Mitigating Language Bias of LMMs in Social Intelligence Understanding with Virtual Counterfactual Calibration

Peng Chen [1], Xiao-Yu Guo [2], Yuan-Fang Li [3],
Xiaowang Zhang [1] *, Zhiyong Feng [1]
[1] College of Intelligence and Computing, Tianjin University
[2] AIML, University of Adelaide [3] Monash University

Abstract

Social intelligence is essential for understanding complex human expressions and social interactions. While large multimodal models (LMMs) have demonstrated remarkable performance in social intelligence question answering (SIQA), they are still inclined to generate responses relying on language priors and ignoring the relevant context due to the dominant prevalence of text-based data in the pre-training stage. To interpret the aforementioned language bias of LMMs, we employ a structure causal model and posit that counterfactual reasoning can mitigate the bias by avoiding spurious correlations between LMMs' internal commonsense knowledge and the given context. However, it is costly and challenging to construct multimodal counterfactual samples. To tackle the above challenges, we propose an output **D**istribution **C**alibration network with **V**irtual **C**ounterfactual (**DCVC**) data augmentation framework. DCVC devises a novel output distribution calibration network to mitigate the impact of negative language biases while preserving beneficial priors. Perturbations are introduced to the output distributions of LMMs to simulate the distribution shifts from counterfactual manipulations of the context, which is employed to construct counterfactual augmented data virtually. Experiments on multiple datasets demonstrate the effectiveness and generalizability of our proposed method.

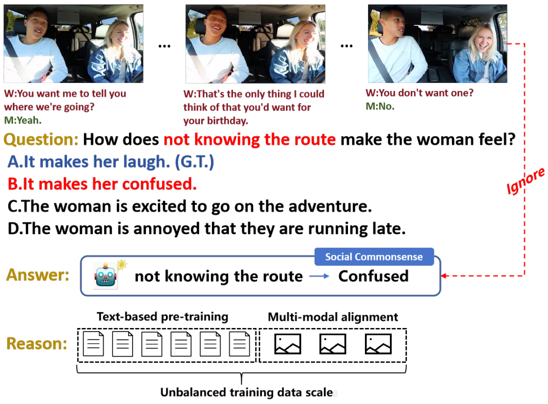

W:You want me to tell you where we're going?
M:Yeah.
W:That's the only thing I could think of that you'd want for your birthday.
W:You don't want one?
M:No.
Question: How does not knowing the route make the woman feel?
A.It makes her laugh. (G.T.)
B.It makes her confused.
C.The woman is excited to go on the adventure.
D.The woman is annoyed that they are running late.
Ignore
Answer:
Social Commonsense
not knowing the route → Confused
Text-based pre-training
Multi-modal alignment
Reason:
Unbalanced training data scale

Figure 1: An example in the Social-IQ-2.0 dataset. The input includes videos along with corresponding audio and subtitles. G.T. stands for the Ground-Truth answer. LMMs tend to select the incorrect answer (option B in red) based on their social commonsense knowledge obtained during pre-training.

1 Introduction

Social intelligence is essential for understanding complex human intentions and social interactions with machine learning models, which has emerged as a nascent area in Natural Language Processing (NLP) and multimodal communities in recent years. A few question-answering (QA) benchmarks have been proposed to evaluate the social intelligence of existing machine learning models (Sap et al., 2019a; Zadeh et al., 2019), including Social-IQ-2.0 (Wilf et al., 2023), a multiple-choice QA dataset with multimodal inputs(videos, audio and subtitles). However, existing works often utilize and optimize small models via modality feature alignment and/or leveraging external knowledge (Xie and Park, 2023). Research on social intelligence employing Large Multimodal Models(LMMs) remains under-explored.

To bridge this gap, we evaluate the performance of two powerful LMMs, Video-LLaVA (Lin et al., 2023) and CREMA (Yu et al., 2024), on the Social-IQ-2.0 dataset. Experimental results (Table 1) show that LMMs demonstrate remarkable performance under the zero-shot setting due to their exceptional cross-modal understanding and reasoning capabilities, achieving accuracy of 61.06% for Video-LLaVA and 63.33% for CREMA. Nevertheless, LMMs are prone to generating content frequently seen during their pre-training stage (corresponding to social commonsense knowledge in the LMMs) due to the different data scales between text-based

*Corresponding authors

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1300–1310
November 12-16, 2024 ©2024 Association for Computational Linguistics

pre-training and multimodal alignment (Pi et al., 2024). As shown in Figure 1, despite the woman in the video ”laughed” (G.T.) in response to her not knowing the route, Video-LLaVA selected the incorrect answer based on the social commonsense acquired during the text-based pre-training stage, which suggests that not knowing the route can “make her confused”. Extra examples are shown in Figure 7 in Appendix B. To further assess the language biases inherent in LMMs, we statistically analyzed the mean output distributions of Video-LLaVA when responding to emotion-related questions: the top 15 words with the highest output probabilities are shown in Figure 2. It is evident that the output distributions with multimodal inputs closely resemble those without context, yet they significantly differ from the answer proportions. To mitigate such biases, Zhang et al. (2024) proposed to detach the output distribution of video-free inputs to ensure that the LMMs generate responses based solely on the visual context. However, beneficial language priors have also been inevitably removed.

To mitigate undesirable language biases while preserving beneficial priors, we propose an output **D**istribution **C**alibration network with **V**irtual **C**ounterfactual data augmentation (**DCVC**). Specifically, we first employ a Structural Causal Model (SCM) (Pearl, 2009) to characterize the causal effect for social intelligence QA, which denotes that the spurious correlation between LMMs and context can be avoided by counterfactual reasoning. Then, an output distribution calibration network is employed to calibrate the output distribution of LMMs adaptively. Furthermore, We expect further to mitigate the language bias of LMMs with counterfactual data augmentation. However, constructing multimodal counterfactual samples is challenging and costly, especially for the complex video modality. To efficiently construct counterfactual samples, we propose a **V**irtual **C**ounterfactual **D**ata **A**ugmentation (**VCDA**) framework to construct virtual counterfactual samples with flipped labels and filter out the high-quality data. Perturbations are introduced to the output distribution of LMMs to simulate the shifts in distributions resulting from counterfactual manipulations of the context.

Overall, our main contributions are as follows:

- We utilize a Structural Causal Model (SCM) to interpret and quantify the language biases in LMMs for the social intelligence QA task.

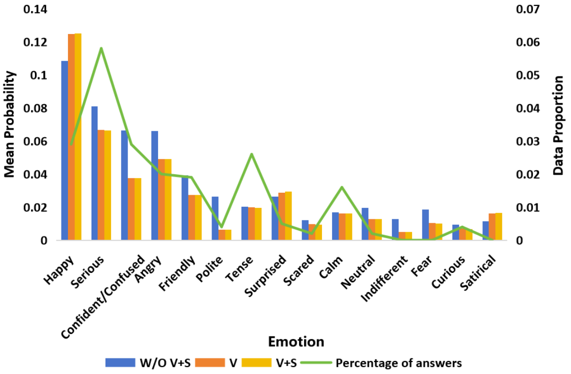


Figure 2: Mean output distributions of Video-LLaVA when responding to emotion-related questions across different inputted modalities, with ’V’ representing video and ’S’ representing subtitles. The proportions of answers are given in the line graph for comparison.

- We employ an output distribution network to adaptively calibrate the output distribution of LMMs, which largely mitigates undesirable language biases and preserves beneficial language priors.

- To efficiently construct multimodal counterfactual samples, we propose a virtual counterfactual data augmentation framework to construct virtual counterfactual samples that simulates the shifts in output distributions resulting from counterfactual manipulations of the context.

2 Related works

Multimodal Question Answering. Multimodal Question Answering aims to answer natural language questions given multiple input modalities, which requires multimodal understanding and commonsense reasoning skills. Previous benchmarks (Antol et al., 2015; Xu et al., 2017; Jang et al., 2017) focus on visual facts such as location and objects/attributes. In recent years, more benchmarks (Lei et al., 2018; Zellers et al., 2019; Sap et al., 2019b; Chen et al., 2024) have tended to tackle commonsense and causal reasoning questions. Regarding the existing methods, while earlier works (Cheng et al., 2023; Yu et al., 2021; Ye et al., 2023) concentrate on multimodal representation learning and modality fusion, large vision-and-language models align the multimodal feature to LLMs by instruction tuning (Ko et al., 2023; Liu et al., 2023; Yu et al., 2024). Different from these works, we further examine the impact of language biases in

LMMs and promote the performance of existing LMMs by adaptively calibrating such biases.

Social Intelligence Learning. Social intelligence is a long-standing research area within sociology and psychology (Andreou, 2006; Daniel Goleman, 2007). In recent years, the study of social intelligence has gained increasing momentum within the machine learning communities. Zadeh et al. (2019) propose a multimodal QA benchmark that requires understanding and reasoning skills of social commonsense and human interaction. Bosselut et al. (2019) conduct an extensive investigation on the automated construction of social commonsense knowledge bases. Furthermore, Xie and Park (2023) propose to leverage emotional cues in social interaction through contrastive learning. While previous work on Social Intelligence has primarily focused on small, fine-tuned models, Our work concentrates on evaluating and enhancing LMMs.

Mitigating Biases in Large Language Models. Studies have been conducted to measure and mitigate political and societal biases of machine learning methods (Zhao et al., 2018; Bender et al., 2021). Recently, with the growing prevalence of large language models, multiple works have examined the biases within these models (Zhou et al., 2023; Li et al., 2024). Zhang et al. (2024) have demonstrated that the outputs of LMMs are primarily influenced by language priors, enabling them to provide confident answers even without visual input. Chen et al. (2024) initially employ fine-tuning based and chain-of-thought based methods to mitigate such bias. Zhang et al. (2024) introduce Visual Debias Decoding (VDD) strategies to redirect the model's focus toward vision information. Our work also advances existing visual decoding strategies, adaptively mitigating language biases in LMMs through calibrated adjustments to the output distribution.

3 Method

In this section, we describe our proposed DCVC framework for mitigating language bias of LMMs. In section 3.1, we introduce the Social Intelligence question-answering task (SIQA). In Section 3.2, a Structural Causal Model (SCM) (Pearl, 2009) is employed to interpret the causal effect for social intelligence QA, which demonstrates that counterfactual reasoning can mitigate the biases by avoiding the spurious correlations between LMMs and context. The next two sections show the specific design of our output distribution-based counterfactual reasoning approach, namely DCVC. In Section 3.3, we introduce a novel calibration network to calibrate output distributions of LMMs adaptively. In Section 3.4, we describe the virtual counterfactual data augmentation method employed to train the calibration network to rectify language biases.

3.1 Preliminary

Given input video v depicting social interaction, as well as corresponding audio a, subtitle s, question and options q, the goal of Social Intelligence QA is to predict a label (i.e., option) $\hat{y} \in \{A, B, C, D, \ldots\}$ corresponding to the right answer.

3.2 Language Bias Analysis

We formalize the causal effect for the Social Intelligence QA task via a Structure Causal Model (SCM) (Pearl, 2009). In Figure 4, an SCM is depicted through a directed acyclic graph $G = (V, E)$, where edges in E represent the causal relationships between key factors in SIQA, which are represented as nodes in V. The key factors include contextual features X (i.e., the content of the input video), knowledge embodied in Large Multimodal Model T, mediator variable M and the prediction Y. The details of SCM are shown as follows:

- $T \to X$. The directed edge between T and X indicates that X is encoded by LMM, and the representation of X inevitably integrates priors derived from T.

- $X \to M \leftarrow T$. M is a mediator variable blended with prior knowledge from LMM T and contextual feature X. The paths among the variables above denote that LMM encodes the contextual feature and integrates prior knowledge of LMM (such as grammar rules or commonsense knowledge) to generate responses.

- $X \to Y \leftarrow M$. The directed path $X \to Y$ denotes that the causal effect between X and Y is not fully represented by the path $X \to M \to Y$. Because the existing LMMs cannot fully represent all information contained in X. Instead, LMM is inclined to generate responses by utilizing social commonsense knowledge, rather than responding faithfully based on the context X. The mediation path $Y \leftarrow M$ is also inevitable due to the aforementioned mechanism of existing LMM.

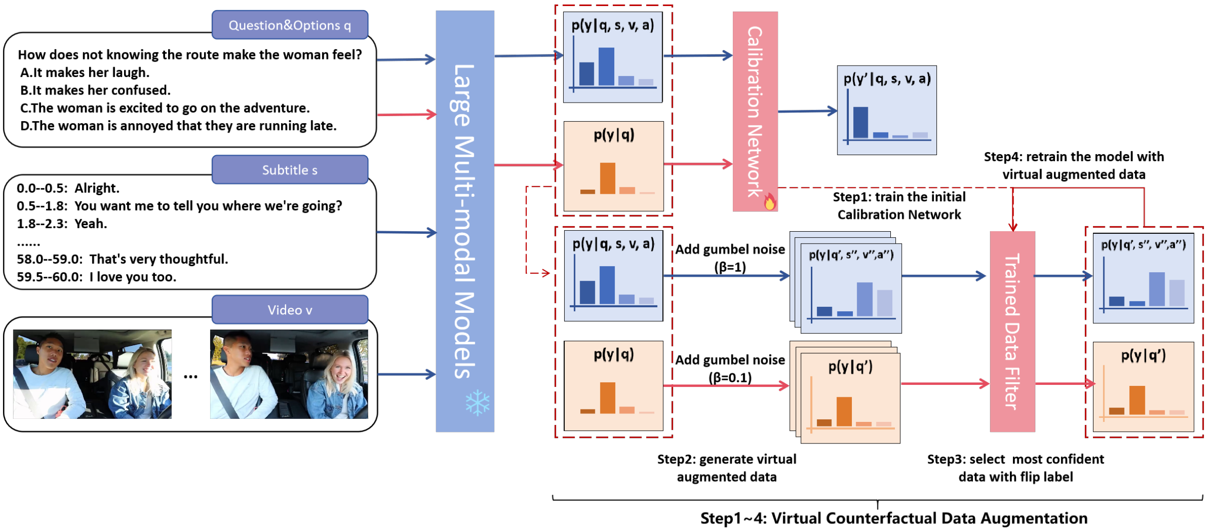

Question&Options q
How does not knowing the route make the woman feel?
A.It makes her laugh.
B.It makes her confused.
C.The woman is excited to go on the adventure.
D.The woman is annoyed that they are running late.
Subtitle s
0.0--0.5: Alright.
0.5--1.8: You want me to tell you where we're going?
1.8--2.3: Yeah.
......
58.0--59.0: That's very thoughtful.
59.5--60.0: I love you too.
Video v
...
Large Multi-modal Models
p(y|q, s, v, a)
p(y|q)
Calibration Network
p(y'|q, s, v, a)
Step1: train the initial Calibration Network
Step4: retrain the model with virtual augmented data
Add gumbel noise (β=1)
Add gumbel noise (β=0.1)
p(y|q', s'', v'',a'')
p(y|q')
Trained Data Filter
Step2: generate virtual augmented data
Step3: select most confident data with flip label
Step1~4: Virtual Counterfactual Data Augmentation

Figure 3: The overall architecture of our proposed output **D**istribution **C**alibration network with **V**irtual **C**ounterfactual data augmentation (**DCVC**). The **DC** adaptively calibrates the output distribution of the LMM to mitigate **undesirable** language biases while preserving **beneficial** priors. Furthermore, virtual counterfactual data augmentation is employed to decouple spurious correlations between the LMM and the context.

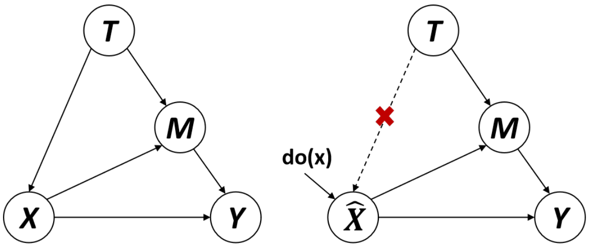


Figure 4: (a) Causal graph for social intelligence QA. (b) Intervene on context X to mitigate spurious correlation related to LMM *T*.

Considering the SCM, it is hard for LMMs to comprehensively capture the true causality between *X* and *Y*, as spurious correlation exits in these two paths: $T \rightarrow X$ and $T \rightarrow M \rightarrow Y$. Specifically, LLMs incorporate prior knowledge while encoding contextual features ($T \rightarrow X$) and generating responses ($T \rightarrow M \rightarrow Y$). While language priors are essential for generating responses, excessive incorporation of prior knowledge when encoding X is prone to lead to misunderstandings or neglect of the context. We propose that the spurious correlations can be avoided by blocking the back-door path $X \leftarrow T \rightarrow M$ via the do($\cdot$) operation:

$$\begin{aligned} P(Y|do(X=\hat{x})) &= \sum_k P(Y|X=\hat{x}, T=t)P(T=t) \\ &= \sum_k P(Y|X=\hat{x}, T=t, M=g(\hat{x},t))P(T=t) \end{aligned} \tag{1}$$

By blocking the back-door path $T \rightarrow X$ by intervening on X, the LMMs become more sensitive to X, thus avoiding over-reliance on the language priors. We will implement the intervention through output distribution-based Virtual Counterfactual Calibration in the next two sections.

3.3 Output Distribution Calibration Network

To mitigate undesirable language biases while preserving beneficial priors, we propose an Output Distribution Calibration Network (DC) to calibrate the output distribution of LMMs adaptively. As shown in Figure 3, DC controls the output distribution of LMMs $p(y|q,s,v,a)$ given the representation of q and language priors $p(y|q)$. Specifically, the question and options q are fed into the pre-trained model for encoding: $h_q = Encoder(q)$. Then, we calculate the element-wise product of the representation for each option with its corresponding output distribution and language priors to obtain the weighted representations for each option:

$$\hat{h}_q = Concat(h_q \circ p(y|q,v,s,a), h_q \circ p(y|q)) \tag{2}$$

where $\hat{h}_q$ denotes the weighted representations for each option, $p(y|q,v,s,a)$ denotes the output distribution of LMM while $p(y|q)$ denotes language priors. Finally, $\hat{h}_q$ is fed into an MLP classifier with softmax for output distribution calibration: $f_{Cal} = softmax(\hat{h}_q \cdot W + b)$, where W and b are learnable parameters.

Through supervised training, DC is capable of assessing the impact of language priors and adaptively mitigate undesirable biases, thereby promoting causal inference:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} y_i \log(f_{Cal}(\hat{h}_q)) \tag{3}$$

where N represents the number of options.

To mitigate the bias of primitive h_q, a Mean Squared Error (MSE) loss function is employed:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}((h_{q_i} \cdot W' + b') - \sigma)^2 \tag{4}$$

where W' and b' are learnable parameters, h_{q_i} are representation of i-th option, $\sigma = \frac{1}{N}\sum_{i=1}^{N}(h_{q_i} \cdot W' + b')$. The MSE loss function is applied to make the output distributions derived solely from the representation of options closer to the average.

The final training objective is:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{MSE} \tag{5}$$

where α is a hyperparameter.

3.4 Virtual Counterfactual Augmentation

To reiterate, the causal intervention operation can block the back-door path $X \leftarrow T \rightarrow M$ and encourage causal inference. Inspired by previous works (Dong et al., 2023; Li et al., 2024), we propose to construct counterfactual augmented data to realize causal intervention, i.e., inverting causal features through slight modifications to reverse the label. Specifically, we would like to construct counterfactual samples by slightly perturbing the input video v, audio a, subtitle s in which way the label is reversed.

However, compared to text-based perturbations, it is exceedingly challenging and costly to construct multimodal counterfactual samples for complex videos. While there have been multiple prior works in data augmentation for videos (Yun et al., 2020; Ding et al., 2022), they focus on the replacement and simple modification of image regions within videos, which is hard to be employed to perform precise adjustments to social interaction in videos. As a result, it remains to be explored how to precisely modify videos for generating counterfactual data.

Inspired by the Virtual Data Augmentation (VDA) technique proposed by Zhou et al. (2021), we propose a Virtual Counterfactual Data Augmentation (VCDA) framework, as shown in Figure 3, to construct virtual counterfactual samples with flipped labels and filter for high-quality data. Instead of being directly introduced to the input context, perturbations are introduced to the output distributions $p(y|q, v, s, a)$ and language biases $p(y|q)$ of LMMs to simulate the shifts in distributions resulting from counterfactual manipulations of the context. This serves as an indirect and virtual counterfactual data augmentation. The augmented data will be employed to train the calibration network to promote the calibration performance of the model further.

Specifically, Gumbel noise is added to $p(y|q, v, s, a)$ and $p(y|q)$ to perform perturbation. The probability density function of the Gumbel distribution is given by:

$$f(x; \mu, \beta) = \frac{1}{\beta}\exp\left(-\frac{x-\mu}{\beta} - \exp\left(-\frac{x-\mu}{\beta}\right)\right) \tag{6}$$

where μ is the location parameter and β is the scale parameter.

We sample a random variable with the same dimension as $p(y|q, v, s, a)$ from the Gumbel distribution, denoted as $Z_{output} \sim \text{Gumbel}(\mu, \beta = 1)$. Similarly, $Z_{priors} \sim \text{Gumbel}(\mu, \beta = 0.1)$ with the same dimension as $p(y|q)$ is sampled. Then, the significantly perturbed distribution $p^{''}(y|q, v, s, a)$ is obtained by shifting the original distribution $p(y|q, v, s, a)$ by Z_{output}, where $''$ denotes significant perturbation. To obtain the slightly perturbed distribution $p^{'}(y|q)$, where $'$ denotes minor perturbation, we shift the original distribution $p(y|q)$ by Z_{priors} with minor scale parameter. Intuitively, $p^{'}(y|q)$ denotes minor perturbations to the question and options q, namely $p(y|q')$. Since the simultaneous perturbation to q is minor, $p^{''}(y|q, v, s, a)$ simulates the effect of applying significant perturbations to the video v, audio a and the subtitle s, namely $p(y|q', v^{''}, s^{''}, a^{''})$.

As the Virtual Counterfactual Augmentation is unsupervised, we employed FlipDA proposed by Zhou et al. (2022) to filter and retain high-quality augmented data. Specifically, we first train the calibration network with original data. Then, virtual augmented data will be generated with the aforementioned method. Next, we apply the trained calibration network as the data filter and select augmented samples with the highest probabilities for flipped labels. Finally, we retrain the DC with the original and counterfactual augmented samples.

4 Experimental Setup

4.1 Datasets

To validate the language bias mitigation performance of our proposed DCVC method, we conduct experiments on two social intelligence understanding QA datasets: Social-IQ-2.0 (Wilf et al., 2023) and DeSIQ-2.0 (Guo et al., 2023). Additionally, NExT-QA (Xiao et al., 2021), a more general-purpose video QA dataset is employed to evaluate the generalizability of DCVC.

Social-IQ-2.0 is an improved version of Social-IQ (Zadeh et al., 2019) with multimodal, multiple-choice questions designed to evaluate the social intelligence understanding capability of machine learning models. The original video about human interactions, the corresponding extracted audio, and automatically generated transcripts are provided. Guo et al. (2023) reveals that Social-IQ, as well as Social-IQ-2.0, contain significant bias in which the distinction between the representations of correct and incorrect choices is readily discernible, regardless of the specific questions or contexts. They introduce DeSIQ and DeSIQ-2.0, two corresponding debiased datasets constructed by applying simple but effective perturbations to the original datasets. Detailed dataset statistics are shown in Appendix A in Table 4.

NExT-QA (Xiao et al., 2021) is a rigorously designed video question answering (VideoQA) benchmark to advance video understanding from the description to the explanation of temporal actions and causal reasoning. Causal questions account for approximately half (48%) of the whole dataset while temporal questions compose 29% of the dataset. Detailed dataset statistics are shown in Appendix A in Table 5.

4.2 Baselines

We compare DCVC with both small and large multimodal language models (LMMs). The fine-tuned small models include **RoBERTa-large** (Liu et al., 2019), **T5-small** (Guo et al., 2023) and **MMTC-ESC** (Xie and Park, 2023). MMTC-ESC proposes to leverage emotional cues in social interactions through contrastive learning and applies the cross-modal attention module to align multimodal representations, which achieves state-of-the-art (SOTA) performance. For video-capable LMMs, we employ two recent, strong models: Video-LLaVA (Lin et al., 2023) and CREMA (Yu et al., 2024) in a zero-shot setting. **Video-LLaVA** (Lin et al., 2023) unifies visual representation into the language feature space to advance the foundational LLM towards a unified LMM and achieves superior performances on a broad range of 9 multimodal benchmarks. **CREMA** (Yu et al., 2024) is an efficient and modular modality-fusion framework for injecting any new modality into video reasoning and achieves better/equivalent performance against strong LMMs with significantly fewer trainable parameters. Additionally, we also fine-tune CREMA as a control. Visual Debias Decoding (**VDD**) Zhang et al. (2024) is a decoding strategy that introduces a calibration step to adjust the output distribution with that of the image-free input. We adapted VDD to make it applicable for social intelligence QA and employed it as a baseline.

4.3 Implementation Details

We utilize the same instructions as Video-LLaVA to obtain output distributions. We set the temperature to 0.1 for Video-LLaVA and set the beam size to 5 for CREMA. For fine-tuning CREMA, Learning rate is set to 5e-5, and max training epoch is set to 10. For our proposed DCVC, we employ RoBERTa-base (Liu et al., 2019) to encode q. The learning rate is set to 1e-5, and the weight decay is set to 1e-2. We apply AdamW as an optimizer with a batch size of 16. Our experiments show optimal results are achieved when α is set to 0.1. For virtual counterfactual data augmentation, we generate ten samples for each original sample. All experiments are conducted on the 2 × NVIDIA 4090 GPUs.

5 Results and Analysis

In this section, we validate the effectiveness of our proposed DCVC through multiple experiments and conduct further analyses. In Section 5.1, the overall performance of DCVC is compared against multiple baselines in Social-IQ-2.0 dataset and DeSIQ-2.0 dataset. In Section 5.2, ablation study is conducted to evaluate the effectiveness of each component. Afterward, we analyze the impact of the type of noise for virtual counterfactual data augmentation in Section 5.3. Finally, we validate the generalizability of the output distribution calibration network in Section 5.4.

5.1 Overall Performance

The overall results are shown in Table 1. It can be seen that our proposed DCVC framework significantly ($p < 0.01$) improves the performance of ”vanilla” LMM Video-LLaVA (by 17.26 points on

Model	Social-IQ-2.0	DeSIQ-2.0
RoBERTa-large (Liu et al., 2019) [q, s]	73.55	81.38
T5-small (Guo et al., 2023) [q, s, v, a]	64.06	74.13
MMTC-ESC (Xie and Park, 2023) [q, s, v, a]	75.94	-
Video-LLaVA (Lin et al., 2023) [q, s, v]	61.06	85.69
Video-LLaVA + VDD (Zhang et al., 2024)	58.23	78.43
Video-LLaVA + DCVC (ours) [q, s, v]	**78.32**	97.04
CREMA (Yu et al., 2024) [q, s, v, a]	63.33	87.62
CREMA + VDD (Zhang et al., 2024)	62.65	84.10
CREMA(fine-tuned) [q, s, v, a]	76.39	**98.29**
CREMA + DCVC (ours) [q, s, v, a]	77.78	97.27

Table 1: Accuracy on the Social-IQ-2.0 and DeSIQ-2.0 development sets. The content in "[ ]" denotes the modalities of the model (q: question and answer options, s: subtitle, v: video, a: audio).

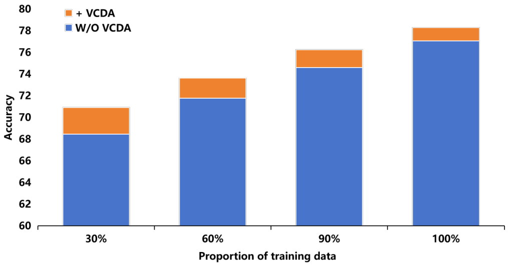


Figure 5: The performance of DCVC under varying proportions of training data (30%, 60%, 90%, 100%) on the Social-IQ-2.0 dataset. The orange segment in the bar chart denotes the performance improvement achieved by incorporating VCDA.

Social-IQ-2.0 and 11.35 points on DeSIQ-2.0) and CREMA (by 14.45 points on Social-IQ-2.0 and 9.65 points on DeSIQ-2.0). Moreover, CREMA, in the zero-shot setting, when coupled with DCVC, achieves comparable performance with dataset-specific fine-tuned results.

As previously mentioned, language biases inherent in the pre-training phase of language models negatively impact LLMs' performance on SIQA. To mitigate the biases, Visual Debias Decoding (VDD) directly detaches the output distribution of video-free inputs to ensure that the LMMs generate responses based solely on the visual context. While excelling in mitigating hallucinations, the rather simplistic calibration of VDD removes not only language biases but also the linguistic priors beneficial for social intelligence reasoning (e.g., basic social commonsense). Consequently, the performance of VDD, when applied to Video-LLaVA, exhibits a moderate decline compared with the baseline. In comparison, our proposed DCVC framework measures the extent of language bias based on the output probabilities. It employs an adaptive calibration network enhanced with virtual counterfactual augmentation, which achieves state-of-the-art (SOTA) performance (78.32% for Video-LLava and 77.78% for CREMA on Social-IQ-2.0).

Surprisingly, Video-LLaVA achieved an accuracy 85.69% on the DeSIQ-2.0 dataset, which is significantly higher than the Social-IQ-2.0 dataset. This experimental result can be attributed to the fact that DeSIQ-2.0 directly replaces the options of the original samples with others from the dataset, rendering the option representations no longer discernible. However, LMMs can easily distinguish the substitute options based on the semantics of the question and options, as the new options, which originate from other samples, often have a lower semantic relevance to the question. Nonetheless, DCVC still demonstrates an improvement of 11.35 points. We leave the construction of an unbiased and more challenging dataset for evaluating LMMs' social intelligence understanding to future work.

5.2 Ablation Study

An ablation study of Video-LLaVA on the Social-IQ-2.0 and DeSIQ-2.0 dataset is conducted to validate the effectiveness of each component. The results are shown in Table 2. The tested modules include: (1) **VCDA**: the virtual counterfactual data augmentation introduced in our work, (2) **MSE Loss**: employed to mitigate the bias of primitive representation of question and options, and (3) **Calibration Network**: our proposed output Distribution Calibration network. As can be seen in the table, with the removal of each component, there

Module	Social-IQ-2.0	DeSIQ-2.0
Video-LLaVA + DCVC	78.32	97.04
- VCDA	77.09	95.64
- MSE Loss	76.33	96.02
- Calibration Network	61.06	85.69

Table 2: Ablation study (Accuracy) on the Social-IQ-2.0 and DeSIQ-2.0 dataset.

is a drop in model performance, demonstrating the effectiveness of each component.

From another perspective, The components are closely interconnected and build upon each other. MSE loss alleviates the inherent biases present in the calibration network. Virtual counterfactual data augmentation, a critical component for mitigating the language biases of LMMs, generates probabilistic augmented data that simulates perturbations in the context. As it is exceedingly difficult to perform actual data augmentation directly on video-related context, our virtual data augmentation approach provides an efficient way to further optimize the calibration network, resulting in better calibration performance.

We also evaluate the performance of DCVC under varying proportions of training data (30%, 60%, 90%, 100%) on the Social-IQ-2.0 dataset. As depicted in Figure 5, the performance of Video-LLaVA with DCVC improved further with increasing training data. Notably, virtual counterfactual data augmentation is more effective with less training data. When only 30% of the training data was utilized, the VCDA module achieved a performance enhancement of 2.48 points. Thus, DCVC is especially beneficial in the low-resource setting.

5.3 Noise Selection Study

We further investigated the impact of different types of noise on the performance of our framework. The tested noise was sampled from three distinct distributions, namely: (1) Gumbel, (2) Logistic, and (3) Gaussian. As depicted in Table 3, all three noises yield comparable performance, with Gumbel noise demonstrating slightly better performance, which could be attributed to its better suitability for sampling from discrete distributions.

5.4 Generalizability Analysis

To evaluate the generalizability of the output distribution calibration network, we further assess its performance on NExT-QA. Figure 6 shows that the calibration network consistently yields performance improvements over the original LMMs. While fine-tuned CREMA already achieves a respectable 71.6% accuracy, the calibration network still results in a 1-point increase. The performance gain is even more pronounced in the zero-shot setting, where the original model performance is lower. Compared to Social-IQ-2.0, the improvements offered by the calibration network are relatively limited on NExT-QA. This experimental result can be partly attributed to the fact that NExT-QA encompasses a more diverse range of question types, making it more challenging for the calibration network to perform uniform calibration.

Types of Noise	Social-IQ-2.0	DeSIQ-2.0
Gumbel	78.32	97.04
Logistic	76.73	96.48
Gaussian	77.86	96.70

Table 3: The effect of different types of noise on the Social-IQ-2.0 dataset and DeSIQ-2.0 dataset.

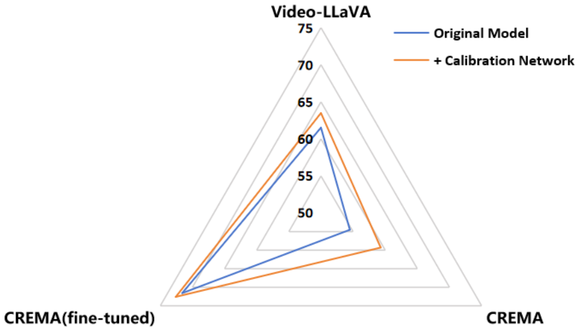


Figure 6: Generalizability analysis of the calibration network on the NExT-QA dataset. The evaluation metric is accuracy.

6 Conclusion

In this paper, we employ a structural causal model to interpret and quantify the language biases of LMMs in the social intelligence question-answering problems. To mitigate the biases while preserving beneficial priors, we propose an output distribution calibration network with virtual counterfactual data augmentation. Experiments on multiple datasets have demonstrated the effectiveness and generalizability of the proposed method. In future work, we will further explore the intrinsic reasons for language bias in LMMs.

7 Limitations

We have only validated the effectiveness of the proposed method on multiple LMMs with 7b parameter scales. Experiments on LMMs of 13b and 33b are expected to be conducted in the future work. In addition, we have analyzed the causal effects of language biases in LMMs through a structural causal model. However, the internal reasons for the existence of biases and other biases in LMMs remain to be explored.

8 Ethics Statement

The datasets and models used in the paper are open-source. This work specifically focuses on a targeted investigation of a particular type of bias, namely language bias of LMMS, not encompassing all forms of bias.

Acknowledgements

This work was supported by the Project of Science and Technology Research and Development Plan of China Railway Corporation (N2023J044) and the DARPA Assured Neuro Symbolic Learning and Reasoning (ANSR) program under award number FA8750-23-2-1016.

References

Eleni Andreou. 2006. Social preference, perceived popularity and social intelligence: Relations to overt and relational aggression. In *School Psychology International*, page 27(3):339–351.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779. Association for Computational Linguistics.

Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. 2024. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. *CoRR*, abs/2403.18346.

Feng Cheng, Xizi Wang, Jie Lei, David J. Crandall, Mohit Bansal, and Gedas Bertasius. 2023. Vindlu: A recipe for effective video-and-language pretraining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 10739–10750. IEEE.

Daniel Goleman. 2007. *Social intelligence*. Random house.

Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. 2022. Motion-aware contrastive video representation learning via foreground-background merging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 9706–9716. IEEE.

Xiangjue Dong, Ziwei Zhu, Zhuoer Wang, Maria Teleki, and James Caverlee. 2023. Co^{2}pt: Mitigating bias in pre-trained language models through counterfactual contrastive prompt tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 5859–5871. Association for Computational Linguistics.

Xiaoyu Guo, Yuan-Fang Li, and Reza Haf. 2023. Desiq: Towards an unbiased, challenging benchmark for social intelligence understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3169–3180. Association for Computational Linguistics.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: toward spatio-temporal reasoning in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1359–1367. IEEE Computer Society.

Dohwan Ko, Ji Soo Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. 2023. Large language models are temporal and causal reasoners for video question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10,* *2023*, pages 4300–4316. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2018. TVQA: localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1369–1379. Association for Computational Linguistics.

Ang Li, Jingqian Zhao, Bin Liang, Lin Gui, Hui Wang, Xi Zeng, Kam-Fai Wong, and Ruifeng Xu. 2024. Mitigating biases of large language models in stance detection with calibration. *CoRR*, abs/2402.14296.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. 2024. Strengthening multimodal large language model with bootstrapped preference optimization. *CoRR*, abs/2403.08730.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019a. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4462–4472. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Socialiqa: Commonsense reasoning about social interactions. *CoRR*, abs/1904.09728.

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

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9777–9786. Computer Vision Foundation / IEEE.

Baijun Xie and Chung Hyuk Park. 2023. Multi-modal correlated network with emotional reasoning knowledge for social intelligence question-answering. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 3067–3073. IEEE.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1645–1653. ACM.

Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. 2023. Hitea: Hierarchical temporal-aware video-language pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15359–15370. IEEE.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 3208–3216. AAAI Press.

Shoubin Yu, Jaehong Yoon, and Mohit Bansal. 2024. CREMA: multimodal compositional video reasoning via efficient modular adaptation and fusion. *CoRR*, abs/2402.05889.

Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, and Jinhyung Kim. 2020. Videomix: Rethinking data augmentation for video classification. *CoRR*, abs/2012.03457.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8807–8817. Computer Vision Foundation / IEEE.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2024. Debiasing multimodal large language models. *CoRR*, abs/2403.05262.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 15–20. Association for Computational Linguistics.

Jing Zhou, Yanan Zheng, Jie Tang, Li Jian, and Zhilin Yang. 2022. Flipda: Effective and robust data augmentation for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8646–8665. Association for Computational Linguistics.

Kun Zhou, Wayne Xin Zhao, Sirui Wang, Fuzheng Zhang, Wei Wu, and Ji-Rong Wen. 2021. Virtual data augmentation: A robust and general framework for fine-tuning pre-trained models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3875–3887. Association for Computational Linguistics.

Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. 2023. Explore spurious correlations at the concept level in language models for text classification. *CoRR*, abs/2311.08648.

Appendix

A Dataset details

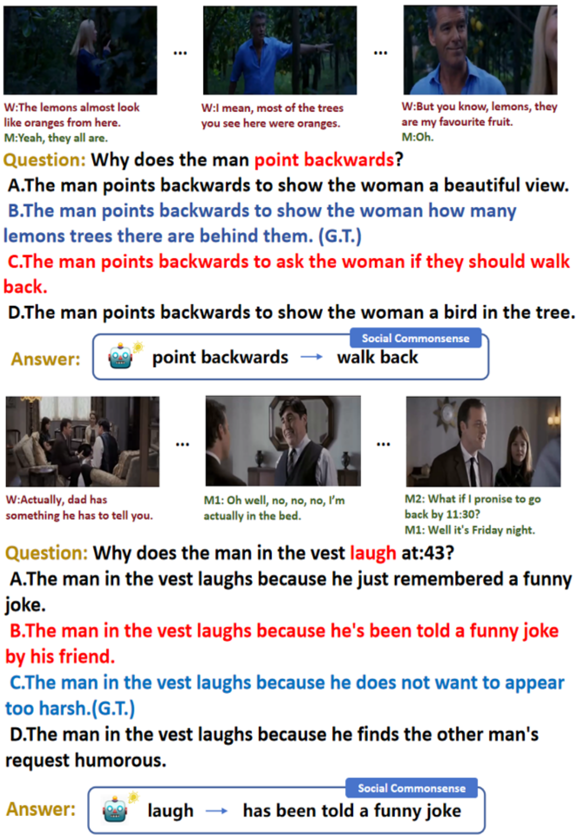
Number	Train	Val	Total
Videos	877	134	1,011
Questions	5,558	881	6,439

Table 4: Statistics of the Social-IQ-2.0 and DeSIQ-2.0 datasets. For each question, there are four options and only one correct answer.

Number	Train	Val	Test	Total
Videos	3,870	570	1,000	5,440
Questions	3,4132	4,996	8,564	47,692

Table 5: Statistics of the NExT-QA dataset. For each question, there are five options and only one correct answer.

B Extra examples of language priors in LMMs on the Social-IQ-2.0 dataset



W:The lemons almost look like oranges from here.
M:Yeah, they all are.

W:I mean, most of the trees you see here were oranges.

W:But you know, lemons, they are my favourite fruit.
M:Oh.

Question: Why does the man point backwards?

A.The man points backwards to show the woman a beautiful view.

B.The man points backwards to show the woman how many lemons trees there are behind them. (G.T.)

C.The man points backwards to ask the woman if they should walk back.

D.The man points backwards to show the woman a bird in the tree.

Answer: point backwards → walk back

W:Actually, dad has something he has to tell you.

M1: Oh well, no, no, no, I'm actually in the bed.

M2: What if I pronise to go back by 11:30?
M1: Well it's Friday night.

Question: Why does the man in the vest laugh at:43?

A.The man in the vest laughs because he just remembered a funny joke.

B.The man in the vest laughs because he's been told a funny joke by his friend.

C.The man in the vest laughs because he does not want to appear too harsh.(G.T.)

D.The man in the vest laughs because he finds the other man's request humorous.

Answer: laugh → has been told a funny joke

Figure 7: Extra two examples in the Social-IQ-2.0 dataset. The input includes videos along with corresponding audio and subtitles. G.T. stands for the Ground-Truth answer. LMMs tend to select the incorrect answer (option B in red) based on their social commonsense knowledge obtained during pre-training.