SUPPLEMENTARY MATERIALS

Anonymous authors
Paper under double-blind review

A RELATED WORK

A.1 SOCIAL REASONING IN LARGE LANGUAGE MODELS

Recent studies have explored the extent to which large language models (LLMs) can perform social reasoning, such as attributing beliefs, recognizing emotions, or inferring intentions (Shapira et al., 2023; Kim et al., 2023). Benchmarks like SocialIQa (Sap et al., 2019) and Social Chemistry 101 (Hwang et al., 2021) provide scenarios where models must understand social norms or moral judgments. However, these benchmarks rely solely on static text, lacking the multimodal and dynamic context typical of real-life social interactions. Recent procedural datasets like BigToM (Gandhi et al., 2023) attempt to evaluate Theory-of-Mind (ToM) reasoning in LLMs, but often in synthetic and text-only settings. These limitations motivate the need for evaluation environments grounded in realistic, multimodal social interactions. Our dataset addresses this gap by offering richly annotated video scenarios designed to test multi-step reasoning over belief, desire, intent, and emotion in context.

A.2 MULTIMODAL SOCIAL INTELLIGENCE AND VISION-LANGUAGE MODELS

Large vision-language models (LVLMs) such as Flamingo (Alayrac et al., 2022), VILA (Chen et al., 2023), and Video-llava (Lin et al., 2023) have shown strong performance in general-purpose video-language tasks. However, their ability to perform fine-grained social reasoning remains limited. Studies have found that these models often over-rely on language priors and fail to leverage visual evidence when answering socially grounded questions (Chen et al., 2024). Datasets like SocialIQ-2.0 (Wilf et al., 2023), VLEP (Lei et al., 2020), and SODA (Wang et al., 2023) introduce social content into video QA, but few offer systematic annotations of mental-state transitions or causal relations. Additionally, models rarely capture nuanced cues like interpersonal gaze, tone, or posture, which are essential for deeper social inference (Wei et al., 2024). Our dataset **R**ead-the-**R**oom **R**easoning for **V**ideo **Q**uestion **A**nswering (*R^3-VQA*) complements these efforts by providing a scalable training set and a fine-grained benchmark with explicit labels for mental-state categories and their causal links in socially rich scenarios.

A.3 THEORY OF MIND IN VIDEO UNDERSTANDING

Theory of Mind (ToM) — the ability to attribute mental states to others — has been a long-standing challenge in AI. Recent work like Watch-and-Help (Puig et al., 2020), NoPa (Puig et al., 2023), and Generative Agents (Park et al., 2023) explores ToM in simulation-based or scripted agent settings. MMToM-QA (Jin et al., 2024) represents a step toward real-world ToM inference, offering multimodal video-based questions about beliefs and desires. Yet, most of these benchmarks focus on short clips, isolated mental states, or handcrafted settings. IntentQA (Li et al., 2023) explores intention inference in video QA, but it does not capture full mental-state causal chains. Our proposed benchmark *R^3-Bench* explicitly models belief, intent, desire, and emotion along multi-step causal paths, enabling a more complete and diagnostic evaluation of ToM-like reasoning in LVLMs under naturalistic, temporally extended scenarios.

B MORE EXAMPLES OF *R^3-VQA*

We provide more examples in fig. 1 for all types of QAs. We choose four QAs for each type. We divide *MSE QA* into *Belief Estimation QA*, *Desire Estimation QA*, *Intent Estimation QA*, and *Emotion Estimation QA*. For more details, please see our website `https://r3-vqa.github.io/`.

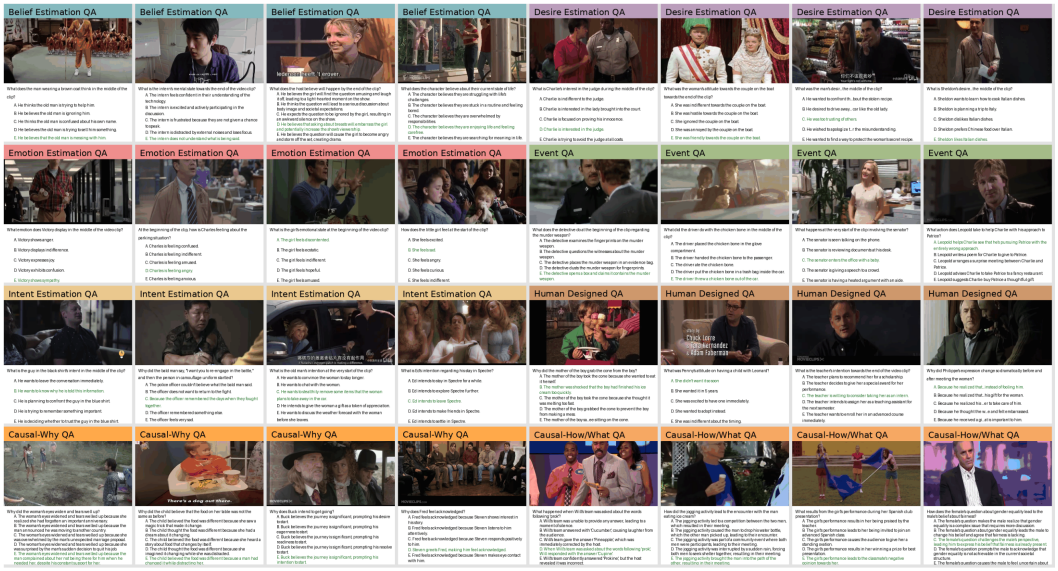

Figure 1: More examples of R^3-*VQA*.

C More Details of QA Generation for R^3-*Bench*

GPT-4o is a powerful large multimodal model. Therefore, we use it to generate questions, answers and incorrect options according to social causal chains annotated by experts. In all prompts, the parts highlighted in dark green are the general templates.

C.1 EU QAs and MSE QAs Generation Prompts

We generate *EU* QAs and *MSE* QAs based on nodes in social causal chains and an example prompt is as follows. We provide an approximate time period for each node to help GPT-4o locate the time in the generated question.

Prompt

The video clip is from(HH:MM:SS) 00:00:30 to 00:00:45.
A reasoning chain is consisted of nodes and sub-chains.
Nodes:

The woman's Belief-1: She couldn't believe the man responded in this way. Duration(HH:MM:SS): 00:00:36-00:00:40

Event-1: The woman said she would not go out with the man. Duration(HH:MM:SS): 00:00:34-00:00:36

Event-2: The man said he didn't invite the woman yet and asked whether she wanted to come. Duration(HH:MM:SS): 00:00:36-00:00:40

Event-3: The woman stared at the man. Duration(HH:MM:SS): 00:00:36-00:00:40

The woman's Intent-1: She wanted to go out with the man. Duration(HH:MM:SS): 00:00:30-00:00:36

The woman's Intent-2: She didn't want to admit her feeling to the man. Duration(HH:MM:SS): 00:00:34-00:00:36

The man's Intent-1: He wanted to invite the woman to go out with him. Duration(HH:MM:SS): 00:00:36-00:00:40

We call question and answer as QA. Each node generates a QA. Note that the node ID should not appear in QA.
If the node is an event, you should generate a event understanding QA.
If the node is a/an belief, desire, intent or emotion, you should generate a mental state estimation QA.
You must generate a event understanding QA according to the following rules:
a. You should summarize the event reasonably.
b. You can refer to the event in the question according to a part of the event and the answer is the rest of the event, and the event can be summarized reasonably.
c. You can refer to the event in the question according to the position of the start and end time in the whole video(at the end of the clip, in the middle of the clip, etc.).
Note that you need to use one of two referential methods when generating a event understanding QA.
You must generate a mental state estimation QA according to the following rules:
a. You should refer to the mental state in the question according to the position of the start and end time in the whole video(at the end of the clip, at the middle of the clip, etc.). If different nodes of the same type (types include Intent, Belief, Desire, and Emotion) of the same person are at the same position in the video, then you must re-divide this position to ensure that there are no nodes of the same type of the same person at it.
b. The answer is the mental state described in the node. You can summarize it reasonably.
Note that you must refine QA after generating them. Think step by step.

C.2 CW QAs AND CH/W QAs GENERATION PROMPTS

We generate *CW* QAs and *CH/W* QAs based on the whole causal chain. An example prompt is as follows.

Prompt

The video clip is from(HH:MM:SS) 00:00:30 to 00:00:45.
A reasoning chain is consisted of nodes and sub-chains.
Nodes:

The woman's Belief-1: She couldn't believe the man responded in this way. Duration(HH:MM:SS): 00:00:36-00:00:40

Event-1: The woman said she would not go out with the man. Duration(HH:MM:SS): 00:00:34-00:00:36

Event-2: The man said he didn't invite the woman yet and asked whether she wanted to come. Duration(HH:MM:SS): 00:00:36-00:00:40

Event-3: The woman stared at the man. Duration(HH:MM:SS): 00:00:36-00:00:40

The woman's Intent-1: She wanted to go out with the man. Duration(HH:MM:SS): 00:00:30-00:00:36

The woman's Intent-2: She didn't want to admit her feeling to the man. Duration(HH:MM:SS): 00:00:34-00:00:36

The man's Intent-1: He wanted to invite the woman to go out with him. Duration(HH:MM:SS): 00:00:36-00:00:40

Sub-Chains:

Sub-Chain-1
- **Reason**: The woman's Intent-1,The woman's Intent-2
- **Result**: Event-1

Sub-Chain-2
- **Reason**: Event-1,The man's Intent-1
- **Result**: Event-2

Sub-Chain-3
- **Reason**: Event-2
- **Result**: The woman's Belief-1

Sub-Chain-4
- **Reason**: The woman's Belief-1
- **Result**: Event-3

We call question and answer as QA. For each sub-chain, generate a "Why" QA and a "How/What" QA. Note that the node ID should not appear in QA.
We define "Why" QA as asking a question about the reasons of a result. It should follow the rules:
a. The question is about sub-chain's result.
b. There may be more than one reasons of a result in one sub-chain. The question and answer should consider all reasons.
We define "How/What" QA as asking a question about a result of the reasons. It should follow the rules:
a. The question should consider and include all reasons.
b. The answer is about the sub-chain's result.
c. The question and answer should focus on how the reasons lead to the results or what are the results of the reasons.
Note that you can summarize the QA appropriately to make it reasonable without changing its meaning. You can also change the question template flexibly without changing the meaning.
Note that you must refine QA after generating them. Think step by step.

C.2.1 Incorrect Options Generation Prompt

We also leverage GPT-4o to generate wrong options. We provide it with the question, the correct answer. We also provide the causal chain as video information to help the model generate options related to the video. The system prompt is *You are a language expert.* and an example prompt is as follows.

Prompt

Please create four additional plausible but incorrect options based on the provided question, answer, and video information. Ensure these options are distinct from the incorrect option I have already given. Below are the question, answer, and video information.

Question: What was the woman's belief about the man's response towards the end of the clip?
Correct Answer: She couldn't believe the man responded in this way.
Video Information:

The video clip is from(HH:MM:SS) 00:00:30 to 00:00:45.
A reasoning chain is consisted of nodes and sub-chains.
Nodes:

The woman's Belief-1: She couldn't believe the man responded in this way. Duration(HH:MM:SS): 00:00:36-00:00:40

Event-1: The woman said she would not go out with the man. Duration(HH:MM:SS): 00:00:34-00:00:36

Event-2: The man said he didn't invite the woman yet and asked whether she wanted to come. Duration(HH:MM:SS): 00:00:36-00:00:40

Event-3: The woman stared at the man. Duration(HH:MM:SS): 00:00:36-00:00:40

The woman's Intent-1: She wanted to go out with the man.. Duration(HH:MM:SS): 00:00:30-00:00:36

The woman's Intent-2: She didn't want to admit her feeling to the man.. Duration(HH:MM:SS): 00:00:34-00:00:36

The man's Intent-1: He wanted to invite the woman to go out with him.. Duration(HH:MM:SS): 00:00:36-00:00:40

Sub-Chains:

Sub-Chain-1
- **Reason**: The woman's Intent-1,The woman's Intent-2
- **Result**: Event-1

Sub-Chain-2
- **Reason**: Event-1,The man's Intent-1
- **Result**: Event-2

Sub-Chain-3
- **Reason**: Event-2
- **Result**: The woman's Belief-1

Sub-Chain-4
- **Reason**: The woman's Belief-1
- **Result**: Event-3

Provide the four wrong options. Let's think step by step.

D MORE DETAILS OF DATA GENERATION PIPELINE OF *R^3-FDT*

We introduce the detailed generation pipeline based on extracted information, as shown in algorithm 1. We describe each part in conjunction with the prompts used.

GenerateChains First, we generate causal chains based on the extracted information. The generated content includes symbolic causal chains and corresponding explanations. For each movie clip, we ask GPT-4o to give the three most meaningful causal chains. We explain the task in detail and provide examples in the system prompt to improve the generation quality. We input the extracted information into the model as the user prompt. The system prompt and the user prompt template used are shown below.

Algorithm 1 *Generation Pipeline*

Input: Extracted Information I
Output: Causal QAs (Q^c, A^c, O^c), Node QAs (Q^n, A^n, O^n), Causal Chains and Corresponding Explanations (C_f, E_f)

Causal Chains Generation
$(C, E) \leftarrow \texttt{GenerateChains}(I)$
$(C^{'}, E^{'}) \leftarrow \texttt{CorrectChains}(C, E)$
$(C_f, E_f) \leftarrow \texttt{FormatChains}(C^{'}, E^{'})$
QA & Options Generation
$(Q^c, A^c) \leftarrow \texttt{GenerateCausalQAs}(C_f, E_f)$
$(Q^n, A^n) \leftarrow \texttt{GenerateNodeQAs}(C_f, E_f)$
$(O^c, O^n) \leftarrow \texttt{GenerateOptions}(Q^c, A^c, Q^n, A^n, I)$
return (Q^c, A^c, O^c), (Q^n, A^n, O^n), (C_f, E_f)

System Prompt for *GenerateChains*

Task: Given some clues: the detailed description, actions, dialogues as well as the script in a movie clip, please provide three of the most reasonable, meaningful and non-overlapping causal chains about which clues can infer other clues and the corresponding explanation.
The causal chains should be faithful to the given clues and include comprehensive and complex mental states of differenet characters. Each causal chain should correspond to a textual explanation.
When generating a causal chain, you can not only use the provided information, but also set up some nodes through reasoning. These nodes can represent some characters' mental states, inlcuding beliefs, intents, desires and emotions. Note that there may also be some mental states in the description and script, which can also be used.
List some challenging causal chains that include more than two rounds of reasoning, ensuring that these chains are naturally coherent and meaningful. Note that dialogues should mainly be used to understand mental states, background knowledge, etc and should not occur heavily in causal chains. Each node in the causal chain should contain multiple nodes.
Regarding the provided clues, there are five key points to note:
1. Dialogues, actions and descriptions are real, while the script may not exactly match those in the movie clip. The script is served as a reference during the filming of the video. However, the overall plot direction will be consistent.
2. In the given dialogues, we don't know who said each sentence. Please refer to the script to determine the speakers and the correct storyline. The dialogues in the script may not exactly match those in the video, but the overall plot direction will be consistent. When generating causal chains using dialogues, use the given "Dialogues" for the specific content of the conversation and the "Script" for its speaker.
3. The provided actions do not specify who is performing them, but you can use the script to identify the person responsible for each action. 4. There may be some information in the script outside of this movie clip. You can only use it to understand background knowledge, but you can not use it to generate causal chains. You should **first** infer the information in the script that belongs to the movie clip and then generate causal chians using only this information and clues other than the script. 5. Be mindful of aligning the timelines among dialogue and actions. Pay attention to the actions of a character when they are speaking, as this can help you determine a more accurate causal chain.
When generating causal chains, you should denote description as D, actions as A, dialogues as L, the script as S, beliefs as B, intents as I, desires as R and emotions as E. A causal chain is divided into multiple subchains, separated by ';'. Adjacent subchains must have at least one node in common. Each subchain includes one or more reasons and a result. The reasons of a subchain are in an "and" relationship, which means that they together lead to the result. The reasons and the result are separated by '->'. The reasons are separated by ','. Here are some causal chains and their explanations for reference. Note that this is only a format example and not an actual causal chain from the video.

Chain1: B1,A1->E2;E2->A2
Explanation1: A woman (B1, the woman believes ...) and the man sits down ... (A1), so the woman is impatient (E2), then she does... (A2).
Chain2: B2->I1;I1,L1->A3;A3->D1
Explanation2: <omitted>
Chain3: S1,D2->E3;E3->L2;L2,R1->B3
Explanation3: <omitted>
...<omitted>

Guidelines For Causal Chain Generation:
- Read the detailed description, the script, dialogues and actions carefully, comprehensively understanding the storyline, paying attention to the content, such as the scene where the movie clip takes place, the main characters, main characters' behaviors and mental states, and the development of the events.
- Infer main characters' meaningful mental states that are not given, explore the causal relationship between events and mental states, between mental states, or between events. - Select the three most meaningful, reasonable and non-overlapping causal chains and generate them in the required format.

The user prompt is:

User Prompt Template for *GenerateChains*

Please generate three causal chains according to the following detailed description, actions, the dialogue and the script:
Descriptions (D):
{Descriptions}

Actions (A):
{Actions}

Dialogues (L):
{Dialogues}

Script (S):
{Script}

CorrectChains To further improve the quality of causal chains, we ask GPT-4o to correct the previously generated causal chains and explanations. We only provide the model with symbolic causal chains and explanations. The system prompt and user prompt templates used are shown below.

System Prompt for *CorrectChains*

You are provided with causal chains and explanations. A causal chain is divided into multiple subchains, separated by ';'. Each subchain includes one or more reasons and a result. The reasons of a subchain are in an "and" relationship, which means that they together lead to the result. The reasons and the result are separated by '->'. The reasons are separated by ','.
An example of a causal chain is: B1, A1 -> E2; E2, A3 -> I2.
Please note that there **must** be a **causal relationship** between reason nodes and a result node in causal chains.
Given causal chains and explanations, Please correct them from the following aspects:
1. Correct the wrong causal relationship between the reasons and the result in each subchain. Please note that the causal relationship must be that the reasons **lead to** the result.

2. Correct the inconsistency between the causal chain and the corresponding explanation.

Guidelines For Correcting Causal Chains and Explanations:
- Check the explanation carefully. First, separate each subchain from the explanation, then check whether there is a causal relationship (**lead to**) between the reason nodes and the result node in each subchain. If not, then please correct the explanation accordingly.
- Check the consistency between the causal chain and the corresponding, corrected explanation, and correct any inconsistencies.
- Finally, output the corrected causal chains and explanations.

User Prompt Template for *CorrectChains*

Chain1: {Symbolic Chain}
Explanation1: {Text description}
Chain2: {Symbolic Chain}
Explanation2: {Text description}
Chain3: {Symbolic Chain}
Explanation3: {Text description}

FormatChains To facilitate QA generation, we extract the content corresponding to each node from the explanation. We provide GPT-4o with node types and symbols, allowing it to extract the content of each node from an explanation and convert it into a complete sentence. The system prompt and user prompt template are shown below.

System Prompt for *FormatChains*

There are eight types of nodes: belief (denoted as B), desire (denoted as R), intent (denoted as I), emotion (denoted as E), description (denoted as D), dialogue (denoted as L), action (denoted as A) and script (denoted as S). Given a description containing the nodes' contents and ids, with the nodes' ids in parentheses, please extract the content of each node (a complete sentence including the character name).

User Prompt Template for *FormatChains*

Description:
{Description}

GenerateNodeQAs Based on the content of each node, we generate a questions and an correct answer. We additionally require GPT-o to recognize whether the node type is a *factual event* or a *mental state*, so as to prompt it to generate high-quality QA. The system prompt and user prompt template used in this part are shown below.

System Prompt for *GenerateNodeQAs*

Task:
Given the nodes in a causal chain, please generate a question-answer pair for each node. The nodes are divided into two types: **factual events** and **mental states**.
Guidelines For Question-Answer Pair Generation For A Node:
- Determine whether the node's type is **factual event** or a **mental state**.
- Generate a question-answer pair for this node.

- Output node type (factual event or mental state), question and answer.

User Prompt Template for *GenerateNodeQAs*

Nodes:
{Node Content}

GenerateCausalQAs We provide the node contents and the causal relationships between them to GPT-4o to generate a *Causal-Why* QA for each subchain. The system prompt and user prompt template used are shown below.

System Prompt for *GenerateCausalQAs*

Given nodes and subchains in a causal chain, please generate a causal-why question-answer pair for each subchain. For each question-answer pair, you need to summarize the answer to keep it as short as possible.

User Prompt Template for *GenerateCausalQAs*

Nodes:
{Nodes}
Subchains:
{Subchains}

GenerateOptions To prevent the generated questions from being answered by models through simple elimination and thus learning shortcuts during training, we provide GPT-4o with extracted information to generate incorrect options. We require GPT-4o to generate options that are (i) correct from common sense but wrong when combined with the video context, (ii) appear in the video but are incorrect. We generate four incorrect options for each QA pair. The system prompt and user prompt template used in this part are shown below.

System Prompt for *GenerateOptions*

For each question-answer pair, please create four additional plausible but incorrect options as distractors based on the provided question, answer and the video information (including description, actions, dialogues as well as the script).
The multiple-choice question consisting of four incorrect options and the question-answer pair should be difficult enough. For each "why" question, every incorrect option should have about the same number of words as the answer.
Each incorrect option should fall into one of the following two categories:
1. Plausible but absent: The option is a reasonable or commonsense answer to the question, but it does not appear in the video information.
2. Mentioned but irrelevant or incorrect: The option is based on information that does appear in the video, but it does not correctly answer the question.

User Prompt Template for *GenerateOptions*

Question-Answer Pairs
{QA Pairs}

Video Information
Descriptions:
{Descriptions}

Actions:
{Actions}

Dialogues:
{Dialogues}

Script:
{Script}

E More Details of GRPO Training

The system prompt and user prompt template used during training are shown below. When training on *R^3-FDT* and testing on *R^3-Bench* and SocialIQ 2.0, we input the content recognized by Whisper to the model as the dialogue. When testing on IntentQA, we input *"not provided"* as the dialogue to the model. The training parameters can be found on `https://anonymous.4open.science/r/F071/`.

System Prompt for GRPO Training

A conversation between User and Assistant. The user asks a question, and the Assistant solves it based on the dialogues (if provided), the provided video. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think><answer> answer here </answer>

User Prompt Template for GRPO Training

{Video Placeholders} Dialogues:
{Dialogues}
Please answer the following questions related to this video:
{Question and Options}

The change curve of the average reward during the training process is illustrated as fig. 2. The model converges on the training data and the change is not drastic in the later period, indicating that a stable training process.

F Screenshots of Human Study

Figure 3 shows a screenshot of the page used for the human study. The page provides necessary promptings, a question, and options. The human subjects select options and submit.

G Details of "Read the Room Challenge"

The informed consent form, tutorial and quiz can be seen in `https://bnupsych.asia.qualtrics.com/jfe/form/SV_eRQaALI4DHIGN7w`.

The screenshot of the main page of "Read the Room Challenge" can be seen in fig. 4. The website is `https://exp.readtheroom.link`.

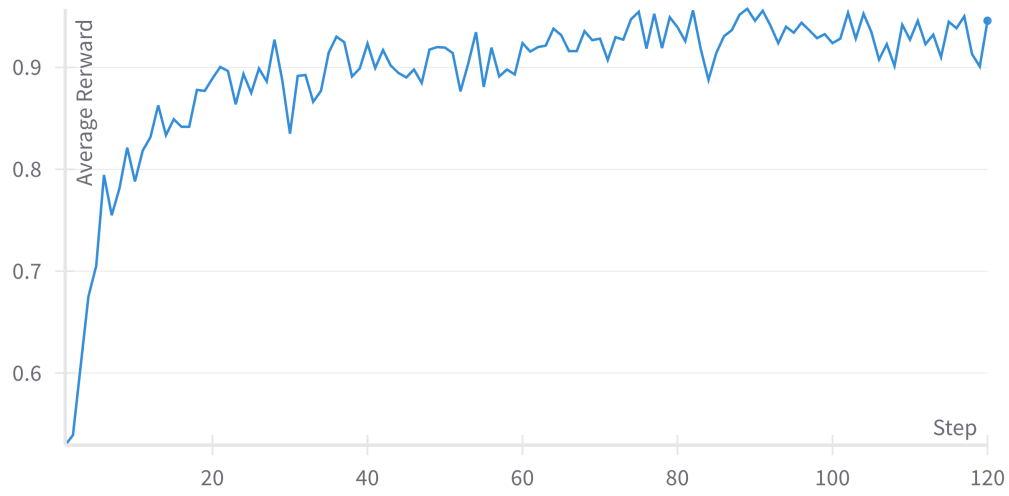

Figure 2: Average reward during training.

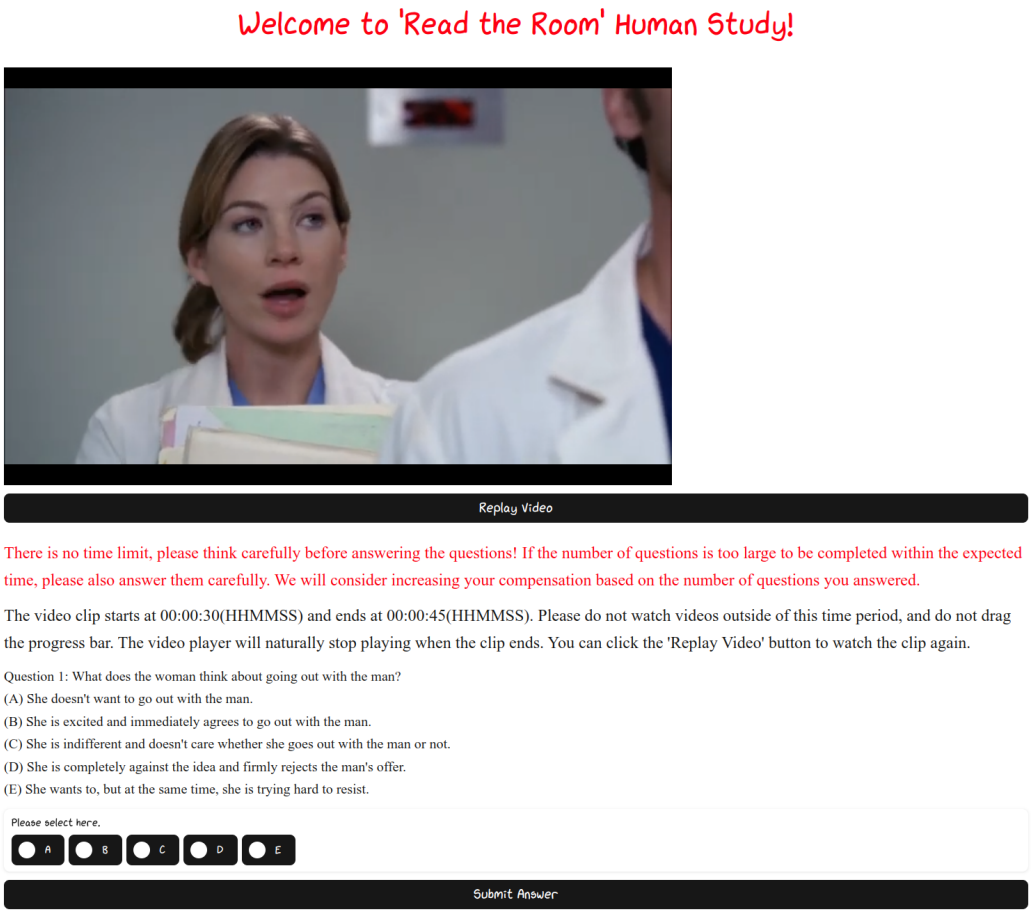

Figure 3: A screenshot of human study.

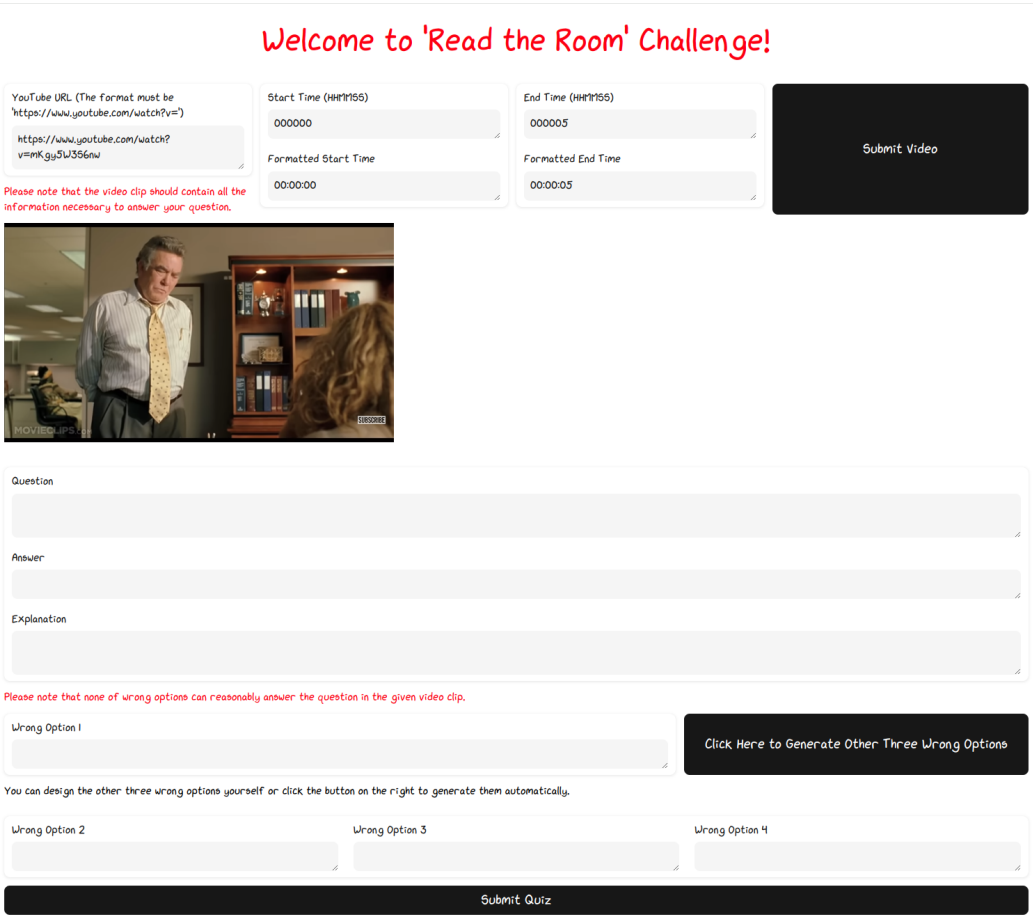

Figure 4: A screenshot of the main page of ”Read the Room" Challenge.

Table 1: Accuracy (%) on four social reasoning benchmarks before and after applying GRPO reinforcement learning. The “Baseline” row reports performance of the pretrained model without fine-tuning, while “+ Ours-FT” shows results after training on R^3-*FDT*.

	SocialIQ-2.0	IntentQA	Ours(Generated QA)	Ours(Human Designed QA)
Qwen2-VL-7B	55.97	82.99	54.88	30.06
Qwen2-VL-7B+ Sub	61.51	82.38	63.47	24.37
Qwen2-VL-7B+Sub+Ours-FT	67.73	88.85	86.59	39.87
Qwen2-VL-7B+Ours-FT				
Qwen2.5-VL-7B+Sub+Ours-FT(FPS=2+kl=0.005)	62.18	86.18	87.33	42.72
Qwen2.5-VL-7B+Sub+Ours-FT(kl=0.005)	63.87	87.96	85.89	41.14
Qwen2.5-VL-7B+Sub+Ours-FT (FPS=2,kl=0.001)	61.51	88.05	87.81	44.30

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Howard Chen, Long Ouyang, Tatsunori B. Hashimoto, Eli Zelikman, and Percy Liang. Vila: Learning human utility-aware agents using language feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Peng Chen, Xiao-Yu Guo, Yuan-Fang Li, Xiaowang Zhang, and Zhiyong Feng. Mitigating language bias of lmms in social intelligence understanding with virtual counterfactual calibration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1300–1310, 2024.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36:13518–13529, 2023.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In *EMNLP*, 2021.

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

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*, 2023.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. *arXiv preprint arXiv:2010.07999*, 2020.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11963–11974, 2023.

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.

Joon Sung Park, Joseph O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration. *arXiv preprint arXiv:2010.09890*, 2020.

Xavier Puig, Tianmin Shu, Joshua B Tenenbaum, and Antonio Torralba. Nopa: Neurally-guided online probabilistic assistance for building socially intelligent home assistants. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7628–7634. IEEE, 2023.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*, 2023.

Ziqiang Wang, Yujia Zhou, Ning Ding, Zhiyuan Liu, and Maosong Sun. Soda: Million-scale dialogue distillation with social commonsense context. In *ACL*, 2023.

Jianan Wei, Tianfei Zhou, Yi Yang, and Wenguan Wang. Nonverbal interaction detection. In *European Conference on Computer Vision*, pp. 277–295. Springer, 2024.

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