MeetingToM: Evaluating Multimodal LLMs on Theory-of-Mind Reasoning in Multi-Party Meetings

Anonymous ACL submission

Abstract

Theory of Mind (ToM), the ability to infer other's beliefs, intentions, and states of knowledge, is central to social interaction, yet remains challenging for current Multimodal Large Language Models (MLLMs), especially in multi-party meetings where cues are distributed across speech and behavior. Existing multimodal ToM benchmarks mainly focus on video-grounded question answering over overt, externally verifiable signals, and provide limited coverage of latent social states and group dynamics. We introduce **MeetingToM**, a benchmark for complex social behavior reasoning in naturalisitc multi-party meetings. MeetingToM targets meeting-specific phenomena such as **pseudoconsensus**, where apparent agreement masks private dissent under social pressure. The benchmark is hierarchically organized to evaluate ToM at increasing levels of social granularity, including (i) subject-level mental state prediction, (ii) dyadic-level addressee understanding, and (iii) group-level consensus reasoning. We provide a unified evaluation protocol and conduct systematic analyses of representative MLLMs, revealing persistent limitations in integrating non-verbal cues, inferring hidden attitudes, and distinguishing genuine consensus from pseudoconsensus. Our results highlight key challenges and establish MeetingToM as a testbed for advancing meeting-grounded ToM in multimodal models.

1 Introduction

Humans continuously infer and update their own and others' beliefs, intentions, and states of knowledge during social interactions. This innate capability, known as Theory of Mind (ToM), is fundamental to social behavior reasoning. Prior work shows that ToM is context-sensitive (Speiger et al., 2025), with goals, plans, and social norms shaping how people reason about others' belief states. Notably, these contextual factors are especially salient in multi-party meetings, where participants bring distinct goals and partial knowledge. Accordingly, in this work we seek to probe computational Theory of Mind in meeting scenarios. An AI system capable of interpreting complex social states in such setting would enable a wide range of practical applications, such as collaborative decision support and socially aware assistants.

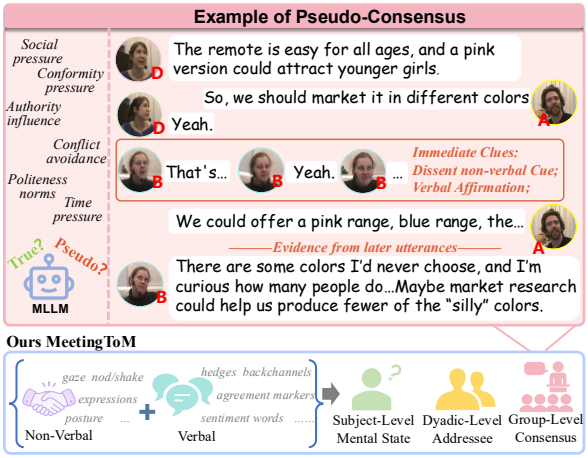


Figure 1: *Hierarchically structured tasks in our MeetingToM benchmark and example of pseudoconsensus.* Our MeetingToM benchmark assesses whether MLLMs can jointly reason over verbal and non-verbal cues to support hierarchical Theory of Mind tasks, spanning subject-level mental state recognition, dyadic-level addressee and attitude inference, and group-level consensus assessment. Our benchmark consider complex phenomena such as *pseudoconsensus*, where verbal agreement masks non-verbal dissent under social pressure.

With the advent of Multimodal Large Language Models (MLLMs), there has been growing interest in probing their capabilities for Theory of Mind from a multimodal perspective, motivated by the fact that beliefs and intentions are often conveyed through both verbal and non-verbal cues. Existing multimodal ToM benchmarks (Jin et al., 2024a; Shi et al., 2025) largely emphasize video-grounded question answering, often focusing on overt, exter-

nally verifiable signals, rather than systematically testing nuanced, latent social states that unfold in realistic group interactions.

Multi-party meetings, however, present a particularly challenging and consequential setting for multimodal ToM. Beyond understanding what is said, models must piece together evidence from speech and behavior, including turn-taking, gaze, posture and how people respond to each other, while considering norms such ad politeness and comformity. Crucially, meetings frequently contain surface-level agreement that msks private reservations: participants may verbally endorse proposals while subtly signaling doubt or disengagement, and disagreements may only become apparent through subsequent actions or indirect cues. Such phenomena call for benchmarks to explicitly evaluate whether models can infer hidden attitudes and group-level social dynamics, including cases where consensus is only apparent.

To this end, we introduce **MeetingToM**, the first benchmark that thoroughly evaluates MLLMs' capabilities for complex social behavior reasoning in naturalistic multi-party meetings. Consider the example in Fig. 1. We present a case in which subject B verbally agrees with another participant, yet their body language suggests otherwise, and subsequent utterances reveal underlying concerns. This social phenomenon, in which a subject conceals dissent due to social pressure, is often referred to as **pseudoconsensus** (Janis, 2008). Concretely, our benchmark tasks are hierarchically organized, with increasing demands on ToM capabilities:

- **Subject-Level Mental State Prediction**: Infers a target participant's mental state from a brief video segment and aligned utterances.
- **Dyadic-Lvele Addressee Understanding**: Involves identifying the referent of "you" using gaze and body orientation cues, followed by inferring the addressee's attitude
- **Gorup-Level Consensus Reasoning**: determine whether group agreement is genuine, reflects pseudoconsensus, or is absent, and to identify any participants who privately dissent.

In our experiments, we first benchmark the performance of open-sourced and proprietary MLLMs. We then provide a systematic analysis of how input modalities affect overall performance and examine how chain-of-thought and emotional chain-of-thought prompting strategies benefit social state understanding with MLLMs.

2 Related Work

Social Intelligence in Meeting Scenarios. Early research in social signal processing seek to quantify group dynamics through observable behaviors such as the exchange of conversational turns, the recognition of roles, and the establishment of dominance hierarchies 1/(Vinciarelli et al., 2008; Pentland, 2008). Recent studies have transitioned toward modeling higher level social constructs like group cohesion via nonverbal cues (Sanchez-Cortes et al., 2012; Lee et al., 2024). For instance, (Vinciarelli et al., 2008) modeled social signaling through prosodic features, while (Jayagopi et al., 2009) utilized visual attention to infer dominance. Recently, (Ouyang et al., 2025) proposed a multi-speaker attention alignment framework to characterize social targets in group-level rapport. However, these approaches mostly focus on direct behavioral signals, without reasoning inherent social dynamics, such as pseudoconsensus, that mask private disagreement under social pressure (Almaatouq et al., 2020; Villa-Cueva et al., 2025). Our benchmark bridges this gap by introducing novel tasks that identify both internal cognitive states and complex social dynamics in meeting scenarios.

Multimodal Social AI Benchmarks. Previous multimodal benchmarks primarily evaluate dyadic social affect recognition (Zadeh et al., 2018; Busso et al., 2008; Poria et al., 2017). Representative examples include Social-IQ (Zadeh et al., 2019), which targets in-the-wild multimodal social question answering, and SIV-Bench (Kong et al., 2025), which focuses on interaction reasoning over social scenes, states, and dynamics. The ToM-oriented resources such as MoMentS (Villa-Cueva et al., 2025) emphasize narrative-centric mental state inference, while MM-Soc (Jin et al., 2024b) targets socially situated online multimodal content. More recent benchmarks introduced in 2025 further push toward evidence-grounded and multi-agent ToM reasoning: Social Genome (Mathur et al., 2025) evaluates grounded social reasoning with human-annotated reasoning traces over multimodal interactions, and MuMA-ToM (Shi et al., 2025) tests embodied multi-agent Theory-of-Mind inference from video and textual contexts. Meeting corpora offer a distinct testbed for multiparty professional collaboration, yet large-scale resources such as MISP 2025 (Gao et al., 2025) are largely limited to audio-visual diarization. In contrast, our benchmark probes the in-depth reasoning ability

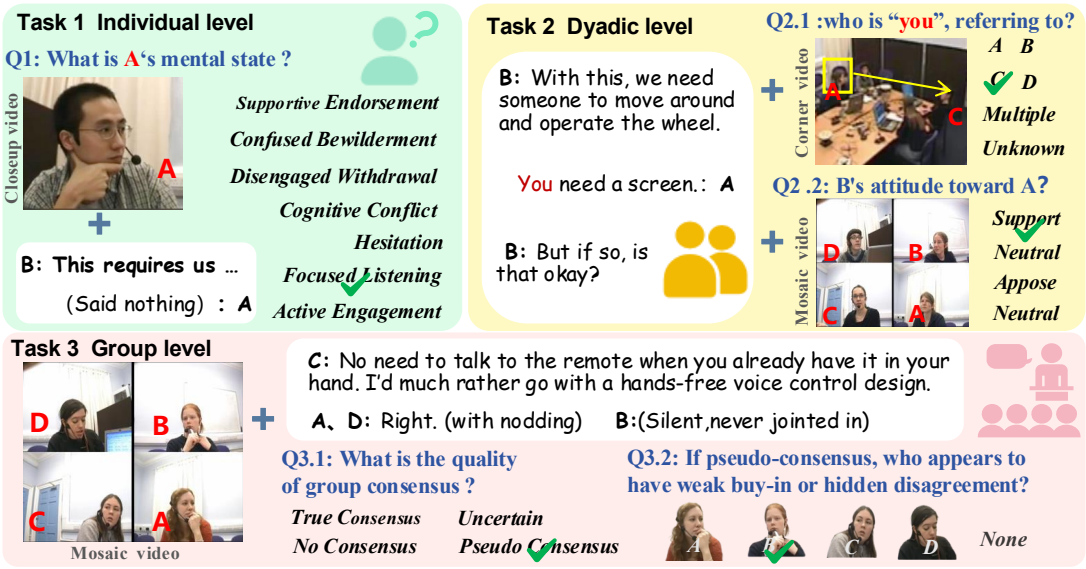


Figure 2: *Overview of the hierarchical tasks in our MeetingToM benchmark.* Individual-level Task assesses mental state prediction in meeting scenarios. Dyadic-level Task evaluates addressee identification and interpersonal attitude inference. Group-level Tasks probes group consensus reasoning, including consensus quality assessment and the identification of hidden disagreement. These three tasks require progressively richer integration of verbal and non-verbal cues for social reasoning in multi-party interactions.

of MLLMs, spanning subject-level internal states, dyadic-level reference understanding, and group-level consensus reasoning.

Theory of Mind in Multiparty Interaction. Theory of Mind (ToM) concerns inferring and tracking latent mental states, including beliefs, intentions, and false beliefs. Early NLP work evaluated ToM mainly via story based QA (e.g., ToM QA Dataset (Nematzadeh et al., 2018)), but later analyses showed that common setups can be addressed by exploiting dataset regularities (Le et al., 2019). Recent LLM benchmarks for computational ToM primarily focus on dialogue-based QA, emphasizing higher-order belief recursion (Wu et al., 2023), broader skill coverage (Chen et al., 2024), and long-form narratives with personality and intention cues (Xu et al., 2024). Broader evaluations with human comparison report notable gaps (Kosinski, 2024; Strachan et al., 2024) and motivate explicit belief tracking and robustness oriented assessment (Sap et al., 2022; Sclar et al., 2023). Beyond text modality, ToM is increasingly tested in video grounded QA (Jin et al., 2024a; Villa-Cueva et al., 2025), interactive dialogue requiring belief updates (Qiu et al., 2024; Yu et al., 2025), and role play or embodied multi-agent settings (Shinoda et al., 2025; Shi et al., 2025). Our work shares the motivation of multimodal ToM modeling, with a specific focus on meeting scenarios, where shared goals and task-driven participant roles provide priors that shape how MLLMs model participants' internal states.

3 MeetingToM Benchmark

3.1 Overview

We introduce **MeetingToM**, a multimodal benchmark designed to evaluate MLLMs' capability to interpret belief states of participants in naturalistic multi-party meeting scenarios. Our benchmark is motivated by three core observations about real-world meetings: (1) participants continuously attribute their own and others' mental states; (2) referential expressions require joint modeling of discourse context, speaker intent, and participant roles for referent disambiguation; and (3) apparent group consensus can exert social pressure that masks individual disagreement. These phenomena demand integrated reasoning over visual cues such as facial expressions, gestures, and posture, as well as verbal content such as transcripts. Concretely, MeetingToM comprises three hierarchically structured tasks with progressively increasing Theory of Mind (ToM) reasoning depth:

- Task 1 targets *subject-level* mental state identification with multimodal signals.
- Task 2 requires *dyadic-level* understanding about referential intent and interpersonal attitudes.
- Task 3 demands *group-level* reasoning about consensus and pseudoconsensus.

State	Closeup View	Brief Description	Non-verbal clues	Utterances
Active Engagement		*Actively driving the task , frequent initiative in speaking, and clear cognitive investment in the conversation.*	• Expressive facial expression	"We are thinking along those lines, just a different shape."
Supportive Endorsement		*Signalling agreement of others' ideas through brief affirmations and light, alignment-oriented elaborations.*	• Repeated nodding • Approving expression	" yep, yep..."
Focused Listening		*Closely attending to others' talk, while selectively processing and mentally integrating the information.*	• Focused expression • Gaze on the speaker	No verbal signals
Cognitive Conflict		*Challenging proposals that clash with one's knowledge, values, or goals, while remaining oriented to the task.*	• Rapid gaze shifts • Pointing at others	"Is it everybody is going to evaluate , or just the Market?"
Confused Bewilderment		*Experiencing cognitive disequilibrium due to overload or ambiguity, leading to partial loss of understanding.*	• Blank expression • Furrowed brows	No verbal signals
Hesitation		*A shortlived state between understanding and action, marked by disfluency or gaze aversion before speaking.*	• Hesitant expression • Scratching his head	"Probably...required to um... mo...most of the displays you see..."
Disengaged Withdrawal		*Pulling back from the meeting, with face compliance, behavioural signs of impatience or restlessness.*	• Distracted Focus • Frozen posture	No verbal signals

Figure 3: *Illustration of participant cognitive state definitions in meeting scenarios.* For each state, we show a representative close-up view, a brief semantic description, characteristic non-verbal cues (e.g., facial expressions, gaze, and gestures), and example utterances when present. These examples highlight how fine-grained mental states manifest through both verbal and non-verbal cues.

In the following sections, we first formally define the MeetingToM benchmark tasks and then detail the data sources and annotation process.

3.2 Task Definition

Input Notations. We formalize MeetingToM as a multiple-choice visual question answering benchmark that requires models to select the correct option based on video inputs captured from specific perspectives and accompanying conversational utterances. We denote conversation utterances as $\mathcal{U}$ and the video segment from view p as $\mathcal{V}^p$, where $p \in \{g, A, B, C, D\}$, where g denotes the global view and A–D denote the close-up views of the four participants, respectively. We also define a mosaic view $\mathcal{V}^m$ by spatially concatenating the synchronized closeup views $\{\mathcal{V}^A, \mathcal{V}^B, \mathcal{V}^C, \mathcal{V}^D\}$ into a 2×2 four-panel video.

Task 1: Subject-level Mental State Prediction. This task evaluates whether models can recognize each participant's mental state within a short temporal window, given the corresponding video segment and utterances. Building on prior work in psychology and cognitive science (Apperly, 2012), we define the following meeting mental states as annotation labels:

- ACTIVE ENGAGEMENT: Actively leading or driving the discussion (e.g., Kahn, 1990)
- SUPPORTIVE ENDORSEMENT: Showing agreement or validation (e.g., Edmondson, 1999)
- FOCUSED LISTENING: Attentive but neutral reception (e.g., Broadbent, 1958)
- COGNITIVE CONFLICT: Experiencing disagreement or skepticism (e.g., De Dreu, 2006)
- CONFUSED BEWILDERMENT: Exhibiting confusion or uncertainty (e.g., Sweller, 1988)
- HESITATION: Pausing before responding or acting (e.g., Clark and Fox Tree, 2002)
- DISENGAGED WITHDRAWAL: Showing low engagement or attention drift (e.g., Kahn, 1990)

Fig. 3 presents a visual illustration of these mental states, and detailed definitions are included in the Appendix A.

Given a short close-up video clip of the target subject A, denoted by $\mathcal{V}^A$, and the utterances $\mathcal{U}$, the model predicts A's mental state:

$$\hat{s}_A = \arg\max_{s \in \mathcal{S}} p\big(s \mid \mathcal{V}^{\mathrm{A}}, \mathcal{U}, \mathcal{Q}^{\mathrm{ms}}\big)\,, \qquad (1)$$

where $\mathcal{S}$ is the set of aforementioned mental state, and Q^{ms} is the mental state probing query. This task requires models to integrate verbal and nonverbal cues such as speech, facial expressions, and gaze for mental state reasoning.

Task2: Dyadic-Level Referential Reasoning. This task considers two aspects of pair-wise social behavior reasoning based on specified visual views. When a speaker A uses the pronoun "you", the model performs **(i) Addressee Identification** to determine who is being addressed and **(ii) Attitude**

Inference to assess the attitude the addressee holds toward A's statement.

Given the global view $\mathcal{V}^{\text{g}}$ and surrounding utterances $\mathcal{U}$, and Addressee Identification query $\mathcal{Q}^{\text{addr}}$, this task is defined as

$$\hat{r} = \arg\max_{r \in \mathcal{R}} p\big(r \mid \mathcal{V}^{\text{g}}, \mathcal{U}, \mathcal{Q}^{\text{addr}}\big)\,, \qquad (2)$$

where $\mathcal{R}$ is the addressee label space including: A, B, C, D, MULTIPLE, and UNKNOWN.

Correspondingly, the Attitude Inference task is defined as follows:

$$\hat{m}_{\hat{r} \to A} = \arg\max_{m \in \mathcal{M}} p\big(m \mid \mathcal{V}^{\text{m}}, \mathcal{U}, \mathcal{Q}^{\text{att}}, \hat{r}\big)\,. \qquad (3)$$

where $\mathcal{M}$ is the attitude label space with four classes: SUPPORT, OPPOSE, NEUTRAL, and UNCERTAIN. $\hat{r} \to A$ refers to the directional relationship relationship from speaker A to addressee, while $\mathcal{Q}_{\text{att}}$ denotes the attitude inference prompt.

These two tasks is particularly challenging. The Addressee Identification task requires models to reason about gaze direction and body orientation grounded in the specified pronoun and aligned utterances. The Attitude Inference task requires models to infer genuine sentiment from verbal signals and nonverbal cues across stitched views.

Task 3: Group-Level consensus Reasoning. This task examines whether MLLMs can identify genuine group agreement or merely surface-level consensus that masks hidden dissent. Specifically, at the group decision-making moment $t^{\star}$,the model performs **(i) Consensus Classification** to assess the quality of agreement and **(ii) Dissent Identification** to identify the dissenting participant when pseudo-consensus is detected, based on the mosaic view $\mathcal{V}^{\text{m}}$ and the utterances $\mathcal{U}$. Consensus Classification is defined as

$$\hat{c} = \arg\max_{c \in \mathcal{C}} p(c \mid \mathcal{V}^{\text{m}}, \mathcal{U}, \mathcal{Q}^{\text{cons}})\,, \qquad (4)$$

where $\mathcal{C}$ is the consensus label space with four classes: TC (True Consensus), PC (Pseudo Consensus), NC (No Consensus), and U (Uncertain).

If $\hat{c} =$ PC, the model identifies the hidden dissenter with dissent identification query $\mathcal{Q}^{\text{diss}}$:

$$\hat{n} = \arg\max_{n \in \mathcal{N}} p\big(n \mid \mathcal{V}^{\text{m}}, \mathcal{U}, \mathcal{Q}^{\text{diss}}\big)\,, \qquad (5)$$

where $\mathcal{N}$ denotes the set of participants, with NONE indicating no dissenter is presented.

This task probes into an interesting social phenomenon: in professional settings, social pressure often leads individuals to verbally agree while their nonverbal behavior reveals reservation or disagreement. Detecting such pseudo-consensus requires identifying subtle incongruence between verbal affirmation and dissent visual cues such as crossed arms, delayed responses, or gaze aversion.

3.3 Data and Annotation

Source Data. MeetingToM is constructed from the AMI Meeting Corpus (McCowan et al., 2005), which contains multi-party meetings recorded with synchronized multi-view video streams (gloabl view and individual closeup view), word-level transcriptions with timestamps, and rich metadata. We focus on the scenario-based meetings, where four participants engage in product design discussions, providing natural opportunities for agreement, disagreement, and consensus-building.

Video Clip Extraction. We extract video clips centered around task-relevant events. For Task 1, we extract 5-second segments focusing on individual participants during moments of social interaction. For Tasks 2 and 3, we around 50 seconds to capture sufficient context for referential resolution and consensus dynamics.

Annotation Pipeline. We adopt a hybrid annotation approach combining automated question generation with human answer annotation:

1. **Candidate Identification**: For Task 1, we sample temporal windows during active social interaction; for Task 2, we extract utterances containing second-person pronouns; for Task 3, we identify decision points and agreement markers based on dialogue act patterns.
2. **Question Generation**: GPT-5 generates task-specific questions based on video content and meeting context, including target participant selection (Task 1), addressee candidates (Task 2), and consensus trigger moments (Task 3). We additionally leverage the original AMI annotations as auxiliary supervision and metadata when constructing and validating these task instances.
3. **Human Annotation**: Two trained annotators independently watch each video clip with transcript access and provide ground-truth answers. Annotators received detailed guidelines with definitions and examples for all label categories.
4. **Disagreement Resolution**: For instances where annotators disagreed, a third annotator adjudicated to produce final gold labels.

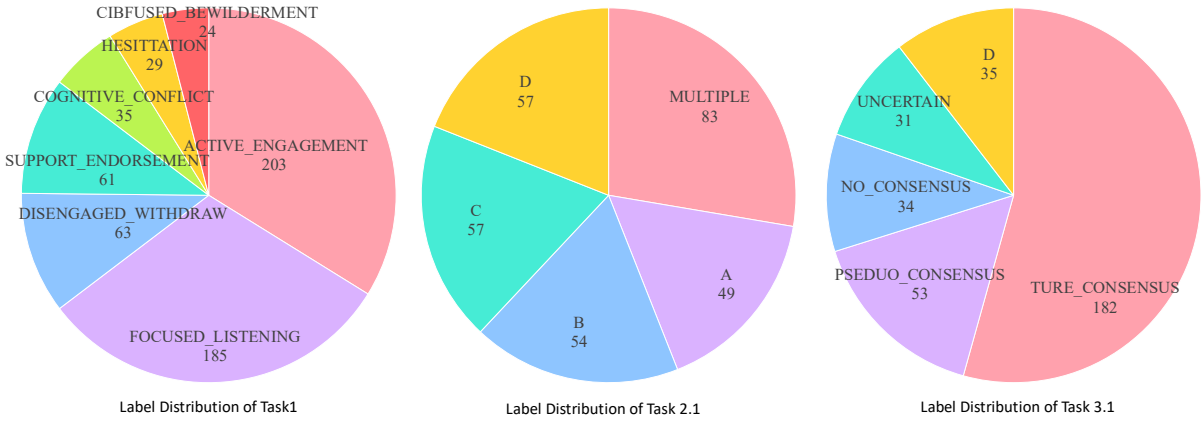


Figure 4: *Per-task label distribution*. We present the label distributions for representative tasks (1, 2.1, and 3.1).

Annotation Quality. We measure inter-annotator agreement using the nominal form of Krippendorff's alpha (Krippendorff, 2018). Task 1 achieves $\kappa = 0.7104$, Task 2 achieves $\kappa = 0.7319$ for addressee and $\kappa = 0.5014$ for attitude, and Task 3 achieves $\kappa = 0.5681$ for consensus type and $\kappa = 0.5729$ for dissenter identification. The moderate-to-substantial agreement indicates reliable annotation quality while reflecting the inherent subjectivity in ToM judgments from observable cues.

Benchmark Statistics. The final benchmark contains 1800 video clips across 60 meeting sessions, composed of 600 for each task, yielding 1800 question-answer pairs in total. Figure 4 summarizes the label distributions for representative Tasks 1, 2.1, and 3.1, revealing both class imbalance and task-specific characteristics. For Task 1,the mental state inference, labels are skewed toward Active Engagement and Focused Listening, indicating that the dataset predominantly captures cooperative and attentive social behaviors, while cognitively demanding or ambiguous states such as Hesitation and Confused Bewilderment are comparatively rare. As expected, the addresses identification in Task 2 exhibits a more balanced distribution across options A–D and Multiple. In contrast, group-level consensus quality classification is dominated by True Consensus, reflecting the inherent sparsity and subtlety of dissent-related social phenomena. Overall, these statistics highlight the varying degrees of imbalance across tasks and underscore the challenging nature of our proposed benchmark

4 Experimental Results

4.1 Experimental Setup

Evaluated Models. To conduct a thorough analysis of the MeetingToM benchmark, we evaluate a diverse range of open-source and proprietary MLLMs, including GPT-4o (OpenAI, 2024), GPT-5 (OpenAI, 2025), Gemini-3 Pro (Pichai et al., 2025), and Qwen-2.5/3 (Bai et al., 2025; Qwen-Team, 2025) series with varying model sizes.

Model	Task1	Task2.1	Task2.2	Task3.1	Task3.2
chance-level	14.29	16.67	25.00	25.00	
Proprietary models					
Gemini-3 Pro	61.00	66.00	47.50	42.50	36.00
GPT-4o	37.00	26.63	46.73	26.53	13.78
GPT-5	56.00	53.00	46.50	41.50	36.50
Open-sourced models					
Qwen2.5-vl-32B	54.50	29.50	45.50	23.50	12.50
Qwen3-vl-7B	44.50	37.00	46.50	27.50	17.00
Qwen3-vl-32B	40.00	57.00	46.50	34.00	24.00

Table 1: *Overall performance of prevailing MLLMs on our MeetingToM benchmark.* Proprietary models consistently outperform open-source models developed during the same time period.

Evaluation Metrics. We evaluate all tasks using exact-match accuracy. We also report a by-chance baseline for each sub-question, defined as the expected accuracy of uniform random guessing over the label set (i.e., $1/K$ for a K-way classification). We do not report a by-chance baseline for Task 3.2 since its scoring is conditioned on answering Task 3.1 correctly.

4.2 Main Results

Overall Performance. The performance of both proprietary and open-source models is reported in Tab. 1. Not surprisingly, the latest Gemini-3-Pro model achieves the best performance across almost all tasks, and models with more parameters generally yield better performance. Another interesting observation is that moving from GPT-4o

Methods	Task1	Task2.1	Task2.2	Task3.1	Task3.2
Video	49.00	28.00	44.50	22.50	10.00
Utter.	26.50	45.23	49.75	26.50	14.00
Video & Utter.	40.00	57.00	46.50	34.00	24.00

Table 2: *Ablation of input modalities using Qwen3-VL-32B*. Both verbal cues from video frames and non-verbal cues from conversational utterances are critical for social behavior reasoning.

to GPT-5 yields substantial performance improvements across all tasks, for example a 29% gain on the mental state prediction task. In contrast, improvements on standard VQA benchmarks are relatively modest, suggesting a potential shift in model development emphasis toward socially aware reasoning capabilities. A similar trend is observed in the transition from Qwen2.5 to Qwen3, as reflected by their comparative performance on our benchmark. In addition, even though open-source models perform on-par with proprietary models on general benchmarks, they lag behind on our social dynamic reasoning tasks, highlighting the necessity of such benchmarks for designing computational social AI systems.

Cross-Task Discussion. Consistent with the hierarchical design of our benchmark, which progressively demands stronger reasoning capabilities, subject-level mental state reasoning is comparatively easier than dyadic-level referential understanding and group-level consensus reasoning. Note that dissent identification (Task 3.2) shows significantly lower performance, as it is evaluated conditionally on the model correctly predicting the consensus classification (Task 3.1).

4.3 Additional Analysis

Ablation on Input Modalities. We conduct ablation studies on input modalities with Qwen3-vl-32B and report the results in Tab. 2. For most tasks, combining visual inputs with conversational utterances yields better performance than using either modality alone. Interestingly, for mental state inference, the video modality alone achieves the best performance, while incorporating conversational utterances degrades performance by 9%. This may be because interjections, such as "hmm" or "emm" can indicate confusion, hesitation, or cognitive conflict, and thereby diluting salient visual signals. For addressee attitude inference in Task 2.2, the joint modality setting underperforms the utterance-only setting but 3.25%. We speculate that this is because MLLMs fail to ground the utterances to the specified target in the stitched visual views.

Benefits from Thinking Prompts. Thinking prompts have been shown to be effective in improving VQA performance. We further analyze how such prompting strategies affect model performance on our benchmark. Specically, we consider the Chain-of-Thought (Wei et al., 2022) and Emotional Chain-of-Thought (ECoT)

The experimental results are summarized in Tab. 3. Notably, both prompting strategies lead to notable performance improvements. Especially for mental state inference in Task 1, CoT and ECoT lead to performance improvements of 19% and 21%, respectively. For Gemini, CoT and ECoT yield the largest gains at the group level, with 14.5% and 9.5% improvements in consensus classification and 19% and 14.5% improvements in dissent identification, respectively.

We further provide visual examples in Fig. 5 demonstrating how EoT and CoT influence the model's predictions. As shown in the left example, EoT prompts the model to focus on expression-related visual cues, resulting in more accurate predictions. In contrast, in the right example, EoT over-emphasizes facial expressions and overlooks other relevant signals, resulting in confusion between hesitation and focused listening, while standard CoT yields a more accurate prediction.

Ablations on Contextual Priors. To examine how much MeetingToM reasoning depends on contextual priors beyond raw visual observations, we conduct an ablation study on Qwen3-VL-32B by seperately adding three sources of structured context: role, phase, and dialogue act (DA). Role specifies each participant's functional position in the meeting (e.g., Project Manager, Marketing Expert, User Interface Designer), providing expectations about responsibilities, goals, and typical behaviors, and thus serves as a prior about what a participant is likely to say or do rather than treating speakers as interchangeable. Phase indicates the coarse topic or stage of the meeting discussion, supplying high-level situational context about what is being negotiated at that moment. DA annotates the communicative function of each utterance, offering a lightweight discourse structure that tests whether the model can leverage interaction patterns instead of relying on surface text alone. We compare the baseline setting (no additional priors) against vari-

Model	Prompt	Task1	Task2.1	Task2.2	Task3.1	Task3.2
Gemini-3 Pro	No Prompt	61.00	66.00	47.50	42.50	36.00
	CoT	66.00 ↑5.00	63.00 ↓3.00	54.00 ↑6.50	57.00 ↑14.50	55.00 ↑19.00
	ECoT	67.50 ↑6.50	66.50 ↑0.50	46.50 ↓1.00	52.00 ↑9.50	50.50 ↑14.50
Qwen3-vl-32B	No Prompt	40.00	57.00	46.50	34.00	24.00
	CoT	59.00 ↑19.00	55.50 ↓1.50	48.50 ↑2.00	26.63 ↓7.37	15.58 ↓8.42
	ECoT	61.00 ↑21.00	51.50 ↓5.50	45.50 ↓1.00	37.00 ↑3.00	30.00 ↑6.00

Table 3: Performance comparison across prompting strategies. Colored arrows indicate absolute performance change relative to the corresponding No Prompt baseline (↑ improvement, ↓ degradation).

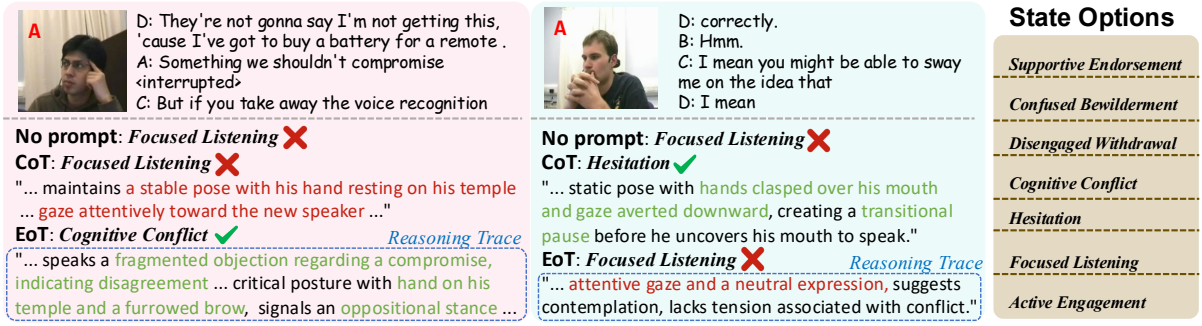


Figure 5: *Qualitative comparison of different prompting strategies for subject-level mental state prediction.* We present representative cases in which EoT or CoT leads to more accurate predictions, respectively. The displayed reasoning traces explain how these prompting strategies improve social behavior understanding.

Methods	Task2.1	Task2.2	Task3.1	Task3.2
Qwen3-vl-32B	57.00	46.50	34.00	24.00
Base Model w. DA	32.62	42.78	29.65	20.60
Base Model w. phase	54.27	46.23	26.63	16.58
Base Model w. role	49.00	50.00	29.50	19.00

Table 4: Prior ablation results on Qwen3-VL-32B. Task 1 is not evaluated in this setting (n=0).

ants that add each prior individually, and report the resulting changes in performance across tasks. This ablation isolates the extent to which structured contextual signals help the model form more accurate belief-state inferences, and whether improvements arise from better social reasoning or from stronger contextual shortcuts.

As shown in Table 4, adding these contextual priors does not yield consistent gains for Qwen3-VL-32B. In fact, performance is largely unchanged or even decreases across several subtasks: incorporating DA leads to a substantial drop on Task2.1 and modest degradations on Task3, while phase and role provide at best marginal improvements on isolated metrics (e.g., a small gain from role on Task2.2) but hurt group-level reasoning. Overall, these results suggest that for Qwen3-VL-32B, the short meeting clips already contain sufficient visual evidence for referential and consensus-related judgments, and simple structured priors do not reliably add useful signal. One plausible explanation is that these priors are either redundant with what the model can infer from the video and transcript, or they introduce additional noise and encourage shortcut reasoning that does not align with the benchmark's multimodal cues.

5 Conclusion

We present MeetingToM, a multimodal benchmark for evaluating theory of mind capabilities in naturalistic meeting scenarios. Our benchmark comprises three task families: mental state recognition from short video segments, dyadic social reasoning about referential intent and interpersonal attitudes, and group-level consensus assessment. Experiments with state-of-the-art MLLMs reveal substantial room for improvement, particularly in detecting subtle nonverbal cues and resolving conflicts between verbal and visual signals. We investigate the benefits of chain-of-thought–style prompting and analyze the pitfall of injecting contextual priors. We hope MeetingToM will catalyze progress toward machines that can genuinely understand the social dynamics of human interaction.

6 Limitations

First, MeetingToM currently draws from English-language meetings, and thus does not measure MLLM's performance in multilingual settings. We plan to expand the sources to multilingual meetings in future versions. Second, constrained by current MLLM's video context limits and the difficulty of annotating long-form interactions, we evaluate models on 5s / 50s key meeting segments rather than full-length meetings (e.g., 30 minutes). Extending this benchmark to longer durations is an important direction as models and tooling improve. Third, while we carefully annotate the data in a systematic and consistent manner, we cannot re-contact the original meeting participants to obtain self-reported mental states, so some labels may still differ from the participants' true private beliefs or intentions at the time, and we will explore improved data collection protocols to better approximate participant-grounded signals in future work.

7 Ethical considerations

MeetingToM is built solely from the publicly available AMI Meeting Corpus, following its documented access and distribution procedures. We do not collect new recordings and do not attempt to re-identify individuals beyond the anonymization already provided by AMI. Our annotations focus on short-lived, episode-bound mental states that are grounded in observable verbal and non-verbal cues within each segment. We explicitly avoid labeling stable personal traits or sensitive attributes. Since private mental states are not directly observable and we cannot obtain participant self-reports, the labels should be interpreted as systematic third-person inferences from available evidence rather than definitive ground truth about participants' internal experiences.

References

Joseph A. Allen and Nale Lehmann-Willenbrock. 2023. The key features of workplace meetings: Conceptualizing the why, how, and what of meetings at work. *Organizational Psychology Review*, 13(4):355–378.

Joseph A. Allen, Nale Lehmann-Willenbrock, and Steven G. Rogelberg, editors. 2015. *The Cambridge Handbook of Meeting Science*. Cambridge University Press, Cambridge.

Abdullah Almaatouq, Alejandro Noriega-Campero, Abdulrahman Alotaibi, P. M. Krafft, Mehdi Moussaid, and Alex Pentland. 2020. Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences of the United States of America*, 117(21):11379–11386.

Ian A. Apperly. 2012. What is "theory of mind"? concepts, cognitive processes and individual differences. *The Quarterly Journal of Experimental Psychology*, 65(5):825–839.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Albert Bandura. 2000. Exercise of human agency through collective efficacy. *Current Directions in Psychological Science*, 9(3):75–78.

Matthew M. Botvinick, Todd S. Braver, Deanna M. Barch, Cameron S. Carter, and Jonathan D. Cohen. 2001. Conflict monitoring and cognitive control. *Psychological Review*, 108(3):624–652.

Donald E. Broadbent. 1958. *Perception and Communication*. Pergamon Press, Oxford.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition*, 84(1):73–111.

Carsten K. W. De Dreu. 2006. When too little or too much hurts: Evidence for a curvilinear relationship between task conflict and innovation in teams. *Journal of Management*, 32(1):83–107.

Edward L. Deci and Richard M. Ryan. 2000. The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4):227–268.

Sidney D'Mello and Arthur Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157.

Amy C. Edmondson. 1999. Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44(2):350–383.

Jean E. Fox Tree and Herbert H. Clark. 1997. Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, 62(2):151–167.

Chris D. Frith and Uta Frith. 2006. The neural basis of mentalizing. *Neuron*, 50(4):531–534.

Ming Gao, Shilong Wu, Hang Chen, Jun Du, Chin-Hui Lee, Shinji Watanabe, Jingdong Chen, Sabato Marco Siniscalchi, and Odette Scharenborg. 2025. The multimodal information based speech processing (misp) 2025 challenge: Audio-visual diarization and recognition. In *Interspeech 2025*.

Irving L Janis. 2008. Groupthink. *IEEE Engineering Management Review*, 36(1):36.

Dinesh Babu Jayagopi, Hayley Hung, Chuohao Yeo, and Daniel Gatica-Perez. 2009. Modeling dominance in group conversations using nonverbal cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):501–513.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer D. Ullman, Antonio Torralba, Joshua B. Tenenbaum, and Tianmin Shu. 2024a. MMToM-QA: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16077–16102, Bangkok, Thailand. Association for Computational Linguistics.

Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, and Srijan Kumar. 2024b. Mm-soc: Benchmarking multimodal large language models in social media platforms. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6192–6210, Bangkok, Thailand. Association for Computational Linguistics.

William A. Kahn. 1990. Psychological conditions of personal engagement and disengagement at work. *Academy of Management Journal*, 33(4):692–724.

Fanqi Kong, Weiqin Zu, Xinyu Chen, Yaodong Yang, Song-Chun Zhu, and Xue Feng. 2025. SIV-Bench: A video benchmark for social interaction understanding and reasoning.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Richard S. Lazarus. 1991. *Emotion and Adaptation*. Oxford University Press, New York, NY.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Sangmin Lee, Minzhi Li, Bolin Lai, Wenqi Jia, Fiona Ryan, Xu Cao, Ozgur Kara, Bikram Boote, Weiyan Shi, Diyi Yang, and James M. Rehg. 2024. Towards social ai: A survey on understanding social interactions. *Preprint*, arXiv:2409.15316.

Nale Lehmann-Willenbrock and Joseph A. Allen. 2018. Modeling temporal interaction dynamics in organizational settings. *Journal of Business and Psychology*, 33(3):325–344.

Christina Maslach and Michael P. Leiter. 1997. *The Truth About Burnout: How Organizations Cause Personal Stress and What to Do About It*. Jossey-Bass, San Francisco, CA.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. 2025. Social genome: Grounded social reasoning abilities of multimodal models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24879–24902, Suzhou, China. Association for Computational Linguistics.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, Sebastien Bourban, Michael Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The ami meeting corpus. In *Proceedings of Measuring Behavior 2005: 5th International Conference on Methods and Techniques in Behavioral Research*, pages 137–140, Wageningen, The Netherlands. Noldus Information Technology. Symposium: Annotating and Measuring Meeting Behavior.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L. Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2392–2400.

OpenAI. 2024. Hello gpt-4o. OpenAI (product announcement). May 13, 2024.

OpenAI. 2025. Introducing gpt-5. OpenAI (release announcement). August 7, 2025.

Liangyang Ouyang, Yifei Huang, Mingfang Zhang, Caixin Kang, Ryosuke Furuta, and Yoichi Sato. 2025. Multi-speaker attention alignment for multimodal social interaction. *Preprint*, arXiv:2511.17952.

Alex Pentland. 2008. *Honest Signals: How They Shape Our World*. MIT Press.

Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. 2025. A new era of intelligence with gemini 3. Google Blog (Products / Gemini). Nov 18, 2025 (The Keyword).

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Shuwen Qiu, Mingdian Liu, Hengli Li, Song-Chun Zhu, and Zilong Zheng. 2024. MindDial: Enhancing conversational agents with theory-of-mind for common ground alignment and negotiation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 746–759, Kyoto, Japan. Association for Computational Linguistics.

Qwen-Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Steven G. Rogelberg. 2019. *The Surprising Science of Meetings: How You Can Lead Your Team to Peak Performance*. Oxford University Press, New York, NY.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Dairazalia Sanchez-Cortes, Oya Aran, Amanzadeh Mastaneh, Shidhore Luxman, and Daniel Gatica-Perez. 2012. Inferring emergent leadership in small groups using nonverbal cues. *IEEE Transactions on Multimedia*, 14(2):354–367.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Klaus R. Scherer. 2001. Appraisal considered as a process of multilevel sequential checking. In Klaus R. Scherer, Angela Schorr, and Tom Johnstone, editors, *Appraisal Processes in Emotion: Theory, Methods, Research*, pages 92–120. Oxford University Press, Oxford.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980, Toronto, Canada. Association for Computational Linguistics.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2025. Muma-tom: Multi-modal multi-agent theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1510–1519.

Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. 2025. Tomato: Verbalizing the mental states of role-playing llms for benchmarking theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Also available as arXiv:2501.08838.

Marie Luise Speiger, Katrin Rothmaler, Ulf Liszkowski, Hannes Rakoczy, and Charlotte Grosse Wiesmann. 2025. Evidence that altercentric biases in a continuous false belief task depend on highlighting the agent's belief. *Cognition*, 256:106055.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Karie Tennant, Ashley Long, and Tammy J. Toney-Butler. 2023. Active listening. In *StatPearls*. StatPearls Publishing, Treasure Island, FL. [Updated 2023 Sep 13].

Kenneth W. Thomas. 1992. Conflict and negotiation processes in organizations. In Marvin D. Dunnette and Leaetta M. Hough, editors, *Handbook of Industrial and Organizational Psychology*, 2 edition, volume 3, pages 651–717. Consulting Psychologists Press, Palo Alto, CA.

Emilio Villa-Cueva, S. M. Masrur Ahmed, Rendi Chevi, Jan Christian Blaise Cruz, Kareem Elzeky, Fermin Cristobal, Alham Fikri Aji, Skyler Wang, Rada Mihalcea, and Thamar Solorio. 2025. Moments: A comprehensive multimodal benchmark for theory of mind. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22591–22611.

Alessandro Vinciarelli, Maja Pantic, Hervé Bourlard, and Alex Pentland. 2008. Social signals, their function, and automatic analysis: A survey. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, ICMI '08, pages 61–68, New York, NY, USA. Association for Computing Machinery.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Howard M. Weiss and Russell Cropanzano. 1996. Affective events theory: A theoretical discussion of the structure, causes, and consequences of affective experiences at work. In Barry M. Staw and L. L.

Cummings, editors, *Research in Organizational Behavior*, volume 18, pages 1–74. Elsevier Science/JAI Press. Available from MIT OpenCourseWare reading archive.

Christopher D. Wickens, William S. Helton, Justin G. Hollands, and Simon Banbury. 2021. *Engineering Psychology and Human Performance*, 5 edition. Routledge, New York, NY.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand. Association for Computational Linguistics.

Fangxu Yu, Lai Jiang, Shenyi Huang, Zhen Wu, and Xinyu Dai. 2025. Persuasivetom: A benchmark for evaluating machine theory of mind in persuasive dialogues. *Preprint*, arXiv:2502.21017.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-IQ: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8807–8817.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-modal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2236–2246.

A Conceptual Background for Meeting Mental States

This section summarizes the conceptual background for the seven meeting-specific mental states used in MeetingToM. The taxonomy is designed to jointly capture affective valence, task engagement, and typical interactional behavior, drawing on work in engagement, cognitive load, affect, and meeting science. Table 5 provides a compact overview.

Engaged Involvement. Engaged Involvement is a dynamic cognitive and affective state in which participants invest substantial attentional and emotional resources in the collaborative process, experiencing the meeting as meaningful and worth their effort. It aligns with psychological engagement (Kahn, 1990), self-determination theory on intrinsic motivation and goal congruence (Deci and Ryan, 2000), and work on psychological safety and collective efficacy in teams (Edmondson, 1999; Bandura, 2000). In meetings, Engaged Involvement typically appears as sustained focus, initiative taking, elaborated contributions, and cooperative problem solving, consistent with human performance models and meeting effectiveness research (Wickens et al., 2021; Allen and Lehmann-Willenbrock, 2023; Allen et al., 2015).

Focused Listening. Focused Listening is a cognitively active but externally restrained state in which participants selectively attend to others' speech, integrate new information, and anticipate intentions without yet taking the floor. It is grounded in selective-attention and cognitive-load accounts of information processing (Broadbent, 1958; Sweller, 1988), as well as work on mentalizing and perspective taking in social cognition (Frith and Frith, 2006). Subtle feedback behaviors such as nodding, gaze alignment, or brief backchannels signal this form of engaged reception in meetings and help maintain task alignment and conversational efficiency (Tennant et al., 2023; Allen and Lehmann-Willenbrock, 2023).

Supportive Endorsement. Supportive Endorsement denotes a positively valenced socio-emotional state of agreement, validation, and affiliative reinforcement toward others' ideas. It arises under conditions of trust, psychological safety, and perceived goal alignment (Edmondson, 1999; Kahn, 1990), and is linked to the broadening effects of positive affect on cooperation (Weiss and Cropanzano, 1996). Cognitively, it reflects social appraisal focused on consistency and feasibility rather than threat (Scherer, 2001); behaviorally, it appears as explicit agreement, praise, and light co-construction that strengthen cohesion and accelerate consensus in meetings (Allen et al., 2015).

Cognitive Conflict. Cognitive Conflict is a negatively valenced yet high-engagement state in which participants actively challenge ideas that contradict their knowledge, values, or goals. Organizational research distinguishes such task conflict from relationship conflict and shows that, when managed well, it can improve decision quality and innovation (Thomas, 1992; De Dreu, 2006). From a cognitive perspective, conflict monitoring and error detection

processes are activated (Botvinick et al., 2001), while appraisal processes evaluate threats to goals and self (Lazarus, 1991). In meetings, Cognitive Conflict manifests as counterarguments, requests for justification, and attempts to correct inconsistencies; under psychologically safe climates it can be productive, but it can also escalate into interpersonal strain if mishandled (Edmondson, 1999; Allen and Lehmann-Willenbrock, 2023).

Confusion & Bewilderment. Confusion & Bewilderment describe a state of cognitive disequilibrium arising when informational demands exceed prior knowledge or when instructions are ambiguous. Cognitive load theory characterizes this overload as a mismatch between task complexity and cognitive resources (Sweller, 1988), while affective-events theory explains the accompanying uncertainty and mild negative affect at work (Weiss and Cropanzano, 1996). Studies of learning show that confusion can be adaptive, triggering metacognitive monitoring and help-seeking when properly scaffolded (D'Mello and Graesser, 2012). In meetings, this state often appears as furrowed brows, delays, or clarification questions, reflecting attentional bottlenecks in processing complex or poorly structured information (Broadbent, 1958; Allen and Lehmann-Willenbrock, 2023).Unlike hesitation, confusion primarily reflects difficulty in understanding the task content rather than uncertainty about expressing one's view.

Hesitation (Transitional). Hesitation (Transitional) is a short-lived cognitive and emotional bridge between uncertainty and overt action, characterized by speech disfluencies, timing delays, and visible self-monitoring before speaking. Psycholinguistic work shows that fillers such as *uh* and *um* signal difficulties in planning and coordination between thought and speech (Clark and Fox Tree, 2002; Fox Tree and Clark, 1997). Appraisal theories interpret hesitation as reflecting real-time checks of adequacy and social-evaluative risk (Scherer, 2001; Lazarus, 1991). In meetings, hesitation often precedes key turns in decision making and is observable in micro-pauses, hedging, and gaze aversion, which interact with temporal interaction patterns over the course of a discussion (Lehmann-Willenbrock and Allen, 2018).Unlike confusion, hesitation can arise even when the content is understood, reflecting uncertainty about the social consequences of speaking up.

Disengaged Withdrawal. Disengaged Withdrawal is a psychologically detached, negatively valenced state in which participants reduce cognitive and emotional investment in the meeting. It builds on engagement and disengagement accounts (Kahn, 1990) and burnout research on chronic strain and cynicism (Maslach and Leiter, 1997). Cognitively, attention drifts from task content toward thoughts about time and external concerns; affectively, boredom and frustration dominate and are located in low-arousal negative regions of the affective circumplex (Russell, 1980). In meeting science, withdrawal is visible as silence, multitasking, or minimal eye contact and is associated with low participation energy and ineffective meeting structures (Allen and Lehmann-Willenbrock, 2023; Allen et al., 2015; Rogelberg, 2019; Weiss and Cropanzano, 1996).

B Mapping Meeting Mental States to AMI Annotations

Meeting mental state	Engagement	Valence	Expression	Key AMI-derived cues
Engaged Involvement	High	Positive	High	Dense *dialogueActs* involving informing, suggesting, and elaborating; frequent nods in *headGesture*; task-oriented *handGesture* such as pointing or writing; *focus* on the current speaker or shared artefacts; forward-leaning patterns in *movement* within the corresponding *segments*.
Focused Listening	High	Positive	Low	Long stretches with few floor-taking *dialogueActs* beyond backchannels; sustained *focus* on the current speaker; subtle nods in *headGesture*; minimal *handGesture* and low-amplitude *movement*, while the content of *words* stays on-topic.
Supportive Endorsement	Medium	Positive	Medium	Short agreement or positive-assessment *dialogueActs*; nodding *headGesture* aligned with others' turns; occasional affiliative *handGesture*; stable *focus* on the proposer during supportive comments; local *segments* that follow another participant's proposal in the same *topic*.
Confusion & Bewilderment	Medium	Negative	Low	Clarification or repair-initiation *dialogueActs*; increased *disfluency* in the *words*; tilts or brief freezes in *headGesture*; unstable *focus* alternating between speaker and materials; often appearing when *topics* shift or new information is introduced.
Hesitation (Transitional)	Medium	Negative	Low	Filled pauses and self-repairs in *disfluency*; delayed response after incoming *dialogueActs*; micro-shifts in *headGesture* and *focus*; small self-adaptive *handGesture* or posture changes within a single *segment*.
Cognitive Conflict	High	Negative	High	Disagreement or rejection *dialogueActs*; overlaps in turn-taking; head shakes in *headGesture*; expansive *handGesture*; rapid *focus* shifts between opponents and shared artefacts around contested *topics*.
Disengaged Withdrawal	Low	Negative	Low	Sparse *dialogueActs*; *focus* on table or objects rather than speakers; minimal task-related *handGesture*; leaning-back or slumped *movement*; off-topic or absent *words*.

Table 5: Meeting mental state taxonomy in MeetingToM. Engagement, valence, and expression describe coarse affective–behavioral profiles; the rightmost column shows how each state is operationalized using existing AMI-style annotation layers.

State prompt

[System]
"You are an evaluation model. Follow the output schema specified in the user prompt strictly.
[Guideline]
You will be given a short meeting clip video showing one target participant in upper body view only, and a transcript for the same time window. The target participant shown in the video is {target_agent}.
Base your decisions primarily on visual behavior in the video. The transcript is optional and may be used only to locate reaction moments and must not override what is visible.

Minimal label meanings:
ACTIVE_ENGAGEMENT means the target is actively contributing or leading, indicated by speaking with directive hand gestures, forward lean, animated expression, or active gaze engagement with others.
SUPPORTIVE_ENDORSEMENT means positive alignment or validation toward someone else's idea, indicated by repeated nodding aligned to others' speech, a warm approving smile, open posture, or quick affirmative reactions after key utterances.
FOCUSED_LISTENING means attentive but neutral receiving with no clear positive or negative stance, indicated by gaze mostly toward the speaker or screen, upright posture, and sparse small nods that are not clearly affirmative.
COGNITIVE_CONFLICT means disagreement rather than confusion, indicated by repeated rejecting movements, tight lips or tense jaw, defensive posture such as arms crossed or leaning back, dismissive gestures, or negative stance.
CONFUSED_BEWILDERMENT means cognitive overload rather than disagreement, indicated by puzzled brow or squint, head tilt, scanning others for clarification, mouth slightly open with a searching look, or self touch such as chin or forehead touch while thinking.
HESITATION means a transitional pause before speaking or acting with tentative turn taking, indicated by delayed response when a turn is available, mouth opens then closes, lip press then release, brief gaze aversion down or away then back, or a half gesture that starts then stops.
DISENGAGED_WITHDRAWAL means low engagement with attention drifting away from the interaction, indicated by gaze away down or to the side, leaning back with long stillness, fidgeting with objects, and a flat checked out affect.

Also provide a rationale (in 2 to 4 sentences) that references only observable cues.
The output format must strictly follow: {{"Q1":"<OPTION>","Q2":"<OPTION>","rationale":"<2-4 sentences>"}}.
[Context]
<target_corner_view.mp4, transcript.txt>

[CoT]
Think step by step to decide the best label.

First assess whether the target is clearly visible, partially visible, poorly visible, or not visible at all. Then identify at least two clear cues across categories such as affect, gaze, posture, gesture, timing, and engagement. If you cannot identify clear cues, treat the evidence as insufficient and choose a conservative label.

Resolve common confusions as follows. Supportive endorsement requires clearly positive signals like a nod smile pattern; otherwise prefer focused listening. Confused bewilderment indicates comprehension difficulty signs; hesitation indicates turn taking pause or abort before speaking. Focused listening implies gaze mostly on the speaker with stable posture; disengaged withdrawal implies sustained gaze away or down with low reactivity.

[EoT]
Follow the Emotional Chain-of-Thought below and think step by step.

Step 1 Understanding context.
Infer what is happening in this moment and the target participant's likely role in the interaction. Use the transcript only to locate timing if needed.

Step 2 Recognizing emotions from visible cues.
From affect, gaze, posture, head movement, engagement level, gesture energy, and response timing, infer the target participant's likely emotional tendency, while tracking uncertainty when cues are weak.

Step 3 Recognizing self emotions.
Notice any immediate interpretive bias or emotional pull you feel while observing, and acknowledge uncertainty internally rather than forcing a confident story.

Step 4 Managing self emotions.
Regulate overconfidence and avoid projecting emotions onto the target. Keep an evidence-first stance and let uncertainty remain when the video does not support a clear inference.

Step 5 Influencing others' emotions and generating the final answer.
Synthesize the emotionally informed visual evidence into the single best option that matches the dominant state across the whole window. If the evidence is weak or mixed, choose the most conservative option and lower confidence.

Consensus prompt

[System]
"You are an evaluation model. Follow the output schema specified in the user prompt strictly.
[Guideline]
You will be given one meeting clip as a 2x2 closeup mosaic video. Both questions refer to the same attached mosaic video. Base your decisions primarily on visual behavior in the video. The transcript is optional and may be used only to locate reaction moments and must not override what is visible.

Minimal label meanings:
TRUE_CONSENSUS: broad, genuine alignment with no notable misalignment signals.
PSEUDO_CONSENSUS: surface agreement, but at least one participant shows persistent subtle misalignment (hesitation/tense/avoidant/skeptical), suggesting compliance rather than true buy-in.
NO_CONSENSUS: clear disagreement or mixed/fragmented signals across participants.
UNCERTAIN: insufficient visual evidence to distinguish the above.

Also provide a rationale (in 2 to 4 sentences) that references only observable cues.
The output format must strictly follow: {{"Q1":"<OPTION>","Q2":"<OPTION>","rationale":"<2-4 sentences>"}}.
[Context]
<mosaic_view.mp4, transcript.txt>

[CoT]
Think step by step to decide Q1 and Q2.

First identify one or two clear reaction moments such as synchronized nods, affect shifts, or clear disagreement signals.

For Q1 choose the option that best matches the visuals. Remember label meanings mentioned above.
For Q2, if Q1 is PSEUDO_CONSENSUS and there is exactly one clearly standout subtle dissenter, then choose that agent. Otherwise if there are multiple candidates or ambiguous(you cannot reliably map from A,B,C and D), then Q2 must be NONE.

[EoT]
Follow the Emotional Chain-of-Thought below and think step by step.

Step 1 Understanding context.
Infer what is being discussed or decided in this moment and when the key reactions occur.

Step 2 Recognizing others' emotions.
From visible cues such as facial affect, gaze, posture, head movement, engagement, and timing, infer the likely emotional tendencies of each visible participant, while tracking uncertainty when cues are weak.

Step 3 Recognizing self emotions.
Notice your own immediate interpretive bias or emotional pull while observing the scene, and explicitly acknowledge uncertainty internally rather than forcing a confident story.

Step 4 Managing self emotions.
Regulate overconfidence and avoid projecting emotions onto participants. Keep a calm, evidence-first stance and let uncertainty remain when the video does not support a clear inference.

Step 5 Influencing others' emotions and generating the final answers.
Synthesize the emotionally informed visual evidence into the best Q1 label describing the group dynamic. For Q2, select a dissenter only if one participant stands out clearly; otherwise NONE.

You prompt

[System]
"You are an evaluation model. Follow the output schema specified in the user prompt strictly.
[Guideline]
You will be given two meeting clips from the same interaction window.
The first video is a corner view, which you must use to answer Q1 about addressee using layout, position, and orientation cues.
The second video is a 2x2 closeup mosaic, which you must use to answer Q2 about attitude using per-person facial and upper-body cues.
Answer primarily from visual evidence from the relevant video. Use the transcript only to help locate timing or turn-taking. If the transcript conflicts with what you see, trust the videos. Also provide a rationale (in 2 to 4 sentences) that references only observable cues, and may specify the originating video source (e.g., Corner or Mosaic).
The output format must strictly follow: {{"Q1":"<OPTION>","Q2":"<OPTION>","rationale":"<2-4 sentences>"}}.
[Context]
<2x2_corner_view.mp4, mosaic_view.mp4, transcript.txt>

[CoT]
Think step by step to decide Q1 and Q2.

For Q1 focus only on the corner view during the addressing moment. Prefer consistent cues such as head and gaze orientation, pointing or directing hand gestures, body orientation and immediate visible reactions right after the addressing moment.

For Q2 focus only on the mosaic view and decide whose attitude is being asked about before judging attitude.
Supportive cues include nodding, smiling, warm approving affect, forward lean, and quick affirmative reactions.
Oppositional cues include head shakes, frowns, brow furrow, lip press, tense jaw, arms crossed, leaning back, dismissive gestures, and clearly negative stance.
Neutral cues include steady attentive gaze, stillness, neutral affect, and stable listening posture.

[EoT]
Follow the Emotional Chain-of-Thought below and think step by step.

Step 1 Understanding context.
Infer what is happening in this interaction window and when the key addressing moment occurs.

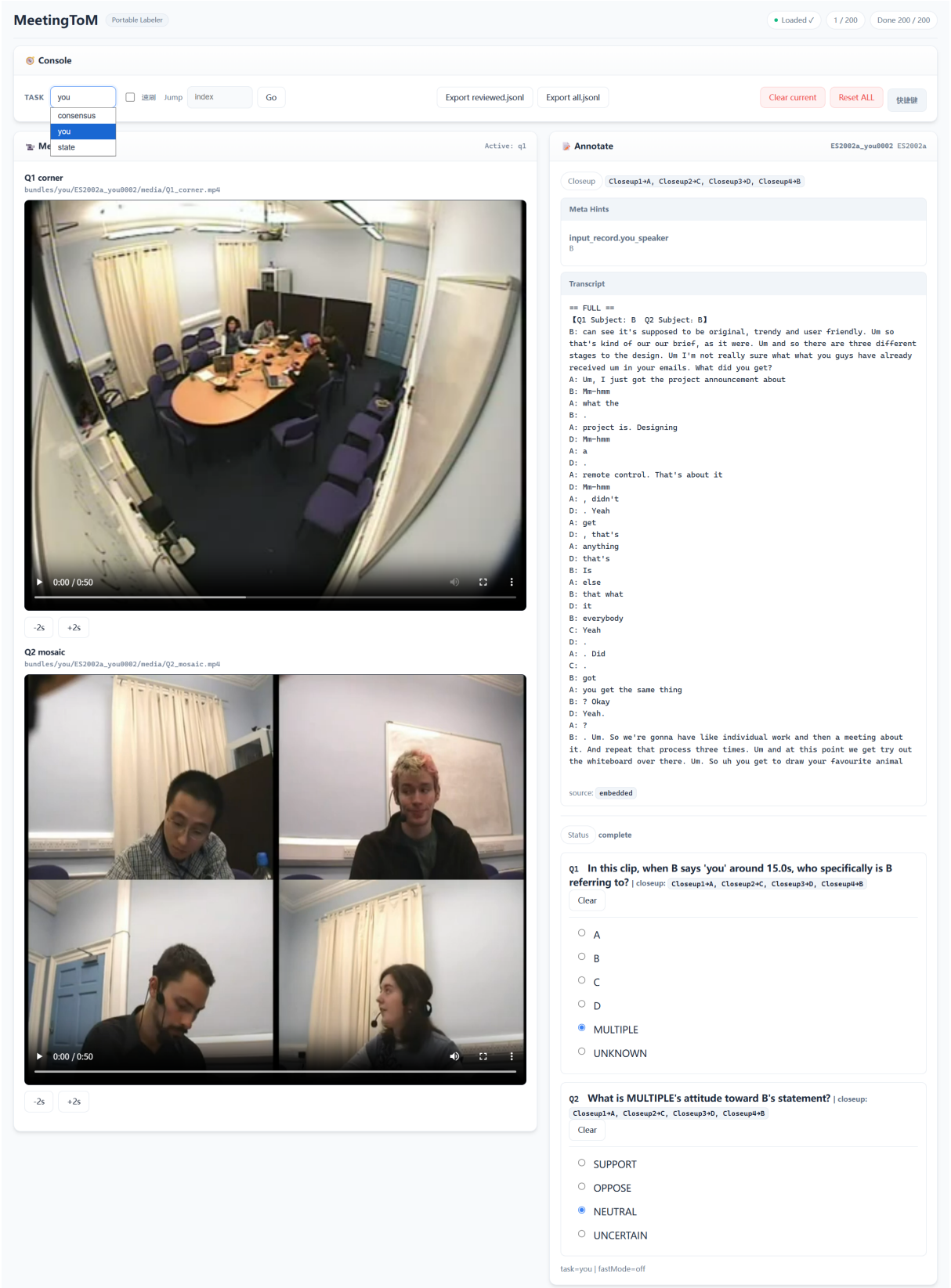
Step 2 Recognizing others' emotions and social signals.
From body orientation, positioning, approach/avoid movements, gaze direction when visible, and turn-taking patterns, infer who is being addressed in the moment, while tracking uncertainty when cues are weak.

Step 3 Recognizing self emotions.
Notice any immediate interpretive bias you feel while watching, and acknowledge uncertainty rather than forcing a confident story.

Step 4 Managing self emotions.
Regulate overconfidence and avoid projecting intentions onto participants. Keep an evidence-first stance and let uncertainty remain when the video does not support a clear inference.

Step 5 Influencing others' emotions and generating the final answers.
For Q1, decide the addressee based on the dominant addressing pattern in the corner view across the whole window.
For Q2, decide the attitude based on the dominant facial and upper-body affect patterns in the mosaic across the whole window.

Figure 6: MeetingToM's CoT EoT Prompt Example


MeetingToM
Portable Labeler
Loaded ✓
1 / 200
Done 200 / 200
Console
TASK
you
consensus
you
state
速刷
Jump
index
Go
Export reviewed.jsonl
Export all.jsonl
Clear current
Reset ALL
快捷键
Active: q1
Q1 corner
bundles/you/ES2002a_you0002/media/Q1_corner.mp4
0:00 / 0:50
-2s
+2s
Q2 mosaic
bundles/you/ES2002a_you0002/media/Q2_mosaic.mp4
0:00 / 0:50
-2s
+2s
Annotate
ES2002a_you0002 ES2002a
Closeup
Closeup1→A, Closeup2→C, Closeup3→D, Closeup4→B
Meta Hints
input_record.you_speaker
B
Transcript
== FULL ==
【Q1 Subject: B Q2 Subject: B】
B: can see it's supposed to be original, trendy and user friendly. Um so that's kind of our our brief, as it were. Um and so there are three different stages to the design. Um I'm not really sure what what you guys have already received um in your emails. What did you get?
A: Um, I just got the project announcement about
B: Mm-hmm
A: what the
B: .
A: project is. Designing
D: Mm-hmm
A: a
D: .
A: remote control. That's about it
D: Mm-hmm
A: , didn't
D: . Yeah
A: get
D: , that's
A: anything
D: that's
B: Is
A: else
B: that what
D: it
B: everybody
C: Yeah
D: .
A: . Did
C: .
B: got
A: you get the same thing
B: ? Okay
D: Yeah.
A: ?
B: . Um. So we're gonna have like individual work and then a meeting about it. And repeat that process three times. Um and at this point we get try out the whiteboard over there. Um. So uh you get to draw your favourite animal
source: embedded
Status complete
Q1 In this clip, when B says 'you' around 15.0s, who specifically is B referring to? | closeup: Closeup1→A, Closeup2→C, Closeup3→D, Closeup4→B
Clear
A
B
C
D
MULTIPLE
UNKNOWN
Q2 What is MULTIPLE's attitude toward B's statement? | closeup: Closeup1→A, Closeup2→C, Closeup3→D, Closeup4→B
Clear
SUPPORT
OPPOSE
NEUTRAL
UNCERTAIN
task=you | fastMode=off
Prev (J)
Active video: q1
Next (K)

Annotation Instructions (MeetingToM)

A. Inputs and Viewing Protocol

Annotators are given a ~50-second meeting clip from AMI with a corner (wide) view and participant close-ups ($A, B, CandD$), plus a transcript with timestamps.

Viewing protocol

- First pass (global context): watch the full clip once in corner view to understand topic flow and interaction structure.
- Second pass (evidence collection): re-watch focusing on the relevant time window (Task-dependent) using the target close-up as primary evidence, and the corner view as secondary evidence for gaze/turn-taking and group alignment.
- No over-interpretation: labels must be justified by observable visual cues; avoid inferring intentions that are not behaviorally supported.

Evidence priority

Visual cues (face, gaze, posture, gesture, timing) > transcript content > inference.
When cues conflict, prioritize stable, repeated, and temporally aligned signals over isolated micro-movements.

B. Task 1 Instructions: Participant Mental-State (Early vs Late)

Goal

For a specified target participant X, assign the participant's dominant mental/interaction state in two windows:

- Early window: 0–5 seconds
- Late window: last 5 seconds ($T - 5$ to T)

Step-by-step

Open the target close-up for participant X; use corner view to verify who is speaking and where attention is directed.
For each window, collect at least two supporting cues (e.g., gaze alignment + posture; facial affect + head movement; gesture + turn-taking role).
Assign one label per window from the Task 1 label set.
If the participant's state changes within a window, label the dominant state; if unclear, label the state most evident at the end of the window.

Key rules

- Label the target participant's state (not the speaker's, and not the group's).
- Use HESITATION only when "about-to-speak but holds back" behaviors dominate (e.g., lean-in then retreat; hand half-raise then drop; pre-speech pauses).
- If visibility is insufficient (face fully occluded/blurry for most of the window), mark as UNCERTAIN (or follow dataset policy for skipping).

C. Task 2 Instructions: "You" Addressee and Addressee Stance

When the speaker uses "you", annotate:

Q1. Addressee Identification

1. Step-by-step:
   Use transcript timestamps to locate the "you" token.
   Inspect the 1–2 seconds around the token in corner view and the speaker close-up.
2. Determine addressee using converging cues:
   Speaker gaze direction
   Pointing/indicating gesture
   Speaker torso orientation
   Which participant reacts (eye contact, nod, posture shift, verbal response)

3. Output one of: A/B/C/D/E, MULTIPLE (plural/group address), or UNKNOWN (insufficient evidence).

Q2. Addressee Stance Toward the Speaker

1. Time window: focus on the addressee's immediate reaction during and right after the "you" usage (typically ±2 seconds).
2. Step-by-step:
   Switch to the addressee close-up and check for stance cues (head movement, facial affect, posture openness/defensiveness, gaze engagement/avoidance).
3. Assign one stance label: SUPPORT / OPPOSE / NEUTRAL / UNCERTAIN.
4. Key rules ("politeness trap"):
   - If verbal content suggests agreement but body language shows resistance (e.g., arms crossed, lean back, stiff face), label based on visual stance.
   - NEUTRAL means attentive but non-committal (not disengaged).
   - Use UNCERTAIN if the addressee is not visible/too blurry or cues are strongly contradictory.

D. Task 3 Instructions: Group Consensus Quality (+ Hidden Dissenter)

Goal: At a decision/convergence moment, label the group consensus quality.

Q1. Consensus Quality

1. Step-by-step:
   Identify the decision/convergence moment (explicit proposal acceptance, "are we agreed?", summarizing decision, or clear topic closure).
   Use corner view to assess group-level alignment (shared attention, synchronized nods, posture openness, affect).
2. Assign one label:
   - TRUE_CONSENSUS: alignment signals across ≥2 participants and no salient dissent cues.
   - PSEUDO_CONSENSUS: surface agreement but one participant shows dissent "leakage" (e.g., stiff face, lip press, minimal/token nod, arms crossed, lean back, gaze away at commitment).
   - NO_CONSENSUS: overt disagreement or fragmented attention (head shakes, negating gestures, side-talk, widespread disengagement).
   - UNCERTAIN: key reactions not observable due to occlusion/quality or no interpretable convergence moment.

Q2. Hidden Dissenter (only if PSEUDO_CONSENSUS)

Locate the participant with the strongest and most temporally aligned dissent leakage at the agreement moment.
Output that participant's ID (A/B/C/D/E). Otherwise output NONE.

E. Quality Control Checklist (Annotator Self-Check)

Before submitting each item, confirm:

- I used the correct time window for the task (Task 1 early/late; Task 2 around "you"; Task 3 at convergence).
- My label is supported by ≥2 visual cues (unless the clip is clearly insufficient, then UNCERTAIN/UNKNOWN).
- I did not use later events in the clip to justify earlier-window labels.
- If cues were weak/contradictory, I chose the conservative option (NEUTRAL/FOCUSED_LISTENING/UNKNOWN/UNCERTAIN per task).

Figure 8: MeetingToM Annotation guideline