 10 seasons of the TV show *Friends*. The reasons we use *Friends* are: (1) it is a sitcom series, which has numerous conversations that contain diverse topics of daily life; (2) Though having as many as 220 episodes, it has a relatively small number of main characters, which is convenient for automatic face labelling and data cleaning; and (3) It's easy to get publicly available resources like high-quality subtitles that are often manually revised and paired perfectly with the video by a large group of TV fans, which greatly reduces manual labour during the data construction process as well as guarantees the data quality.

Content and speaker of each utterance are extracted from transcripts and subtitles [1]. Faces and their character names in each frame are automatically detected and labelled for the train set (Season 1, 2, 4-10), and are manually labelled for the test set (Season 3) to ensure its accuracy.

3.1 Construction Process

Figure 2 shows the overall construction process of the dataset. Now we introduce every step in details:

Frame Selection. Each utterance is paired with one frame as the visual context. For all frames of the video clip corresponding to one utterance, we detect all faces per frame using an off-the-shelf face detector (Zha, 2017). Following Kalogeiton and Zisserman (2020), we merge the faces in adjacent frames into face tracks and thus remove the faces that are not in any track to clean out false positive detection. Finally, we select the frame with the most detected faces as the paired visual context of this utterance.

Character Face Prototype Construction. C1C (Kalogeiton and Zisserman, 2020) is a dataset with human-labelled face tracks for season 3 of *Friends*. We choose a set of 18 main characters, manually select 20 faces in different viewing angles for each main character, and encode them using Facenet-512 (Schroff et al., 2015) to get facial representation prototypes for each character.

Automatic Face Labelling. Then we automatically label the detected faces with character names by finding their nearest neighbour in the encoded embedding space. For each detected face per frame, we encode it with Facenet-512 and calculate the cosine similarity between its feature and all prototypes. If the largest cosine similarity is greater than a threshold $t = 0.5$ (this threshold is set to maximize accuracy described in the following paragraph), we label this face with the corresponding character name, otherwise we think this face does not belong to any of the main characters and discard it from the detected faces list.

To verify the accuracy of the automatic face labelling process, we use the same method to detect and label faces in season 3 and compare the results with human-annotated ones from C1C. The rule of verification is as follows: if the IoU of bounding boxes of an automatically labelled face and a human-annotated face is greater than 0.5, we identify them as a pair of identical faces. Given this threshold, 95% of all pairs of identical faces are labelled with correct names, which verifies the effectiveness of our automatic face labelling method.

For the test set, we directly use the human-annotated faces in C1C to ensure the accuracy of face labelling, serving as a high-quality ground-truths for this test set. Moreover, in order to comply with the setting of imperfect face recognition results in real-world applications and stay consistent with the training set, we also created a test-hard set by randomly removing 20% labelled faces.

Session Selection with Sliding Windows. We use a sliding window of size m to select m adjacent utterances if the following conditions are met: (1) all speakers are in the main character set; (2) the time intervals between all adjacent utterances are shorter than 8 seconds, which is a heuristic rule to prevent selecting utterances from different scenes. Therefore, we use $m = \{5, 8\}$ to create 2 datasets with different context lengths. Note that different dialogue sessions may contain the same utterances, as they belong to different contexts and thus the preceding or succeeding textual and visual contents differ. We use the accuracy of each turn in each session as the evaluation metric.

3.2 Dataset Statistics

Dataset statistics are shown in Table 1. Apart from the basic statistics, we also count the proportion of speakers whose faces are not detected in the corresponding frame or the entire session with m faces. Note that the test-hard set includes a significantly larger number of speakers not in current frame (24.31 for 5 turns, 25.32 for 8 turns) than the test-easy set (6.52 for 5 turns, 6.43 for 8 turns).

[1] https://my-subs.co/showlistsubtitles-610-friends; https://fangj.github.io/friends

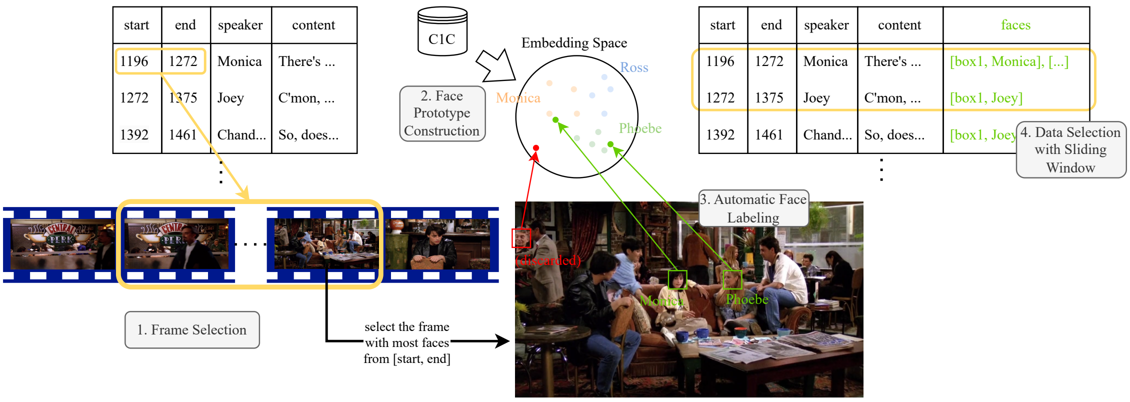
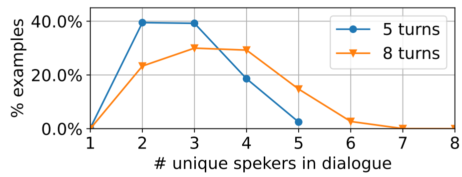


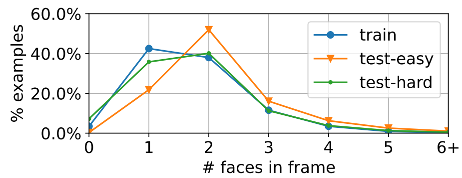
Figure 2: An overview of the construction process of Friends-MMSI dataset.

This situation is more difficult for speaker identification task, as the candidate model needs to find out more clues from the context rather than only the corresponding frame to infer who is the real speaker.

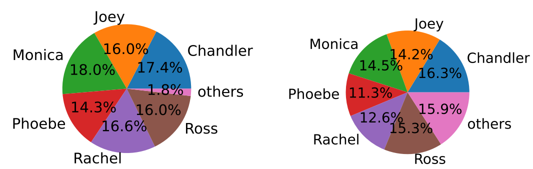
In addition, the test-hard set includes significantly more numbers (2.91 for 5 turns, 1.59 for 8 turns) where the speaker is not even appearing in all frames of a session, than that of the test-easy set (1.01 for 5 turns, 0.42 for 8 turns). It perfectly matches the real-world scenarios where a speaker is talking outside the camera, or like the voice-over technique. We believe this dataset thus can serve as a better simulation of the real situated conversations and a valuable evaluation even for the industrial use. More detailed data distribution regarding the number of unique speakers, labelled faces, and the main characters are shown in Figure 3. Note that the test-hard set includes slightly fewer faces per frame, since we remove 20% labelled faces.



(a) Number of unique speakers in each dialogue categorized by context length



(b) Number of labelled faces in each frame categorized by data splits



(c) Distribution of the main characters in speakers

(d) Distribution of the main characters in face labels

Figure 3: Detailed data distribution of Friends-MMSI.

4 Model

Our proposed benchmark dataset raises an increased demand on how to leverage both visual and context contexts to address this multi-modal multi-party speaker identification problem. In this section, we describe our baseline method, which consists of a CNN-based face recognition model to recognize speaking faces, a Transformer-encoder based model to analyse multi-speaker relations based on dialogue contexts, and a quadratic binary optimization problem solver to combine their results and thus identify the speaker of each utterance. Figure 4 shows the overview of our proposed baseline method, and we introduce each module in the following sections.

4.1 CNN-based Speaking Face Recognition Visual Model

We fine-tune a CNN model M_1 to predict the probability of each face in each frame belongs to the speaker of the corresponding utterance: $p_{face} = M_1(face) \in (0, 1)$, where $face$ is an image region acquired by cropping the image img using the bounding box b. The speaking label of this face y_{face} is set to 1 if the character name c of this face is identical to the speaker name y, and 0 otherwise: $y_{face} = \mathbf{1}[c = y]$. We use the cross-entropy

	5 turns			8 turns		
	train	test-easy	test-hard	train	test-easy	test-hard
# session	13584	2017	2017	8730	1325	1325
# frames	21092	3069	3069	16990	2480	2480
# words in utterance	18.87	20.28	20.28	18.71	20.42	20.42
# faces per frame	1.73	2.19	1.76	1.73	2.20	1.78
# speakers in each session	2.83	2.85	2.85	3.43	3.47	3.47
% speakers not in current frame	24.07	6.52	24.31	23.69	6.43	25.32
% speakers not in all frames	6.53	1.01	2.91	3.39	0.42	1.59

Table 1: Dataset Statistics of Friends-MMSI.

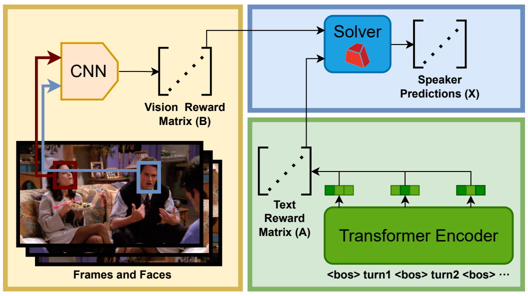


Figure 4: Model Overview.

classification loss as the training objective.

4.2 Transformer-Encoder Based Speaker Relation Text Model

We fine-tune a transformer encoder model M_2 to predict whether every two utterances in a dialogue are spoken by the same speaker. The intuitive reason behind it is that for some utterances, it is hard to identify the speaker solely from the corresponding frame by M_1. We thus try to conjecture its speaker by finding whether it likely shares the same speaker with another utterance, for which we have confidences or prior knowledge to infer its speaker.

Given a dailogue session consists of m utterances, we prepend an `<eos>` token to each utterance as the input of M_2 as like: `<eos>`$u_1 \cdots$`<eos>`u_m. We use the last hidden state of each `<eos>` h_i as the representation of each utterance, and use a head layer to calculate the similarity of every two representations:

$$p_{sim}^{ij} = \sigma(W_2 \mathrm{GeLU}(W_1[hi; hj; |hi - hj|] + b_1) + b_2)$$

where $i, j = 1, \cdots, m$, and (W_1, b_1, W_2, b_2) are learnable parameters. σ is the sigmoid activation function, and $p_{sim}^{ij} \in (0, 1)$ is a scalar that denotes the probability of two utterances spoken by the same person. The loss function is defined as:

$$\mathcal{L}_{M_2} = MSE(p_{sim}, y_{sim}) + MSE(p_{sim}, p_{sim}^T)$$

where $y_{sim} \in \{0, 1\}^{m \times m}$ is the ground truth label of whether any two utterances are from the same speaker, and MSE denotes mean squared error loss.

4.3 Speaker Identification Problem Solver

In order to leverage both visual and context contexts, we need to integrate the outputs of both models for speaker identification. For each dialogue session in the dataset, we first obtain a candidate speaker set by recording all faces appeared in every frame: $\mathbf{C} = \{c_1, \cdots, c_l\}$. We construct a reward matrix $\mathbf{B} \in \mathbb{R}^{l \times m}$ of selecting a character c_i as the speaker of the utterance u_j. If the face of c_i appears in the frame v_j, b_{ij} is set to the probability of that face as a speaking face predicted by M_1, otherwise $b_{ij} = 0$. However, $\mathbf{B}$ can only express those situations that the speaker appears in the corresponding frame. It is indeed a limitation of visual models since it can only view what can be viewed. To address those problems, the dialogue context is necessary to conjecture the speaker, we then construct another reward matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ of measuring the probability of assigning the same speaker to two utterances u_i and u_j. We first pass all utterances into the model M_2 as described in the previous subsection to get the similarity matrix p_{sim}. However, if we simply use this similarity matrix p_{sim} as the reward matrix $\mathbf{A}$, since all elements in p_{sim} are larger than 0, the optimization solver tends to assign the same speaker to every utterance in order to get the maximum rewards. To avoid which, we subtract the similarity matrix with the mean value of its elements, *i.e.*, $\mathbf{A} = p_{sim} - \mathrm{mean}(p_{sim})$.

With $\mathbf{A}$ and $\mathbf{B}$ in hand, the task of multi-modal multi-party speaker identification can be represented by a quadratic binary optimization problem:

$$\begin{aligned} \text{Maximize} \quad & f(X) = (1-\alpha) X^T A X + \alpha X B \\ \text{s.t.} \quad & X \in \{0,1\}^{m \times l}, \\ & \sum_{j=1}^{l} X_{ij} = 1, \quad i = 1, 2, \ldots, m \end{aligned}$$

where α is a hyperparameter to control the weight of two rewards and is selected according to the performance on validation set. By now, this problem can be easily solved using optimization problem solvers like Gurobi (Gurobi Optimization, LLC, 2023), which adaptively makes decisions based on the output of M_1 and M_2. As discussed in Section 5, the reason we use an optimization solver instead of an end-to-end pre-trained model is that this task of MMSI still remains challenging to use the general attention mechanism of pre-trained models like Violet (Fu et al., 2021) to fuse different modalities. Therefore, we have to design a better task-specific method than existing pre-trained multi-modal or single-modal models.

5 Experiment

5.1 Implementation

We use an Inception model (Szegedy et al., 2014) pre-trained on VGGFace2 (Cao et al., 2017) as M_1, and a DeBERTa-v3-large (He et al., 2021) as M_2. M_2 is first fine-tuned on Ubuntu Dialogue Corpus (Hu et al., 2019) and then on Friends-MMSI. Reward matrix weight α is set to 0.8. Both training and inference are conducted on a single GeForce RTX 3090 GPU and 5 CPUs in a few hours.

We conduct experiments in three different settings: (1) only using visual context; (2) only using textual context; and (3) using visual-text multi-modal context. In the visual only setting (1), we only use the model M_1 to predict one face from all detected faces in a frame as the speaker of the utterance. If there are no faces in the frame, we randomly choose a character from the candidate speaker list. We also report $M_1^\dagger$, which means the speaking face is always correctly identified as long as it appears in the frame, as an upper bound performance of using visual information only.

In the text only setting (2), we evaluate the performance of model M_2 and 3-shot ChatGPT [2] with in-context learning. For model M_2, although it is good at judging whether two utterances are said by the same speaker, it is not trained to identify the speaker for a single utterance. Therefore, it can only make guesses according to the relations between sentences. ChatGPT, however, possesses some ability to understand the candidate speaker list and identifying names from utterances, so it can make full use of the contextual information to provide more accurate reasoning. See appendix for details of the experiments using ChatGPT.

In the visual-text multi-modal setting (3), the entire $M_1 + M_2$ model is used together with a quadratic binary optimization solver, and we also try to replace the output of M_1 or M_2 with ground truth labels (*i.e.,* $M_1^\dagger$ for M_1 model and $M_2^\dagger$ for M_2 model) to explore bottlenecks and possible improvement directions. We also fine-tune a strong baseline of video-text multi-modal pre-trained model Violet (Fu et al., 2021), by constructing the below sequence of tokens as input: `[frame patches, candidate speakers, [CLS], utterance 1, [CLS], utterance 2, ...]`, and calculate the cosine similarity between the representation of each utterance (*i.e.,* the last hidden state of the `[CLS]` before it) and each speaker (*i.e.,* the last hidden state of the speaker name in `candidate speakers`).

We also report the human performance of this task. For the human experiment, we randomly sample 80 dialogue sessions from each (5 turns / 8 turns) test-easy set, provide dialogue contents, frames, face bounding boxes & labels to participants, and ask them to select a speaker for each utterance from the candidate speaker set (*i.e.,* the characters that appear in all frames). All participants are recruited from Chinese undergraduate and graduate students who are proficient in English and not familiar with *Friends*. This prerequisite is to guarantee the fair experimental results, since they have no prior knowledge with *Friends*. This process requires intensive efforts from humans, according to their post-interview, as the task of selecting speakers requires careful observation and a thorough understanding of the dialogue contents. We thus only perform the human studies on the test-easy set, since we believe the human performance on the test-hard set should be apparently worse.

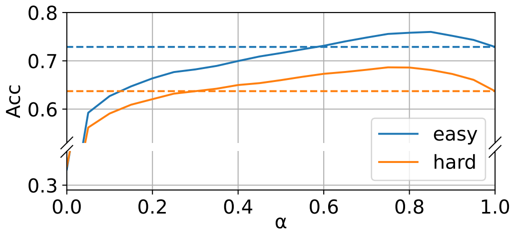
5.2 Main Results

According to the listed results in Table 2, we obtain the following observations: (1) visual information acquired by the vision model, including which face appears in the frame and looks like a speaking face, provides the most critical clues, shown by the performance of M_1 and $M_1^\dagger$. It can be concluded that this speaker identification task is still vision dominant. (2) Speaker relations acquired by the text model also play a vital supporting role to make an

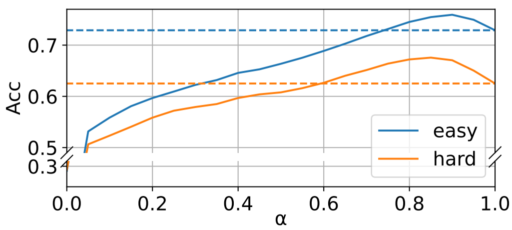
[2] `https://chat.openai.com`

		5 turns		8 turns	
		easy	hard	easy	hard
0	random (std.dev.)	31.82 (0.25)	32.61 (0.47)	28.54 (0.49)	29.03 (0.27)
Visual Information Only					
1	M_1	72.88	63.72	72.90	62.51
2	$M_1^\dagger$	94.97	82.09	94.96	81.70
Text Information Only					
3	M_2	33.24	33.85	29.09	29.33
4	ChatGPT	37.21	37.24	33.35	32.81
Multi-Modal Information					
5	Violet	32.66	33.16	27.73	28.86
6	M_1 + M_2	75.81	68.61	74.53	67.21
7	Human	82.25	-	84.49	-
8	M_1 + $M_2^\dagger$	84.90	78.01	90.80	83.93
9	$M_1^\dagger$ + M_2	96.40	87.46	96.86	87.34

Table 2: Accuracy on the test-easy and test-hard set of Friends-MMSI. † indicates that we use ground truths instead of outputs by that model, to serve as upper bounds.



(a) Test set performance on 5 turns dataset.



(b) Test set performance on 8 turns dataset.

Figure 5: The change of accuracy with respect to α. The dotted horizontal line shows the performance of only using the visual model.

improvement of $3\% \sim 5\%$ from M_1 to M_1 + M_2. The textual contexts benefit this task not only by understanding dialogue contents, but also for more real scenarios where the speaker does not appear in the frame; (3) Comparing the difference from $M_1^\dagger$ to $M_1^\dagger$ + M_2, and from M_1 to M_1 + M_2, we notice that the speaker relation benefits our model more when the paired visual model turns more accurate. Logically, it has to accurately identify speakers of some utterances before it is able to identify other utterances using this speaker relation information. (4) Comparing M_1 + M_2 with human performance (line 7) and models with ground truths (lines 8 and 9) as upper bounds, we find that both the visual and text model still have room for improvement. (5) Directly fine-tuning a multi-modal pre-trained model (line 5) may not even reach convergence even if many attempts were made to choose the best input format or training objectives, as the heterogeneous aspects that are essential to solve this problem remain difficult to be understood by the model. It also may be due to the reason that this speaker identification task is difficult to be formatted as the proper and shorten input of the model and thus to be easily learned. (6) The strong LLM model ChatGPT only has slightly better few-shot performance than random, indicating that it is non-trivial to apply pre-trained language models to this task, thus task-specific techniques needs to be developed, especially with the help of visual modality. It indicates that our proposed task and benchmark are still challenging and far away from a solution.

5.3 Analysis of Reward Weights α

Figure 5 shows the change of accuracy with respect to α. Note that when $\alpha = 0$, the task reduces to using only the text model M_2, and when $\alpha = 1$, the task reduces to using only the visual model M_1. In a considerable range of α values, introducing the results of M_2 improves the overall accuracy, compared with using M_1 only. It verifies that textual contexts certainly contribute to this speaker identification task.

6 Conclusion

In this paper, we propose multi-modal multi-party speaker identification (MMSI), an important prerequisite of multi-modal multi-party dialogue understanding, and discuss the definition and application of this task. We construct Friends-MMSI, the first benchmark for MMSI, from the famous TV Series *Friends*. We propose a simple yet effective baseline method, consisting of a CNN-based visual model, a transformer-based text model, and an optimization problem solver. We conduct extensive experiments on Friends-MMSI with various models for validation. Results indicate our newly proposed task and benchmark are still challenging and require more elaborate solutions from the community and industry. Finally, we discuss limitations and possible future directions of this work.

7 Limitations

In this section, we discuss the limitations, as well as possible future directions:

Increasing the diversity of speakers. In real-world application scenarios, such as conversational agents or understanding meeting recordings, interlocutors may not be limited to a specified set of main characters like Friends-MMSI do. Therefore, we wish the model to be person-agnostic: it should be able to identify speakers directly from the dialogue structure and their expressions or behaviours in the visual context, rather than learning from shortcuts such as characters' speaking habits. Although our dataset has already included the complicated scenario where speakers may not appear in the frame, we are still considering constructing benchmarks with more speakers, or even with open-ended ones (e.g., a pre-defined character list is not presented).

Utilizing more multi-modal information for speaker identification. The baseline model we introduced in Section 4 only makes use of facial appearance and dialogue content, and neglects other potential information such as face bounding boxes, gestures, background in the visual context, etc. Utilizing those visual information requires ingenious model structure and training methods, which are non-trivial to design. We leave this exploration to the future work, as well as welcoming more contributions from the community.

8 Ethical Concerns

Since the proposed dataset Friend-MMSI is collected from *Friends*, an English-language TV series filmed in the United States, and most of the actors/actresses are white, models trained on this dataset may contain bias and are not representative of scenes with other languages, races, and cultural backgrounds. Readers should be aware of this ethical concern when analyzing or quoting the findings of this work.

References

2017. S3fd: Single shot scale-invariant face detector. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 192–201.

Huda AlAmri, Vincent Cartillier, Abhishek Das, Jue Wang, Stefan Lee, Peter Anderson, Irfan Essa, Devi Parikh, Dhruv Batra, Anoop Cherian, Tim K. Marks, and Chiori Hori. 2019. Audio visual scene-aware dialog. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7550–7559.

Juan Leon Alcazar, Fabian Caba Heilbron, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbeláez, and Bernard Ghanem. 2020. Active speakers in context. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12462–12471.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. 2017. Vggface2: A dataset for recognising faces across pose and age. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2016. Visual dialog. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1080–1089.

Gourav Datta, Tyler Etchart, Vivek Yadav, Varsha Hedau, Pradeep Natarajan, and Shih-Fu Chang. 2022. Asd-transformer: Efficient active speaker detection using self and multimodal transformers. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4568–4572.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2023. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *Annual Meeting of the Association for Computational Linguistics*.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. 2021. Violet : End-to-end video-language transformers with masked visual-token modeling. *ArXiv*, abs/2111.12681.

Jia-Chen Gu, Tianda Li, Quan Liu, Xiaodan Zhu, Zhenhua Ling, Zhiming Su, and Si Wei. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Jia-Chen Gu, Zhenhua Ling, QUAN LIU, Cong Liu, and Guoping Hu. 2023. Gift: Graph-induced fine-tuning for multi-party conversation understanding. In *Annual Meeting of the Association for Computational Linguistics*.

Jia-Chen Gu, Chongyang Tao, and Zhenhua Ling. 2022. Who says what to whom: A survey of multi-party conversations. In *International Joint Conference on Artificial Intelligence*.

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. Mpc-bert: A pre-trained language model for multi-party conversation understanding. In *Annual Meeting of the Association for Computational Linguistics*.

Gurobi Optimization, LLC. 2023. Gurobi Optimizer Reference Manual.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. Gsn: A graph-structured network for multi-party dialogues. In *International Joint Conference on Artificial Intelligence*.

Vicky S. Kalogeiton and Andrew Zisserman. 2020. Constrained video face clustering using1nn relations. In *British Machine Vision Conference*.

Okan Köpüklü, Maja Taseska, and Gerhard Rigoll. 2021. How to design a three-stage architecture for audio-visual active speaker detection in the wild. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1173–1183.

Ran Le, Wenpeng Hu, Mingyue Shang, Zhenjun You, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019. Who is speaking to whom? learning to identify utterance addressee in multi-party conversations. In *Conference on Empirical Methods in Natural Language Processing*.

Juan Le'on-Alc'azar, Fabian Caba Heilbron, Ali K. Thabet, and Bernard Ghanem. 2021. Maas: Multi-modal assignation for active speaker detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 265–274.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wen Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *ArXiv*, abs/2305.06355.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *ArXiv*, abs/2304.08485.

Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2020. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. *ArXiv*, abs/2012.15015.

Kyle Min, Sourya Roy, Subarna Tripathi, Tanaya Guha, and Somdeb Majumdar. 2022. Learning long-term spatial-temporal graphs for active speaker detection. In *European Conference on Computer Vision*.

N. Mostafazadeh, Chris Brockett, William B. Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *ArXiv*, abs/1701.08251.

Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and response selection for multi-party conversation. In *Conference on Empirical Methods in Natural Language Processing*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, E. Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *ArXiv*, abs/1810.02508.

Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew C. Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, and Caroline Pantofaru. 2019. Ava active speaker: An audio-visual dataset for active speaker detection. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-chat: Engaging grounded conversations. In *Annual Meeting of the Association for Computational Linguistics*.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, D. Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

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. *Proceedings of the 29th ACM International Conference on Multimedia*.

Shuhe Wang, Yuxian Meng, Xiaoya Li, Xiaofei Sun, Rongbin Ouyang, and Jiwei Li. 2021. Openvidial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts. *ArXiv*, abs/2109.12761.

Yuxuan Wang, Zilong Zheng, Xueliang Zhao, Jinpeng Li, Yueqian Wang, and Dongyan Zhao. 2023. VSTAR: A video-grounded dialogue dataset for situated semantic understanding with scene and topic transitions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada,* *July 9-14, 2023*, pages 5036–5048. Association for Computational Linguistics.

Abudukelimu Wuerkaixi, You Zhang, Zhiyao Duan, and Changshui Zhang. 2022. Rethinking audio-visual synchronization for active speaker detection. *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 01–06.

Jun Xiong, Yu Zhou, Peng Zhang, Lei Xie, Wei Huang, and Yufei Zha. 2022. Look&listen: Multi-modal correlation learning for active speaker detection and speech enhancement. *ArXiv*, abs/2203.02216.

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. *ArXiv*, abs/2108.01453.

Yinhe Zheng, Guanyi Chen, Xin Liu, and Ke Wei Lin. 2022. Mmchat: Multi-modal chat dataset on social media. *ArXiv*, abs/2108.07154.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv*, abs/2304.10592.

A Experiments of ChatGPT

We use in-context learning to perform 3-shot inference with ChatGPT. Instruction, input and expected target we use is as follows:

You are listening to a conversation among a group of people. You will be provided with a name list and the content of conversation, and need to guess which people in the name list speaks each turn of the conversation. Answer one name for each turn in the dialogue, `[num turns]` comma-seperated names in all.
Name list: `[candidate 1], [candidate 2],...`
Conversation (one turn per line):
`[turn1]`
`[turn2]`
...
Answer: `[speaker 1],[speaker 2],...`

`[several more examples]`

Name list: `[candidate 1], [candidate 2],...`
Conversation (one turn per line):
`[turn1]`
`[turn2]`
...
Answer: `[speaker 1],[speaker 2],...`

If ChatGPT generates more than `[num turns]` names, we only keep the first `[num turns]` names as its predictions. If ChatGPT generates less than `[num turns]` names, or generates names not in the candidate list, we pad its prediction / replace the name not in the candidates list with names randomly selected from the candidate list.