MoMentS: A Comprehensive Multimodal Benchmark for Theory of Mind

Emilio Villa-Cueva[1], S M Masrur Ahmed[2], Rendi Chevi[1], Jan Christian Blaise Cruz[1], Kareem Elzeky[1], Fermin Cristobal[1], Alham Fikri Aji[1], Skyler Wang[3], Rada Mihalcea[4], Thamar Solorio[1]

[1]MBZUAI, [2]University of Houston, [3]McGill University, [4]University of Michigan

github.com/villacu/MoMentS

Abstract

Understanding Theory of Mind is essential for building socially intelligent multimodal agents capable of perceiving and interpreting human behavior. We introduce MoMentS (Multimodal Mental States), a comprehensive benchmark designed to assess the ToM capabilities of multimodal large language models (LLMs) through realistic, narrative-rich scenarios presented in short films. MoMentS includes over 2,300 multiple-choice questions spanning seven distinct ToM categories. The benchmark features long video context windows and realistic social interactions that provide deeper insight into characters' mental states. We evaluate several MLLMs and find that although vision generally improves performance, models still struggle to integrate it effectively. For audio, models that process dialogues as audio do not consistently outperform transcript-based inputs. Our findings highlight the need to improve multimodal integration and point to open challenges that must be addressed to advance AI's social understanding.

1 Introduction

Throughout our lives, we continuously generate hypotheses about other people's emotions, knowledge, and a range of other mental states; these hypotheses guide how we understand and interact with others. This ability, known as Theory of Mind (ToM) (Premack and Woodruff, 1978), is essential for interpreting behavior at the individual level and fundamental to coherent human social interaction (Byom and Mutlu, 2013).

Humans rely on more than just language to express their mental states. Gaze, facial expressions, body posture, gestures, and vocal cues all play an important role in communicating how we feel and what we think. This combination of verbal and non-verbal cues provides relevant multimodal information to infer mental states of others (Byom and Mutlu, 2013; Bayliss and Tipper, 2006; De Sonneville et al., 2002).

For artificial agents, this information can serve as multimodal input that enhances socially intel-

Figure 1: Overview of MoMentS questions.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22591–22611
November 4-9, 2025

ligent behavior, empowering users across a wide range of applications: from facilitating communication and enhancing collaboration to offering companionship. A robust ToM enables such systems to anticipate intentions, understand desires and emotions, and detect knowledge gaps, to adapt their behavior to support users more effectively (Oguntola et al., 2021). Importantly, this requires not only inferring individual mental states, but doing so in context—accurately "reading the room" by processing these signals to interpret human behavior in socially situated settings (Williams et al., 2022).

Most existing benchmarks proposed to measure ToM in artificial agents predominantly center around belief-tracking tasks within text-based narratives or simplified multimodal settings (Chen et al., 2025a). While these approaches evaluate models' ability to reason about *who knows or believes what*, they frequently neglect the interplay of emotions, intentions, pragmatic communication, and social contexts that characterize genuine human interactions. Consequently, a clear gap exists between existing evaluations and the richer, socially grounded reasoning required in realistic scenarios.

To support the development of socially intelligent multimodal agents and assess current models' ToM in realistic, socially grounded scenarios, we introduce **MOMENTS (Multimodal Mental States)**, a comprehensive multimodal video question-answering benchmark designed to evaluate ToM across seven abilities derived from the ATOMS taxonomy (Beaudoin et al., 2020): Intentions, Desires, Beliefs, Knowledge, Percepts, Non-literal Communication, and Emotions. The dataset comprises 2,335 *human-annotated* questions and 9,340 candidate answers sourced from 168 long-form videos, annotated with short and long context windows, multimodal cue markers, and adversarially-generated distractors to minimize biases.

To the best of our knowledge, MOMENTS is the first benchmark to evaluate multimodal ToM in real-world videos with human actors, framing it explicitly as a socially situated ability. Our contributions are as follows:

- **MOMENTS**: A novel multimodal benchmark with over 2,300 questions from real-world, long-form video data, explicitly structured to assess diverse ToM abilities.
- An **LLM-in-the-loop annotation framework** designed to produce challenging distractors and mitigate bias in answer sets.
- A baseline evaluation of multimodal LLMs, highlighting that although visual information improves performance, current models still predominantly rely on textual cues, underscoring the need for improved multimodal integration throughout the reasoning process.

2 Related Work

Prior benchmarks for ToM broadly fall into two categories: text-only and multimodal. Traditional text-only benchmarks, such as TOMI (Le et al., 2019) and HI-TOM (Wu et al., 2023), predominantly focus on probing models' ability for nested belief tracking and logical inference through text stories lacking realistic social context. TOMBench (Chen et al., 2024) expands beyond belief tracking alone, incorporating a broader taxonomy of social and pragmatic ToM tasks (e.g., faux-pas detection, persuasion, hidden emotions, desires) within everyday textual scenarios. Despite this richer coverage, it remains constrained by its purely textual format, lacking multimodal information critical to human social understanding (Byom and Mutlu, 2013).

Multimodal approaches such as MMToM-QA (Jin et al., 2024) present procedurally-generated videos of single actors in household tasks, primarily evaluating goal and belief inferences without meaningful social interaction or emotional complexity. Similar to the text-only evaluations discussed above, this setup fails to reflect the depth and nuance of genuine human social behavior, limiting its applicability in evaluating socially intelligent AI systems (Chen et al., 2025a).

From the social intelligence perspective, Social IQa (Sap et al., 2019) probes social and emotional intelligence of models through multiple choice questions that require reasoning about social motivations, reactions, and actions based on specific situations. SOTOPIA (Zhou et al., 2023) evaluates how models navigate complex social scenarios and achieve social goals. EmoBench (Sabour et al., 2024) measures emotional intelligence by assessing models' ability to understand and apply emotional knowledge in complex social scenarios. However, these works are again limited to text-only evaluations and do not measure ToM directly.

Social Genome (Mathur et al., 2025) (based on SocialIQ2 (Wilf et al., 2023)) addresses the evalua-

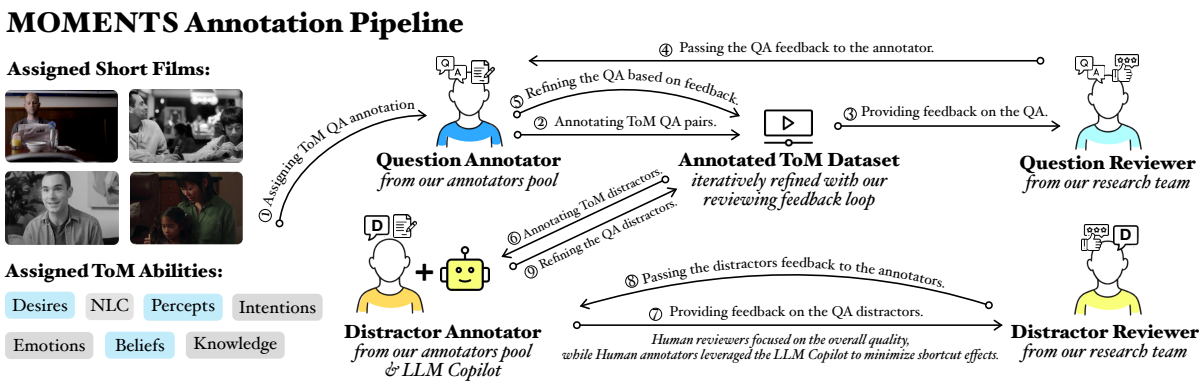


Figure 2: MOMENTS Annotation Pipeline. Different colored t-shirts represent different annotators/reviewers.

tion of social interaction understanding in VLMs through video-based multiple-choice questions, but videos are limited to 60 second clips, and evaluation is not designed to evaluate ToM. Furthermore, Guo et al. (2023) observed a strong bias in the representations of correct and incorrect answer candidates, where LLMs can achieve high accuracy with no context at all.

Given the limitations in prior work, there is a need for evaluating ToM within realistic multimodal settings, capturing authentic social interactions beyond goals and beliefs alone (Chen et al., 2025a).

3 Dataset Design

Recognizing the limitations of previous benchmarks, we design MOMENTS based on two core principles: (1) an *established taxonomy* of socially relevant ToM abilities –Emotions, Non-Literal Communication, Desires, Intentions, Knowledge, Percepts, and Beliefs– to evaluate ToM beyond the commonly addressed belief and goal probing abilities, and (2) *long-form videos* with real human actors that provide sufficient context and multimodal signals (e.g., facial expressions, gaze, body language, speech tone) to characterize interpersonal dynamics and mental states. This section outlines our taxonomy for probing different ToM abilities, the video selection process, and the annotations included in each question.

3.1 Taxonomy and Question Design

We adopt the **ATOMS taxonomy** (Abilities in Theory of Mind Space) introduced by Beaudoin et al. (2020) from their meta-analysis of ToM studies and proposed as a systematic framework for model evaluation by Ma et al. (2023). ATOMS categorizes ToM into seven distinct abilities: Knowledge, Emotions, Desires, Beliefs, Intentions, and Non-literal Communication (NLC). We describe and exemplify each ability in Table 1. This taxonomy supports precise question formulation and provides a detailed framework for systematically evaluating specific ToM abilities in models. We design annotation guidelines (See Appendix A.10) around it.

3.2 Video Selection

Existing datasets contain synthetic videos or minute-long clips that provide short temporal context. We instead propose to use *short films* as these contain more complex characterizations and longer temporal contexts, while having a self-contained narrative. Our videos come from the SF20K dataset (Ghermi et al., 2025), which contains a curated collection of short films from the YouTube channel *Omeleto*. Ghermi et al. (2025) verified that these films exhibit minimal information leakage to state-of-the-art language models compared to other common video sources like the sitcom *Friends*. Additionally, the videos are high-quality, vary in length (10 to 20 minutes), and provide complete, cohesive stories.

Not all short films have scenarios suitable for evaluating ToM. To filter these out, we prompted GPT-4o with film synopses to identify ones that likely contain interesting social dynamics. We then select videos with the highest likelihood of generating meaningful question-answer pairs and assign each annotator a subset of these to annotate.

3.3 Data Organization

In line with prior work, we adopt a multiple-choice question-answer (MCQA) format, where each question includes one correct answer and three plausible

ToM Ability	Overview	Example Q	Example A
Knowledge	Understanding what a person knows or does not know based on their sensory access.	Why is the soldier interested in the boy's bottle?	The soldier does not know what is inside the bottle and wants to find out.
Emotions	Identifying and reasoning about emotional responses, their evolution, and when emotions are hidden or complex.	How do the old woman and the young woman feel in this conversation?	The younger woman feels annoyed, and the older woman feels angry.
Desires	Situations that involve preferences, conflicting desires, or actions driven by desire.	What does the girl want after walking past the group and reading the sign?	She wants to go in to the establishment, finding it appealing.
Beliefs	Comprehending true and false beliefs and how beliefs influence actions.	What does the woman with the ponytail think of the man who is watching TV?	She thinks the man watching TV is aggressive.
Intentions	Understanding goals, motivations, and the underlying reasons for actions.	Why does the old man give a beer to the bearded man and leave the cabin?	He wants the bearded man to follow him so they can talk outside.
Percepts	Reasoning about what a character can or cannot perceive through their senses.	Why didn't the woman with the long hair protect herself from the man?	Because the man came up behind her and she didn't see him.
NLC	Interpreting humour, sarcasm, deception, and other speech that goes beyond literal meaning.	Why does the young man in white ask the man in blue if he likes his work?	He's being sarcastic and wants to annoy the man in blue.

Table 1: Overview for ATOMS abilities with example question/answer pairs extracted from MOMENTS

but incorrect distractors. Figure 1 exemplifies two items from MOMENTS, and more representative examples are presented in Appendix A.1. Below, we describe the structure and annotations included in each data point:

Questions are derived from specific scenes in the short films and must probe one or more ToM abilities as defined in the ATOMS taxonomy.

Answer Set includes one correct option and three distractors. Annotators are instructed to write distractors that are as plausible as possible, such that only a nuanced understanding of the context can lead to the correct answer. We paid special attention to the distractors, see Section 4.2 for more details on this.

Tags for ToM Abilities specify which ToM abilities (See Table 1) are targeted by the question. Questions may be annotated with multiple abilities, acknowledging that these often intersect in various scenarios.

Timestamps mark the start and end of the video segment relevant to the question. Each question is annotated with two context windows:

- **Full Context Window** $[t_0, t_j]$: A longer segment starting from the beginning of the video, intended to provide full narrative context useful for understanding character backgrounds, motivations, and evolving social dynamics.
- **Focused Context Window** $[t_i, t_j]$: A shorter segment containing only the immediate context required to answer the question. This window excludes broader narrative information, focusing instead on the specific scene being queried.

During evaluation, we explicitly instruct models that the question refers to the end of the provided interval (t_j). This approach minimizes reliance on temporal references in the question that may hint at the correct answer, which requires understanding the interaction. If leveraged effectively, the Full Context Window provides all the information required to understand characters, providing better insights into their mental states and interpersonal dynamics.

Multimodal Cue Tags indicate whether answering the question relies on interpreting specific nonverbal or auditory signals. These tags were optionally marked by annotators and are present only when such cues were deemed necessary for understanding the interaction. The possible cues include: *"Facial Expressions or Gaze"*, *"Body Language"*, and *"Speech-related"*.

Statistic	Length
Question Length	12.64 ± 4.2
Correct Answer Length	14.62 ± 7.8
Distractor Length	14.97 ± 7.7
$[t_i, t_j]$ length (s)	42.44 ± 55.5
$[t_0, t_j]$ length (s)	388.47 ± 262.3
Number of Videos	168
Video length (m)	14.56 ± 4.65

Table 2: **Top:** Mean length $\pm$ SD of questions, correct answers, and distractors (in words), together with the average duration of the Focused $[t_i, t_j]$ and Full $[t_0, t_j]$ Context Windows (in seconds). **Bottom:** Number of Videos and average duration (in minutes)

4 Annotation Methodology

Creating multiple-choice questions for this task is challenging. Annotators must understand different ToM abilities, find relevant moments in short films, and write clear questions. Making good distractors is also difficult because humans often create distractors that models can easily guess without seeing the video context, as observed by Guo et al. (2023) in other multimodal social understanding datasets.

To address these challenges, we conducted two pilot annotation rounds (see Appendix A.2) before launching the main annotation phase. Findings from the pilots helped us refine our pipeline to address the cognitive demands of ToM question creation, reduce annotation biases, and ensure question quality. The final methodology included carefully structured annotation phases, refined guidelines, and a custom-built platform to support robust distractor generation.

4.1 Annotation Pipeline

Annotation guidelines were centered around the ATOMS taxonomy and the specific goals of the benchmark. They included illustrative examples, key indicators (*what to look for*) for each ToM ability, and clearly defined criteria for both acceptable and problematic question types. We iteratively refined the guidelines based on feedback from our expert sociologist and from the annotators themselves during the pilot runs.

The main annotation phase spanned six weeks and involved 16 annotators who collectively produced 2,335 questions. This phase followed the methodology developed during the second pilot and incorporated several design choices aimed at improving quality and reducing bias (see Figure 2 for an overview):

- Annotators were asked to watch the full short film before writing questions to ensure understanding of character motivations and social dynamics.
- Each was assigned 2–3 ToM abilities to specialize in, promoting category-specific expertise.
- The schedule alternated weekly: a week focused on writing questions, the next on creating distractors for peers' questions. During the distractor-creation stage, annotators flagged poorly written or overly subjective questions, adding a layer of peer-based quality control.
- A custom platform integrated an LLM for real-time distractor feedback, flagging biased sets automatically (Section 4.2).
- We provided weekly feedback based on a review of the submitted material. For questions, we emphasized clarity, appropriate ToM category assignment, and avoidance of overly subjective QA pairs. For distractors, we focused on ensuring that none of the distractors could be considered a "technically correct" answer.
- We provided bonuses for early submissions and for the annotators who produced the highest-quality questions.

This approach encouraged focused annotation, peer-based quality control, and robust distractor generation, resulting in the final MOMENTS evaluation dataset. Table 2 report statistics about the dataset. In Appendix A.7, we report the demographics of annotators, cost of the annotations, and number of questions associated to each ToM ability.

4.2 Framework for Distractor Creation

Models frequently rely on subtle biases to guess correctly; our initial pilot batch showed this issue with models consistently achieving non-trivial performance by identifying correct answers without the required context. Creating high-quality distractors remains challenging for annotators despite providing them with guidelines; even subsequent re-annotation of distractors by us similarly demonstrated persistent biases.

While various post-hoc strategies exist to mitigate distractor bias (Ye and Kovashka, 2021; Guo et al., 2023), we integrate bias prevention directly into the annotation workflow. We designed a custom annotation platform embedded with an LLM acting as an on-the-fly evaluator for newly proposed distractor sets.

Given a question with one correct answer and

three proposed distractors, the platform evaluates potential biases distractor as described in Algorithm 1.

Algorithm 1: Distractor Set Assessment

Input: Question Q, correct answer a^*, distractors $D = \{d_1, d_2, d_3\}$, trials N, threshold k.
Output: Indicator of biased distractors

```
c ← 0;
for i ← 1 to N do
    A ← shuffle({a*} ∪ D);
    a ← LLMAnswer(Q, A);
    if a = a* then
        c ← c + 1;
if c ≥ k then
    return flag biased;
```

We establish empirically determined $k = 5$ and $N = 6$ to balance reliability and computational efficiency. A distractor set is flagged as biased if the model identifies the correct answer k or more times out of N trials. We initially employed GPT-4o-mini as the LLM for the first 800 questions; as we observed that the cost was relatively low, we decided to use GPT-4o for the remaining annotations.

5 Experimental Evaluations

We conduct experiments to evaluate the performance of current multimodal models in inferring mental states and to identify the factors that influence their performance. Specifically, we aim to answer: (i) How well do these models perform overall and across different ToM abilities? (ii) To what degree does visual information and context length impact performance? and (iii) How effective is our LLM-in-the-loop distractor creation platform at mitigating answer set biases? To this end, we report model accuracies on MOMENTS, ablate the effect of the visual modality and context window length, and assess performance in a no-context setting against baselines lacking bias-mitigation mechanisms.

5.1 Experimental Setup

We evaluate three types of multimodal LLMs: Video, Audiovisual, and Speech-based.

Video LLMs These models process visual and text inputs. We evaluate LLaVA-Video 7B and 72B (Zhang et al., 2024b), InternVL 2.5 8B and 78B (Chen et al., 2025b), LongVA 7B (Zhang et al., 2024a), and Qwen2.5 VL 7B (Bai et al., 2025). Each model receives 64 uniformly sampled frames per question (see Appendix A.5 for frame count ablations).

Model	$[t_0, t_j]$		$[t_i, t_j]$	
	T	VT	T	VT
● LLaVA-Video-7B	47.0	49.36 (+2.3)	45.6	**52.01** (+6.4)
● InternVL2.5 8B	46.0	46.63 (+0.6)	45.4	51.79 (+6.4)
● LongVA-7B-DPO	41.0	44.24 (+3.3)	41.5	44.5 (+3)
● Qwen2.5 VL 7B	41.3	38.05 (+-3.3)	38.4	44.33 (+5.9)
● LLaVA-Video-72B	63.2	65.96 (+2.8)	62.1	**67.66** (+5.6)
● InternVL2.5 78B	53.3	61.09 (+7.8)	52.2	61.48 (+9.3)
	T	VT	T	VT
▲ Qwen2.5-Omni-7B	46.8	53.41 (+6.6)	45.8	**56.19** (+10.4)
▲ VideoLLaMA2-7B-AV	37.6	40.96 (+3.3)	38.4	43.13 (+4.7)
▲ MiniCPM-o 2.6 (8B)	47.8		47.1	50.17 (+3.1)
	A	VA	A	VA
▲ Qwen2.5-Omni-7B	44.41	52.69 (+8.3)	48.46	**55.59** (+7.1)
▲ VideoLLaMA2-7B-AV	34.22	42.32 (+8.1)	34.22	43.6 (+9.4)
▲ MiniCPM-o 2.6 (8B)	39.5		39.6	48.25 (+8.6)
	A		A	
▼ Kimi-Audio-7B	31.7		**48.6**	
▼ Qwen2-Audio-7B	34.5		35.5	
	VA			
*Human (average of 3)	86.33 ± 1.15			
*Human (majority-vote)	91.0			

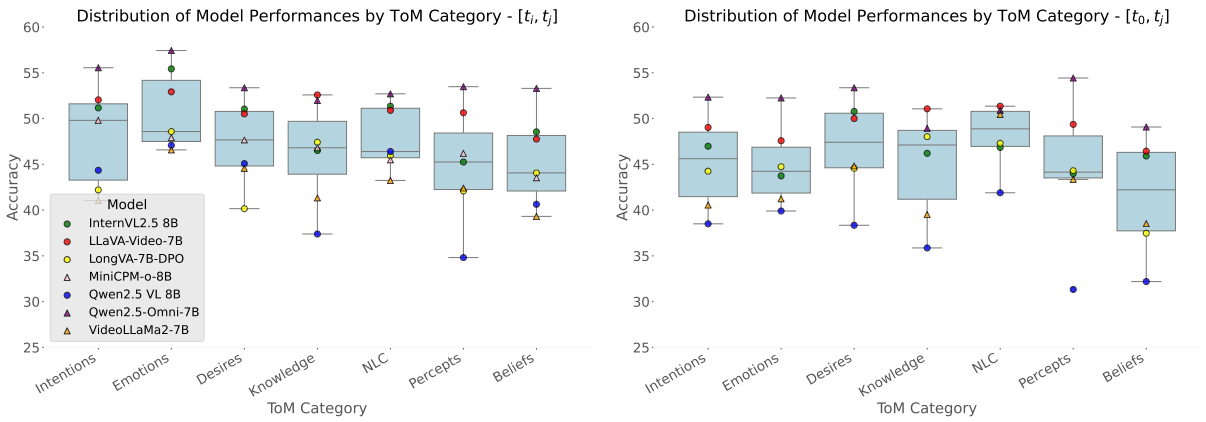
Table 3: Accuracy of different models on MOMENTS, reported for both Full $[t_0, t_j]$ and Focused $[t_i, t_j]$ context windows. For Video LLMs (●), we show performance with transcripts only (T) and video+transcripts (VT). For Audiovisual LLMs (▲), we additionally report audio (A) and video+audio (VA) inputs. For Speech LLMs (▼), we report performance using audio only (A). *Human evaluation is carried out in a sample of 100 randomly drawn items.

Since Video LLMs process only vision and text, we transcribe dialogues using ASR through WhisperX using Whisper *large-v2* as the backbone model (refer to Appendix A.4 for an evaluation on the performance of the ASR system).

Audiovisual LLMs These models process visual, text, and audio inputs. We evaluate Qwen2.5-Omni 7B (Xu et al., 2025), VideoLLaMa2 7B (Cheng et al., 2024), and MiniCPM-o 2.6 (8B) (Yao et al., 2024). We observed that MiniCPM-o 2.6 yielded errors in long-form videos; therefore, for this model we only report results on the Focused Context Window.

Speech LLMs These models process audio and text inputs. We evaluate Kimi-Audio 7B (KimiTeam et al., 2025) and Qwen2 Audio (Chu et al., 2024).

All models are evaluated using the *transformers* library (Wolf et al., 2020) with temperature set to 0. Under 70B parameters models run on a single NVIDIA A100 GPU, whereas larger models



(a) Boxplot of accuracies in the Focused Context Window.

(b) Boxplot of accuracies in the Full Context Window.

Figure 3: Boxplots comparing accuracies across different models (at ∼7B parameter scale) and ToM abilities. Results for Audiovisual LLMs (▲) are reported using video and audio inputs (VA), while results for Video LLMs (●) use video and transcript inputs (VT).

(>70B) run on three NVIDIA H100 GPUs.

5.2 LLM Evaluation

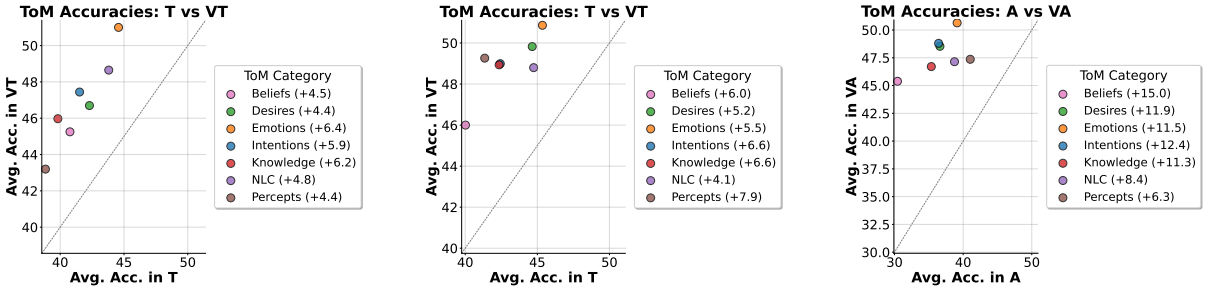
We evaluate models under different input conditions: textual or audio dialogues only (Transcripts (T) or Audio (A)), and combined inputs (Vision+Transcripts (VT) or Vision+Audio (VA)). Additionally, we compare model performance across two temporal contexts: the Full Context Window $[t_0, t_j]$ and the Focused Context Window $[t_i, t_j]$.

Global Accuracy Table 3 reports the global accuracy on MOMENTS; we observe that video input improves performance in most cases. However, the gains are modest, indicating that current models may underutilize visual cues. Performance tends to drop when using the longer Full Context Window, we attribute this to the fact that long video understanding is still challenging for open models.

As a reference, we conducted human evaluations on a subset of 100 randomly sampled questions with three different evaluators on the Full Context Window and the VA setting. Individual accuracies were 87.0/85.0/87.0 (an average of 86.3). Majority-vote accuracy with ties marked as incorrect was 91.0. We observed a percent agreement of 0.80, and a Fleiss κ of 0.733, which indicates substantial agreement between evaluators when selecting an answer. In Table 3, we report both majority-vote (marking ties as incorrect) and the average of individual accuracies.

Accuracy by ToM Ability Figure 3 presents boxplots with per-model scatter points, showing model accuracy across different ToM abilities under two context window conditions. As observed in the global accuracies, models perform better using shorter Focused Context Windows, though the impact of context length varies by ability. For example, accuracy is notably higher for Emotions and Intentions questions with shorter contexts, suggesting that these tasks rely more on immediate cues. Among results within the longer Full Context Window, Knowledge, Desires, and Non-literal Communication (NLC) questions perform relatively better, suggesting that longer context may be beneficial for understanding characters' background and effectively answering these questions. For both of the context settings, Percepts and Beliefs remain the most challenging abilities. Future work should investigate how context window length affects human performance in this task.

In Figure 4, we compare the effect of visual input for Video and Audiovisual LLMs. Vision contributes positively across all abilities, although the improvement varies by model type. For Video LLMs, the gains are relatively consistent, while Audiovisual LLMs show greater variation, especially when dialogues are provided as audio: some abilities improve by over 12 points, others by only 6-8. This larger gain with vision is mainly due to the weaker performance of the audio-only setting, where models struggle on abilities such as Beliefs or Intentions that likely require more complex information that the visual channel can provide. Notably, Non-literal Communication shows the smallest improvement, suggesting stronger reliance on dialogue than visual cues compared to



(a) Video LLMs (∼7B scale) (b) Audiovisual LLMs (VT vs. T) (c) Audiovisual LLMs (VA vs. A)

Figure 4: Comparison between average accuracies across the evaluated Video LLMs and Audiovisual LLMs with and without vision across different ToM abilities (Focused Context Window). The number in parentheses refers to the improvement due to the visual modality.

	$\Delta_{Focus-Full}$	Δ_{VT-T}	Δ_{VA-A}
Body Language	3.71	7.92	8.05
F. Exp. and Gaze	2.01	6.5	6.79
Speech-related	1.77	3.62	6.32

Table 4: Effect of context length and visual input on questions marked as reliant on multimodal cues. $\Delta_{Focus-Full}$ is the average accuracy difference between Focused and Full Contexts, with positive values indicating better performance in shorter intervals. Δ_{VT-T} and Δ_{VA-A} represent average accuracy gains from adding visual input (VT vs. T across Video and Audiovisual LLMs; VA vs. A across Audiovisual LLMs), both reported for the shorter context interval.

Model	SIQ2-dev	M-P1	MOMENTS
Qwen2.5 VL 7B	52.49	60.59	36.05 (-24.54)
LongVA-7B-DPO	53.38	58.49	34.85 (-23.63)
LLaVA-Video-7B	56.2	59.48	40.10 (-19.38)
InternVL2.5 8B	51.43	55.39	36.26 (-19.13)

Table 5: Accuracy by *guessing* the correct answer, where models are not provided with any context about the question. M-P1 refers to our first pilot study, and SIQ2-dev to the development set of SocialIQ2 (Wilf et al., 2023).

the other abilities.

Multimodal Cues We further analyze performance on questions requiring multimodal understanding (facial expression or gaze, body language, and speech-related cues). As shown in Table 4, incorporating visual input and using a shorter context window generally improves performance, particularly for questions involving Body Language and Facial Expressions or Gaze. Speech-related questions benefit less from visual input when dialogues are provided as transcripts (VT), confirming stronger reliance on text. Interestingly, Audiovisual models show larger gains in Speech-related questions when dialogues are provided as audio (VA), suggesting vision is more beneficial in this setting.

5.3 Evaluation on Answer Set Bias

In this section, we evaluate the effectiveness of using an LLM-in-the-loop design during the annotation pipeline, specifically for distractor creation. For MCQA-style ToM evaluation to be meaningful, questions should not be answerable without access to some form of context such as video, audio, or transcripts. However, as observed in our initial pilot and in prior work (Guo et al., 2023), models often exploit biases in question-answer sets to guess the correct answer even without contextual input.

We assess the extent of this issue by comparing MOMENTS to two baselines: our initial pilot (which did not use LLM assistance for distractor creation) and SocialIQ2, a similar video MCQA dataset. We prompt models with only the questions and answer options (without context) and measure their accuracy. As shown in Table 5, our proposed LLM-assisted distractor generation substantially reduces answer-set bias and lowers model accuracy by over 20 percentage points, highlighting the effectiveness of our approach.

By reducing biases in the answer sets, we create greater headroom for models to improve through reasoning based on the provided context.

6 Open Challenges for Future Model Development

Our evaluations on MOMENTS suggest that current limitations in multimodal ToM performance may stem not only from the reasoning capabilities of large language models, but also from how these systems access and process multimodal evidence.

Our findings point to several technical factors that likely limit models' ability to reason about mental states in socially rich scenarios. In this section, we outline four open challenges that, if addressed, could foster progress toward building better social multimodal agents.

Capturing Prosody and Ambient Sound in Audio Transcripts alone omit environmental sounds and paralinguistic cues (speaker prosody, intonation), which support accurate inferences about *Percepts*, *Emotions*, *Intentions*, and *Non-literal Communication*. In addition, errors in the ASR propagate downstream. While some audio-native models like Kimi-Audio and Qwen2.5-Omni (audio-only setting, A) demonstrate slight advantages over transcripts when vision is not provided, most transcript-only models perform comparably or better when visual input is present (see Table 3). Future research should focus either on integrating rich audio descriptors into existing video-text pipelines or improving the current audio-processing capabilities of audiovisual models, especially for understanding longer conversational contexts.

Precise Vision–Speech Alignment Prior work has shown that state-of-the-art models struggle to attribute utterances to speakers in multimodal conversations (Chang et al., 2025). Answering *Who said what, when?* requires time-synchronized links between each utterance, the speaking character, and the surrounding visual context. Without such alignment, models cannot track which speakers possess which knowledge, nor can they exploit gaze, facial expressions, or body language that modulate dialogue meaning. The small gains we observe from adding vision (Table 3), and the limited improvements on questions marked as reliant on visual cues (Table 4), also indicate that existing pipelines underutilize this channel.

Human-Centered Frame Selection Uniform frame sampling risks missing short yet meaningful signals while wasting computation on redundant content. Simply increasing the frame rate is expensive and, as our ablation in Appendix A.5 shows, does not improve performance. Specialized frame sampling strategies that prioritize human-salient events (faces, hands, gaze shifts) are needed to capture the cues that observers actually rely on.

Structured Reasoning over Multimodal Evidence Reasoning improves a wide range of *text-only* tasks, including ToM benchmarks. However, as Mathur et al. (2025) reports, asking VLMs to reason neither boosts accuracy nor yields human-aligned explanations for social MCQA in videos. We argue that effective multimodal reasoning may be bottlenecked by the three challenges above: inadequate audio representations, weak vision–speech alignment, and sub-optimal frame selection. Until models receive richer, better-organized evidence, additional reasoning steps are unlikely to help.

7 Conclusion

We introduced MOMENTS, a benchmark that probes seven ToM abilities in realistic, long-form videos. It contains over 2,300 human-annotated MCQA items with substantially reduced biases in answer sets compared to prior datasets. From baseline experiments with Video, Audiovisual, and Speech LLMs we observe: (i) visual input offers consistent yet modest gains, indicating underutilization of visual cues; (ii) using audio inputs does not yield a noticeable improvement over transcripts, suggesting that current models still have struggle to effectively integrate this modality; and (iii) performance tends to drop on extended context windows, highlighting limitations in long-range video reasoning.

Based on these results, we identify several open challenges that likely constrain progress on multimodal ToM tasks, ranging from multimodal alignment and audio processing to frame selection and reasoning over multimodal evidence. Addressing these issues will be essential for developing AI systems capable of truly understanding, predicting, and responding to human mental states in real-world social settings.

Limitations

We adopt a multiple-choice QA format in MOMENTS to streamline annotation and ensure consistent evaluation. While this design supports scalable benchmarking, it limits analysis of lower-level behavioral cues such as turn-taking, speech acts, or gesture dynamics as we do not provide fine-grained annotations on them. Investigating the relation between these cues and specific ToM abilities remains an important direction for future work. Additionally, MOMENTS uses static video data, which does not capture model performance in interactive or dynamic social environments. Extending evaluation to such settings is a promising but currently chal-

lenging task, as it would require reliably simulating complex, multimodal human behaviors. Finally, although using multiple annotators per question could reduce subjectivity, resource constraints limited us to one annotator per question. To mitigate this, we incorporated peer-checking during distractor creation and conducted multiple rounds of author review to ensure data quality and consistency.

References

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. *Preprint*, arXiv:2502.13923.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *Preprint*, arXiv:2303.00747.

Andrew P Bayliss and Steven P Tipper. 2006. Predictive gaze cues and personality judgments: Should eye trust you? *Psychological science*, 17(6):514–520.

Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H Beauchamp. 2020. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10:2905.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Lindsey J Byom and Bilge Mutlu. 2013. Theory of mind: Mechanisms, methods, and new directions. *Frontiers in human neuroscience*, 7:413.

Kent K. Chang, Mackenzie Hanh Cramer, Anna Ho, Ti Ti Nguyen, Yilin Yuan, and David Bamman. 2025. Multimodal conversation structure understanding. *Preprint*, arXiv:2505.17536.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. 2025a. Theory of mind in large language models: Assessment and enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31539–31558, Vienna, Austria. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

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.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

LMJ De Sonneville, CA Verschoor, C Njiokiktjien, V Op het Veld, N Toorenaar, and M Vranken. 2002. Facial identity and facial emotions: speed, accuracy, and processing strategies in children and adults. *Journal of Clinical and experimental neuropsychology*, 24(2):200–213.

Ridouane Ghermi, Xi Wang, Vicky Kalogeiton, and Ivan Laptev. 2025. Long story short: Story-level video understanding from 20k short films. *Preprint*, arXiv:2406.10221.

Xiao-Yu Guo, Yuan-Fang Li, and Reza Haf. 2023. DeSIQ: Towards an unbiased, challenging benchmark for social intelligence understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3169–3180, Singapore. Association for Computational Linguistics.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B Tenenbaum, and Tianmin Shu. 2024. Mmtom-qa: Multimodal theory of mind question answering. *arXiv preprint arXiv:2401.08743*.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *Preprint*, arXiv:2504.18425.

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.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory

of mind in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1011–1031, Singapore. Association for Computational Linguistics.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. 2025. Social genome: Grounded social reasoning abilities of multimodal models. *Preprint*, arXiv:2502.15109.

Ini Oguntola, Dana Hughes, and Katia Sycara. 2021. Deep interpretable models of theory of mind. *Preprint*, arXiv:2104.02938.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssouf Kebe, Paul Pu Liang, and Louis-Philippe Morency. 2023. Social-iq 2.0 challenge: Benchmarking multimodal social understanding. `https://github.com/abwilf/Social-IQ-2.0-Challenge`.

Jessica Williams, Stephen M Fiore, and Florian Jentsch. 2022. Supporting artificial social intelligence with theory of mind. *Frontiers in artificial intelligence*, 5:750763.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

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*, pages 10691–10706, Singapore. Association for Computational Linguistics.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Keren Ye and Adriana Kovashka. 2021. A case study of the shortcut effects in visual commonsense reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3181–3189.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024a. Long context transfer from language to vision. *Preprint*, arXiv:2406.16852.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024b. Video instruction tuning with synthetic data. *Preprint*, arXiv:2410.02713.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

A Appendix

A.1 Samples from MOMENTS Across ToM Abilities

Figure 5 presents representative samples of MOMENTS questions covering different ToM abilities. Each example includes the question, the full answer set (one correct option and three distractors), the targeted ToM abilities, and any multimodal cues identified by annotators as relevant for answering the question.

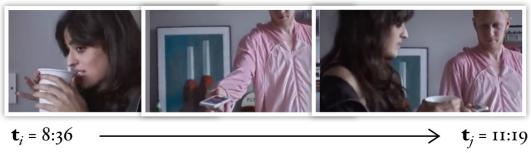
A.2 Pilot Annotations

We conducted two pilot annotation phases prior to the main annotation batch to identify challenges and refine our pipeline.

First Pilot Annotation We recruited annotators through Prolific, selecting participants who were native English speakers with a university degree. Each annotator was asked to create both questions and distractors covering all seven ToM abilities. This pilot produced 268 question–answer sets. From analyzing submissions from this annotation batch, we identified the following issues:

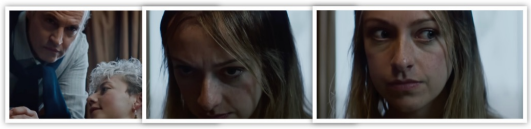
Question Why does the boy turn to look at the man with the glasses?
ToM Abilities Percepts
Multimodal Cues Body Language, Face Expression & Gaze
Answers
- ○ The boy turns to the man with the glasses for help opening his pencil case.
- ○ The boy looks at him because he wants him to leave him alone.
- ● The boy looks at him because he hears the man's phone ring.
- ○ The boy looks at him because the man calls his name.

Question Why did the woman in the red coat have that reaction when she saw the crib?
ToM Abilities Knowledge
Multimodal Cues Face Expression & Gaze
Answers
- ○ She reacted that way because she thought the crib had been given away already.
- ○ She reacted that way because she assumed no one would keep something so old.
- ● She reacted that way because she didn't know the crib would be there.
- ○ She reacted that way because she thought the room had been cleared out entirely.



Question How does the woman feel when her phone rings?
ToM Abilities Emotions
Multimodal Cues Body Language, Speech-related
Answers
- ○ The woman is curious to know who is calling her.
- ○ The woman is nervous that the call might be bad news.
- ○ The woman is excited that someone is calling.
- ● The woman is annoyed by the call.

Question Why didn't the woman in the white blouse respond to the friend who greeted her?
ToM Abilities Desires, Non-Literal Communication
Multimodal Cues Face Expression & Gaze
Answers
- ○ Because she didn't recognize the voice and wasn't sure who was speaking.
- ○ Because she was distracted and didn't realize someone was speaking to her.
- ○ Because she wasn't in the mood to talk and chose to ignore it.
- ● Because she wanted to hide the fact that she wasn't deaf.



Question What does the blonde-haired girl assume when she sees people talking?
ToM Abilities Beliefs, Percepts
Multimodal Cues Face Expression & Gaze
Answers
- ○ She assumes they are planning a surprise for her since she is new.
- ○ She believes they are talking about drama within the family.
- ● She believes that they are probably judging her.
- ○ She believes they want to steal her strategy for playing the game.

Question Why does the woman begin to move the containers?
ToM Abilities Intentions, Beliefs
Multimodal Cues N/A
Answers
- ○ She thinks that shifting the containers will create a more practical living space and make daily routines easier for both of them
- ○ She thinks that moving the containers will make the apartment feel less oppressive and help her regain a sense of personal control.
- ● She thinks that reducing the clutter in the apartment will demonstrate her commitment to improving her condition in the eyes of the man.
- ○ She thinks that rearranging the containers will distract the man from his frustrations long enough to keep him from leaving.

Figure 5: Samples from MOMENTS Representing Each ToM Abilities.

- Many questions were low quality, some had grammatical issues, others focused on plot rather than ToM.
- Models achieved over 50% accuracy without context, pointing to biases in the distractor sets (see Table 5).
- Annotators often mislabeled the ToM ability, indicating limited understanding of the categories.

We traced these problems to the following causes:

- Time constraints imposed by Prolific's system created pressure that negatively impacted annotation quality.
- Prolific communication channels made direct communication with annotators difficult, as they did not communicate their questions effectively.
- Tasking annotators with all seven categories was overwhelming, leading to overall misclassification.
- Most effort was spent on writing questions, resulting in weaker distractors.
- Models could exploit biases in seemingly good distractors, without needing any context to answer.

Second Pilot Annotation To address these issues, we made the following changes:

- We directly hired seven undergraduate students from psychology and social sciences and used group messaging for better communication.
- Each annotator was assigned only 2–3 ToM abilities to help them specialize.
- Annotation was split into two phases: creating questions in the first week and distractors in the second. This was done to help annotators concentrate their efforts on writing high-quality questions first, then shift their focus to creating high-quality distractors.
- A custom annotation platform with an LLM was introduced to automatically flag biased distractors (see Section 4.2).
- Annotators were encouraged to spread their work throughout the week to reduce low-quality submissions due to pressure in last-minute submissions.
- We provided weekly reviews and feedback to improve consistency and quality.

This second pilot resulted in 350 high-quality questions. Most of the design choices from this phase were carried over to the main annotation batch.

A.3 Prompt For Video Filtering

```
You are a film critic and psychologist
    with expertise in Theory of Mind (
    ToM) as described by the ATOMS
    taxonomy. Your task is to analyze
    the movie synopsis and captions
    below to determine how likely it is
    that the movie includes themes or
    questions related to Theory of Mind.

Theory of Mind involves understanding
    and attributing mental states to
    oneself and others. Consider the
    following key components:
1. Knowledge: Recognizing that
    characters hold organized
    information and mental
    representations that shape their
    understanding.
2. Emotions: Identifying complex
    emotional responses, including mixed
     or evolving emotions.
3. Desires: Understanding that
    characters may have varied and
    sometimes conflicting desires
    driving their actions.
4. Beliefs: Discerning true versus false
     beliefs and recognizing higher-
    order beliefs (beliefs about others'
     beliefs).
5. Intentions: Inferring characters'
    goals and the reasoning behind their
     actions.
6. Percepts: Noting how characters
    perceive their world differently
    based on their sensory experiences.
7. Non-literal Communication:
    Interpreting subtleties such as
    sarcasm, humor, or metaphors that
    imply meanings beyond the literal
    words.

Using this framework, please analyze the
     following content:

Movie Synopsis: {synopsis}
Movie Captions: {caption}

Based on your analysis, provide a
    probability (as an integer
    percentage between 0 and 100)
    indicating how likely it is that
    this movie involves Theory of Mind
```

Model	English				Non-English			
	$[t_0, t_j]$		$[t_i, t_j]$		$[t_0, t_j]$		$[t_i, t_j]$	
	T	VT	T	VT	T	VT	T	VT
● LLaVA-Video-7B	47.59	49.11	46.53	51.39	43.90	50.68	40.38	55.28
● InternVL2.5 8B	46.33	46.84	45.06	50.84	44.17	45.53	47.43	56.91
● LongVA-7B-DPO	41.37	43.85	41.11	44.71	38.75	46.34	43.36	43.36
● Qwen2.5 VL 7B	40.71	38.23	37.92	44.15	44.72	37.13	40.92	45.26
● LLaVA-Video-72B	62.99	66.08	62.58	67.59	64.23	65.31	59.35	68.02
● InternVL2.5 78B	53.72	61.11	52.35	60.46	51.22	60.98	51.49	66.94
● Average	48.78	50.87	47.59	53.19	47.83	50.99	47.15	55.96
	T	VT	T	VT	T	VT	T	VT
▲ Qwen2.5-Omni-7B	46.38	52.71	45.57	54.99	49.32	57.18	47.15	62.60
▲ VideoLLaMA2-7B-AV	36.76	40.66	38.58	42.99	42.28	42.55	37.40	43.90
▲ MiniCPM-o 2.6 (8B)	47.70		46.99	49.37	48.51		47.43	54.47
▲ Average	43.61	46.68	43.71	49.11	46.70	49.86	43.99	53.66
	A	VA	A	VA	A	VA	A	VA
▲ Qwen2.5-Omni-7B	44.30	52.15	48.91	54.28	44.99	55.56	46.07	62.60
▲ VideoLLaMA2-7B-AV	33.77	41.92	33.82	42.99	36.59	44.44	36.31	46.88
▲ MiniCPM-o 2.6 (8B)	39.54	**	40.00	48.10	39.02	**	37.67	49.05
▲ Average	39.21	47.04	40.91	48.46	40.20	50.00	40.02	52.85
	A		A		A		A	
▼ Kimi-Audio-7B	31.95		48.41		30.08		49.59	
▼ Qwen2-Audio-7B	35.54		35.65		29.00		34.42	
▼ Average	33.75		42.03		29.54		42.01	

Table 6: Global accuracy of different Video LLMs (●), Audiovisual LLMs (▲), and Speech LLMs (▼) across English and non-English videos.

```
related questions or themes. Your
answer should be only the integer
value with no additional commentary.
 choose the best number that seems
appropriate based on the data.
```

A.4 Evaluation on ASR quality

In this subsection, we describe our audio processing pipeline, present, and report its ASR performance on a subset of human-annotated transcripts.

ASR Pipeline We use WhisperX (Bain et al., 2023) to transcribe the short films. Its multilingual capabilities make it suitable for both English and non-English videos in our dataset. For speaker diarization, we employ PyAnnote (Bredin, 2023).

ASR Quality Evaluation We evaluate the ASR pipeline using different base Whisper models on a subset of 50 human-transcribed videos, reporting global **Word-Error Rate (WER)** and **Diarization Error Rate (DER)**. For global WER we concatenate each file's reference and ASR transcripts lower-casing and punctuation removal and computing WER $= (S + I + D)/N$, where S, I, and D are the numbers of substituted, inserted, and deleted words, and N is the total number of reference words. For DER, we evaluate only within spans where the reference marks speech. The score is DER $= (T_{\text{missed}} + T_{\text{confusion}})/T_{\text{ref}}$, where T_{missed} is reference speech with no ASR coverage, $T_{\text{confusion}}$ is overlapped speech attributed to the wrong mapped speaker, and T_{ref} is the total duration of speech in the reference annotation. We report these in Table 7, while Whisper large-v3 scores the lowest average global WER, in practice we notice that it failed to transcribe some of the videos. This does not happen with *large-v2*, whose DER is the lowest; because of this, we opted for the latter as the chosen model for transcribing audio for the Video LLMs.

	global-WER	DER
base	36.2	41.2
large-v2	20.6	36.3
large-v3	16.6	40.9

Table 7: Comparison of average WER and DER across the three evaluated models.

A.5 Ablation on number of frames

Increasing the number of video frames increases computational cost, as most Video LLMs embed frame patches significantly extending the context

Model	$[t_0, t_j]$ T	VT-64	VT-96	$[t_i, t_j]$ T	VT-64	VT-96
LLaVA-Video-7B	46.33	47.7 (**+1.4**)	47.3 (+1.0)	44.45	50.7 (**+6.3**)	49.7 (+5.3)
LongVA-7B-DPO	40.94	45.5 (**+4.5**)	42.6 (+1.6)	41.19	42.9 (+1.8)	44.6 (**+3.4**)
InternVL2.5 8B	45.58	45.6 (+0.1)	48.2 (**+2.6**)	44.45	51.7 (**+7.3**)	50.4 (+6.0)

Table 8: Global accuracy on a subset of 1,500 MOMENTS samples using only transcripts (T), and transcripts plus 64 or 96 frames (VT-64 and VT-96). Results are reported for both the Full ($[t_0, t_j]$) and Focused ($[t_i, t_j]$) Context Windows. We mark in bold the highest increase over T between 96 and 64 frames.

length processed by the language model. To assess the tradeoff between context length and performance, we evaluate three models on 1,500 randomly selected MOMENTS entries using 64 and 96 frames.

As shown in Table 8, increasing the number of frames does not lead to consistent improvements. In several cases, performance actually drops, likely due to redundancy or context saturation. Based on these results, we use 64 frames for all main Video-LLMs evaluations in the paper.

A.6 Performance comparison in English and Non-English videos.

As noted in Appendix A.7, MOMENTS includes a subset of non-English videos. Table 6 compares performance on English-only and non-English clips. We find no substantial drops in accuracy for non-English videos; in fact, models with visual inputs often perform better in this setting. A possible explanation for this is that most of the non-English videos include subtitles in the frames, which may support the temporal grounding of dialogues.

A.7 Dataset Statistics and Annotation Cost

The main annotation batch involved 16 participants: 12 undergraduate students in psychology and social sciences, two computer scientists, and two clinical psychologists. 12 of them were female and 4 male, all of them between 20 and 30 years old. Twelve of the annotators were from Canada, and the remaining were from Mexico. All participants were explained the purpose of their annotations in an onboarding session. The annotation process received approval from the MBZUAI Ethics Review Board.

Annotation Cost Annotators were compensated at a rate of 17 CAD per hour through UpWork. To encourage steady progress, a weekly bonus of 10 USD was provided to those who completed at least half of their assignments by midweek. An additional performance-based bonus of 150 USD was awarded to annotators who produced the highest-quality annotations. The total cost of the MOMENTS main annotation effort amounted to 8,745 USD.

Language	Number of Videos
English	144
Russian	6
Spanish	5
French	3
Persian	3
Italian	1
Arabic	1
Swedish	1
Korean	1
Danish	1
Hindi	1
Japanese	1
Total	**168**

Table 9: Number of videos per language.

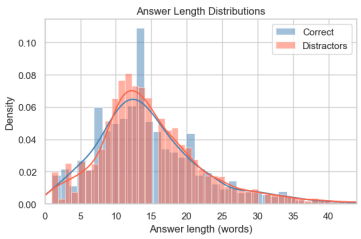
Dataset Statistics MOMENTS contains 2,335 questions across 168 short films, the majority of which are in English (144). We also include a subset of 24 films in other languages. Table 9 reports the number of videos per language.

In Table 2, we report the average question length, average answer lengths, and durations of the full and Focused Context Windows. We also display the distributions of lengths for answers, Focused, and Full context windows in Figures 6a, Figure 6b, and Figure 6c, respectively.

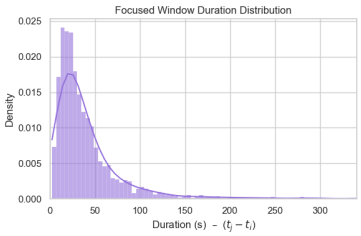
A.8 Copyright and License

We release MOMENTS annotations under a **CC BY-NC-SA 4.0** license (Attribution-NonCommercial-ShareAlike 4.0 International), intended only for academic research purposes.

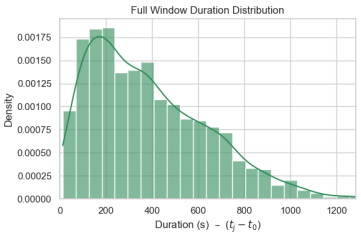
Following Ghermi et al. (2025) and Wilf et al. (2023), we do not distribute the video content directly. We provide URLs linking to the original videos on YouTube, complying with YouTube's Terms of Service (`https://www.youtube.com/static?template=terms`).



(a) Distribution of lengths for correct answers and distractors.



(b) Length distribution of the Focused Context Windows.



(c) Length distribution of the Full Context Windows.

Figure 6: Histograms of different the statistics reported in Table 2.

ToM Ability	# Questions
Emotions	599
Beliefs	379
Desires	386
Intentions	1026
Percepts	316
Knowledge	329
NLC	222

Table 10: Number of questions associated with each ToM ability.

A.9 Ethical Considerations

Representation and Bias Most of MOMENTS videos are in English and reflect Western cultural norms. Additionally, annotators were from Canada and Mexico, which may influence interpretations of emotions, intentions, or non-literal communication.

Potential Misuse MOMENTS is designed to evaluate models' ability to infer mental states in socially grounded scenarios to foster progress in socially intelligent AI. However, ToM capabilities could also be misused to simulate deceptive, manipulative, or persuasive behavior in artificial agents. To mitigate this risk, we license the dataset for academic research only under a CC BY-NC-SA 4.0 license, and we strictly stand against any use in applications that exploit it for unethical purposes.

Personally Identifying Information or Offensive Content Questions and answer sets do not contain personally identifying information as they use descriptors to refer to the characters. Since questions ask about character's mental states, they do not contain offensive content.

A.10 Guidelines for Question and Distractor Annotation

The following pages contain the annotation guidelines provided to annotators during the first annotation batch. Separate documents were provided for the question creation and distractor creation stages to reflect the specific goals and challenges of each.

Question Annotation Guidelines

Your Task

As an annotator, you will watch short films with self-contained stories and identify relevant moments that display these ToM abilities. Using the context from these interactions, you will create questions and answers that will serve to evaluate AI systems' reasoning skills about the observed behaviors of characters in the context of the video.

You will be provided with a set of short films and **must create a total of 150 questions about the Theory of Mind abilities you were assigned.** You can distribute these questions across the videos based on which ones you find more interesting or provide richer material for question creation. **For each question, you will provide a correct and precise answer**.

For each question you create, **you must select the ToM ability that the question is testing and the multimodal signal required to answer it**. While selecting multiple ToM abilities is possible if the question genuinely tests multiple aspects or is an intersection of different abilities, we encourage selecting only the most relevant ability when possible. This helps maintain clarity in what each question is measuring. The specific ToM abilities (knowledge, emotions, beliefs, desires, intentions, percepts, and non-literal communication) will be explained in detail below.

Please read the annotation guidelines in this document.

General Guidelines for Question Creation

Questions should be about the character mental states or interactions, not about the film's plot: The questions should ask about the character's mental states and how they influence actions, and interactions, not about the plot of the film or the message that the film is trying to convey.

✅ "Why is the girl with gray shirt ignoring the man?"
✅ "Why did the boy suddenly become quiet?"
❌ "Why did the team lose the championship?"
❌ "What is the film trying to convey about friendship?"

Character References and Question Framing: When writing questions, use descriptive terms instead of character names.

✅ "Why is the blonde girl faking a smile?"
✅ "What the emotional progression of the man with the green vest?"
❌ "Why is Ana faking a smile?"
❌ "What the emotional progression of Ahmed?"

Video Timestamps and Context: For each question, you must mark a timestamp in the annotation platform such that the interval contains the necessary context to answer the question:

- The **end of the marked timestamp serves as the temporal reference for the question**. (This means that the question refers to whatever happened **right before** the end of the interval).
- Make sure questions are understandable with the interval up to the selected timestamp. If the video contains a plot-twist after the timestamp end which may change the interpretation of a character's behavior, we assume the observer (human or AI) does not know it.

Multimodal Signal Toggles: When we observe people interacting, we don't just rely on their words to understand what they're thinking or feeling. We also pay attention to a range of **Multimodal Cues** that provide crucial context for interpreting mental states. For each question you create, identify if any of these signals are relevant (if any) for correctly answering the question. These signals often reveal information that contradicts or enriches what characters explicitly say. In particular you will mark a toggle for each of the signals.

- **Facial Expressions & Gaze**: Eye contact, facial emotions, shared attention
- **Body Language**: Posture, proximity/touch, meaningful gestures
- **Speech-Related**: Vocal tone, nonverbal sounds (sighs, laughter), backchanneling

When creating your questions, ***consider which signals are essential for correctly inferring the characters' mental states***. A simple way of checking this is "Do I need to take into account this social signal to answer the question?".

Questions that require integrating multiple signals or noticing contradictions between verbal and nonverbal communication often make for more challenging and insightful Theory of Mind assessments.

We provide a few examples on these on the example section, and in the last section we also show a set of questions to detect whether you should mark these toggles for your question.

Question Difficulty and Complexity: We aim to create challenging questions that require deeper analysis whenever possible. Some ToM abilities may inherently result in easier questions, but try to incorporate additional context or complexity when possible.

When designing questions, consider:

Information Sources:

- Verbal cues (dialogue content, speech tone)
- Non-verbal cues (facial expressions, body language, gestures)

Context:

- Pay attention to the full interaction context (history, relationships, setting)
- Observable behavior may not reflect true mental states which should be interpreted in their context

! What makes a good-complex question?

- **Higher-order Theory of Mind.**
  - Questions about what one person thinks/believes about another person's mental state
  - Chains of social reasoning involving multiple people's perspectives
- **Multiple modality integration**
  - Questions requiring synthesis of verbal and visual/speech cues.
- **Social Dynamics Understanding**
  - Questions about complex group dynamics and power relationships
  - Situations involving implicit social rules or cultural norms
- **Temporal Reasoning**
  - Understanding how mental states or emotions evolve over time.
  - Connecting past events to current behaviors or reactions.
- **Conflicting Mental States**
  - Questions about mixed or contradictory emotions/desires within a person
  - Conflicts between stated intentions and actual behavior

For example, given the following scene:

> *A business meeting where a junior employee nervously presents while their manager interrupts with subtle criticisms. (ToM abilities → Percepts; Intentions; Emotions. Signals → Gaze & Facial Expressions)*
>
> - Good Example: Why does the other team member keep making eye contact with the presenter after each interruption?
> - Poor Examples: "Why is the presenter feeling nervous?", "Is the presenter nervous?"

! What makes a bad-easy question?

- Can be answered just by extracting information from the dialogues.
- Is written in a way that hints to the correct solution.
- Asks about the plot, or characters actions independent of their mental states or their interactions.
- Can be answered with Yes/No.

Guidelines for specific Theory of Mind categories

Below are descriptions of each ToM category and their associated sub-abilities. **Annotators must create questions and categorize them into the 7 main abilities, not the sub-abilities**. Sub-abilities should only refine annotators' understanding of categories and help them better identify them. The question does not have to belong to a particular sub-ability and can be an intersection.

Knowledge

- **Overview**:

  Understanding what a person knows or doesn't know based on their senses or their access to information. Moments where participants' actions depend on what they know (the information they hold about the world or others).

- **What to Look For:**
  - **Percepts-Knowledge Links**: Identifying when senses (seeing, hearing, etc.) influence knowledge.
  - **Information-Knowledge Links**: Recognizing knowledge gaps in actors caused by missing information.
  - **Knowledge-Attention Links**: Understanding that novelty draws more attention than familiarity.
- **Example Question-Answer pairs:**
  - **Q:** Why is the old man more interested in the box on the table than the old book?
    - **A (correct):** The box is unfamiliar to him, therefore is more interesting.
    - ***A (distractor):** The box is more visually appealing, therefore is more interesting.

👁 Note: In this first annotation stage, **you are not tasked to create distractors**, you only have to write the correct question . But we include them so you understand how the data is going to look like in the end.

Emotions

- **Overview**:

  Identifying and reasoning about emotional responses, their evolution, and when emotions are hidden or complex. Annotate both explicit emotional expressions and subtler emotional cues. Emotions should belong to the following subset:

  - admiration
  - amusement
  - anger
  - annoyance
  - approval
  - caring
  - confusion
  - curiosity
  - joy
  - desire
  - disappointment
  - disapproval
  - disgust
  - embarrassment
  - excitement
  - fear
  - gratitude
  - grief
  - love
  - nervousness
  - optimism
  - pride
  - realization
  - relief
  - remorse
  - sadness
  - surprise
  - neutral

  Note: If the emotion is not on the list (or cannot be described with any entry from the list) it is possible to use another one, but try to stick to this taxonomy as much as possible.

- **What to Look For**:
  - **Typical Emotional Reactions**:
    Common responses to situations (for example, smiling when happy).
  - **Atypical Emotional Reactions**:
    Unexpected responses to situations (for example, laughing in a sad context).
  - **Mixed Emotions**:
    Recognizing conflicting or simultaneous emotions.
  - **Hidden Emotions**:
    Identifying when emotions are hidden or disguised.

- **Emotion Regulation**:
  Identifying strategies used to manage emotions.

Important Notes: When you assign your question as an "Emotion" ToM ability, you must:

- Actually include emotion labels within your answer.
- The emotion labels **must be from the above pre-defined categories**; e.g. emotional state or experience such as "guilty", "frustration" is not allowed to answer your question because they don't belong to the above pre-defined categories.
- You may provide additional details of the speaker's emotional state or experience.
- You may use derivative words of the emotional categories; e.g. sad → saddening, saddened, joy → joyous, joyful.

Take a look of below examples for clarity. **Example of Emotion Ability Question-Answer pairs**:

- **Q:** How is the girl feeling after hearing the news?
  - ✅ **A (Correct Answer):** She is saddened but hides her true feelings with smile.
  - ✅ **A (Distractor):** Her smile shows optimism about the situation.
  - ❌ **A (Correct Answer):** She is frustrated but hides her true feelings with smile.
  - ❌ **A (Distractor):** Her smile shows optimism about the situation.
- **Q:** What is the emotional progression of the blonde guy during his wedding?
  - ✅ **A (Correct Answer):** Nervous → Joyous → Surprise
  - ✅ **A (Distractor):** Nervous → Joyous → Fearful
  - ❌ **A (Correct Answer):** Nervous → Guilty → Surprise
  - ❌ **A (Distractor):** Nervous → Joyous → Hatred
- Notice that you are expected to always stick with the emotional labels or keywords that we have pre-defined in the above section.
- The inclusion of emotional labels that are arbitrary or out-of-category are ***highly discouraged*** and can led to ***rejection***.

Desires

- **Overview**: Understanding situations that involve preferences, conflicting desires, or actions driven by desire. Desires can be displayed through verbal and non-verbal cues; keep an eye out for both.
- **What to Look For:**
  - **Discrepant Desires**: Different people want different things.
  - **Multiple Desires**: Identifying when a person has coexisting or successive desires.
  - **Desires Influence on Emotions and Actions**: Understanding how a person's desires affect their behavior.
  - **Desire-Action Contradiction**: Explaining when actions conflict with stated desires.
- **Example Question-Answer pairs**:
  - **Q:** Why does the student hesitate to eat dessert?
    - **A (correct):** He wants to eat the dessert but is concerned about calories.
    - **A (distractor):** He does not want to eat the dessert feels pressured to try it.
  - **Q:** Why is the dark haired man making such an effort to defend his friend's actions?
    - **A (correct):** He knows his friend is capable of stealing, but he feels attracted to him and wants to be on his good side.
    - **A (distractor):** He does not believe his friend is capable of stealing, so taking this position is right thing to do as his friend.

Beliefs

- **Overview**: Understanding belief states, especially those involving true, false, or second-order beliefs. Keep an eye on moments where an actor bases their actions on incorrect or incomplete beliefs.
- **What to Look For:**
  - **Beliefs Influence on Emotions and Actions**: Linking actions or emotions to belief states.
  - False Beliefs: Understanding incorrect beliefs about reality, including:
    - Objects or their contents (e.g., a box containing something unexpected)
    - Locations of entities (e.g., misunderstanding where something was moved)
    - Identity of objects or people (e.g., mistaking something based on appearance)
  - **Second-Order Beliefs**: Beliefs about someone else's beliefs.
  - **Sequence False Beliefs**: Understanding how beliefs are formed and disrupted in scenarios where an expected sequence of events is interrupted by an unexpected event.
- **Example Question-Answer Pairs**:
  - **Q**: Why does the mother pretend to be surprised by the birthday cake?
    - **A (correct):** Because she thinks her daughter believes the cake is a surprise, even though she helped plan it.
    - **A (distractor):** Because she knows that is the most polite response, even though she helped plan it.
  - **Q:** What does the man thinks his son thinks of him after witnessing him fight.
    - **A (correct):** He probably thinks his son believes he is a violent man.
    - **D (distractor):** He probably thinks his son believes he is strong and fearless.

Intentions

- **Overview**: Understanding goals, motivations, and the underlying reasons for actions. Look for contextual cues that reveal underlying goals or motivations, especially in the context of conversations.
- **What to Look For:**
  - **Discrepant Intentions**: Recognizing different intentions behind similar actions.
  - **Intention Attribution**: Recognizing motivations behind actions (for example, why did an actor interrupt another one?).
  - **Intention Explanation**: Explaining reasons for past or current intentions.
- **Example Question-Answer pairs**:
  - **Q:** Why does the woman interrupt the conversation?
    - **A (correct):** She wants to change the topic.
    - **A (distractor):** She feels left out of the conversation.
  - **Q:** Why does the guy with the suit move closer to the door during the conversation?
    - **A (correct):** He intends to leave soon and is trying to escape the conversation.
    - **A (distractor):** He wants to end the conversation and this is a cue for the other person to leave.

Percepts

- **Overview**: Understanding sensory perspectives and their influence on actions or understanding. Focus on scenarios where participants rely on or lack their senses to interpret events.
- **What to Look For:**
  - **Simple Visual Perspective Taking**: Recognizing that others see different things.
  - **Complex Visual or Auditory Perspective Taking**: Adopting another person's visual or auditory perspective in complex scenarios.
  - **Percept-Action Link**: Connecting actions to specific sensory perceptions.
- **Example Question-Answer pairs**:
  - **Q:** Why does the woman states her partner is cheating on her?
    - **A (correct):** She only heard part of a conversation that makes her husband look guilty.
    - **A (distractor):** She observed his husband flirtatious attitude with his colleague, which hints that he is cheating.
  - **Q:** Why does the old man not try to catch the falling vase?
    - **A (correct):** He didn't see it, and therefore didn't try to catch it.
    - **A (distractor):** He knew that even if he tried to catch it it would be too late.

Non-Literal Communication

- **Overview**: Understanding indirect, non-literal expressions in communication. Keep an eye on tone, timing, and reactions of actors to identify indirect meanings.
- **What to Look For:**
  - **Sarcasm/Irony**: Recognizing irony or sarcasm, when someone says the opposite of what they mean.
  - **Egocentric Lies**: Identifying deliberate, self-serving lies motivated by personal interest rather than considering others' feelings or social harmony.
  - **White Lies**: Understanding when an actor lies to protect others' feelings.
  - **Involuntary Lies**: Recognizing when someone unintentionally conveys incorrect information.
  - **Humor**: Understanding jokes or humor as non-literal communication.
  - **Faux Pas**: Recognizing unintentional social errors, such as when someone accidentally says (or does) something that offends or embarrasses another person because they are unaware of the social or emotional implications of their action.
- **Example Question-Answer pairs**:
  - **Q:** Why does the man says he is stuck in traffic?
    - **A (correct):** He's lying to avoid a conflict with his boss.
    - **A (distractor):** He's lying to avoid worrying his boss and colleagues.
  - **Q:** Why does the woman says "Great! Another meeting"?
    - **A (correct):** She's expressing frustration because she doesn't want another meeting
    - **A (distractor):** She's glad about another meeting since the last one went wrong and she can emend her mistakes.

Examples on short films.

We provide some examples on videos along with their designed Theory of Mind abilities and Social Markers. Note that some questions may not have these markers.

Your are NOT tasked to write the explanations. We include them so you better understand the task.

Video 1: Alex

- Q: Why is the girl with a gray t-shirt looking at her friend?
  - A: To hint that she expects to have support from her in the current situation.
  - Timestamp: `5:19`
  - ToM ability: **Intentions**
    - Why his ToM ability? The question asks about the intentions of the girl with a gray t-shirt when looking at her friend. Not how she feels (Emotions), or what she wants from her (Desires)

- Multimodal Signals: **Face Expression**
  - Why these Signals?: The facial expression on the girl in gray shows her outrage at the situation and a potential call for support from her friend.
- Q: Why is the girl with a hat picking that many clothes for her friend to try?
  - A: She hopes this can help her get a job in that store by leaving a good impression with the shopkeeper.
  - Timestamp: `3:05`
  - Multimodal Signals: **Intentions**, **Desires**
    - Why his ToM ability? The question asks about the reason for the actions of the girl with a hat. We must understand what motivates her (Desires) and how she tries to obtain this (Intentions).
  - Multimodal Signals: **Face Expression, Body Language**
    - Why these Signals?: Her facial expression and body language convey information about her excitement about the potential opportunity of working in that store.
- Q: What does the girl with a gray t-shirt believe is the reason her friend is asking her to try different clothes?
  - A: She believes her friend wants to make a good impression on the shop clerk.
  - Note: This example is different from the previous one because we are asking about how one of the characters reads the other (second-order Theory of Mind)
  - Timestamp: `3:18`
  - ToM ability: **Beliefs**
    - Why his ToM ability? The question asks about the beliefs of the girl with a gray t-shirt about her friend's behavior. We are not asking about the friend's behavior but what the girl with a gray t-shirt believes (Beliefs) about it.
  - Multimodal Signals: **Face Expression, Body Language**
    - Why these Signals?: Same as before
- Q: Which two emotions is the shop clerk probably feeling after the bra is snatched from her hand?
  - A: Embarrassment and remorse
  - Timestamp: `8:03`
  - ToM ability: **Emotions**
    - Why his ToM ability? We are directly asking about emotions.
  - Multimodal Signals: **Facial Expression**
    - Why these Signals?: We can answer this by looking at the context of the interaction (in which she falsely blames the customer of stealing) which leads to the customer "snatching" the product from her hands and the clerk reacting with an embarrassed expression.

Video 2: Fault

- Q: Why do the player with a white cap and the player with a black t-shirt turn to their friend after the coach asks them to pay his fees?
  - A: They realized their friend had lied to them by hiring someone to play with them.
  - Timestamp: `4:33`
  - ToM ability: **Non-literal communication, Knowledge**
    - Why his ToM ability? Understanding the interaction to answer the question requires understanding that the man with a green t-shirt lied to his friends to beat them at tennis (egocentric lies - Non-literal communication), that they were unaware of this and that they have realized this now (Knowledge) .
  - Multimodal Signals: **Facial Expression (Gaze)**
    - Why these Signals?: By observing the faces of the other players and their head movements, we know they realized they had been lied to.
- Q: Why are the other tennis players calling out the player's name with purple shirt.
  - A: They are unaware he has passed away and believe he can't hear them.
  - Timestamp: `1:35`
  - ToM ability: **Knowledge, Percepts, Beliefs**
    - Why his ToM ability? Answering this requires understanding that the friends are unaware that their friend is dead (Knowledge), and thus, they believe he is alive and can't hear them (Beliefs, Percepts). However, we know this is not the case and he can't hear them (Percepts).
  - Multimodal Signals: N/A
    - None of the defined social signals are required to answer this question.
- Q: **Why does the man with a white cap ask the man in a blue t-shirt to let him know when he is ready?**
  - A: To pressure him to lace his shoelaces faster and get back into the game.
  - Timestamp: `3:32`
  - ToM ability: **Intentions**, **Non-literal communication**
    - Why his ToM ability? In this case, the comment is said to achieve something (Intentions) to make the other player hurry. However, this is said in a slightly sarcastic/joking way (Non-literal communication), not directly asking him to hurry but asking to let them know when he is done lacing his shoes (something that should be done really quickly).
  - Multimodal Signals: **Speech-related, Body language**
    - Why these Signals?: We can tell this is a comment to pressure him, given the body language of the other players (eager to resume the game) and the tone of voice in which the man with a white cap says this.

Suggested Workflow for Annotation

1. **First Watch: Understanding the Story**
   - Watch the short film and focus on understanding the overall narrative and context
   - Pay attention to key characters and their relationships
   - Note the general emotional tone and setting
2. **Identifying Moments Linked to ToM Abilities**
   - Mark the cues that may be related to ToM abilities assigned to you, for example:
     - Significant character interactions
     - Changes in emotional states
     - Moments of misunderstanding or revelation
     - Non-verbal cues (gestures, expressions)
     - Important dialogue exchanges
   - Mark timestamps of these moments
3. **Question Creation**
   - For each identified moment and marked timestamp:
     - Create a challenging question that requires understanding the context.
     - Write the correct answer based on observable evidence
     - Mark the toggles of the multimodal signals required to answer (if any)
4. **Quality Check:** Verify each question:
   - Can it be answered using only the interval?
   - Does it test the assigned ToM ability?
   - Is the question as challenging as it could be?
   - Compare with provided examples for reference.

Questions to Detect Multimodal Signals in Question-Answer Pairs

Marking the toggles should not take much of your time; you can use these simple questions to quickly determine wether you should mark the toggle for each of the annotation cues.

- Gaze & Facial Expressions
  - Does answering this question require noticing where characters are looking or their eye contact patterns?
  - Does this question involve interpreting facial emotions or expressions?
  - Is shared visual attention (multiple people focusing on the same thing) important for answering this question?
- Body Language
  - Does answering require interpreting a character's posture or stance?
  - Is physical proximity or touching between characters relevant to the answer?
  - Are hand gestures, head movements, or other body movements essential for understanding the situation?
- Speech-Related Cues
  - Does tone of voice (rather than just the words) matter for answering correctly?
  - Are nonverbal vocalizations (sighs, laughter, gasps) important to the question?
  - Do listening behaviors (nods, "uh-huh," etc.) play a role in understanding the interaction?

DISTRACTOR Annotation Guidelines

Your Task

As an annotator, you will be given a set of short films and corresponding questions about Theory of Mind with their correct answers. For each question, your job is to create three **distractors**.

> A **distractor** is an incorrect answer choice in a Multiple-Choice Question (MCQ) that appears plausible at first glance but is ultimately wrong. Good distractors should confuse/mislead a test-taker who relies only on superficial cues. They should force the test-taker analyze relevant context—whether visual, auditory, or textual—to identify the correct answer.

While creating these distractors, you are allowed to modify the original question or answer if you find inconsistencies or errors, or if modifications would help you to generate more coherent and challenging distractors.

The end goal is to ensure that the evaluated system is understanding the interactions the video to distinguish the correct answer from the distractors, rather than relying on shortcuts or superficial hints. By doing so, we aim to test Theory of Mind understanding while making the questions as challenging as possible without making them unsolvable.

A central challenge in this process is avoiding **shortcut effects (**situations where an AI system can guess the correct answer without analyzing the video, purely by exploiting superficial patterns in the question or answer choices and *guessing)*. For this, you will also have a copilot AI that can detect possible shortcut effects in your proposed distractors, helping you refine them until they are both plausible and ultimately incorrect.

In the annotation platform, you will also find three toggles:

- ☐ Invalid Question (Mark if the question is not asking about Theory of Mind abilities)
- ☐ Invalid Answer (Mark if the answer for the question is incorrect)
- ☐ Incorrect timestamps (Mark if the timestamps are incorrect)

Since you will be creating distractors for the QA pairs of other annotators, we provide this so you can mark the samples that are incorrect.

Below, we provide guidelines to create hard distractors and reduce shortcut effects.

Please read the annotation guidelines in this document as well as the provided examples.

General Guidelines for Distractor Creation

What makes hard distractors?

- **Requires Understanding of Visual Cues / Speech Tone:** The distractor should only be disproven by paying attention to how a character speaks, reacts, or what is shown in the scene.
- **Requires Understanding the Larger Context:** The distractor should require linking details across larger timeframes.
- **Is Partially Correct but Has a Key Inaccuracy:** The distractor may describe the correct setting or actions but changes a key detail.

Things to avoid when creating distractors.

- Distractors that could be considered "technically correct" - they must be definitively wrong.
  - Example:

    Why does the woman feel ignored while talking to the man?
    - (correct answer) Because he is too focused on his phone.
    - ❌ (distractor) Because she believes he more interested in something else.
      - This is not a good distractor, as it is technically correct!
    - ✅ (distractor) Because he is only waiting for her to finish speaking so he can speak.
      - On the other hand, this distractor is definitely not correct, but could be plausible. The only way to know this would be to observe the interaction.
- Distractors that imply mean or insensitive behavior in humans (AI systems tend to be biased toward predicting empathetic behavior).
  - Example:
    What does the man believes his daughter is trying to achieve by blocking the door?
    - (correct answer) He believes she is trying to get him to change his mind about grounding her.
    - ❌ (distractor) He believes she is doing this because she wants him to be late for work.
      - This is not a good distractor (unless it makes sense because of the particular context). Models will quickly discard this as it implies that the father believes his daughter means harm for him.
    - ✅ (distractor) He believes she is doing this because she wants him to spend more time with her than he currently does.
- Distractors that can be answered by extracting the answer from the dialogues. In this case we are not really evaluating interaction understanding but extractive QA skills.

Understanding Shortcut Effects

- Shortcut effects occur when annotators create distractors that are easy to eliminate without actually understanding the video context. This is very common and leads to models being able to guess the correct answer by exploiting patterns in the answer choices rather than truly understanding the scene.

Tips for creating good distractors and avoiding shortcut effects.

- **Use the copilot!** You will be provided with an annotation platform which immediately catches if your answers contain shortcut effects. Use it to remove them from your questions (but be careful to not over-rely on this to create unanswerable questions with no correct answer).

- Start by copying the correct answer and modify it in ways that make it incorrect while maintaining similar length and style.
- Ensure all distractors are equally plausible at first glance.
- Match the complexity level and detail of the correct answer.

Guidelines for specific Theory of Mind abilities

Refer to the ***Question Creation Guidelines*** for a specific description of each of the Theory of Mind abilities.

Examples of short films.

Here, we extend the examples from the question creation guidelines example with distractors and explain why the distractors here are considered *valid distractors (incorrect)*.

Distractors should sound like possible correct answers if you look at the question, answer, and distractors. However, once we watch the video, it must be clear that they are incorrect. Here, we provide several examples of different distractors per question and explain why they are incorrect.

Your are NOT tasked to write the explanations. We include them so you better understand the task.

Video 1: Alex

- Q: Why is the girl with a gray t-shirt looking at her friend?
  - A: To hint that she expects to have support from her in the current situation.
  - Timestamp: `5:19`
  - ToM ability: **Intentions**
  - Distractor 1: To hint that she knows that the current situation is her fault that the bra disappeared.
    - Explanation: There is no evidence in either the body language or the context that makes us think that the girl with a gray t-shirt is putting the blame on her friend. Therefore, this option is incorrect.
  - Distractor 2: To let her know that she knows that she is the one who has stolen the bra, but, as her friend, she should cover her.
    - Explanation: There is no evidence in the video of the friend stealing the bra or the girl with a gray t-shirt observing something that makes her think this. Therefore, this option is incorrect.
  - Distractor 3: To let her know that she will support her no matter what.
    - Explanation: In this case, the look is *asking* for support *from* her friend instead of conveying support to her friend. In this case, the direction of who is asking for support from whom is wrong.
- Q: Why is the girl with a hat picking that many clothes for her friend to try?
  - A: She hopes this can help her get a job in that store by leaving a good impression with the shopkeeper.
  - Timestamp: `3:05`
  - ToM ability: **Intentions**, **Desires**
  - Distractor 1: She hopes this will make her forget about the situation in her workplace.
    - Explanation: In this context, we see that she finds out they are hiring in the store, and then she puts much more effort into picking more clothes to try. Since her behavior changed after discovering this, we can discard this option. Therefore, this option is incorrect.
  - Distractor 2: She hopes this will convince her to buy a nicer bra than her current one.
    - Explanation: Same explanation as above.
  - Distractor 3: She hopes this will help her make a good impression with the store's owner.
    - Explanation: The store owner is not in the context and cannot possibly observe her actions. Therefore, this option is incorrect.
- Q: What does the girl with a gray t-shirt believe is the reason her friend is asking her to try different clothes?
  - A: She believes her friend wants to make a good impression on the shop clerk.
  - Note: This example is different from the previous one because we are asking about how one of the characters reads the other (second-order Theory of Mind)
  - Timestamp: `3:18`
  - ToM ability: **Beliefs**
  - Distractor 1: She believes her friend wants to make a good impression on the store's owner.
    - Explanation: The store owner is not even present in the store. Therefore, the friend could not possibly make a good impression on them. This option is incorrect.
  - Distractor 2: She believes her friend wants to buy her something nice since her job has been going well.
    - Explanation: Given the facial cues and body language of the friend *after* her interaction with the shop clerk, as well as the comment, "I want to make a good impression", it's very unlikely that the girl in the gray t-shirt believes that her friend is doing this because her job has been going well (also, if we look at the interaction, things aren't great at her job, making this distractor even more incorrect).
  - Distractor 3: She believes her friend wants her to buy herself something nice because she rarely does it.
    - Explanation: Given the facial cues and body language of the friend *after* her interaction with the shop clerk, as well as the comment, "I want to make a good impression", it's very unlikely that the girl in the gray t-shirt believes that her friend is doing this because she wants her to buy herself something nice because she rarely does it. Also nowhere in the video is it mentioned that she rarely buys herself nice things.
- Q: Which two emotions is the shop clerk probably feeling after the bra is snatched from her hand?

- A: Embarrassment and remorse
- Timestamp: `0:03`
- ToM ability: **Emotions**
- Distractor 1: Surprise and fear
  - Explanation: The face does not show fear nor surprise but clear embarrassment and remorse. Also, fear would not make sense in the context.
- Distractor 2: Surprise and disgust
  - Explanation: The face does not show disgust nor surprise but clear embarrassment and remorse. Also, disgust would not make sense in the context.
- Distractor 3: Sadness and disgust
  - Explanation: The face does not show sadness nor disgust but clear embarrassment and remorse. Also, neither disgust nor sadness would make sense in the context.

Video 2: Fault

- Q: Why do the player with a white cap and the player with a black t-shirt turn to their friend after the coach asks them to pay his fees?
  - A: They realized their friend had lied to them by hiring someone to play with them.
  - Timestamp: `4:33`
  - ToM ability: **Non-literal communication**, **Knowledge**
  - Distractor 1: They know he is the one who hired him, and they were expecting him to pay his fees.
    - Explanation: Given their conversation, we can tell they were unaware he had hired the couch. They likely expect him to pay for his fees, but they were unaware that the player was hired, which makes this option incorrect.
  - Distractor 2: They are confused because they initially thought he would not charge them anything.
    - Explanation: Same as the previous, if we had the information they knew from before the player was hired, this answer could be possible, but since they didn't know, we can immediately discard it. This option is incorrect.
  - Distractor 3: They know he is the one who hired him, but they want to be the ones who pay the fees out of gratitude.
    - Explanation: In this case, their expression does not convey gratitude, and they were unaware that the player was hired. This option is incorrect.
- Q: Why are the other tennis players calling out the player's name with purple shirt.
  - A: They are unaware he has passed away and believe he can't hear them.
  - Timestamp: `1:35`
  - ToM ability: **Knowledge**, **Percepts**, **Beliefs**
  - Distractor 1: They are not aware that he is dead, and they believe he may be ignoring them.
    - Explanation: Given the context of the video, in which at the beginning we can see that the player with the purple shirt is a bit deaf and the others are aware of this, it is much more likely that "they believe he can't hear them" is "they believe he can't hear them". Therefore, this option is incorrect.
  - Distractor 2: They are terrified that he is dead, and they are calling out his name out of grief and desperation.
    - Explanation: The fact that they are calling him from afar lets us know that they are not close enough to observe (and realize) that he has passed away. Also, the tone of voice in which they call his name does not convey grief or desperation. Therefore, this option is incorrect.
  - Distractor 3: They are terrified that he may be dead, and they are calling out his name out of grief and desperation in hopes that some help will arrive.
    - Explanation: The fact that they are calling him from afar lets us know that they are not close enough to observe (and realize) that he has passed away. Also, the tone of voice in which they call his name does not convey grief or desperation.
- Q: **Why does the man with a white cap ask the man in a blue t-shirt to let him know when he is ready?**
  - A: To pressure him to lace his shoelaces faster and get back into the game.
  - Timestamp: `3:32`
  - ToM ability: **Intentions**, **Non-literal communication**
  - Distractor 1: To implicitly ask him to play easy on him as he is much older than **the man in a blue t-shirt**.
    - Explanation: By reading the players' body language, we cannot see signs of the man with a white cap feeling intimidated by the man in a blue t-shirt. On the contrary, he seems eager to start playing. Therefore, this question is incorrect.
  - Distractor 2: So he is not surprised by the man in a blue t-shirt when he returns to the game.
    - Explanation: We can see that **the man in the blue t-shirt** is taking an unusual amount of time to lace his shoes; while this happens, the other players look eager to begin the game; judging by their body language, it is very unlikely that they this he will surprise them, so this option is definitely incorrect.
  - Distractor 3: To distract him while he laces his shoelaces in hopes of getting an edge in the tennis match.
    - Explanation: This distractor is wrong because the person who is lacing the shoes is not the same person asking the question.

Suggested Workflow for Annotation

For each question, we estimate that it should take you an average of 6 minutes to create distractors (depending on the complexity of the question). Questions are ordered by video, so you will create distractors for each video simultaneously.

Here is the proposed pipeline for annotating each question:

1. **Review**
   - Read the question and correct answer
   - If you need, rewatch the marked video interval
   - Verify the question and answer are correct; if not, modify accordingly
2. **Create Distractors**
   - Create 3 variations that are plausible but incorrect
   - Keep a similar length and style to the correct answer
   - Ensure distractors relate to the same Theory of Mind ability as the question
3. **Refine with Copilot**
   - Evaluate your distractors for shortcuts using the system
   - Revise and recheck with copilot until the distractors pass the test
   - You can use the copilot system as many times as needed
4. **Final Check**
   - Read all options together (correct answer + distractors)
   - Verify there is only one clearly correct answer
   - Submit your completed annotation