Beyond Utterance: Understanding Group Problem Solving through Discussion Sequences

Zhuoxu Duan
Rensselaer Polytechnic Institute
Troy, New York, USA
duanz2@rpi.edu

Zhengye Yang
Rensselaer Polytechnic Institute
Troy, New York, USA
yangz15@rpi.edu

Brooke Foucault Welles
Northeastern University
Boston, Massachusetts, USA
b.welles@northeastern.edu

Richard J. Radke
Rensselaer Polytechnic Institute
Troy, New York, USA
rjradke@ecse.rpi.edu

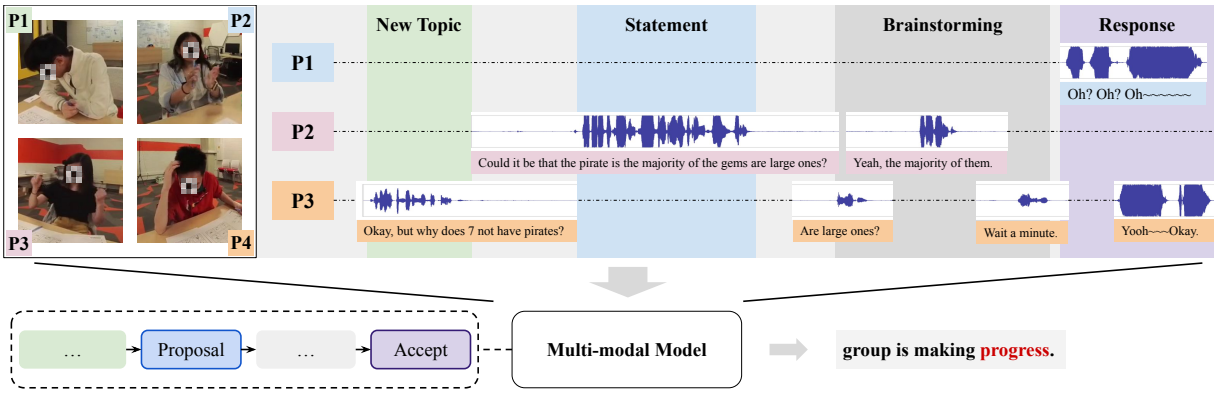

Figure 1: An example discussion segment formed by a sequence of dialogue acts.

Abstract

Automatically understanding and facilitating effective group collaboration remains a core challenge across social science and computational research. While prior work has focused on fine-grained social cues or coarse behavioral patterns, understanding the intermediate structure of dialogue—how sequences of utterances (discussion segments) reflect evolving group knowledge—is critical. This paper introduces a novel discussion segmentation framework and taxonomy for modeling collaborative problem-solving (CPS) processes, classifying segments into categories such as "task progress", "task attempt", and "grounding". We collected and annotated over 1,700 multi-modal discussion segments from 21 group discussions, both in-person and online, based on this taxonomy. We further propose a baseline model that integrates audio, visual, and textual signals to classify discussion segments with an average F1 score of 69.3%. Notably, this lightweight expert model achieves performance comparable to, and sometimes exceeding, proprietary state-of-the-art multimodal large language models. These findings highlight the promise of sequence-level discourse analysis for automated facilitation and human-agent collaboration.

CCS Concepts

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; **HCI theory, concepts and models**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

Multi-modal Learning, Social Interaction Modeling, Meeting Analysis, Collaborative Problem Solving, Human-AI Collaboration, Group Behavior Analysis

ICMI '25, Canberra, ACT, Australia

ACM ISBN 979-8-4007-1499-3/2025/10
https://doi.org/10.1145/3716553.3750758

ACM Reference Format:
Zhuoxu Duan, Zhengye Yang, Brooke Foucault Welles, and Richard J. Radke. 2025. Beyond Utterance: Understanding Group Problem Solving through Discussion Sequences. In *Proceedings of the 27th International Conference on Multimodal Interaction (ICMI '25), October 13–17, 2025, Canberra, ACT, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3716553.3750758

1 Introduction

Meetings play a crucial role in decision-making, collaboration, and problem-solving across various domains. Effective facilitation improves meeting efficiency and team coordination, leading to significant economic and social benefits. Developing a mediator agent capable of automatically facilitating meetings has long been a goal in the research community [2, 6, 28, 30, 38].

Achieving agent-driven facilitation requires not just surface-level behavior detection, but an understanding of how group discussions evolve. Prior work has explored this from two extremes: low-level cues like gaze or turn-taking [22, 29], and high-level summaries like engagement or role estimation [26, 49, 51]. However, these approaches often miss the intermediate structure—the sequence of utterances that form meaningful discussion units and reflect evolving group understanding. Despite its importance, this level remains underexplored.

In this work, we focus on the intermediate discourse structure in meetings by modeling how group problem-solving unfolds through sequences of utterances. Rather than attempting a universal solution across all meeting types, we ground our study in Collaborative Problem Solving (CPS)—a goal-driven, structured setting where discussion patterns are more observable and analyzable.

We introduce a taxonomy for classifying CPS discussion segments into types such as task progress, failed attempts, and grounding, based on sequential patterns of utterances. To support this, we present a multimodal model that integrates text, audio, and visual signals through feature-level fusion and sequence modeling, illustrated in Figure 1. We further benchmark our approach against recent large language models (LLMs), highlighting current limitations in their ability to capture these mid-level discourse patterns.

Our key contributions include:

- A new taxonomy for modeling utterance sequences in collaborative problem-solving settings, oriented towards automatic meeting understanding;
- A multi-modal model that integrates text, audio, and visual features to classify variable-length utterance sequences;
- Empirical comparisons with GPT-4o variants, showing both the promise and current limitations of LLMs in structured social understanding tasks.

2 Related Work

This section surveys related work across three research areas that inform our study: (1) artificial mediators in meetings, (2) computational modeling of social interactions at different granularities, and (3) conversation analysis and grounding in social psychology. Together, they form a foundation for our focus on collaborative problem solving in group meetings.

2.1 Towards Artificial Mediators in Meetings

Automatic meeting analysis was formally introduced by McCowan et al. [24], aiming to recognize common meeting activities such as discussion, presentation, and note-taking. However, the dataset used consisted of "scripted meetings," where events were artificially well-distributed and clear-cut, failing to capture the complexity and messiness of real-world interactions. Moreover, these broad action labels offer limited insight into deeper contextual understanding, making them insufficient for enabling downstream interventions like facilitation.

While scattered efforts have attempted to interpret group communication, they often overlook contextual richness. In the rush to explore the promise of "automatic facilitation," researchers have tended to sidestep technical challenges, resulting in overly simplified recognition pipelines. Consequently, agent interventions are reduced to basic attention-balancing or rigid coordination routines. For example, Bohus and Horvitz [6] presented one of the earliest efforts to explore automatic facilitation via an embodied agent that synchronizes gaze, gestures, and speech to manage turn-taking. Similarly, Schiavo et al. [38] proposed an instrumented table that estimates participants' attention from cameras and redirects attention to less involved individuals through animations on screens.

To move towards artificial mediators that can act like humans to positively affect human interactions, Park and Lim [30] highlighted the role of user expectations in shared human/agent contexts. Building on this, Müller et al. launched the MultiMediate challenges beginning in 2021 [25–28]. These challenges address key communication tasks such as next speaker prediction, backchannel detection, and engagement estimation.

2.2 Modeling Social Interactions

Tasks introduced in the MultiMediate challenges fall under the broader scope of social interaction modeling, where the aim is to automatically recognize and interpret multimodal behaviors in human communication. Depending on task objectives, these behaviors can be modeled at different levels of granularity.

At the fine-grained level, researchers have explored low-level social signals such as gaze and visual focus of attention, which help interpret engagement and attentiveness in group settings [29, 50]. Other signals like backchannels, subtle listener responses to a speaker, are used for downstream tasks such as agreement detection [1]. Turn-taking dynamics, including predicting end-of-turn moments, have also received attention in recent work [22].

At the utterance level, much work has focused on analyzing the function, emotion, or intent of a given utterance. Emotion recognition is addressed in both human-human and human-agent settings [16, 19], while intent modeling is enhanced through contrastive multimodal techniques [40]. Social deduction games provide a testbed for more structured utterance-level tasks like speaking target identification and utterance function classification, as demonstrated by recent multimodal benchmarks [20, 23].

At the summary level, researchers have examined social roles and personality traits by analyzing visual co-occurrences and turn patterns in group meetings. These approaches aim to infer higher-order attributes such as dominance or extraversion from group dynamics [49, 51].

2.3 Conversation Analysis in Social Psychology

The challenge of defining appropriate units of analysis has long been central to research in social and psychological domains. As Russell and Staszewski [37] point out, the chosen unit—be it a word, utterance, or sequence—depends heavily on the research question, yet this variability limits cross-study comparability. This issue carries over directly into social interaction modeling, where

inconsistent action definitions can hinder generalization and fair evaluation of detection methods.

To address this, Russell and Czogalik [36] advocate for sequence-based analysis in conversation research. Rather than isolating events, they argue that transitions and contingencies between utterances are critical to understanding interactional meaning. Langewitz et al. [21] and Russell [35] further formalize this by identifying recurring structures, such as "exchanges" (pairs of utterances forming a unit of interaction) and "rounds" (longer, thematic segments), as useful analytic constructs. These concepts directly inspire our proposed strategy for segmenting discussions into utterance sequences, capturing conversational flow and group dynamics more meaningfully than isolated utterance-level labels.

In computational contexts, Waibel et al. [44] recognized the importance of dialogue structure for meeting summarization and retrieval, highlighting that conversations offer richer semantic features than traditional bag-of-words representations. However, their work did not explore how to structure or segment dialogue for downstream analysis.

Renals and Ellis [34] explored topic segmentation in meetings, assuming that meetings consist of multiple evolving discussions. Their approach used the Bayesian Information Criterion (BIC) to segment based on speaker turns, but results diverged from human-marked topic boundaries, suggesting that speaker change alone is insufficient to capture the true flow of discussion or topic shifts.

Building on these insights, we narrow our scope from general meetings to the more tractable case of collaborative problem solving (CPS)—a structured, goal-oriented setting where group dynamics and knowledge construction processes are more clearly observable. Kapur et al. [18] conceptualize problem-solving in CPS as a complex, evolutionary process, emphasizing the need for dynamic models that trace reasoning and coordination over time. Dillenbourg and Traum [11] highlight grounding, the continuous process of establishing mutual understanding, as a core mechanism in CPS. These perspectives underscore the importance of sequencing and interpretation not just at the utterance level, but across extended, multimodal interactions where shared understanding evolves.

2.4 Dialogue Segmentation and Act Detection

Ang et al. [2] proposed to study automatic dialog act (DA) segmentation and classification in meeting environments, which could help meeting understanding in tasks like retrieval, question answering, and summarization at a higher level than words. Firdaus et al. [13] further combined DA with intent detection and slot filling to help understand conversations. Qin et al. [32] applied graphs to model interactions between speakers to help predict dialogue acts and sentiments. However, since DAs are defined at a level equal to or shorter than utterances, they lack the capability to convey information of a complete dialogue.

Dialogue segmentation has been widely used in Natural Language Processing (NLP) to summarize conversations [14, 15, 45, 47]. These methods rely on clean text input and detect shifts in linguistic features such as utterance cohesion [15, 47]. However, in real-world meetings, where interruptions and non-verbal cues shape discussions, text-only segmentation fails to capture the dynamics of collaborative problem-solving. In contrast, our approach models utterance sequences using multimodal features to track the evolution of meeting states.

3 Task Formation

In this work, we aim to understand collaborative problem solving (CPS) progress by modeling how discussions evolve during group meetings. Specifically, we define a discussion segment as a continuous sequence of utterances that addresses a single topic or subproblem within the broader meeting context. Understanding these utterance sequences is crucial for tracing how group knowledge states develop over time.

Given a sequence, our task is to classify it into a category that reflects its contribution to the group's problem-solving process (e.g., task progress, failed attempt, grounding). This setup falls under the domain of action recognition, where a trimmed segment is assigned a single label. Although it would be ideal to perform action detection—predicting both the temporal boundaries and labels of discussion segments from continuous recordings—this problem remains significantly more complex and underexplored, especially for social and multimodal signals where "actions" (topic shifts) are subtle and context-dependent [9, 31, 42]. Therefore, we focus on the more tractable classification formulation in this study, laying foundations for future work in detection.

3.1 Discussion Segmentation

Before classification, we need to extract discussion segments from continuous meeting recordings. This segmentation relies on human judgment, following structured rules to ensure consistency.

Inspired by prior work on topic segmentation [34], we define the boundaries of a discussion segment based on topic change events. A new sequence typically begins with:

- An utterance starting from silence; or
- A successful turn-taking act, where a participant initiates a new focus of conversation.

Similarly, a sequence ends when:

- Participants lapse into silence, indicating closure; or
- Another participant shifts the focus through an overt turn-taking signal, redirecting the group to a new topic.

Our approach is visually inspired by the end-of-turn illustration proposed in [22], as shown in Figure 2. Here, while a participant (in blue) is speaking, another participant (in orange) interrupts using strong turn-taking cues, such as increased volume and excited tone. After gaining the group's attention, the orange participant formally proposes a new idea, shifting the discussion topic from "size of gems" to "space between gems".

This segmentation approach ensures that each discussion sequence is thematically coherent, setting a clear and consistent unit of analysis for classification.

3.2 Class Definitions

Figure 1 illustrates an example discussion segment consisting of a series of dialog acts: a new topic starts from silence, followed by a proposal; the group assesses it and finally agrees with excited tones of voice. We can roughly split sequences into three parts: a starting

Figure 2: Illustration of turn-taking leading to a topic change. Figure style adapted from [22]. Utterances before and after this segment are omitted for simplicity.

statement, intermediate discussion, and final reaction. We propose a taxonomy involving different combinations of acts in each part, as shown in Figure 3 and discussed further below. This taxonomy was derived during annotation, inspired by team function classes [24] and common dialog acts.

Progress sequences start with a constructive proposal or statement (which is usually well-worded), and ends with some responses, showing clear intention of accepting the proposal (such as "I think it's correct," "Yeah, I agree with you," or repeating the statement, instead of open-ended short responses like "Yeah," "Uh-huh," or "Maybe"). *Progress* patterns clearly reflect the evolution of problem-solving state changes and are the most important category for our study.

Attempt is similar to *Progress*, but the response is a rejection by the group based on counterargument or silence. This class shows that the team tried to come up with a solution, but failed. An occasional or short series of *attempts* indicates that the team has some direction and is brainstorming towards a *progress* result.

Grounding encompasses other types of constructive discussions that do not fall into *progress* or *attempt*, generally including task confirmations and coordination between team members. In social psychology studies about communication, "grounding" is the process of building and maintaining group knowledge (e.g., establishing common ground) [8].

Unhelpful covers non-constructive communications that do not benefit the task. Patterns may include making jokes, talking about irrelevant topics, or team members talking to themselves.

Interruption is a special class that simulates potential automatic meeting mediation. In our experiments, an external voice occasionally interjects hints into the discussion. These utterances, together with participants' reactions to them, are considered *interruption* sequences.

4 Method

In this section, we formally describe our proposed multimodal framework for classifying sequences. The overall pipeline is illustrated in Figure 4 and consists of several stages: feature extraction and synchronization, augmentation, multimodal fusion, sequence pattern modeling, and classification.

In real-world face-to-face communication, understanding an utterance requires more than just the words being spoken. Text captures the core semantic content, making speech-to-text (STT) a crucial first step. However, audio provides additional context through prosody, tone, and emphasis, which are essential for interpreting the speaker's intent and emotional nuance. Complementing these, visual cues such as facial expressions, gestures, and body posture

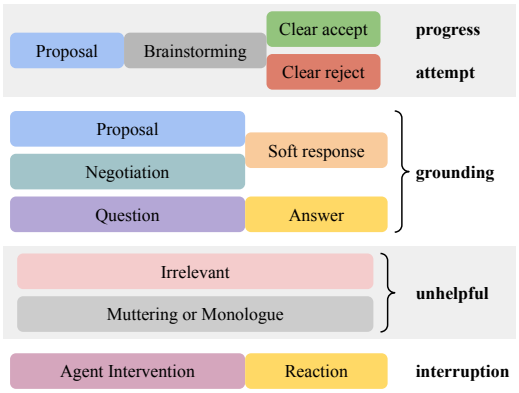

Figure 3: Visualization of proposed class definitions, based on utterance patterns.

serve as powerful indicators of listener reactions and conversational dynamics. Together, these modalities offer a more comprehensive understanding of human communication.

4.1 Feature Extraction & Synchronization

In modeling collaborative discussions, it is essential to capture the semantic, verbal, and non-verbal signals embedded in communication. We extract three main modalities: text, audio, and visual features. However, these modalities operate at different temporal granularities, creating a synchronization challenge that must be addressed before multi-modal fusion.

Audio and Text Extraction. We process the raw audio stream using Voice Activity Detection (VAD) to segment continuous recordings into utterance-level chunks. This segmentation is necessary both to match text and audio representations and to manage model limitations. We applied the Silero VAD method [41] supported in PyTorch.

Each utterance's audio data is passed through the wav2vec-BERT 2.0 encoder [5], generating audio embeddings $\mathbf{A}_{\text{raw}} \in \mathbb{R}^{L \times T_a \times D_a}$, where L is the number of VAD segments, T_a is the number of samples within a segment (which varies between segments), and D_a is the embedding dimension.

Concurrently, we transcribe each utterance into text using the Whisper model [33], subsequently extracting textual embeddings $\mathbf{T} \in \mathbb{R}^{L \times D_t}$ via Jina embeddings [39]. To align dimensions between text and audio, we must aggregate audio embeddings within each VAD segment. Although an RNN or transformer-based structure is a possibility, these can easily overfit to our limited data. Instead, we take the average along the T_a axis:

$$\mathbf{a}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} \mathbf{a}_{l,i}, \quad l = 1, \ldots, L \tag{1}$$

where n_t is the number of audio frames within the t-th VAD segment. Thus, the final audio embeddings, $\mathbf{A} \in \mathbb{R}^{L \times D_a}$, are synchronized with the text embeddings, $\mathbf{T} \in \mathbb{R}^{L \times D_t}$.

Visual Feature Extraction. For visual representation, we extract both upper-body pose and facial cues. Pose sequences are extracted

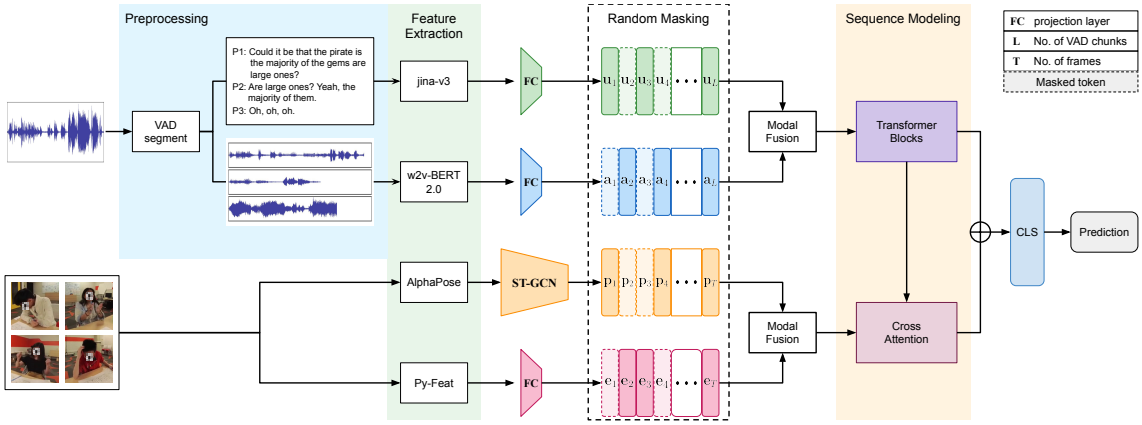

Figure 4: Overview of our proposed multimodal modeling pipeline. The input stage includes raw audio/video data and transcripts. Multimodal feature extraction is performed using wav2vec-BERT, Whisper, AlphaPose, and Py-Feat. The synchronized multimodal embeddings then undergo sequence modeling via a transformer encoder, followed by gated fusion strategies before classification.

using AlphaPose [12] at the original video frame rate (25 fps). We extracted Facial Action Units (AUs) and basic emotion categories with Py-Feat [7], sampled at 1 fps to match available high-confidence face detections.

We aim to further capture motion patterns like nodding and shaking heads with pose sequences, using a Spatial-Temporal Graph Convolutional Network (ST-GCN) [48]. Taking advantage of the convolutional structure, we gradually downsample the temporal resolution while increasing the spatial resolution. This aligns the output visual embeddings ($\mathbf{V} \in \mathbb{R}^{T \times D_v}$) to a fixed 1 fps temporal rate, independent of VAD chunking.

4.2 Random Sequence Masking

Due to the limited availability of annotated multimodal data, we found augmentation methods to be necessary to reduce overfitting and enhance model robustness. Inspired by successful masking strategies from BERT [10] and MAE [17], we implemented random masking at the sequence level. For a sequence of length L, each embedding vector (audio-text VAD segment or visual frame) is randomly masked during training with probability p. This masking is independently applied to each modality after synchronization and projection. The choice of mask ratios for each modality is discussed in Section 5.2.

4.3 Feature Level Fusion

Text-Audio Fusion. Given synchronized text and audio embeddings, we project each to a common dimension D via linear layers, obtaining $\mathbf{T}', \mathbf{A}' \in \mathbb{R}^{L \times D}$. To fuse these embeddings, we employ a gated addition mechanism [3]:

$$\mathbf{G}_{ta} = \sigma(\mathbf{W}_g[\mathbf{T}'; \mathbf{A}'] + \mathbf{b}_g) \tag{2}$$

$$\mathbf{F}_{ta} = \mathbf{G}_{ta} \odot \mathbf{T}' + (1 - \mathbf{G}_{ta}) \odot \mathbf{A}' \tag{3}$$

where σ is the sigmoid function, and $\mathbf{W}_g, \mathbf{b}_g$ are learnable parameters in a fully connected layer. A layer normalization [4] and residual connection are applied to stabilize and enhance feature integration:

$$\mathbf{F}_{ta}^{norm} = \text{LayerNorm}(\mathbf{F}_{ta} + \mathbf{T}') \tag{4}$$

Visual embeddings from pose and facial features are fused in the same way.

Text-Audio Sequence Pattern Modeling. The fused text-audio embeddings $\mathbf{F}_{ta}^{norm}$ contain rich sequential information essential for recognizing sequence patterns. To model temporal dependencies, we apply a Transformer encoder [43] with positional embeddings. Specifically, we prepend a learnable classification token (CLS) to the input sequence:

$$\mathbf{F}_{enc} = [\text{CLS}; \mathbf{F}_{ta}^{norm}] + \mathbf{P} \tag{5}$$

where $\mathbf{P}$ denotes positional embeddings. The same operation is also performed on the visual sequence. The transformer encoder outputs a sequence representation:

$$\mathbf{H} = \text{TransformerEncoder}(\mathbf{F}_{enc}) \tag{6}$$

The CLS token output $\mathbf{h}_{CLS}$ serves as the aggregated feature representation for classification.

Late Fusion with Visual Clues. Visual features often provide intermittent but informative signals (e.g., nodding, facial expressions) that do not always align temporally with speech content. To leverage these visual cues effectively, we introduce a cross-attention module between text-audio sequence embeddings and visual embeddings. Specifically, the output sequence $\mathbf{H}$ from the transformer encoder serves as the query (Q), and visual embeddings $\mathbf{V}$ serve as the key (K) and value (V):

$$\mathbf{H}_{\text{attn}} = \text{Softmax}\left(\frac{\mathbf{H}\mathbf{W}_Q(\mathbf{V}\mathbf{W}_K)^\top}{\sqrt{d_k}}\right)\mathbf{V}\mathbf{W}_V \tag{7}$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable parameter matrices, and d_k is the dimension of the query and key vectors. This way, the CLS token $\mathbf{h}_{\text{attn}}$ from $\mathbf{H}_{\text{attn}}$ brings learned visual information to the text-audio sequence.

Then, we apply a residual connection to $\mathbf{h}_{CLS}$ followed by a layer norm to get the final feature:

$$\mathbf{h}_{\text{final}} = \text{LayerNorm}(\mathbf{h}_{CLS} + \mathbf{h}_{\text{attn}}) \tag{8}$$

Finally, $\mathbf{h}_{\text{final}}$ is fed into a fully connected layer for classification:

$$\hat{y} = \text{Softmax}(\mathbf{W}_{fc}\mathbf{h}_{\text{final}} + \mathbf{b}_{fc}) \tag{9}$$

where $\hat{y}$ is the predicted class distribution. We use the standard cross-entropy loss between the predicted distribution $\hat{y}$ and the ground-truth one-hot label y:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{10}$$

where C is the number of classes, y_c is the ground-truth indicator for class c, and $\hat{y}_c$ is the predicted probability for class c.

5 Experiments

5.1 Dataset

We collected a new dataset to study collaborative discussions. Groups of 3 or 4 participants were recruited to play a puzzle game called *Cursed Treasure*, in which participants must determine a secret word by inferring the rules of several "curses" that relate to the pattern of colored gems inside a set of treasure chests. Successfully completing the task involves solving 7 subproblems that can be accomplished in any order.

A total of 21 sessions were recorded. There were 75 participants in total, and each person only played the game once. 12 of the sessions were conducted and recorded over Zoom, and 9 of the sessions were conducted in-person and recorded with an Insta360 camera. Session durations ranged from 20 to 45 minutes (the maximum time allowed), depending on how quickly teams solved (or failed to solve) the puzzle. In total, we obtained approximately 840 minutes of video recordings. Based on the segmentation rules introduced in Section 3, each session was divided into 48 to 129 discussion segments. Two trained annotators labeled a total of 1,747 segments, of which 1,548 were deemed valid after quality control. Segment durations ranged from 2.16 to 99.85 seconds, with the majority falling within 15–40 seconds. The class distribution is also long-tailed. Annotation consistency was measured using Inter-Rater Reliability (IRR), achieving a high agreement score of 92.9%. More details about the dataset statistics and the complete coding manual are given in the supplementary material.

5.2 Implementation Details

Model Parameters. We convert raw input modalities into pre-extracted feature sequences as described in Section 4.1. For the model structure, we applied a universal embedding dimension of 256 across all modalities to make the following fusion stage easy to align. To be specific, all the fully-connected layers mentioned in Figure 4 are one-layer. Text and audio embeddings are directly mapped to 256, while pose and facial features from 4-person views are first concatenated before projection.

The random masking ratios are 75% for text and audio features, and 95% for pose and facial features during training. This produced the best performance in practice during our experiments, which are discussed in detail in Section 5.5. In the sequence modeling part, we use a transformer encoder with 4 heads and 2 layers; the cross-attention module only applies 1 head for extracting visual cues without introducing too many parameters.

Training Settings. During training, we use the AdamW optimizer with learning rate 5×10^{-5} and cosine annealing. The maximum training epoch is set to twice the actual epoch. Models are trained with batch size 32 and 100 epochs. A weighted random sampler is applied to balance the ratio of samples between each class. We monitor the overfitting of each model, and do early stopping after the validation loss stops decreasing. Depending on the complexity of the models in Section 5.5, the stopping epoch varies from 30 to 60.

5.3 Evaluation

We evaluate our method as a classification task by reporting both prediction accuracy (ACC) and macro-averaged F1 score (F1) across the five classes. Due to the long-tailed nature of the class distribution, macro F1 is particularly important for assessing balanced performance. To mitigate group-level bias, we perform five-fold cross-validation. In each fold, four group sessions are selected for testing (one fold includes five sessions due to uneven group numbers), ensuring that all sessions are evaluated. Furthermore, to mitigate the distribution shifts between Zoom and in-person sessions, stratified k-fold splitting is used to ensure that each fold maintains a balanced proportion. Each group session contains a unique set of participants, so there is no overlap between training and test data across folds.

Our best-performing model, which uses text and audio modalities fused via a gated addition strategy, achieves an average accuracy of 68.9% and a macro F1 score of 68.4% across the five folds. The detailed class-wise performance is shown in the confusion matrix in Figure 5. Here, prediction results and ground truth labels from all folds are concatenated and normalized by the support (sample count) of each class, allowing the matrix to reflect overall dataset performance. Consequently, diagonal elements directly represent the average per-class accuracy.

As observed, *Interruption* segments are identified with high accuracy, likely due to their distinctive intervention-response structure. In contrast, some confusion exists between *grounding* and other classes. This is consistent with expectations: although *progress*, *attempt*, and *unhelpful* each have distinct definitions, they often share conversational patterns with *grounding*, making precise differentiation challenging. We note that our model rarely confuses *progress* and *attempt*, which would be important for a facilitation agent.

5.4 Comparison to LLMs

We compare our best-performing model against GPT-4o, the flagship OpenAI model at the time of submission. We investigated several prompting strategies including (1) zero-shot text prompting, where discussion segment transcriptions are directly passed as input; and (2) one-shot text prompting, where one labeled example for each defined class is included to guide the model. Additionally, we evaluated the GPT-4o-audio variant, which supports both audio and text inputs. For these experiments, audio recordings of discussion segments are provided alongside text instructions requesting class label predictions. As a further baseline, we also supplied the

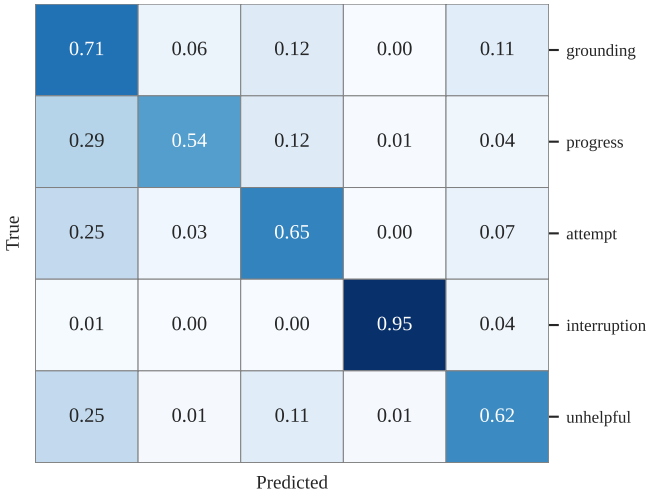

Figure 5: Normalized Confusion Matrix (5-fold concatenated) of the best performance model.

```
Classify the type of discussion from a problem solving task with a
group of 3 to 4 people. For a given discussion, only assign one class.

Class Definitions:   {class_definition}
Examples: {n_shot_samples}

(Discussion audio is provided in the 'input_audio')

Discussion Transcription:\n{discussion}
Predicted Class:
```

:{class_label}

Figure 6: GPT prompting method.

transcript together with audio input. Our general prompting structure is illustrated in Figure 6. All interactions use the OpenAI API, with temperature set to 0 and maximum token length limited to 10, to encourage stable one-word label outputs.

Table 1 summarizes the results. Note that GPT is tested on the full dataset, while our method uses five-fold cross validation for a fair comparison. In the zero-shot setting, GPT struggles to fully interpret the task based on class definitions alone, leading to relatively low but balanced performance. Providing one labeled example per class (one-shot) significantly improves both accuracy and F1, suggesting the model benefits from minimal context adaptation. Although GPT-4o-audio variants exhibit higher raw accuracy compared to zero-shot text, their F1 scores remain low, indicating biased predictions, often defaulting to the dominant *grounding* class. This highlights a limitation of current multimodal LLMs: despite accepting audio inputs, they struggle to extract detailed discourse patterns, a trend also observed in recent studies on vision-language models [46]. In contrast, our proposed model, while lightweight and domain-specific, consistently achieves higher or comparable performance across evaluation metrics.

Table 1: Comparison of our model with GPT-4o under various prompting and input settings. Accuracy (ACC) and F1 scores are reported. Bolded results denote the best overall performance. Underlined results highlight the best among GPT-4o variants. For our model, the performance is presented as mean ± standard deviation over multiple runs.

Model	ACC (%)	F1 (%)
GPT-4o (Zero-shot)	41.26	45.04
GPT-4o (One-shot)	65.20	65.31
GPT-4o-audio w. Text Prompt	54.22	43.41
GPT-4o-audio + transcript	59.83	43.83
Ours (Best Model)	**70.3± 3.3**	**69.3±3.2**

5.5 Ablation Studies

We conducted several ablation studies to better understand the design choices in our framework. Unlike the main evaluation, these studies use a fixed validation set of four group sessions to observe relative performance changes across model variants.

Impact of Random Masking Ratio. We examined the effect of different mask ratios, ranging from 0% to 95%, using only text embeddings as input. As shown in Figure 7, applying non-zero masking significantly reduces overfitting. However, extremely high masking (e.g., 95%) leads to unstable loss behavior due to excessive information loss, while lower ratios (25% or 50%) are insufficient for strong regularization. Based on these findings, we adopt a 75% masking rate for text and audio modalities. For visual sequences, we found that a 95% mask rate on pose and facial features is still stably helping our final model. This is possibly due to the nature of visual patterns: most motions or steady states are irrelevant, with occasional short hints like nodding or smiling.

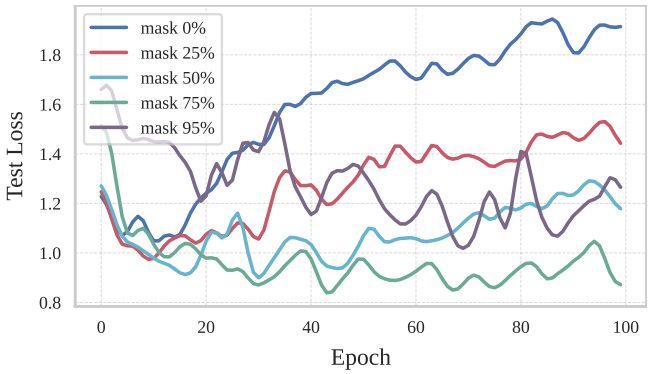

Figure 7: Test losses with different random mask ratios.

Fusion Strategies. Leveraging voice activity detection (VAD), our model operates on aligned text and audio utterance embeddings, as well as frame-level pose and facial features. Given this alignment, simple fusion methods like element-wise addition or concatenation are feasible alternatives to more complex cross-attention mechanisms. In addition, we compare different fusion strategies, with and

Table 2: Class-wise Accuracy (ACC) and F1 scores for each modality configuration in percentage (%). Checkmarks (✓) indicate the modality used. Average columns report macro averages across all classes.

Text	Audio	Visual	Progress		Attempt		Grounding		Unhelpful		Interruption		Average	
			ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
✓			52.94	56.25	52.50	49.41	72.02	73.55	63.93	61.42	90.48	88.37	67.28	65.80
	✓		47.06	35.56	30.00	26.97	52.98	58.75	60.66	60.16	80.95	79.07	52.78	52.10
✓	✓		**64.71**	**61.97**	**55.00**	**57.14**	75.00	76.36	**75.41**	71.32	95.24	97.56	72.84	**72.87**
✓		✓	61.76	56.00	40.00	43.24	75.00	72.83	50.82	55.36	95.24	97.56	66.05	65.00
✓	✓	✓	55.88	56.72	27.50	38.60	**86.90**	**79.13**	70.49	**75.44**	95.24	97.56	**73.77**	69.49

Table 3: Effect of different fusion strategies and gating mechanisms during training (random masking applied).

Fusion Method	Gate	ACC (%)	F1 (%)
Add	–	67.59	66.05
Concat	–	68.52	65.56
Add	✓	**72.84**	**67.99**
Concat	✓	64.51	62.97

without gating operations that dynamically control the contribution of each modality. Strategies are applied to both speech and visual fusion stages after random masking. Table 3 reports the results. Addition with gating yields the best performance, achieving 73% accuracy and 68% F1 score. While concatenation achieves similar results without gating, it introduces more parameters, making the model prone to earlier overfitting, especially under limited training data. Moreover, gating does not benefit concatenation, likely because concatenated representations require more complex interactions that a simple gating mechanism cannot effectively model.

Modality Ablations. Finally, we study the contributions of different modalities under fixed random masking and gated-add fusion settings. To maintain consistency, unused modalities are initialized but omitted from the forward pass during training and evaluation.

Table 2 presents the detailed results. Text is clearly the dominant modality, achieving strong performance on its own. Audio embeddings alone are insufficient, but they significantly enhance text features when combined. The role of visual features is more complex: while they slightly improve recognition for *grounding* and *unhelpful* classes, they harm the detection of *progress* and *attempt*. Overall, we were surprised to find the visual features to provide limited benefit. This may be due to the characteristics of our specific task, in which participants often look down at paper materials during thinking phases, generating substantial noise in pose and facial signals. These factors make extracting reliable visual cues such as nodding or facial expressions particularly challenging for simple models. More ablations about sequential model choice and different facial feature frame rates can be found in the supplementary.

6 Safe and Responsible Innovation Statement

All data used was collected with informed consent under IRB protocols, ensuring participant privacy and ethical handling. While automation in social settings offers promise, it raises risks related to surveillance, misinterpretation, and fairness. We acknowledge potential biases introduced by limited cultural and task diversity in our dataset and call for broader validation. Our model is designed for interpretability and remains lightweight to promote transparent and controllable deployment.

7 Conclusions and Discussion

We proposed a structured approach to classifying discussion segments based on utterance sequences, bridging the gap between low-level social signals and high-level behavioral summaries. Our model achieves competitive performance with LLMs like GPT-4o while remaining lightweight and efficient for practical use. On the other hand, LLMs show strong potential for generalization across diverse tasks with minimal tuning. These complementary strengths suggest a promising future where task-specific models provide robust baselines, while LLMs offer flexible scaffolding for broader, less-constrained social interaction scenarios.

Several challenges remain. Our framework assumes coherent, single-topic segments, limiting its application to more fragmented group interactions. The taxonomy focuses on major CPS states but does not capture transitional or ambiguous phenomena, such as failed grounding or divergent individual reasoning, which could offer richer insights into group processes. Moreover, segmentation itself depends on human judgment, highlighting the need for scalable automation. Our dataset is also single-task, introducing language priors from the puzzle-solving context; extending to diverse collaboration tasks is needed for broader generalization.

Looking ahead, two directions are key: universal feature encoders that align text, audio, and visual modalities for more flexible modeling, and multi-scale sequence models to capture social acts across temporal levels. Unified models supporting these aspects will better facilitate dynamic, real-world meetings.

Acknowledgments

This research was supported by a grant from the Strengthening Teamwork in Novel Groups' Collaborative Research Alliance of DEVCOM ARL (U.S. Army Research Lab) under Grant No. W911NF-19-2-0135. Thanks to Samuel Westby (Northeastern University) for collecting data from Zoom meetings that contributed to this study. Jay Lorch and Michelle Teague created the Cursed Treasure puzzle. ChatGPT was used to polish the wording of this paper. All the ideas we present are original.

References

[1] Ahmed Amer, Chirag Bhuvaneshwara, Gowtham K Addluri, Mohammed M Shaik, Vedant Bonde, and Philipp Müller. 2023. Backchannel detection and agreement estimation from video with transformer networks. In *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[2] Jeremy Ang, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Vol. 1. IEEE, I–1061.

[3] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992* (2017).

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[5] Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation. *arXiv preprint arXiv:2312.05187* (2023).

[6] Dan Bohus and Eric Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. 1–8.

[7] Jin Hyun Cheong, Eshin Jolly, Tiankang Xie, Sophie Byrne, Matthew Kenney, and Luke J Chang. 2023. Py-feat: Python facial expression analysis toolbox. *Affective Science* 4, 4 (2023), 781–796.

[8] Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In *Perspectives on socially shared cognition*, Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley (Eds.). American Psychological Association, Washington, DC, USA, 127–149.

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*. 720–736.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[11] Pierre Dillenbourg and David Traum. 2006. Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *The Journal of the Learning Sciences* 15, 1 (2006), 121–151.

[12] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2022. AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[13] Mauajama Firdaus, Hitesh Golchha, Asif Ekbal, and Pushpak Bhattacharyya. 2021. A deep multi-task model for dialogue act classification, intent detection and slot filling. *Cognitive Computation* 13 (2021), 626–645.

[14] Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. *arXiv preprint arXiv:2109.09777* (2021).

[15] Qinyu Han, Zhihao Yang, Hongfei Lin, and Tian Qin. 2024. Let Topic Flow: A Unified Topic-Guided Segment-Wise Dialogue Summarization Framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).

[16] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*. 1122–1131.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.

[18] Manu Kapur, John Voiklis, and Charles K Kinzer. 2017. Problem solving as a complex, evolutionary activity: A methodological framework for analyzing problem-solving processes in a computersupported collaborative environment. In *Computer Supported Collaborative Learning 2005*. Routledge, 252–261.

[19] Nikola Kovacevic, Christian Holz, Markus Gross, and Rafael Wampfler. 2024. On Multimodal Emotion Recognition for Human-Chatbot Interaction in the Wild. In *Proceedings of the 26th International Conference on Multimodal Interaction*. 12–21.

[20] Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James Rehg, and Diyi Yang. 2023. Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games. *Association for Computational Linguistics: ACL 2023* (2023).

[21] Wolf Langewitz, Matthias Nübling, and Heidemarie Weber. 2003. A theory-based approach to analysing conversation sequences. *Epidemiology and Psychiatric Sciences* 12, 2 (2003), 103–108.

[22] Meng-Chen Lee and Zhigang Deng. 2024. Online Multimodal End-of-Turn Prediction for Three-party Conversations. In *Proceedings of the 26th International Conference on Multimodal Interaction*. 57–65.

[23] Sangmin Lee, Bolin Lai, Fiona Ryan, Bikram Boote, and James M Rehg. 2024. Modeling multimodal social interactions: new challenges and baselines with densely aligned representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14585–14595.

[24] L McCowan, Daniel Gatica-Perez, Samy Bengio, Guillaume Lathoud, Mark Barnard, and Dong Zhang. 2005. Automatic analysis of multimodal group actions in meetings. *IEEE transactions on pattern analysis and machine intelligence* 27, 3 (2005), 305–317.

[25] Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Anna Penzkofer, Dominik Schiller, François Brémond, Jan Alexandersson, Elisabeth André, et al. 2024. MultiMediate'24: Multi-Domain Engagement Estimation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11377–11382.

[26] Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Dominik Schiller, Mohammed Guermal, Dominike Thomas, François Brémond, Jan Alexandersson, et al. 2023. Multimediate'23: Engagement estimation and bodily behaviour recognition in social interactions. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9640–9645.

[27] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Hali Lindsay, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2022. MultiMediate'22: Backchannel Detection and Agreement Estimation in Group Interactions. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7109–7114.

[28] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Guanhua Zhang, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2021. Multimediate: Multi-modal group behaviour analysis for artificial mediation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4878–4882.

[29] Kazuhiro Otsuka, Keisuke Kasuga, and Martina Köhler. 2018. Estimating Visual Focus of Attention in Multiparty Meetings using Deep Convolutional Neural Networks. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 191–199. doi:10.1145/3242969.3242973

[30] Sunjeong Park and Youn-kyung Lim. 2020. Investigating user expectations on the roles of family-shared AI speakers. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[31] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. 2024. An outlook into the future of egocentric vision. *International Journal of Computer Vision* 132, 11 (2024), 4880–4936.

[32] Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2021. Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 13709–13717.

[33] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.

[34] Steve Renals and Dan Ellis. 2003. Audio information access from meeting rooms. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, Vol. 4. IEEE, IV–744.

[35] Robert L Russell. 1988. A new classification scheme for studies of verbal behavior in psychotherapy. *Psychotherapy: Theory, Research, Practice, Training* 25, 1 (1988), 51.

[36] Robert L Russell and Dietmar Czogalik. 1989. Strategies for analyzing conversations: Frequencies, sequences, or rules. *Journal of Social Behavior and Personality* 4, 3 (1989), 221.

[37] R. L. Russell and C. Staszewski. 1988. The Unit Problem: Some Systematic Distinctions and Critical Dilemmas for Psychotherapy Process Research. *Psychotherapy: Theory, Research, Practice, Training* 25, 2 (1988), 191–200. doi:10.1037/h0085333

[38] Gianluca Schiavo, Alessandro Cappelletti, Eleonora Mencarini, Oliviero Stock, and Massimo Zancanaro. 2014. Overt or subtle? Supporting group conversations with automatically targeted directives. In *Proceedings of the 19th international conference on Intelligent User Interfaces*. 225–234.

[39] Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual Embeddings With Task LoRA. arXiv:2409.10173 [cs.CL] https://arxiv.org/abs/2409.10173

[40] Kaili Sun, Zhiwen Xie, Mang Ye, and Huyin Zhang. 2024. Contextual augmented global contrast for multimodal intent recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26963–26973.

[41] Silero Team. 2024. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. https://github.com/snakers4/silero-vad.

[42] Elahe Vahdani and Yingli Tian. 2022. Deep learning-based action detection in untrimmed videos: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4302–4320.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. *Advances in neural information processing systems* 30 (2017).

[44] Alex Waibel, Michael Bett, Florian Metze, Klaus Ries, Thomas Schaaf, Tanja Schultz, Hagen Soltau, Hua Yu, and Klaus Zechner. 2001. Advances in automatic meeting record creation and access. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, Vol. 1. IEEE, 597–600.

[45] Jinxiong Xia, Cao Liu, Jiansong Chen, Yuchen Li, Fan Yang, Xunliang Cai, Guanglu Wan, and Houfeng Wang. 2022. Dialogue topic segmentation via parallel extraction network with neighbor smoothing. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2126–2131.

[46] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2024. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13204–13214.

[47] Linzi Xing and Giuseppe Carenini. 2021. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. *arXiv preprint arXiv:2106.06719* (2021).

[48] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[49] Lingyu Zhang, Indrani Bhattacharya, Mallory Morgan, Michael Foley, Christoph Riedl, Brooke Welles, and Richard Radke. 2020. Multiparty visual co-occurrences for estimating personality traits in group meetings. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2085–2094.

[50] Lingyu Zhang, Mallory Morgan, Indrani Bhattacharya, Michael Foley, Jonas Braasch, Christoph Riedl, Brooke Foucault Welles, and Richard J Radke. 2019. Improved visual focus of attention estimation and prosodic features for analyzing group interactions. In *2019 International Conference on Multimodal Interaction*. 385–394.

[51] Lingyu Zhang and Richard J Radke. 2020. A multi-stream recurrent neural network for social role detection in multiparty interactions. *IEEE Journal of Selected Topics in Signal Processing* 14, 3 (2020), 554–567.