Towards Online Multimodal Social Interaction Understanding

Anonymous authors
Paper under double-blind review

Abstract

In this paper, we introduce a new problem, Online-MMSI, where the model must perform multimodal social interaction understanding (MMSI) using only historical information. Given a recorded video and a multi-party dialogue, the AI assistant is required to immediately identify the speaker's referent, which is critical for real-world human-AI interaction. Without access to future conversational context, both humans and models experience substantial performance degradation when moving from offline to online settings. To tackle the challenges, we propose Online-MMSI-VLM, a novel framework based on multimodal large language models. The core innovations of our approach lie in two components: (1) multi-party conversation forecasting, which predicts upcoming speaker turns and utterances in a coarse-to-fine manner; and (2) socially-aware visual prompting, which highlights salient social cues in each video frame using bounding boxes and body keypoints. Our model achieves state-of-the-art results on three tasks across two datasets, significantly outperforming the baseline and demonstrating the effectiveness of Online-MMSI-VLM.

1 Introduction

Multimodal Social Interaction Understanding (MMSI) plays a critical role in advancing human-AI social interaction, aiming to interpret social behaviors by jointly leveraging verbal and non-verbal cues (Lee et al., 2024b; Lai et al., 2023; Lee et al., 2024a). Recent work has explored tasks such as speaking target identification, pronoun coreference resolution, and mentioned player prediction. One common goal is to identify a speaker's referent among multiple participants using recorded dialogues and video streams. For example, as illustrated in Figure 1 (1), when a user asks *Which player is the speaker referring to?* at a moment, the model analyzes the multimodal input and responds, "Player0." MMSI can enable applications such as AI-powered assistants for social support and collaborative tasks. Smart AR glasses, for instance, can help autistic individuals better understand social cues (Haber et al., 2020; Elsherbini et al., 2023), while intelligent assistants can actively participate in social settings (Breazeal et al., 2016). As a result, MMSI has attracted growing interest from the social AI research community (Lee et al., 2024b; Li et al., 2024a; Feng et al., 2025).

Prior MMSI studies focused on *offline* setting, conducting referent identification by leveraging both past and future context at a given moment (Lee et al., 2024a). They rely on extended transcripts and video data, producing responses with inherent delays. However, real-world AI assistants are expected to respond in real time, interpreting and responding to social dynamics without access to future information (i.e., *online* setting). To bridge this gap, we introduce Online-MMSI, a new problem formulation where models must perform MMSI tasks using only historical information. This setting is crucial for building responsive, real-time intelligent systems that are practical and deployable in real-world social environments.

Yet, Online-MMSI presents substantial and unique challenges compared with traditional offline settings. As shown in Figure 1 (2), performance drops significantly when both models and humans shift from offline to online settings across three social tasks. First, online models lack access to delayed and explicit cues, such as future transcripts or the referent's responses. For example, the referent may respond directly to the initial speaker, or other participants might clarify or reiterate the referent in later turns. Traditional offline MMSI model (Lee et al., 2024a) rely on seeing the future social cues directly, making it not an ideal solution for real-

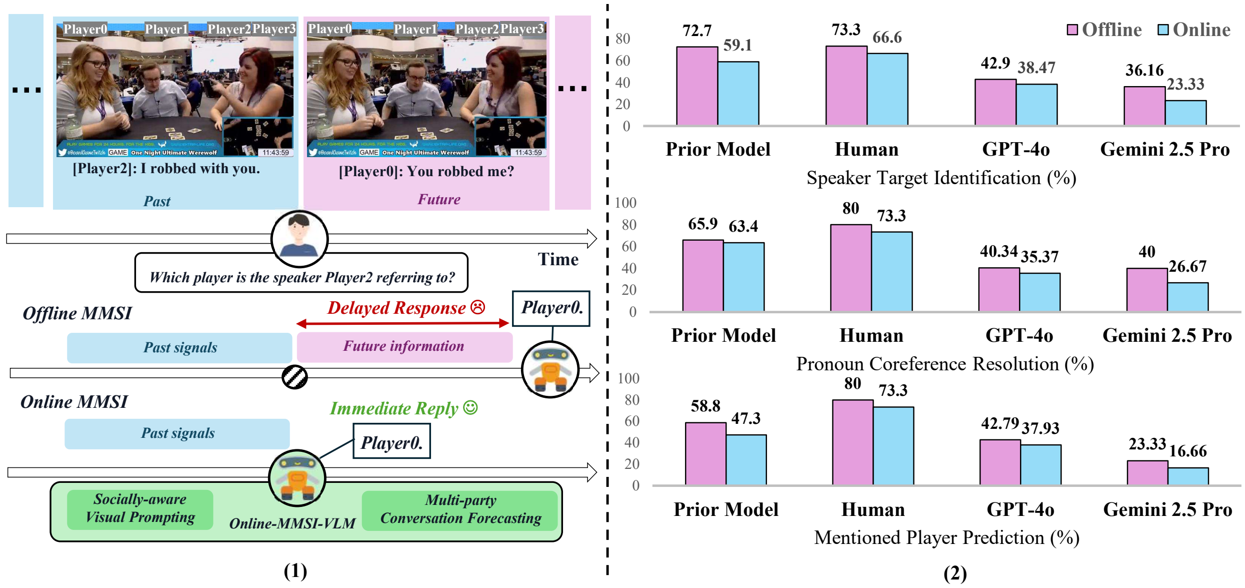


Figure 1: (1) Existing MMSI studies depend on both past and future context (offline setting), limiting their practicality in real-world scenarios that require immediate interpretation and response. To bridge the gap, we introduce a new problem, Online-MMSI, where the model must perform MMSI using only historical information (online setting). To address the challenge, we propose Online-MMSI-VLM, a novel multimodal large language model-based framework that integrates multi-party conversation forecasting and socially-aware visual prompting. (2) Performance drops significantly when both models and humans shift from offline to online settings across three social tasks on the YouTube dataset (Lai et al., 2023).

time scenarios where such cues are unavailable. Second, online models must rely entirely on immediate and past signals, where subtle social cues–such as pointing gestures and posture, head turns, or body orientation–are crucial for resolving referents. Figure 1 (1) illustrates such a case: the speaker is gesturing toward the referent with the index finger, serving as critical disambiguating contexts for referent resolution. Since social scenes often involve multiple participants engaged in dynamic interactions, it is difficult to detect such subtle yet informative social signals from raw RGB video frames alone without explicit guidance.

To address these challenges, we propose Online-MMSI-VLM, a novel multimodal large language model-based framework that integrates multi-party conversation forecasting and socially-aware visual prompting. For the first challenge, cognitive studies suggest that humans improve online interpretation by anticipating upcoming social interactions (Epperlein et al., 2022; Ma et al., 2024; Hadley & Culling, 2022). This motivates us to leverage multimodal large language models to anticipate future conversations to enrich social context. Our coarse-to-fine strategy first predicts the identity of the upcoming speaker, followed by generating their likely utterance. For the second challenge, we propose to enhance the representation of past video by using socially-aware visual prompting, which employs off-the-shelf detectors to extract and highlight the positions, postures, and gestures of participants. This facilitates the model to attentively interpret subtle and complex social interactions in the previous video. By jointly forecasting future dialogue and enriching past visual context, our framework can effectively tackle the challenges of Online-MMSI. We evaluate our approach using two VLM models on three referent identification tasks within two social datasets (Lai et al., 2023). The results indicate that Online-MMSI benefits from multi-party conversation forecasting and socially-aware visual prompting, and Online-MMSI-VLM achieves state-of-the-art performance.

In summary, our contributions are threefold: 1) We propose a new task, Online-MMSI, where the AI assistant must interpret multimodal social interactions and provide immediate feedback using only historical dialogues and videos. 2) We propose Online-MMSI-VLM, a multimodal large language model-based framework that integrates multi-party conversation forecasting and socially-aware visual prompting. To the best of our knowledge, it is the first work to use multimodal large language models for advancing MMSI. 3) Extensive experiments, for three tasks on two social datasets, demonstrate the effectiveness of our approach and show Online-MMSI-VLM establishes a strong benchmark for Online-MMSI.

2 Related Works

2.1 Multimodal Social Interaction Understanding

Multimodal Social Interaction Understanding (MMSI) aims to interpret complex interactions among multiple participants by leveraging both verbal and non-verbal cues, which has attracted increasing attention in the machine learning and social AI communities. For non-verbal social cues, some studies have attempted to perceive social meaning from visual behaviors such as body gestures (Benitez-Garcia et al., 2021; Liu et al., 2021; Chen et al., 2023a; Kapitanov et al., 2024), gaze patterns (Jindal & Manduchi, 2023; Chong et al., 2020; Grauman et al., 2022; Tafasca et al., 2023), and facial expressions (Zhang et al., 2021; Savchenko, 2023; Zhao et al., 2024; Li et al., 2024b). For verbal social cues, prior works have explored social understanding from linguistic signals such as game-theoretic agents (Feng et al., 2025), speaker intent (Malhotra et al., 2022; Chen et al., 2023c), and dialogue sentiment (Feng et al., 2021). For multimodal social cues, several studies integrate verbal and non-verbal modalities to holistically interpret social interactions, such as recognizing emotions (Cheng et al., 2024a; Lian et al., 2025; 2024), conversational dynamics (Raman et al., 2022; Ryan et al., 2023; Hou et al., 2024), and social situations (Hyun et al., 2023; Guo et al., 2025).

Recently, Li et al. (2024a) integrate the perception capability of vision models and the reasoning capability of LLMs for social relation recognition. Gupta et al. (2024) introduce a novel framework to jointly predict the gaze target and social gaze label for all people in the scene. Lee et al. (2024b) conduct a comprehensive survey, identifying three keys for effective social understanding: multimodal social cues, multi-party dynamics, and beliefs. Lai et al. (2023) introduce a multimodal dataset for modeling persuasion behaviors. Cao et al. (2025) introduce a large-scale dataset for multi-person gesture analysis. Lee et al. (2024a) introduce three tasks to model the fine-grained dynamics between multiple people: speaking target identification, pronoun coreference resolution, and mentioned player prediction. To address these tasks, Lee et al. (2024a) introduce a densely-aligned multimodal framework to capture social cues from transcript and video.

Despite these advances, current research typically focuses on the *offline* setting, where the model can have access to entire video frames and dialogues to make predictions. However, this setup does not align well with real-time requirements in interactive AI assistant systems, where the model can only leverage historical information (i.e., the *online* setting). Although methods from recent research (Lee et al., 2024a) can theoretically be adapted to online settings through data rearrangement, they do not adequately account for the challenge of missing future social cues and exploiting historical information in the online setting. In our work, we investigate the online setting for MMSI and propose novel methods to address the challenge.

2.2 Multimodal Large Language Models

Multimodal large language models (MLLMs) have demonstrated outstanding performance in vision-language tasks (Liu et al., 2023; Team et al., 2024; Team, 2025; Zhang et al., 2023a; Zhu et al., 2023; Zhang et al., 2023b; Chen et al., 2023b; Lin et al., 2023; Chen et al., 2024b; Cheng et al., 2024b; Zhang et al., 2024b; Thawakar et al., 2025; Zhang et al., 2024c). This success is attributed to their strong perception and reasoning capabilities, which benefit addressing dynamic multimodal social interaction understanding.

Online scenarios attract growing attention (Wang et al., 2021; Zhao & Krähenbühl, 2022; Girdhar & Grauman, 2021; Zhao et al., 2023). While several online MLLMs have been developed for training and deployment (Chen et al., 2024a; Liu et al., 2024; Fu et al., 2025), our work aims to conduct Online-MMSI.

Conversation forecasting has been explored for conversation agents (Hassan et al., 2024; El Hattami et al., 2023), predicting derailment (Chang & Danescu-Niculescu-Mizil, 2019), and harmful behaviors (Sicilia & Alikhani, 2024). These studies mainly focus on conversation between agents and users. Instead, we focus on understanding multi-party conversations of multimodal social events.

Visual prompting has also proven effective in enhancing MLLM performance (Wu et al., 2025; Cai et al., 2024; Wu et al., 2024; Lei et al., 2024; Yang et al., 2023; Xu et al., 2024). Some studies show that attentively leveraging the history is crucial for improving online video understanding in noisy and large-volume visual data (Yao et al., 2025; Zhang et al., 2024a; Huang et al., 2024; Chandrasegaran et al., 2024; Patil et al., 2024). In this work, we propose to leverage visual prompting to enhance historical video for Online-MMSI.

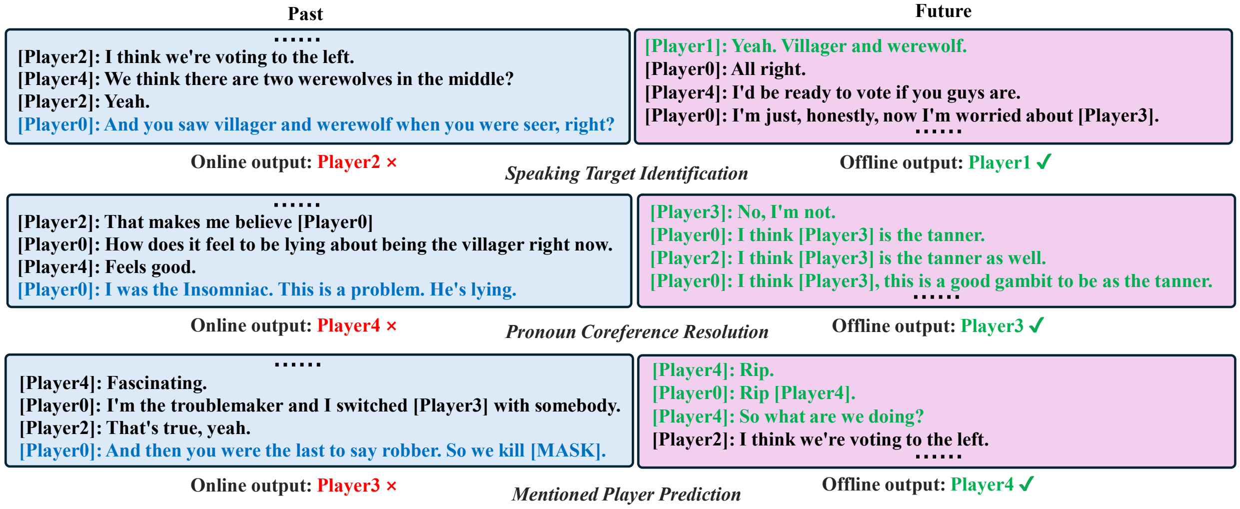


Figure 2: Illustration of the challenge of missing future information when transitioning from the offline setting to the online setting. In the online setting, only immediate (blue font) and past data (black) is available, whereas the offline setting additionally includes useful future context (green), such as the referent's responses. As a result, the online model may be confused and provide an incorrect answer.

3 Problem Formulation and Challenges

Following Lee et al. (2024a), we focus on three social interaction understanding tasks: speaking target identification, pronoun coreference resolution, and mentioned player prediction, using video and its corresponding transcript. The online task formulation is as follows: At timestep t, the system receives:

- **User prompt** (P_t): a query related to the specific task. The prompts for the three social tasks are "Identify which player the speaker is talking to?", "Determine which player a pronoun refers to?", and "Predict which player is mentioned by name?", respectively.
- **Historical dialogues** ($\mathcal{D}_{(t-d):t}$): the most recent d time steps of dialogue history. Each entry consists of both the speaker identity and the corresponding utterance. E.g., "[Player0]: So, you're a mason? [Player0]: I'm the troublemaker. [Player3]: Did you swap me with anybody?"
- **Recorded video frames** ($\mathcal{V}_{(t-d):t}$): the most recent d time steps of video frames involving multiple participants. These frames include dynamic non-verbal social cues, such as gestures and gazes.

The AI assistant is required to provide an immediate reply:

- **Response** (R_t): the answer to the referent identification query. For instance, "Player0".

Given P_t, $\mathcal{D}_{(t-d):t}$, and $\mathcal{V}_{(t-d):t}$, the objective is to optimize the model to generate an accurate R_t.

In the offline setting, the model has access to both past and future data–P_t, $\mathcal{D}_{(t-d):(t+d)}$, and $\mathcal{V}_{(t-d):(t+d)}$–and generates a response $R_{(t+d)}$ at timestep $t+d$. In contrast, the online setting restricts access to only historical video and dialogue, cutting off any future context. To quantify the impact of this constraint, Figure 1 (2) compares performance across the three social tasks under both settings using the YouTube dataset (Lai et al., 2023). We evaluate the previous method (Lee et al., 2024a), human participants, GPT-4o (OpenAI, 2024), and Gemini 2.5 Pro (Comanici et al., 2025). To ensure fairness in the user study, each participant first viewed the online input, followed by the offline version. This design mitigates potential bias by preventing participants' online judgments from being influenced by future context seen in the offline setting.

As shown in the results, there is a considerable drop in accuracy across all three tasks when both humans and models when transitioning from offline to online scenarios. The prior model (Lee et al., 2024a) suffers 13.6% in speaking target identification, 2.5% in pronoun coreference resolution, and 11.5% in mentioned player

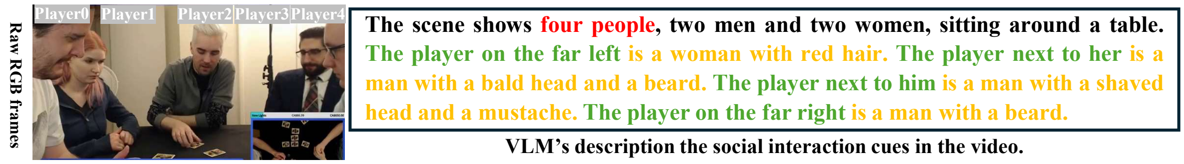


Figure 3: Illustration of the challenge in relying solely on immediate and past context when transitioning from offline to online settings, where subtle social cues become increasingly critical. In complex multi-party interactions, such cues are often too nuanced to be reliably inferred from raw RGB video alone, especially in the absence of explicit visual prompting. When asked to describe social interactions, the VLM frequently produces coarse spatial references (green), generic appearance descriptions (yellow), or even hallucinated content (red), while failing to capture the subtle and intricate social signals in multi-person dynamics.

prediction. The average accuracy of humans, GPT-4o, and Gemini 2.5 Pro drops by around 7%, 5%, and 10%, respectively, confirming that the performance gap stems from task difficulty. The evaluation results of GPT-4o and Gemini 2.5 Pro underscore that current LLMs still struggle with understanding multimodal multi-party social interaction compared with humans (Inoue et al., 2025; Tan et al., 2023).

This performance degradation arises from two fundamental limitations. First, online models lack access to delayed and explicit cues–such as future transcripts or the referent's responses–that are available in offline settings. The referent frequently replies to the speaker or is clarified by other participants in subsequent turns. For instance, as seen in Figure 2, in speaking target identification, after the speaker says "[Player0]: And you saw villager and werewolf when you were seer, right?", the referent later replies "[Player1]: Yeah. Villager and werewolf." Offline models can leverage such useful future context to correctly identify "Player0" as the referent. In contrast, online models must rely solely on historical information, which may be ambiguous. For example, in mentioned player prediction, before the speaker says "[Player0]: And then you were the last to say robber. So we kill [MASK].", a previous utterance states "[Player0]: I'm the troublemaker and I switched [Player3] with somebody." The online model may be confused and incorrectly answer "Player3".

Second, online models must resolve social references based entirely on immediate and past signals, where subtle social cues–such as pointing gestures, posture, head turns, and body orientation–become essential. However, in complex interactions involving multiple participants, such signals are often too subtle to interpret from raw RGB video alone, especially without structured visual guidance. Experimental results, as shown in Figure 3, indicate that VLMs fail to capture detailed and accurate social cues from raw RGB video. When prompted to describe the social interaction cues in the video, the VLM typically returns only coarse attributes–such as the relative positions and appearances–using phrases like "on the far right" and "shaved head and a mustache," and occasionally produces hallucinations, such as incorrectly stating "four people."

4 Methodology

To address these challenges, we introduce Online-MMSI-VLM, a framework designed to respond to social tasks using only historical input. As illustrated in Figure 4, the model takes as input the user prompt, historical utterances, and video frames, and produces a real-time response under the online setting. To compensate for the lack of future information and enhance the quality of historical signals, we propose two key components: (i) *multi-party conversation forecasting*, which anticipates future dialogue turns; and (ii) *socially-aware visual prompting*, which guides the model to attend to socially relevant visual cues.

4.1 Multi-party Conversation Forecasting

To enable forecasting multi-party conversations, we append a forecasting query after the task prompt, which is: "Predict the upcoming speakers' turns and then predict the upcoming utterance of each speaker." After answering the Online-MMSI task, the model generates the next few speaker labels and utterances. Directly predicting the full utterances of each future speaker in a single step can be overly complex and may lead to sub-par generation. Instead, we propose a coarse-to-fine approach with the following two phases:

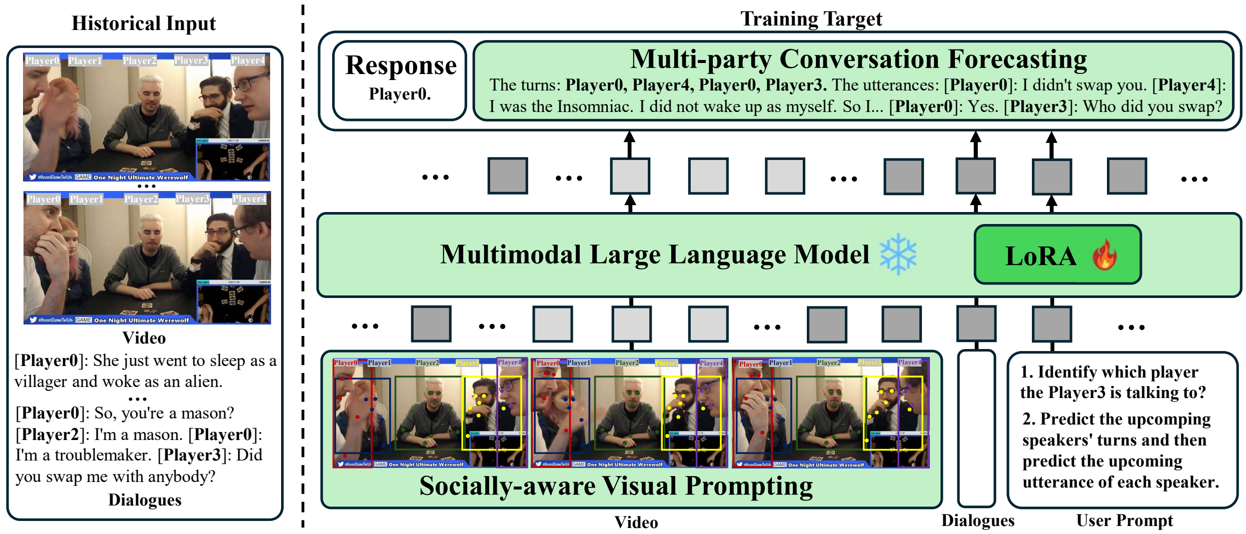


Figure 4: The training pipeline of Online-MMSI-VLM. The model takes the user prompt, historical dialogues, and recorded video as input and generates an immediate response. To tackle the challenges of Online-MMSI, we introduce two techniques: (i) *multi-party conversation forecasting* to enrich language context, and (ii) *socially-aware visual prompting* to facilitate historical social cues modeling.

- **Speaker-turn prediction**: The model first generates a sequence of four upcoming speaker identities, e.g., "The upcoming speakers' turns: Player0, Player4, Player0, Player3."
- **Utterance refinement**: For each predicted speaker, the model then produces a fine-grained utterance. For instance: "The upcoming utterances: [Player0]: I didn't swap you. [Player4]: I was the Insomniac. I did not wake up as myself. So I... [Player0]: Yes. [Player3]: Who did you swap?"

This strategy mimics human conversation patterns, where we first anticipate who might speak and in what order, and then try to guess what they would say. The above process can be formulated as:

$$\hat{S}_{t+1:t+K} = f_\theta^{\mathrm{VLM}}(\mathcal{D}_{(t-d):t}), \tag{1}$$

$$\hat{U}_{t+1:t+K} = f_\theta^{\mathrm{VLM}}\big(\hat{S}_{t+1:t+K}, \mathcal{D}_{(t-d):t}\big), \tag{2}$$

where $\hat{S}_{t+1:t+K}$ denotes the predicted sequence of speakers over the next K turns, and $\hat{U}_{t+1:t+K}$ represents the generated sequence of corresponding utterances for each speaker.

4.2 Socially-aware Visual Prompting

Since online models must rely on immediate and past information, capturing subtle visual cues is crucial for resolving referents. Social scenes often involve multiple participants engaged in dynamic non-verbal interactions, making it difficult to localize and interpret key visual cues such as posture and gestures. Without explicit guidance, the model may fail to attend to the socially salient regions in raw RGB image input, especially in cluttered situations. To address this, we propose socially-aware visual prompting, which explicitly highlights important social cues in the historical video frames. Specifically, we annotate all visible participants with: (a) speaker labels, aligning players with their spoken utterances; (b) bounding boxes, indicating each participant's spatial position; and (c) upper body keypoints, capturing posture, gaze direction, facial expressions, and gestures. These annotations are generated using AlphaPose (Fang et al., 2022).

As shown in Figure 4, we integrate these annotations by overlaying them directly onto the video frames. To distinguish between individuals, we assign a unique color to each person's annotation and include a system prompt specifying the mapping: "The red, blue, green, yellow, purple, and orange colors correspond to Player0, Player1, Player2, Player3, Player4, and Player5, respectively." The annotated frames $\hat{\mathcal{V}}_{(t-d):t}$ are fed into the visual encoder of the VLM, providing a more semantic representation of the social scene. This representation is then concatenated with the tokenized historical dialogues $\mathcal{D}_{(t-d):t}$ and the user prompt P_t. The final multimodal embedding is processed by the language head to produce a response R_t:

$$R_t \;=\; f_\theta^{\mathrm{VLM}}\big([\hat{\mathcal{V}}_{(t-d):t},\, \mathcal{D}_{(t-d):t},\, P_t]\big), \tag{3}$$

where R_t represents the predicted speaking target, the referent of a pronoun, or the name of a mentioned player. By visually highlighting key social signals, this prompting approach helps the model focus on the most relevant regions, improving its ability to interpret past visual context in the absence of future cues.

4.3 Supervised Instruction Tuning

To enhance the VLM's performance on specific social tasks, we first construct instruction–output pairs using existing social datasets (Lee et al., 2024a), incorporating future utterances from transcripts into these pairs. We then fine-tune the model via supervised learning with two objectives: (a) task-specific objective: the model learns to produce correct answers $\hat{y}_t$ for a specific Online-MMSI task; (b) forecasting objective: the model learns to anticipate future dialogue turns $\hat{S}_{t+1:t+K}$ and corresponding utterance content $\hat{U}_{t+1:t+K}$.

Let $\mathcal{L}_{\text{mmsi}}$ and $\mathcal{L}_{\text{forecast}}$ denote the loss for Online-MMSI and forecasting. Our training objective is:

$$\mathcal{L}_{\text{total}} \;=\; \mathcal{L}_{\text{mmsi}} \;+\; \mathcal{L}_{\text{forecast}}. \tag{4}$$

This multi-task formulation enables the model to not only answer current queries but also to anticipate future interactions, leading to better context understanding and improved performance in online social scenarios.

4.4 Implementation

We select **LLaMA-3.2-Vision-11B** (Dubey et al., 2024) and **Qwen2.5-VL-7B** (Team, 2025) as our multimodal large language models, both of which are representative state-of-the-art models. For Qwen2.5-VL-7B, we dynamically sample video at a rate of 1 frame per second (fps) as visual input. For LLaMA-3.2-Vision-11B, we sample six frames from each video clip evenly and arrange them into a single 3×2 grid-like image, following the suggestion from prior work (Kim et al., 2024). We apply instruction tuning with LoRA (Hu et al., 2022), targeting the query and value projection layers, to further specialize these models for our tasks. Specifically, we use an alpha value of 16, a dropout rate of 0.05, and a rank of 512 for the LoRA configuration. We train all new or adapted parameters with a unified cross-entropy loss. The learning rate is set to 1×10^{-4} for the speaker identification and mentioned player prediction tasks, and 1×10^{-3} for the pronoun coreference task. The batch size is set to 1, with gradient accumulation steps of 4. Each model is trained for 5 epochs on all three tasks. The historical window length d is set to 10 dialogue turns, and the forecasting horizon K is set to 4 utterance turns. The average duration of each dialogue turn is approximately 2.8 seconds. The tracking and keypoints for prompting are generated by AlphaPose (Fang et al., 2022) and curated using the reference frame from Lee et al. (2024a). More implementations are provided in the supplementary material.

4.5 Datasets

Experiments are conducted on the Werewolf Among Us dataset, which comprises two subsets (YouTube and Ego4D) of social deduction games (Lai et al., 2023). We follow the dataset and task setup of Lee et al. (2024a), while cutting the future video and transcript in each sample for the online setting.

YouTube contains 151 games of One Night Ultimate Werewolf, which corresponds to 151 separate videos with 14.8 hours and transcripts comprising 20,832 utterances. It has 3,255 samples for speaking target identification, 2,679 for pronoun coreference resolution, and 3,360 for mentioned player prediction.

Ego4D has 40 games of One Night Ultimate Werewolf and 8 games of The Resistance: Avalon. It contains 101 separate videos with 7.3 hours and transcripts containing 5,815 utterances, with 832 samples for speaking target identification, 503 for pronoun coreference resolution, and 472 for mentioned player prediction.

Evaluation. We follow Lee et al. (2024a) to report the overall accuracy of the predicted referent.

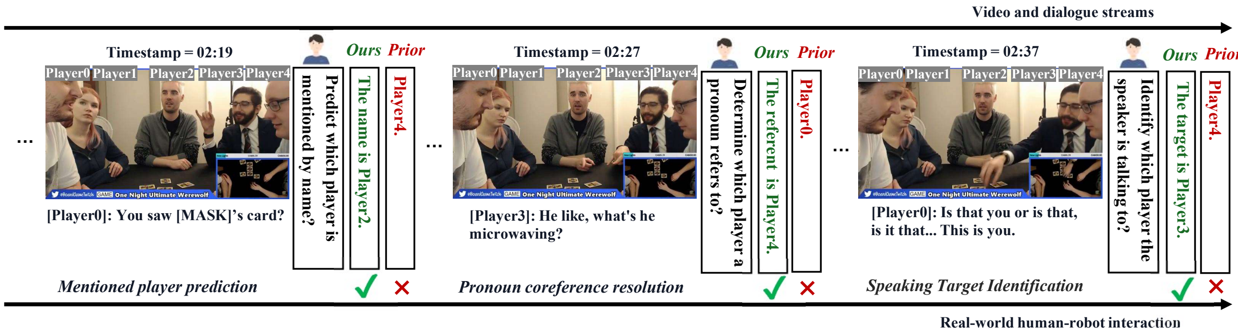


Figure 5: Qualitative results of Online-MMSI-VLM on three social tasks. Compared to the prior work (Lee et al., 2024a), our method provides immediate and accurate referents based on historical video and dialogue.

Table 1: Performance comparison (%) of different methods on three social tasks across the YouTube and Ego4D datasets. Our proposed method, Online-MMSI-VLM, achieves state-of-the-art performance across all tasks. **Bold** indicates the best result for each task, and ↑ denotes improvement over the baseline.

YouTube / Ego4D	Lee et al. (2024a)	Qwen (Ours)	LLaMA (Ours)
Speaking Target Identification	59.1 / 59.6	64.6 / 61.7	**64.9** / **65.1**
Pronoun Coreference Resolution	63.4 / 55.4	**68.8** / **60.7**	68.5 / 59.8
Mentioned Player Prediction	47.3 / 41.0	47.2 / 39.5	**48.4** / **43.4**
Average Accuracy	56.6 / 52.0	60.2↑ / 53.9↑	**60.6**↑ / **56.1**↑

4.6 Performance Comparison

Table 1 presents the performance of our proposed Online-MMSI-VLM framework with two MLLM backbones–Qwen and LLaMA–compared to the existing method (Lee et al., 2024a) across three social understanding tasks: speaking target identification, pronoun coreference resolution, and mentioned player prediction. Evaluations are conducted on both the YouTube and Ego4D subsets. Both variants of Online-MMSI-VLM consistently outperform the baseline across all tasks and datasets. Notably, the LLaMA-based model achieves the highest average accuracy on YouTube (60.6%) and Ego4D (56.1%). Specifically, on the YouTube subset, Online-MMSI-VLM outperforms the baseline by up to 5.78% in speaking target identification, 3.9% in pronoun coreference resolution, and 1.13% in mentioned player prediction. On the Ego4D subset, improvements reach up to 2.08%, 5.36%, and 2.40% respectively for the same tasks.

Across tasks, we observe that pronoun coreference resolution generally achieves the highest accuracies (up to 68.8%), followed by speaking target identification, while mentioned player prediction remains the most challenging, with accuracies typically below 50%. This pattern can be explained by the differences how each task depends on context. Both pronoun coreference and speaking target identification often involve referents located in adjacent utterances–either the preceding or following turn–making them more amenable to forecasting and short-range historical modeling. In contrast, we find mentioned player prediction typically requires tracking entities introduced several turns earlier, as the referent is often not mentioned in the immediate context. This temporal gap increases the difficulty of reasoning mentioned player.

Figure 5 illustrates qualitative comparison examples of the three social tasks in streaming video, demonstrating how our method leverages only past dialogue turns and annotated frames to accurately identify social references. For user queries such as "Predict which player is mentioned by name," "Determine which player a pronoun refers to," and "Identify which player the speaker is talking to," our model consistently provides immediate and accurate referents based solely on prior context. In contrast, the prior method (Lee et al., 2024a) often struggles in online settings due to a lack of tailored design for these challenges.

To assess Online-MMSI-VLM's real-time applicability, we evaluate the model's latency and throughput. Experiments show the model achieves a median latency of 300 ms and a throughput of 1.99 samples per second when running on a single A6000 GPU, supporting deployment in natural conversational scenarios.

Table 2: Ablation study of proposed components for three social tasks on the YouTube and Ego4D datasets. Both socially-aware visual prompting and multi-party conversation forecasting consistently improve performance over the baseline. Their combination yields the best results across all three tasks and both datasets.

Model	Forecasting	Prompting	Speaking Target Identification		Pronoun Coreference Resolution		Mentioned Player Prediction	
			YouTube	Ego4D	YouTube	Ego4D	YouTube	Ego4D
Qwen	-	-	60.76	57.71	62.88	48.21	45.83	32.89
Qwen	-	✓	61.45	58.10	64.43	58.93	46.24	35.52
Qwen	✓	-	63.96	60.00	65.20	54.46	46.52	36.84
Qwen	✓	✓	**64.58**	**61.71**	**68.83**	**60.71**	**47.20**	**39.47**
LLaMA	-	-	63.20	60.20	65.39	54.46	46.61	38.15
LLaMA	-	✓	64.02	64.00	67.87	55.35	47.33	42.10
LLaMA	✓	-	64.43	61.71	67.30	55.35	47.34	40.78
LLaMA	✓	✓	**64.88**	**65.14**	**68.45**	**59.82**	**48.43**	**43.42**

Table 3: Performance impact of forecasting components on three social tasks using Qwen on the YouTube dataset. Detailed future utterance generation consistently enhances results, while adding speaker-turn prediction yields further improvement, demonstrating the strength of a coarse-to-fine strategy.

Speaker Turns	Detailed Utterances	Speaking Target Identification	Pronoun Coreference Resolution	Mentioned Player Prediction
-	-	60.76	62.88	45.83
✓	-	60.31	64.63	46.39
-	✓	62.69	64.92	45.93
✓	✓	**63.96**	**65.20**	**46.52**

4.7 Effects of Conversation Forecasting

Table 2 shows the effects of multi-party conversation forecasting, where forecasting consistently improves performance across all tasks. For example, in the pronoun coreference resolution task on the YouTube dataset, applying forecasting boosts Qwen's accuracy from 62.88% to 65.20%. Similarly, for speaking target identification, forecasting improves Qwen's performance from 60.76% to 63.96%. These results confirm that anticipating future conversational turns and utterances helps the model compensate for the lack of future context in the online setting. The improvement for mentioned player prediction is relatively smaller, possibly because this task, mentioning someone in a dialogue, relies less on future context.

Table 3 further analyzes the impact of the two components in our coarse-to-fine forecasting framework: speaker-turn prediction and detailed utterance generation. Results indicate that enabling both components yields the highest overall performance. In particular, direct detailed utterance generation consistently enhances accuracy by enriching the context. Moreover, integrating speaker-turn prediction provides additional gains, suggesting that predicting turn structure first facilitates the generation of relevant utterances.

Table 4 explores the effects of different forecasting turn lengths (i.e. 2, 4, or 8 future turns). Empirically, a 4-turn forecasting length provides the most reliable performance gains across the three social tasks on both datasets. From the observation in 4.9, multi-party conversation forecasting is challenging due to its uncertainty nature. Extending the forecasting length further increases the risk of hallucinated, question-irrelevant content and adds computational complexity, while contributing little additional useful context. Thus, longer forecasts do not improve performance linearly; instead, they exhibit diminishing returns. For these reasons, forecasting 4 turns strikes the most practical balance between accuracy and complexity.

4.8 Effects of Visual Prompting

Table 2 also shows the effects of socially-aware visual prompting, where prompting further enhances model performance by effectively enhancing past information. For example, in the pronoun coreference resolution

Table 4: Impact of conversation forecasting length on three social tasks using Qwen across YouTube and Ego4D datasets. Forecasting 4 future turns typically yields the highest performance gains across all tasks.

Dataset	YouTube (%)			Ego4D (%)		
Forecast Length (Turns)	2	4	8	2	4	8
Speaking Target Identification	62.44	**64.58**	61.98	**62.03**	61.71	59.43
Pronoun Coreference Resolution	68.26	**68.83**	67.11	56.07	**60.71**	55.54
Mentioned Player Prediction	45.98	**47.20**	46.25	34.26	**39.47**	39.47

Table 5: Effects of the different visual prompting on the performances for three social tasks. We adopt the Qwen model and conduct evaluations on the YouTube dataset. The improvement illustrates that highlighting players' position and gestures helps the model better understand the social scene.

Visual Cues	Speaking Target Identification	Pronoun Coreference Resolution	Mentioned Player Prediction
Raw RGB frames	63.96	65.20	46.52
+ Bounding boxes	64.13	68.13	46.71
+ Upper body keypoints (on top)	**64.58**	**68.83**	**47.20**

task, adding visual prompting raises Qwen's accuracy from 62.88% to 64.43% and LLaMA's from 65.39% to 67.87% on the YouTube dataset. This result verifies that visual prompting can serve as highlighting the crucial areas in the video, enabling VLMs to capture subtle and informative social cues in historical video.

We investigate the impact of incorporating different types of visual annotations–namely, bounding boxes and upper body keypoints–on model performance. In the baseline setting, raw RGB frames are annotated only with speaker identities to align visual input with the corresponding utterances. As shown in Table 5, augmenting the visual input with bounding boxes yields consistent performance gains across all tasks. Adding upper body keypoints on top of bounding boxes further enhances performance, achieving the highest results on all three tasks. These improvements demonstrate that explicitly highlighting players' spatial positions and gestures enables the model to better interpret social dynamics within the scene.

The tracking and keypoints are provided by Lee et al. (2024a), where annotations are generated by AlphaPose (Fang et al., 2022) followed by human curation with the reference frame. To further investigate model robustness to inaccurate tracking and keypoints, we conducted a preliminary study for visual prompting. During inference, we explored 20% dropout to keypoints or 10% dropout to tracking, simulating keypoint jitter (motion/occlusion) and short-term tracking loss. Our Qwen-based approach maintains strong performance on the speaker target identification task, with 64.5% on Youtube and 61.7% on Ego4D.

Furthermore, as shown in Table 2, combining socially-aware visual prompting and multi-party conversation forecasting consistently yields the best performance across all three tasks and both datasets. This demonstrates the complementary strengths of forecasting future conversations and enriching visual understanding.

4.9 Evaluation of Generated Conversations

Figure 6 presents two examples of generated future speaker turns and utterances, where green text indicates accurate and consistent conversations with the ground truth and yellow denotes different but reasonable utterances. Although the predicted turns and wording may differ slightly from the ground truth, the model successfully produces plausible speakers and content that align with the surrounding context and human social reasoning. This helps the model build a more coherent understanding of the ongoing dialogue while missing future context. In the first example, the generated utterances reinforce the exchange between Player1 and Player5 regarding the Insomniac role. In the second, the forecasting captures a multi-party interaction among Player0, Player1, and Player2 concerning the observation of another player's card.

To further measure the quality of the generated conversations, we employ Macro F1 (Pedregosa et al., 2011) and BERTScore (Zhang et al., 2019) as evaluation metrics. Macro F1 measures the accuracy of predicted speaker turns, while BERTScore evaluates the semantic similarity between generated and ground-truth utterances. We exclude BLEU (Papineni et al., 2002) from our evaluation due to its reliance on

Previous

[**Player5**]: For real.
[**Player4**]: Ah, that's a convenient card.
[**Player1**]: I was the Insomniac.
[**Player5**]: What were you now?

Ground Truth

[**Player1**]: The Insomniac.
[**Player5**]: That's funny because I'm the Insomniac.
[**Player1**]: That's funny because I'm the Insomniac.
[**Player4**]: They have the same story.

Forecast

[**Player1**]: I was the Insomniac.
[**Player5**]: So you're saying he's lying?
[**Player1**]: Yeah.
[**Player5**]: And you're saying you're the Insomniac?

Previous

[**Player2**]: He could be a drunk.
[**Player1**]: Which ones did you look at again? I can't remember.
[**Player0**]: I looked at just his card.
[**Player1**]: Just his card.

Ground Truth

[**Player0**]: Yes.
[**Player1**]: And it said it was a werewolf?
[**Player2**]: I suddenly don't think you're a seer.
[**Player1**]: His said werewolf?

Forecast

[**Player0**]: Yeah.
[**Player1**]: Okay. What did you see?
[**Player0**]: I saw a werewolf.
[**Player2**]: That's a little suspicious.

Figure 6: Samples of ground truth future conversations and generated forecasts. Although the predicted turns and wording may differ slightly from the ground truth (yellow), the model successfully produces plausible speakers and content that align with the surrounding context and human social reasoning (green).

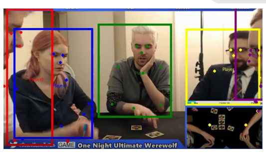
History		Ground Truth Future	Forecast
	**[Player0]: [Player4] could. [Player4]: I was. [Player2]: Okay. You confirmed your story. You switched her and him.**	**[Player0]: These two. Yeah. [Player1]: Yeah. [Player0]: So we're a nice, happy town after that last yelling around.**	**[Player4]: Yeah. [Player2]: Okay. So we have two people who were the robber. [Player4]: I was the robber.**

Figure 7: Failure case of the forecasting module on the STI task. The predictions align with historical context but deviate from the ground truth, illustrating the challenge of anticipating uncertain social dynamics. Notably, the goal of forecasting is to enhance social understanding, rather than exact future prediction.

exact n-gram matches, which makes it less reliable for assessing short and diverse conversational responses, where minor paraphrasing or changes in word order can disproportionately penalize the score. Our approach achieves a Macro F1 of 0.68 and a BERTScore of 0.86, indicating accurate upcoming speaker prediction and strong semantic alignment with the ground truth. These results validate the effectiveness of conversation forecasting in enhancing linguistic context for Online MMSI.

Figure 7 shows a failure case of forecasting where forecasts are consistent with history but diverge from ground truth, highlighting the challenge of anticipating uncertain social dynamics. Through these experiments and evaluation, we observe that multi-party conversation forecasting remains a challenging task due to the inherent unpredictability of real-world dialogue and the current limitations of LLMs in complex social reasoning. Nevertheless, our goal is not to achieve precise forecasting, but to use the process as a means of enhancing social understanding. Interestingly, we observe that even inaccurate forecasts can improve performance–possibly because forecasting encourages the model to role-play the historical dialogue, thereby deepening its understanding of social context and interaction dynamics.

4.10 Furthure Analysis of Visual Prompting

We explored how the model understands the social interaction in the historical video under different visual prompting. We provide Qwen2.5 VL with videos containing different visual promptings and the query: "Describe the social interaction cues in the video." The generated descriptions are presented in Figure 8, where green text indicates accurate social descriptions, red highlights misleading or incorrect descriptions, and black denotes general, non-specific descriptions. The results demonstrate that, by annotating each participant with bounding boxes and body keypoints, the VLM produces more accurate and detailed descriptions of social interactions. For example, the model generates the description: "The person in the yellow box is also smiling and looking at the speaker." Compared to other prompting's generation, such an description has more details about body movement and facial expression. This suggests that socially-aware visual prompting

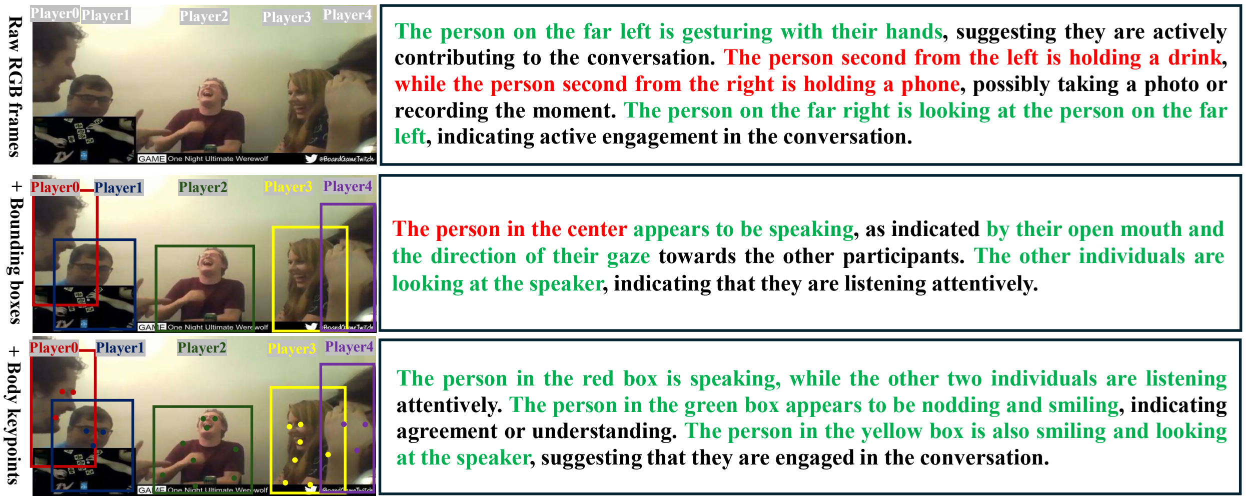


Figure 8: VLLM-generated descriptions under different visual prompting settings. The results show that using raw RGB frames yields generic descriptions with limited reference to social interactions. In contrast, when participants are annotated with bounding boxes and body keypoints, the VLLM generates more accurate and detailed accounts of social dynamics, such as "nodding and smiling" and "looking at the speaker".

effectively emphasizes critical regions of interest, thereby improving the model's ability to interpret subtle social dynamics in noisy and large historical video. We also noticed that, in raw RGB frames input, the description not only fails to capture detailed social interaction cues but also contains hallucinations: "The person second from the left is holding a drink, while the person second from the right is holding a phone." The visual prompting shows capability of suppressing inaccurate description.

To explore more auxiliary signals for visual prompting, we also extracted gaze signal from videos. However, integrating gaze through visual prompting results in performance drops, i.e., 0.8%, 0.3%, and 0.5% on speaker target identification, pronounce coreference recognition, and mentioned player prediction, respectively. It might be due to inaccurate gaze estimations which are derived from GazeFollowing (Lian et al., 2018). Besides, some studies show existing LLMs might struggle to incorporate gaze information within the context of multi-party dialogues (Inoue et al., 2025), indicating more sophisticated methods are needed.

5 Conclusion

We introduce a new task, Online Multimodal Social Interaction Understanding, which requires models to interpret social interactions using only historical information. This setting better aligns with real-world human-AI interaction scenarios, where the AI assistants are required to reply immediately. Since future context is unavailable, the performance of existing models, human, and advanced AI tools degrade significantly. To address this challenge, we propose **Online-MMSI-VLM**, a novel framework built upon recent advanced MLLMs, integrating techniques: (1) *multi-party conversation forecasting*, which anticipates future dialogue turns to enrich linguistic context, and (2) *socially-aware visual prompting*, which highlights critical visual cues for more accurate social reasoning. Extensive experiments across multiple benchmarks demonstrate the effectiveness of our method in online social understanding tasks. Further ablation studies validate the individual contributions of each proposed component. We believe this work takes an important step toward enabling practical and robust social AI systems capable of operating in realistic environments.

Limitations. While our method achieves strong performance on practical Online-MMSI, several limitations remain. First, the approach relies on external preprocessing steps as previous approaches (Lee et al., 2024a), such as visual tracking and speech transcription, which may introduce errors and affect overall performance. Second, the capabilities of the model are bounded by the social reasoning abilities of current LLMs, which may struggle with complex or subtle social dynamics. Future work may explore end-to-end solutions and more advanced perception and reasoning mechanisms to further enhance online social interaction understanding.

References

Gibran Benitez-Garcia, Jesus Olivares-Mercado, Gabriel Sanchez-Perez, and Keiji Yanai. Ipn hand: A video dataset and benchmark for real-time continuous hand gesture recognition. In *2020 25th international conference on pattern recognition (ICPR)*, pp. 4340–4347. IEEE, 2021.

Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. Social robotics. *Springer handbook of robotics*, pp. 1935–1972, 2016.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12914–12923, 2024.

Xu Cao, Pranav Virupaksha, Wenqi Jia, Bolin Lai, Fiona Ryan, Sangmin Lee, and James M Rehg. Socialgesture: Delving into multi-person gesture understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19509–19519, 2025.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2024.

Jonathan P Chang and Cristian Danescu-Niculescu-Mizil. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. *arXiv preprint arXiv:1909.01362*, 2019.

Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *International Journal of Computer Vision*, 131(6):1346–1366, 2023a.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18407–18418, 2024a.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023b.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 39(1):btac817, 2023c.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853, 2024a.

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

Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5396–5406, 2020.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Amine El Hattami, Issam H. Laradji, Stefania Raimondo, David Vazquez, Pau Rodriguez, and Christopher Pal. Workflow Discovery from Dialogues in the Low Data Regime. *Transactions on Machine Learning Research (TMLR)*, 2023.

MM Elsherbini, Ola Mohammed Aly, Donia Alhussien, Ohamed Amr, Moataz Fahmy, Mahmoud Ahmed, Mohamed Adel, Mohamed Fetian, Mahmoud Hatem, Mayar Khaled, et al. Towards a novel prototype for superpower glass for autistic kids. *International Journal of Industry and Sustainable Development*, 4(1): 10–24, 2023.

Theresa Epperlein, Gyula Kovacs, Linda S Oña, Federica Amici, and Juliane Bräuer. Context and prediction matter for the interpretation of social interactions across species. *Plos one*, 17(12):e0277783, 2022.

Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7157–7173, 2022.

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gašić. Emowoz: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. *arXiv preprint arXiv:2109.04919*, 2021.

Xiachong Feng, Longxu Dou, Minzhi Li, Qinghao Wang, Haochuan Wang, Yu Guo, Chang Ma, and Lingpeng Kong. A Survey on Large Language Model-Based Social Agents in Game-Theoretic Scenarios. *Transactions on Machine Learning Research (TMLR)*, 2025.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025.

Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13505–13515, 2021.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Hongcheng Guo, Shaosheng Cao, Boyang Wang, Lei Li, Liang Chen, Xinze Lyu, Zhe Xu, Yao Hu, Zhoujun Li, et al. Sns-bench: Defining, building, and assessing capabilities of large language models in social networking services. In *Forty-second International Conference on Machine Learning*, 2025.

Anshul Gupta, Samy Tafasca, Arya Farkhondeh, Pierre Vuillecard, and Jean-Marc Odobez. Mtgs: A novel framework for multi-person temporal gaze following and social gaze prediction. *Advances in Neural Information Processing Systems*, 37:15646–15673, 2024.

Nick Haber, Catalin Voss, and Dennis Wall. Making emotions transparent: Google glass helps autistic kids understand facial expressions through augmented-reaiity therapy. *IEEE Spectrum*, 57(4):46–52, 2020.

Lauren V Hadley and John F Culling. Timing of head turns to upcoming talkers in triadic conversation: Evidence for prediction of turn ends and interruptions. *Frontiers in Psychology*, 13:1061582, 2022.

Md. Mahadi Hassan, Alex Knipper, and Shubhra Kanti Karmaker Santu. Introducing "Forecast Utterance" for Conversational Data Science. *Transactions on Machine Learning Research (TMLR)*, 2024.

Yuqi Hou, Zhongqun Zhang, Nora Horanyi, Jaewon Moon, Yihua Cheng, and Hyung Jin Chang. Multimodal gaze following in conversational scenarios. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1186–1195, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: A comprehensive benchmark and memory-augmented method. *arXiv preprint arXiv:2501.00584*, 2024.

Lee Hyun, Kim Sung-Bin, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. Smile: Multimodal dataset for understanding laughter in video with language models. *arXiv preprint arXiv:2312.09818*, 2023.

Koji Inoue, Divesh Lala, Mikey Elmers, Keiko Ochi, and Tatsuya Kawahara. An llm benchmark for addressee recognition in multi-modal multi-party dialogue. *arXiv preprint arXiv:2501.16643*, 2025.

Swati Jindal and Roberto Manduchi. Contrastive representation learning for gaze estimation. In *Gaze Meets Machine Learning Workshop*, pp. 37–49. PMLR, 2023.

Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Kraynov, and Andrei Makhliarchuk. Hagrid–hand gesture recognition image dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4572–4581, 2024.

Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a vlm. *IEEE Access*, 2024.

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

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

Sangmin Lee, Minzhi Li, Bolin Lai, Wenqi Jia, Fiona Ryan, Xu Cao, Ozgur Kara, Bikram Boote, Weiyan Shi, Diyi Yang, et al. Towards social ai: A survey on understanding social interactions. *arXiv preprint arXiv:2409.15316*, 2024b.

Yiming Lei, Jingqi Li, Zilong Li, Yuan Cao, and Hongming Shan. Prompt learning in computer vision: a survey. *Frontiers of Information Technology & Electronic Engineering*, 25(1):42–63, 2024.

Wanhua Li, Zibin Meng, Jiawei Zhou, Donglai Wei, Chuang Gan, and Hanspeter Pfister. Socialgpt: Prompting llms for social relation reasoning via greedy segment optimization. *Advances in Neural Information Processing Systems*, 37:2267–2291, 2024a.

Xinpeng Li, Teng Wang, Jian Zhao, Shuyi Mao, Jinbao Wang, Feng Zheng, Xiaojiang Peng, and Xuelong Li. Two in one go: Single-stage emotion recognition with decoupled subject-context transformer. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9340–9349, 2024b.

Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pp. 35–50. Springer, 2018.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, et al. Ov-mer: Towards open-vocabulary multimodal emotion recognition. *ICML 2025*, 2024.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *ICML 2025*, 2025.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Jihao Liu, Zhiding Yu, Shiyi Lan, Shihao Wang, Rongyao Fang, Jan Kautz, Hongsheng Li, and Jose M Alvare. Streamchat: Chatting with streaming video. *arXiv preprint arXiv:2412.08646*, 2024.

Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10631–10642, 2021.

Ning Ma, Norihiro Harasawa, Kenichi Ueno, Kang Cheng, and Hiroyuki Nakahara. Decision-making with predictions of others' likely and unlikely choices in the human brain. *Journal of Neuroscience*, 44(37), 2024.

Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pp. 735–745, 2022.

OpenAI. Gpt-4o, 2024. URL `https://openai.com/index/gpt-4o`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Vaidehi Patil, Yi-Lin Sung, Peter Hase, Jie Peng, Tianlong Chen, and Mohit Bansal. Unlearning Sensitive Information in Multimodal LLMs: Benchmark and Attack-Defense Evaluation. *Transactions on Machine Learning Research (TMLR)*, 2024.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL `http://jmlr.org/papers/v12/pedregosa11a.html`.

Chirag Raman, Jose Vargas Quiros, Stephanie Tan, Ashraful Islam, Ekin Gedik, and Hayley Hung. Conflab: A data collection concept, dataset, and benchmark for machine analysis of free-standing social interactions in the wild. *Advances in Neural Information Processing Systems*, 35:23701–23715, 2022.

Fiona Ryan, Hao Jiang, Abhinav Shukla, James M Rehg, and Vamsi Krishna Ithapu. Egocentric auditory attention localization in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14663–14674, 2023.

Andrey Savchenko. Facial expression recognition with adaptive frame rate based on multiple testing correction. In *International Conference on Machine Learning*, pp. 30119–30129. PMLR, 2023.

Anthony Sicilia and Malihe Alikhani. Eliciting uncertainty in chain-of-thought to mitigate bias against forecasting harmful user behaviors. *arXiv preprint arXiv:2410.14744*, 2024.

Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20935–20946, 2023.

Chao-Hong Tan, Jia-Chen Gu, and Zhen-Hua Ling. Is chatgpt a good multi-party conversation solver? *arXiv preprint arXiv:2310.16301*, 2023.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Qwen Team. Qwen2.5-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025.

Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, and Nong Sang. Oadtr: Online action detection with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7565–7575, 2021.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*, 2024.

Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. Controlmllm: Training-free visual prompt learning for multimodal large language models. *Advances in Neural Information Processing Systems*, 37:45206–45234, 2025.

Jiarui Xu, Yossi Gandelsman, Amir Bar, Jianwei Yang, Jianfeng Gao, Trevor Darrell, and Xiaolong Wang. IMProv: Inpainting-based Multimodal Prompting for Computer Vision Tasks. *Transactions on Machine Learning Research (TMLR)*, 2024.

Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36:24993–25006, 2023.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, et al. Timechat-online: 80% visual tokens are naturally redundant in streaming videos. *arXiv preprint arXiv:2504.17343*, 2025.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024a.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023b.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024b.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34:17616–17627, 2021.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal Chain-of-Thought Reasoning in Language Models. *Transactions on Machine Learning Research (TMLR)*, 2024c.

Chenxi Zhao, Jinglei Shi, Liqiang Nie, and Jufeng Yang. To err like human: Affective bias-inspired measures for visual emotion recognition evaluation. *Advances in Neural Information Processing Systems*, 37:134747–134769, 2024.

Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos? *arXiv preprint arXiv:2307.16368*, 2023.

Yue Zhao and Philipp Krähenbühl. Real-time online video detection with temporal smoothing transformers. In *European Conference on Computer Vision*, pp. 485–502. Springer, 2022.

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

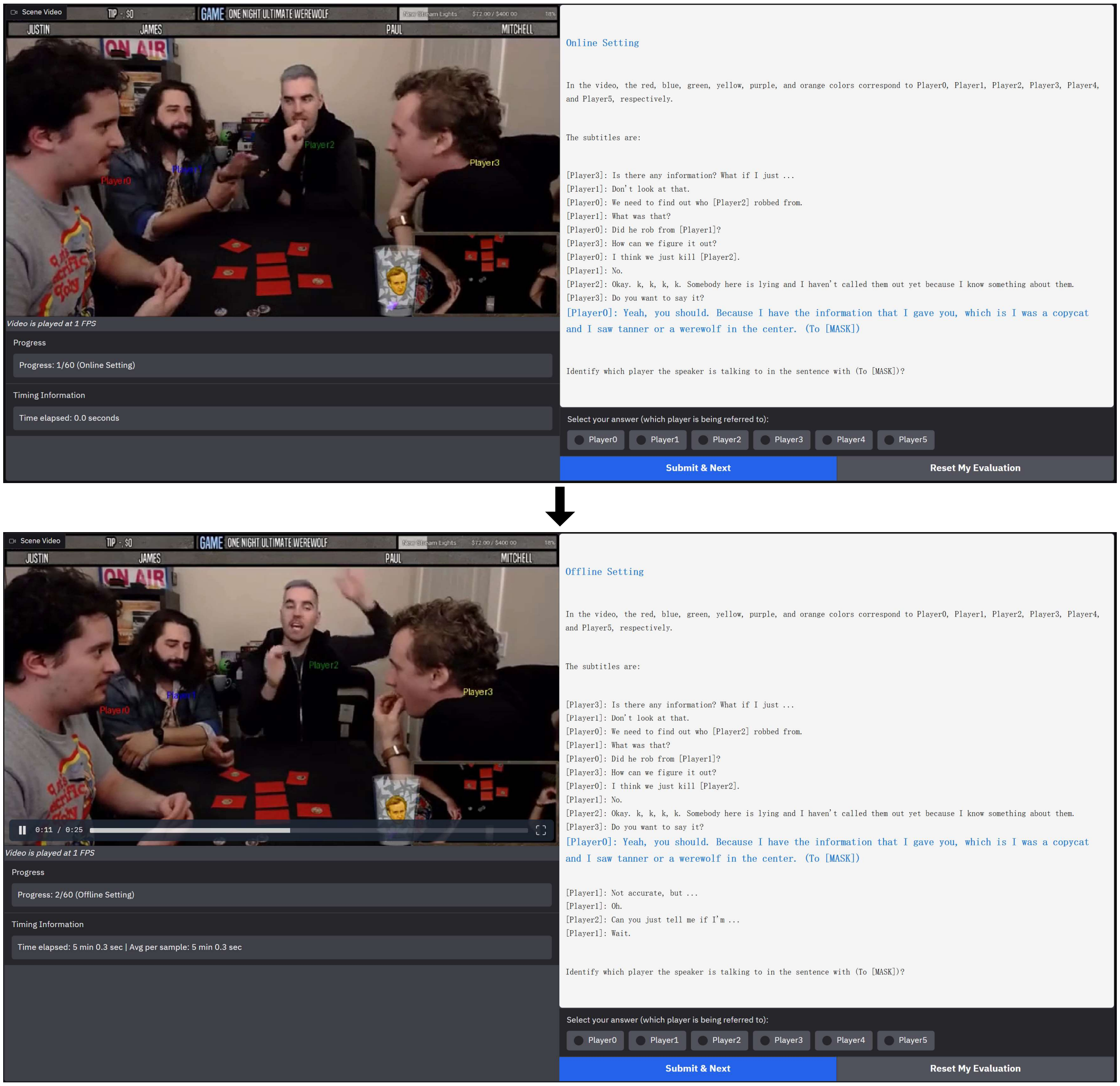


Figure 9: Illustration of human study on online setting and then offline.

A Appendix

A.1 Implementation Details

The training of our video-based model is built upon the Qwen2.5-VL-7B-Instruct[1] (Team, 2025) or LLaMA-3.2-11B-Vision[2] (Dubey et al., 2024) architecture with image splitting disabled. The model operates with bfloat16 precision and employs FlashAttention-2 (Dao et al., 2022) for efficient memory usage. For Qwen2.5-VL-7B-Instruct, the videos are sampled at 1.0 FPS and a resolution constraint of 36 $\times$ 42 $\times$ 10 pixels per frame; for LLaM-3.2-11B-Vision, we transfer the video into a 3 $\times$ 2 grid-like image and set the maximum image size as 1120 $\times$ 1120 pixels. The max new tokens is set as 1024 in both models. A LoRA-based (Hu et al., 2022) fine-tuning strategy is implemented, with LoRA alpha set to 16, a dropout rate of 0.05, a rank of 512, and modifications applied to the query and value projection layers. The training configuration includes

[1]https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
[2]https://huggingface.co/meta-llama/Llama-3.2-11B-Vision

a batch size of 1 per device, gradient accumulation steps of 4, and a total of 5 epochs. The optimizer used is AdamW with a fused implementation, and the learning rate follows a linear decreasing scheduler strategy. The regularization weight decay of 0.01 and a gradient clipping threshold of 0.3.

A.2 Human Study

We select 30 samples randomly for each participant and show the online input followed by its offline version, as shown in 9. In this way, the participant gets rid of being affected by the prior offline information.