V-ALPHASOCIAL: Benchmark and Self-Reflective Chain-of-Thought Generation for Visual Social Commonsense Reasoning

Zongyu Lin[1*], **Zhikun Xu**[2*], **Xiaohan Song**[1*], **Yixin Wan**[1], **Xingcheng Yao**[1],
Tsung-Han Lin[1], **Selina Song**[1], **Pranav Subbaraman**[1], **Ben Zhou**[2],
Kai-Wei Chang[1†], **Yizhou Sun**[1†]
[1]University of California, Los Angeles, [2]Arizona State University
`linzongy21@cs.ucla.edu`, `kwchang@ucla.edu` `yzsun@cs.ucla.edu`

Abstract

Social commonsense reasoning naturally involves both the verbal and non-verbal cues of a social interaction. It is important for Large Vision-Language Models (VLMs) to leverage both textual and visual information in performing tasks like social understanding and reasoning. However, while current LLMs have shown good social reasoning capabilities in textual context, whether they can effectively incorporate visual information in social comprehension remains under-explored. To narrow the gap, we first construct and propose a benchmark: V-SOCIAL, featuring well-aligned text and visual content, tailored to assess visual social commonsense for multimodal foundation models. Through experimenting with V-SOCIAL, we find that even the most advanced VLM, GPT-4o, often falls short in social commonsense reasoning. This highlights the critical need to enhance the social grounding of VLMs. One major obstacle for improving this is the lack of high-quality data with good reasoning process. To overcome this obstacle, we introduce V-ALPHASOCIAL, a novel method that generates high-quality chain-of-thought reasoning paths from unlabeled data. We design a visual reasoning reward model to improve VLM, and then iteratively refine both the VLM and the reward model. Our extensive analysis showcases how our method enhances social commonsense reasoning, proposing an effective approach that facilitates deeper exploration into field. [*]

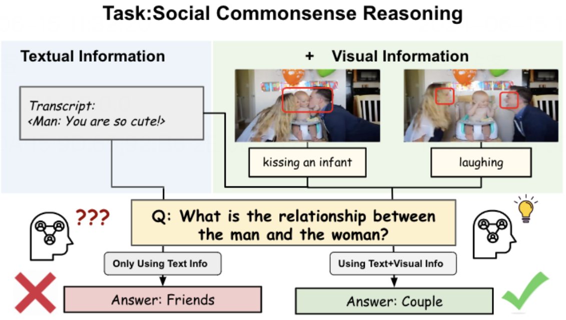


Figure 1: Illustration of visual social commonsense reasoning. Visual information is necessary to solve real-world social commonsense reasoning.

1 Introduction

Recent advances in Large Language Models (LLMs) have significantly enhanced their ability to understand social commonsense and mimic humans in generating socially acceptable response (Ku and Li, 2021; Sap et al., 2019). However, **social interactions are inherently multimodal** (Chaturvedi et al., 2019; Ekman and Oster, 1979), incorporating not only textual but also visual cues such as gestures, facial expressions, and actions. Integrating these features from visual modalities into social commonsense reasoning tasks is crucial for understanding and improving models' social commonsense reasoning holistically.

By incorporating visual encoder with LLM, Vision-Language Models has demonstrated decent performance on a wide range of tasks such as image & video captioning and understanding (Wang and Zhao, 2023; Lin et al., 2023). These models show potential in processing nuanced and context-rich social interactions. However, there is a lack of comprehensive benchmarks specifically designed to evaluate their abilities in visual social commonsense reasoning. Current benchmarks are often constrained by high-quality and aligned multi-modal data (Sap et al., 2019; Sabour et al., 2024). Moreover, they do not fully capture the nuanced context and reasoning depth necessary for a comprehensive understanding of social commonsense. (Wilf et al., 2023; Zadeh et al., 2019). To narrow the gap, we propose a comprehensive benchmark: V-SOCIAL, covering six curated dimensions of social commonsense reasoning. Our dataset consists of data with well-aligned text and visual modality. To

---

[*] Equal Contribution

[*] Corresponding authors.

[*] The benchmark is released in `https://github.com/Rafa-zy/VALPHASOCIAL`.

Findings of the Association for Computational Linguistics: ACL 2025, pages 19025–19047
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

simulate complex social situations, we collect long videos engaging multiple characters and propose questions that require deep reasoning from interactions between characters, which makes the task more challenging. After evaluations of the state-of-the-art VLMs on V-SOCIAL, we find that there are many failure cases such as reasoning errors and social understanding errors. These findings highlight the critical need to extend the general reasoning abilities of current VLMs to nuanced social commonsense reasoning tasks, which has been illustrated in Figure 1.

Moreover, addressing this need presents two key challenges: **the lack of high-quality training data with social reasoning processes**, and **reasoning errors in complex social situations**. As such, we propose a novel framework: V-ALPHASOCIAL, which leverages data with no reasoning path to generate high-quality reasoning processes and augments existing data with "thought process" like human beings. Our approach starts by using a single in-context data sample, grounded in social commonsense with a complete reasoning path, to generate reasoning processes on unlabeled data. A multi-modal discriminator is then trained to provide rewards for selecting optimal reasoning processes. To enhance the diversity of the reasoning process generated by VLMs, we explicitly augment the context with visual information using surrogate models such as BLIP and facial expression recognition (FER) models. This method not only enriches the training process by ensuring high fidelity in the generated reasoning across modalities.

We conduct comprehensive experiments on V-SOCIAL and SocialIQ2 (Wilf et al., 2023) to demonstrate the effectiveness of the proposed V-ALPHASOCIAL, achieving a performance gain of 14% compared to the baselines on V-SOCIAL. Additionally, we find that incorporating visual information significantly improves the performance of social commonsense reasoning, underscoring the importance of this task.

In summary, our contributions are threefold: (1) We curate V-SOCIAL, a novel and high-quality benchmark comprising aligned text, image, and video data to comprehensively evaluate multimodal foundation models' ability to conduct visual social commonsense reasoning. (2) We introduce V-ALPHASOCIAL, an innovative self-evolving approach to iteratively synthesize high-quality reasoning processes from unlabeled multimodal data, boosting the performance of visual social commonsense reasoning for video language models. (3) We conduct comprehensive experiments to demonstrate the effectiveness of our method and provide detailed analysis to cast insights into how visual cues are vital to social commonsense reasoning and how our method mitigates different types of errors by generating high-quality reasoning processes. Our work addresses previous research gaps through these contributions and sets the stage for future advancements in multi-modal understanding in social contexts.

Table 2: Key statistics of V-SOCIAL dataset.

Statistic	**Number**
Q&A Pairs	956
Average Number of Characters	2.46
Average Dialogue Length (Turns)	10.43
Total Videos	128
Average Video Duration (seconds)	90.73
Taxonomy	6
Emotion Understanding	334
Social Relationship Reasoning	190
Social Norm Understanding	201
Conflict Resolution	79
Persuasion and Influence	78
Sense of Humor	74
Task Complexity	2
Easy	703
Hard	253

2 V-SOCIAL

2.1 Taxonomy of Social Commonsense Reasoning

Inspired by previous work on social science and social intelligence for AI (Liu et al., 2023; Li et al., 2024; Yang et al., 2024; Ziems et al., 2024), our annotation taxonomy for V-SOCIAL is represented by the following six dimensions: emotion understanding, social relationship reasoning, social norm understanding, conflict resolution, persuasion and influence, and sense of humor.

To describe the logical relationships and hierarchy among the six dimensions of social commonsense in VLMs (also suitable for LLMs), we categorize them into two broad functionalities: social commonsense understanding and social commonsense expression. These categories reflect the capabilities of VLMs to understand and interpret

Table 1: Comparison Between Different Benchmarks

Benchmark	Task Coverage		Modality for Contexts			Characteristics of Data	
	Emotion	Social	Text	Image	Video	Rich Context	Complex Reasoning
EQBench (Paech, 2024)	✓	✗	✓	✗	✗	✗	✗
EmoBench (Sabour et al., 2024)	✗	✓	✓	✗	✗	✗	✗
VCR (Zellers et al., 2019)	✗	✓	✓	✗	✓	✗	✓
SocialIQA (Sap et al., 2019)	✗	✓	✗	✗	✗	✗	✗
SocialIQ (Zadeh et al., 2019)	✗	✓	✓	✓	✓	✗	✗
SocialIQ2 (Wilf et al., 2023)	✗	✓	✓	✓	✓	✗	✗
V-SOCIAL	✓	✓	✓	✓	✓	✓	✓

social cues and contexts firstly, and secondly, to actively participate in social interactions by generating responses or actions. The following subsections present a breakdown of these categories.

2.1.1 Social Commonsense Understanding

Understanding is the foundation of social commonsense (Kihlstrom and Cantor, 2000). Without a robust understanding of emotions, social relationships, and norms, a VLM cannot effectively engage in social interactions.

Emotion Understanding refers to the ability of VLMs to recognize, interpret, and respond to human emotions expressed through text and visual information (Moore, 2006; Xenos et al., 2024). This involves identifying explicit statements of emotion and inferring underlying feelings from context, tone, facial expressions, and choice of words.

Social Relationship Reasoning is the capacity of VLMs to understand the dynamics and nuances of different social relationships (e.g., familial, professional, casual, communal, etc.) through the process of communications between different parties and how these relationships influence communication and expectations (Evans and Aceves, 2016).

Social Norm Understanding is the ability to grasp and apply unwritten rules and behaviors that are considered acceptable within a society or group (Bicchieri, 2005). This includes understanding politeness, formality levels, and cultural norms. Moreover, social norms are not rigid truths or simple behavior mappings as they are fluid with the times and environments (Ziems et al., 2023).

2.1.2 Social Commonsense Expression

Expression builds on the foundational understanding. These skills allow VLMs to participate actively and appropriately in social exchanges, applying their understanding to enhance interactions.

Conflict Resolution is the skill of navigating disagreements or conflicts in a way that seeks to find a resolution that is acceptable to all parties involved (Behfar et al., 2008). It involves understanding the perspectives of each side, mediating the discussion, and suggesting compromises.

Persuasion and Influence refers to the ability to construct arguments or narratives that can change someone's beliefs, attitudes, or behaviors (Singer, 1989). This involves understanding the audience's values and leveraging rhetorical techniques effectively. For example, understanding utterance-level persuasion behaviors and predicting deduction outcomes with multimodal input is still challenging for current models (Lai et al., 2022).

Sense of Humour is the capability to understand, generate, and appropriately use humor (Graham, 1995). This includes recognizing different types of humor (e.g., sarcasm, wit, puns) and knowing when humor is appropriate in social interactions.

2.2 Problem Formulation

With our taxonomy, we formalize the problem of visual social commonsense reasoning as follows. The input data tuple (v, c, q) consists of: v, representing video content that includes visual scenes depicting multi-party interactions, diverse social settings, and dynamic social exchanges; c, the textual context, which includes detailed speaker information and transcripts extracted from audio data, synchronized with the visual content to provide a cohesive context; and q, a question designed to probe the model's capacity for reasoning about social interactions and commonsense knowledge. The objective for the models, parameterized by p_θ, is to first generate a reasoning process r, and then derive an answer a based on the reasoning.

2.3 Dataset Construction

Traditional benchmarks on social commonsense reasoning focus on the textual part (Sap et al., 2019; Ziems et al., 2023). However, the real challenges faced by these models in practical social applications are far more intricate, often involving multi-modal and complex questions, and nuanced social reasoning. We compare V-SOCIAL and existing social commonsense reasoning benchmarks in Table 1 and summarize key statistics of V-SOCIAL in Table 2. For task coverage, the "Emotion" refers to questions directly asking about emotional expressions, which SocialIQ2 often mixes implicitly with questions.

Specifically, our consistent annotation process, across videos, images, and texts, achieves the well-aligned multi-modal input in formulating complicated social contexts, by which models use to respond appropriately. The features of V-SOCIAL could be summarized as **multi-model input**, **richer context**, **deeper reasoning paths**, and **multi-party, multi-turn interactions**, which is further explained in Appendix A. For annotation quality, seven undergraduate annotators are recruited, and each is assigned with about 30 videos. The raw annotations are 1139 question-answer pairs. However, after careful filtering by the authors, which are researchers at undergraduate or graduate level, data points that lack enough speaker information or clear transcripts, have ambiguous questions, or pair with trivial candidate choices without the need for visual information are filtered out. The final version of V-SOCIAL consists of 956 question-answer pairs, resulting in an 83.9% retention rate. More details of our dataset statistics and annotation process are shown in Appendix B.

3 V-ALPHASOCIAL

3.1 Overview

Due to the scarcity of video-to-text data with high-quality reasoning process, directly fine-tuning a visual language model to address broad social commonsense reasoning is challenging. Consider human beings: we don't explicitly train people on numerous cases in diverse social scenarios to handle issues. Instead, we enhance our social commonsense through a self-evolutionary process. Therefore, our V-ALPHASOCIAL employs a self-training approach, composed of three major components: self-training, tool-augmented context augmentation, and the multi-modal discriminator. The framework is illustrated in Figure 2. In detail, we prompt our video language model with tool-augmented context augmentation to generate reasoning paths on the ground truth data with question and final answer, and then pick out the one that matches the final answer and filtered by the multi-modal discriminator.

3.2 Social Commonsense Enhanced Self-Training

We propose a self-training pipeline on the video language model for social commonsense reasoning. We first utilize the in-context learning capabilities of visual language models to generate data with the reasoning format aligned with few-shot exemplars.

Given two types of common errors in existing models: inconsistent reasoning, and unawareness of common social norms (see Section 4.2) —we use two methods to enhance and ground the reasoning processes of V-ALPHASOCIAL in complex social contexts: (1) First, we leverage chain-of-thought prompting (CoT) (Wei et al., 2022), which is most commonly used to boost the reasoning ability of models. (2) Then, we borrow the idea from *Constitutional AI* (Bai et al., 2022) and proposed a variant of CoT: Social-of-Thought (SoT) by grounding the reasoning process with our handcrafted principles in social commonsense, as shown in Table 15.

Then, we train the basic policy model: VSocial-Policy with the backbone of VideoLLaVA (Lin et al., 2023) on the generated reasoning paths which match the correct final answer, which can be done in an iterative manner. In the generation process, we also use tools to augment the context and train a reward model to filter out high quality reasoning paths, which will be introduced as follows.

3.3 Tool-based Context Augmentation

Just as LLMs struggle with low-level computational tasks (Schick et al., 2024), VLMs often fail to capture fine-grained visual cues, such as body movements or facial expressions, which compromises the quality of reasoning. Therefore, it is logical to employ surrogate models as tools (Shen et al., 2024) to enhance contextual information, thereby improving the overall reasoning process. Additionally, we find that policy models tend to produce homogeneous data in practice, which can stifle the diversity of generated content and limit the model's creative potential. To boost both the **quality and diversity** of the reasoning process, we propose the following surrogate models to augment video-text

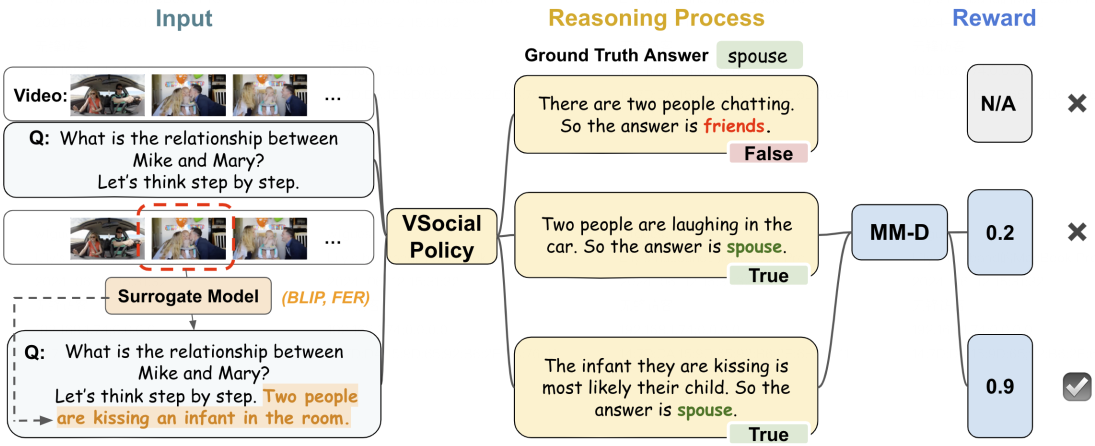

Input
Reasoning Process
Reward
Ground Truth Answer
spouse
Video:
...
Q: What is the relationship between Mike and Mary? Let's think step by step.
...
Surrogate Model
(BLIP, FER)
Q: What is the relationship between Mike and Mary? Let's think step by step. Two people are kissing an infant in the room.
VSocial Policy
There are two people chatting. So the answer is friends.
False
Two people are laughing in the car. So the answer is spouse.
True
The infant they are kissing is most likely their child. So the answer is spouse.
True
MM-D
N/A
0.2
0.9
✗
✗
✓

Figure 2: The overall framework of our V-AlphaSocial. Given a visual social commonsense problem, a video and text are input into our VSocial-Policy model to generate reasoning process and then feed into the multimodal discriminator to derive the reward. If the answer matches with the ground truth, we pick out the data samples that have higher reward; otherwise we just drop the data sample.

pairs. (1) BLIP (Li et al., 2023): We utilize BLIP to caption for each frame and then combine them together as a whole context. (2) Facial Expression Recognition Model: We leverage the off-the-shelf facial expression recognition model (Zhang et al., 2023) to extract the emotions from key characters in the videos. We then concatenate the information as prompts in the context for generating more high quality reasoning process.

3.4 Multi-modal Discriminator

The correctness of the final answer serves as a critical indicator of proper reasoning, as the paths leading to incorrect answers often explicitly highlight errors in the reasoning process. Additionally, assessing the consistency and correctness of the solution is generally more straightforward than generating an entire solution process (Xu et al., 2022; Cobbe et al., 2021). Therefore, to more effectively measure the quality of generated reasoning paths, we train a multi-modal discriminator model, M_D, to optimize a value function that accurately judges the final answer. This approach uses the final answer as a robust signal, essentially representing it as the reward for the entire reasoning process. With the selected reasoning paths, we retrain the VSocial Policy iteratively. Finally, the complete algorithm is outlined in Algorithm 1. The experimental details can be found in the Section 4.1.

4 Experiments

For this part, we conduct extensive experiments to answer the following questions: Q1: Does visual modality brings more information that can help VLMs to perform better in social commonsense reasoning (see Section 4.2) Q2: How does different model perform on our benchmark (see Section 4.2)? Q3: How about the effectiveness of V-AlphaSocial? Does our model generalize well to our collected V-Social? (see Section 4.2) Q4: What are the failure modes of the baseline VLMs and how our method mitigates them? (see Section 4.2)

4.1 Experimental Setup

We leverage the training data of SocialIQ2 (Wilf et al., 2023) with 6,159 Question-Answer pairs, and then utilize **VideoLLaVA** (Lin et al., 2023) as the original policy model to generate reasoning paths, and collect positive and negative samples according to the ground truth of final label. Then, we train our policy model on positive pseudo-labeled data, and train our discriminator model on both positive and negative samples. This can be done in an iterative way. We report both the single-turn V-AlphaSocial and multi-turn results V-AlphaSocial* in our main table. Moreover, we evaluate other state-of-the-art vision language models, including **GPT-4 (Vision, Omni)** (OpenAI, 2023), **GEMINI-1.5-Pro** (Team, 2024), **Qwen-VL-Max** (Bai et al., 2023), **Phi-3-Vision** (Abdin et al., 2024), **GLM-4V** (Wang et al., 2023), **LLaVA-NEXT** (Liu et al., 2024), on our V-Social and SocialIQ2. We also provide human score by our annotator.

4.2 Evaluation on Visual language Models

In this section, we want to answer both Q1 and Q2, including the performance of different models and the effectiveness of our method. We test on both SocialIQ2 as well as our V-Social, which is more out-of-distribution compared with the training data. We have several observations: (1) Comparing two benchmarks, we can see that zero-shot model performs better on our benchmark, mainly because our benchmark provided well-aligned transcript information as important textual context, while SocialIQ2 only relies on visual information. (2) We can see that open-source models still fall behinds close-source models. GPT4-O and GPT4-V rank the top, coming with other close-sourced model such as GEMINI-1.5-Pro and Claude3. Also, we observe the decent performance of open-sourced model like GLM-4V and LLaVA-NEXT. (3) Our method: V-AlphaSocial improves VideoLLaVA by 14.4% on V-Social and 28.5% on SocialIQ2, showing that our method of generating reasoning process is effective. Since our training data is from SocialIQ2, so finetuning VideoLLaVA on SocialIQ2 brings more performance gain. (4) Last but not least, even the state-of-the-art model falls behind human score, emphasizing the necessity of improving VLMs in visual social commonsense reasoning.

Table 3: Performance of different models on our collected V-Social and SocialIQ2.

Model	#param	V-Social	SocialIQ2
	Closed-Source		
GPT4-O	-	82.1	72.8
GPT4-V	-	78.5	70.5
GEMINI-1.5Pro	-	71.7	65.1
	Open-Source		
Phi-3-Vision	4.2B	45.1	49.2
GLM-4V	9B	45.9	42.8
LLaVA-NEXT	7B	40.9	41.8
QwenVLMax	-	68.0	66.2
GLM4V	-	54.2	60.1
QwenVL-Plus	-	63.7	65.5
Hunyuan-Vision	-	59.6	56.3
VideoLLaVA	7B	37.8	39.7
V-AlphaSocial	7B	50.6	68.2
V-AlphaSocial* [†]	7B	52.2	70.1
Human	-	92.5	93.8

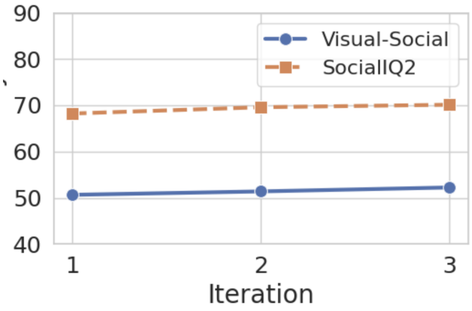


Figure 3: Multiple iterations of V-AlphaSocial on V-Social and SocialIQ2.

Multi Iteration Training We showcase the results of V-AlphaSocial with multiple iterations in Figure 3, and see that the performance of V-AlphaSocial can be improved as the iteration increases on both datasets, and most of the gain comes from the first and the second iteration.

Ablation Study We conduct ablation studies to prove the effectiveness of each component in V-AlphaSocial. As seen in Table 4, by removing each component:M_D, Social-of-Thought (SoT) and tool-based context augmentation (ContextAug), the performance of our method will decrease, showing that each component matters in the whole framework.

Table 4: Ablation Study of V-AlphaSocial

Method	SocialIQ2	V-Social
V-AlphaSocial	70.1	52.2
- M_D	68.3	50.5
- ContextAug	67.1	49.1

Performance Across Different Dimensions We showcase the performance of Video-LLaVA and V-AlphaSocial across six dimensions covered by our taxonomy. As displayed in Figure 4, the performance of all dimensions have been boosted, among which the dimension of conflict resolution has been greatly improved, while the dimension of sense of humor does not yield explicit performance gain, indicating questions about sense of humor are more challenging than conflict resolutions.

The influence of Visual Information To address Q4, we conduct experiments to investigate the influence of different modalities including text and video. Noted that the textual part can be further divided into dialogues between different characters

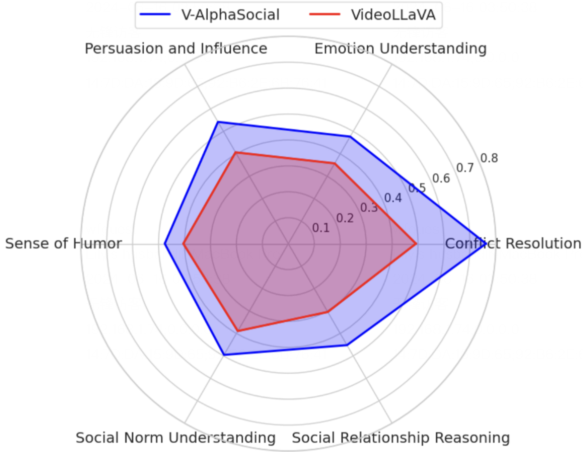


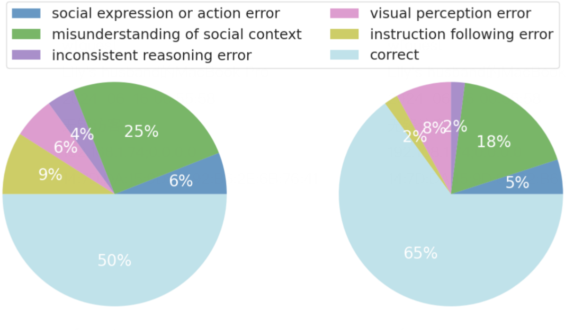
Figure 4: Performance on V-SOCIAL across different dimensions.

and question. We compare on Video-LLaVA using both video and text V.S. Video-LLaVA using only text part. We provide the results on both zero-shot setup and our self-trained setup in Table 5. We observe that both dialogue information and visual information are important to tackle social commonsense reasoning.

Table 5: The influence of visual information using zero-shot GPT-4O and VideoLLaVA.

Modality	GPT-4O	VideoLLaVA
Question Only	65.8	47.4
+ Text	78.8	50.2
+ Text + Visual	82.1	52.2

Error Analysis To give a in-depth analysis about the failure mode lying in the state-of-the-art zero-shot baselines, we choose 100 samples to evaluate on using chain-of-thought prompting on zero-shot Video-LLaVA and V-ALPHASOCIAL and manually label their failure modes in Figure 5. Basically, these failures can be divided into following five modes: (1) instruction following error; (2) visual perception error; (3) misunderstanding of social context; (4) inconsistent reasoning error; (5) social expression or action error. We can observe that the social expression error, misunderstanding of social context error, reasoning error and instruction following error have been mitigated by 4.25% on average, showing the effectiveness of our method. More details of these failure modes are illustrated in Appendix I.



(a) VideoLLaVA

(b) V-AlphaSocial

Figure 5: Error Analysis for VideoLLaVA and V-ALPHASOCIAL.

Task Complexity We also study how the complexity of tasks influence the performance of visual language model. As shown in Figure 16, the accuracy shrinks when increasing the difficulties of the task on both social commonsense understanding and expression, also pointing out the challenges of open-source vision language models to perform well on hard-level cases on V-SOCIAL. Also, as illustrated in Figure 7, there is a noticeable decline in performance as the number of characters in the video increases. This suggests that the current model struggles to effectively handle questions involving a large number of characters.

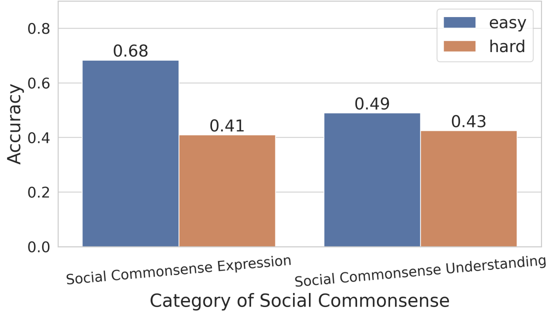


Figure 6: Performance of V-ALPHASOCIAL on V-SOCIAL across different difficulties.

5 Related Work

Social Commonsense Reasoning Benchmarks Compared with general multimodal commonsense benchmark such as *BLINK* (Fu et al., 2024), which implicitly incorporates general commonsense reasoning in images and texts, we center on social commonsense reasoning with rich social interactions in videos. For coverage of social commonse, *EmoBench* (Sabour et al., 2024) consists of multiple choice questions with textual contexts and fo-

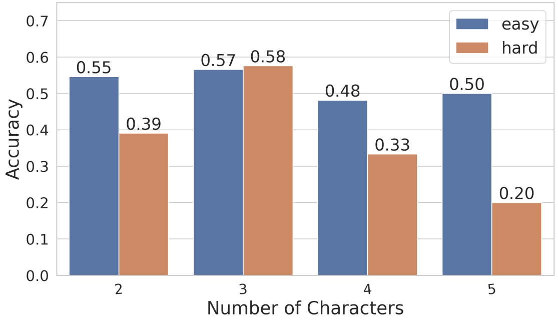


Figure 7: Performance of V-ALPHASOCIAL on V-SOCIAL with different number of characters.

cuses on implied causes of emotion understanding and application only while *EQBench* (Paech, 2024) concentrates on predicting the intensity of a emotion the subject was likely to feel. On contrary, we cover wider range of social commonsense perspectives in six dimensions. Moreover, *VCR* (Zellers et al., 2019) creates question answer pairs with images from movie scenes, challenging models by reasoning with images and textual contexts. Similarly, another two social-related multi-modal video benchmarks, i.e. *SocialIQ* (Zadeh et al., 2019) and *SocialIQ2* (Wilf et al., 2023), are built with open-ended social intelligence questions in varying difficulties, targeting the why and how for complex social commonsense reasoning. Our V-SOCIAL is most similar to them but uniquely provide transcripts, longer video context, and changing scenes to measure advanced social reasoning.

Self-training in LLMs and VLMs Self-training has long been a widely researched and publicly discussed topic (Li et al., 2019; Wang et al., 2020; Karamanolakis et al., 2021). To mitigate the negative influence of error propagation in self-training, (Sohn et al., 2020; Xie et al., 2020) simply set a fixed threshold to filter samples with low confidence. Moreover, Wang et al. (2020) uses meta learning for adaptive sample re-weighting to mitigate error propagation from noisy pseudo-labels. Zhou et al. (2022) utilizes cross-task augmentation with pretrained T0 to do self-training on instruction data. Wang et al. (2022) designs a data-driven filtering approach to boost the performance of self-training on text classification. *ReST* (Singh et al., 2023) proposes a self-training pipeline to progressively enhances LLMs' performance by iteratively generating new traces and selectively learning from those that yield high rewards. For self-training in VLMs, Yang et al. (2023); Kang et al. (2023) augment the training data through diverse self-generated content and achieve stronger performance on tasks such as visual question answering and visual dialog. *Self-Training on Image Comprehension (STIC)* (Deng et al., 2024) constructs a preference dataset for image descriptions with unlabeled images and further refines the VLM's reasoning abilities through a description-infused fine-tuning process. In our work, we generate good quality reasoning process for visual social commonsense problem in a self-training way.

6 Conclusion

Our research addresses the critical challenge of integrating multimodal data for the enhancement of social commonsense reasoning in Visual Language Models (VLMs). By introducing a dataset, V-SOCIAL, we provide the necessary protocols for more effective benchmarking of visual social commonsense reasoning. Furthermore, our novel method, V-ALPHASOCIAL, leverages self-training with context augmentation and a multimodal discriminator to significantly improve the generation of high-quality reasoning paths. Through our extensive analysis, we demonstrate the efficacy of V-ALPHASOCIAL in enhancing the social commonsense capabilities of VLMs. For future work, we plan to generalize our method to other video question answering and explore more in-depth alignment algorithm for boosting VLMs with low-resource data.

7 Acknowledgement

This work was partially supported by NSF 2331966 NSF 2211557, NSF 2119643, NSF 2303037, NSF 2312501, ONR N00014-23-1-2780, SRC JUMP 2.0 Center, Amazon Research Awards, Google and Snapchat Gifts. KW has a financial COI with Google and Amazon.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Singh Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio C'esar Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allison Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xianmin Song, Olatunji Ruwase, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Cheng-Yuan Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv*, abs/2404.14219.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Kristin Behfar, Randall Peterson, Elizabeth Mannix, and William Trochim. 2008. The critical role of conflict resolution in teams: A close look at the links between conflict type, conflict management strategies, and team outcomes. *The Journal of applied psychology*, 93:170–88.

Cristina Bicchieri. 2005. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Iti Chaturvedi, Ranjan Satapathy, Sandro Cavallari, and Erik Cambria. 2019. Fuzzy commonsense reasoning for multimodal sentiment analysis. *Pattern Recognition Letters*, 125:264–270.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, James Zou, Kai-Wei Chang, and Wei Wang. 2024. Enhancing large vision language models with self-training on image comprehension. *Preprint*, arXiv:2405.19716.

Paul Ekman and Harriet Oster. 1979. Facial expressions of emotion. *Annual review of psychology*, 30(1):527–554.

James Evans and Pedro Aceves. 2016. Machine translation: Mining text for social theory. *Annual Review of Sociology*, 42.

Lifeng Fan, Manjie Xu, Zhihao Cao, Yixin Zhu, and Song-Chun Zhu. 2022. Artificial social intelligence: A comparative and holistic view. *CAAI Artificial Intelligence Research*, 1(2).

Andrew H. Farkas, Rebekah L. Trotti, Elizabeth A. Edge, Ling-Yu Huang, Aviva Kasowski, Olivia F. Thomas, Eli Chlan, Maria P. Granros, Kajol K. Patel, and Dean Sabatinelli. 2021. Humor and emotion: Quantitative meta analyses of functional neuroimaging studies. *Cortex*, 139:60–72.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. *Preprint*, arXiv:2404.12390.

Elizabeth E. Graham. 1995. The involvement of sense of humor in the development of social relationships. *Communication Reports*, 8(2):158–169.

Gi-Cheon Kang, Sungdong Kim, Jin-Hwa Kim, Donghyun Kwak, and Byoung-Tak Zhang. 2023. The dialog must go on: Improving visual dialog via generative self-training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Giannis Karamanolakis, Subhabrata Mukherjee, Guoqing Zheng, and Ahmed Hassan Awadallah. 2021. Self-training with weak supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 845–863, Online. Association for Computational Linguistics.

John F Kihlstrom and Nancy Cantor. 2000. Social intelligence. *Handbook of intelligence*, 2:359–379.

Lun-Wei Ku and Cheng-Te Li, editors. 2021. *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, Online.

Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James M. Rehg, and Diyi Yang. 2022. Werewolf among us: A multimodal dataset for modeling persuasion behaviors in social deduction games. *Preprint*, arXiv:2212.08279.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Minzhi Li, Weiyan Shi, Caleb Ziems, and Diyi Yang. 2024. Social intelligence data infrastructure: Structuring the present and navigating the future. *Preprint*, arXiv:2403.14659.

Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. 2019. Learning to self-train for semi-supervised few-shot classification. *Advances in Neural Information Processing Systems*, 32:10276–10286.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M. Dai, Diyi Yang, and Soroush Vosoughi. 2023. Training socially aligned language models on simulated social interactions. *Preprint*, arXiv:2305.16960.

John D. Mayer, Peter Salovey, and David R. Caruso. 2004. Emotional intelligence: Theory, findings, and implications. *Psychological Inquiry*, 15(3):197–215.

Chris Moore. 2006. *The Development of Commonsense Psychology*.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Samuel J. Paech. 2024. Eq-bench: An emotional intelligence benchmark for large language models. *Preprint*, arXiv:2312.06281.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*.

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

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36.

Joseph William Singer. 1989. Persuasion. *Michigan Law Review*, 87(8):2442–2458.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Brian Warkentin, Yundi Qian, Ethan Dyer, Behnam Neyshabur, Jascha Narain Sohl-Dickstein, and Noah Fiedel. 2023. Beyond human data: Scaling self-training for problem-solving with language models. *ArXiv*, abs/2312.06585.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of contexts. *Preprint*, arXiv:2403.05530.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah. 2020. Adaptive self-training for few-shot neural sequence labeling. *arXiv preprint arXiv:2010.03680*.

Yuqing Wang and Yun Zhao. 2023. Gemini in reasoning: Unveiling commonsense in multimodal large language models. *arXiv preprint arXiv:2312.17661*.

Zhihao Wang, Zongyu Lin, Peiqi Liu, Guidong ZHeng, Junjie Wen, Xianxin Chen, Yujun Chen, and Zhilin Yang. 2022. Learning to detect noisy labels using model-based features. *arXiv preprint arXiv:2212.13767*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Alexandros Xenos, Niki Maria Foteinopoulou, Ioanna Ntinou, Ioannis Patras, and Georgios Tzimiropoulos. 2024. Vllms provide better context for emotion understanding through common sense reasoning. *Preprint*, arXiv:2404.07078.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.

Haike Xu, Zongyu Lin, Jing Zhou, Yanan Zheng, and Zhilin Yang. 2022. A universal discriminator for zero-shot generalization. *arXiv preprint arXiv:2211.08099*.

Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S. Bernstein, and John Mitchell. 2024. Social skill training with large language models. *Preprint*, arXiv:2404.04204.

Xiaofeng Yang, Fengmao Lv, Fayao Liu, and Guosheng Lin. 2023. Self-training vision language berts with a unified conditional model. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):3560–3569.

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

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Saining Zhang, Yuhang Zhang, Ye Zhang, Yufei Wang, and Zhigang Song. 2023. A dual-direction attention mixed feature network for facial expression recognition. *Electronics*, 12(17):3595.

Jing Zhou, Zongyu Lin, Yanan Zheng, Jian Li, and Zhilin Yang. 2022. Not all tasks are born equal: Understanding zero-shot generalization. In *The Eleventh International Conference on Learning Representations*.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. Normbank: A knowledge bank of situational social norms. *Preprint*, arXiv:2305.17008.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Preprint*, arXiv:2305.03514.

A Features of Our Benchmark

Multi-modal Input Human communication is inherently multi-modal, incorporating visual cues, text, and often audio components. To emulate this, our dataset includes images, videos, and text, requiring models to synthesize information across modalities to respond appropriately. This integration tests the ability of LLMs and VLMs to process and understand a richer spectrum of communication signals, thus preparing them for real-world applications where multi-modal data is the norm.

Richer Context Effective communication often depends on understanding extended discourse. Our dataset provides scenarios with longer text passages and extended video clips, challenging models to maintain context over longer stretches of information. This feature is critical for developing models capable of engaging in meaningful conversations, where references to earlier points in the dialogue are commonplace and essential for coherence.

Deeper Reasoning Paths To navigate complex social interactions, models must be able to engage in deeper reasoning—beyond surface-level understanding. Our tasks require models to infer motivations, predict outcomes, and synthesize disparate pieces of information. This is achieved through scenarios that involve abstract reasoning, problem-solving, and prediction tasks, all of which are essential for higher-level cognitive processing.

Multi-party, Multi-turn Interactions Real-life social interactions often involve multiple individuals and are rarely confined to single exchanges. To mirror this, our dataset features scenarios involving multiple characters who interact across several turns. This setup tests the ability of models to track different speakers' perspectives and contributions over time, a key component of effective multi-party communication.

Implicit Reasoning To avoid models relying on simple heuristics or shortcuts, our dataset deliberately omits explicit cues that could lead to superficially correct answers. Instead, tasks are designed to require implicit reasoning, compelling models to understand underlying concepts and relationships without direct prompts. This approach encourages a deeper level of understanding and interaction, moving away from rote responses to engaging with the material in a thoughtful and analytical manner.

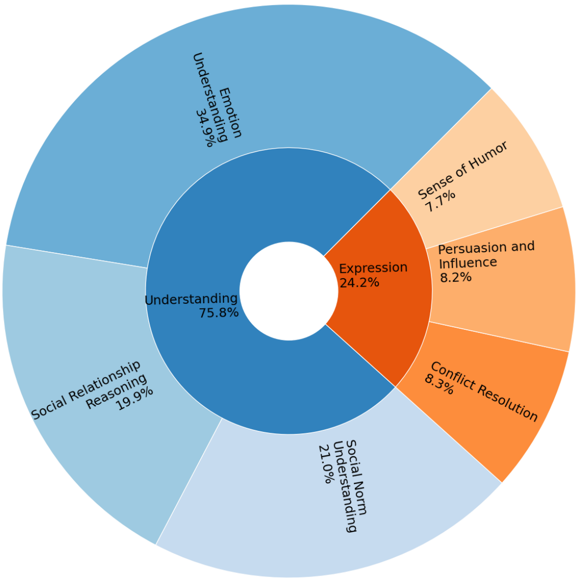


Figure 8: Data distributions according to our predefined taxonomy. The six dimensions (Emotion Understanding, Social Relationship Reasoning, Social Norm Understanding, Conflict Resolution, Persuasion and Influence, Sense of Humor) are categorized into two broad functions, i.e. Understanding & Expression.

B Other Statistics and Comparison of V-SOCIAL

The V-SOCIAL dataset consists of 956 Q&A pairs where 214 videos are needed to correctly answer them. Except for key stastistics in Table 2, we also visualize the detailed distribution graphs in Figure 8, 9, and 10 for *predefined social commonsense taxonomy*, *number of characters*, and *number of dialogue turns* respectively. Moreover, Figure **??** and 13 show the performance of different models under this taxonomy.

Moreover, as we also use the same video source as SocialIQ2 on YouTube, we compare the question answer pair generated from the same video clip (video_id: 2MrFWB__GIA), which has been illustrated in Figure 11, 12. From these, we can see that our data point covers a longer context in the given video and proposes questions with competitive choices. However, in SocialIQ2, the question is more tricky as the option A) even has nothing to do with it and only B) and C) are trying to answer the question, which drastically reduces the difficulty just by using the information in the question part before watching the video.

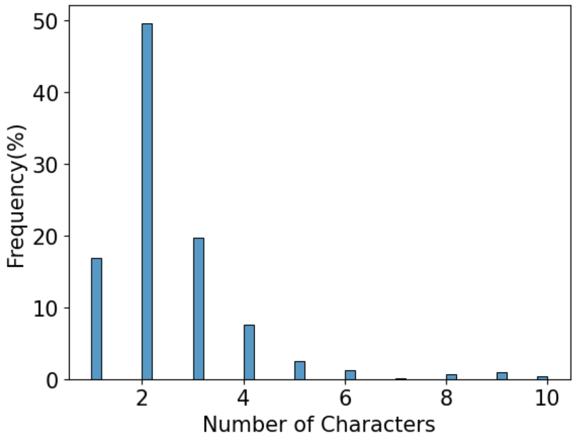


Figure 9: The data distributions for different numbers of characters in videos

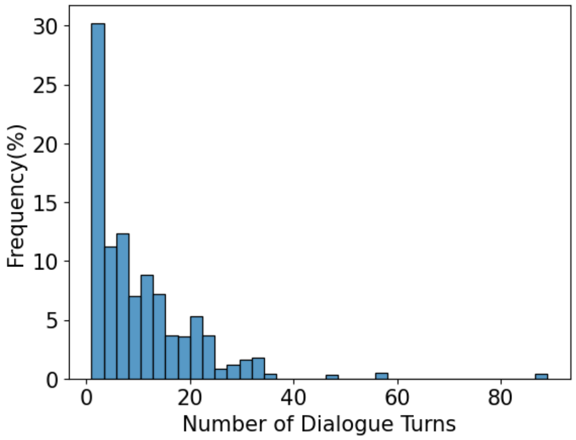


Figure 10: The data distributions for different numbers of dialogue turns in videos

C Data annotation process

Our annotation starts with the Youtube videos from SocialIQ2 (Wilf et al., 2023) and a list of videos are sampled, each labeled with a unique video id and a corresponding .vtt file with the transcript among our team of annotators.
The annotation could be described in following steps:

1. Annotators watch the video clip and select time ranges that have explicit conflicts, social interactions, or emotional scenes.
2. A python script is used to automatically generate the raw transcript (e.g., dialogue between the characters in the videos) for each video.
3. The utterances in the dialogue are manually assigned to the speakers and the speakers information (e.g.,“s1”:“A man in blue T-shirt",“s2”: “A woman in white dress") is also annotated.
4. Multiple choice question-answer pairs are constructed by annotators based on the above information in videos and their understanding. Q&A pairs are curated by different dimensions of our predefined social commonsense taxonomy and difficulties. For example, a question will be labeled as "Hard" if it requires a strong understanding of emotional nuances that are not covered in the transcript alone.

D License for Artifacts

This work involves the following artifacts:

Datasets *SocialIQ2* (Wilf et al., 2023) distributed under MIT license.

Software We use *transformers* (Wolf et al., 2020) and *deepspeed* (`https://github.com/microsoft/DeepSpeed`) for model training and inference, both distributed under Apache-2.0 license.

Models **BLIP** (Li et al., 2023) distributed under BSD-3-Clause license; **FER model** (Zhang et al., 2023) without license specified; **VideoLLaVA** (Lin et al., 2023) distributed under Apache 2.0 license, **Phi-3-Vision** (Abdin et al., 2024) distributed under MIT license, **GLM-4V** (Wang et al., 2023) distributed under the glm-4-9b license, **LLaVA-NEXT** (Liu et al., 2024) distributed under Llama-1/2 community license.

E Further Details of Experimental Setup

To conduct our main experiment on V-SOCIAL and SocialIQ2, we use the following prompts, which are demonstrated in Figure 14 and 15, to prompt V-ALPHASOCIAL by explicitly instruct model to generate good reasoning paths and then gives a final answer for verification.

E.1 Training Configuration

Both our policy model and discriminator model was trained over three epochs with a batch size of 16. The initial learning rate was set at 2×10^{-5}, employing a cosine learning rate scheduler. A warm-up ratio of 0.03 was utilized to gradually introduce the learning rate at the onset of training. The training process was configured to handle a maximum sequence length of 2048 tokens, and it was executed using bfloat16 precision.

Speaker Information: Speaker 1: lady with wax \nSpeaker 2: her fiance Jay
Context: Speaker 1: okay you guys so here is my fiance Jay and we are going to be doing a test on his arm to see if this wax thing actually works or not and he has he has reluctantly agreed to do this he really doesn't want to but he's a good sport he's a good guy so he's going to try it out I put just a tiny bit on his hand one oh failure didn't even do anything okay oh we got a couple hairs actually we got a couple hairs we need to try this again so this is not hard waxing this is um soft waxing 3 two one that oh oh Cy we got some hair especially like right over here let's see your hand a your poor hand's red a see so it didn't totally get everything out it did it did kind of work for his hand a little bit it made his hand red and like did that hurt \nSpeaker 2: no I'm didn't hurt at all\nSpeaker 1: so do you think that this is a hack or is this totally whack\nSpeaker 2: totally whack\nSpeaker 1: totally whack you think it's whack okay thank you babe I love you thank you to Bean for being a participant the audience we're not going to wax bean bag okay so I don't know if my technique is off or something maybe I'm not that great at waxing people so I'm going to try um I'm going to try it on my arm three two 3 two 1 3 2 1 well it did take it off of my hand in the air as you can see and like I don't know if you guys can see this but there are some hairs like right here it feels nice and soft okay so now I'm going to pull it off in three two one oh my god oh that works so you can there's a bunch of blonde hairs okay so there's a couple black hairs and that's for my little Chihuahua Willie somehow his hairs got on my shirt welcome to the life of dog ownership you have dog hairs everything on everything I don't know if you guys can see this see there's kind of like black hairs that's from my chihuahua but there are like blonde hairs on it so I'm
Time Range: 2:00-4:50
Question: How does speaker 2 feel toward speaker 1? A) Hates her B) Loves her C) Indifferent D) They're siblings
Answer: B

Figure 11: Example from V-SOCIAL

Time Range: 0:00-0:48.716016
Question: Does the man love the dog? A) The woman with white hair is a professional hair stylist. B) The man loves the dog because he enjoys the dog's presence. C) No, because he is angry with the dog. D) The man is allergic to dogs.
Answer: B

Figure 12: Example from SocialIQ2

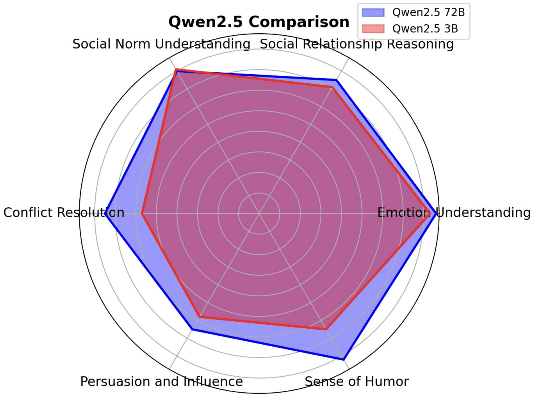


Figure 13: Comparison of Qwen2.5-VL 72B and Qwen2.5-VL 3B Under The Taxonomy. This radar chart compares the performance of two models across six dimensions. The 72B model outperforms the 3B model in all dimensions except for Social Norm Understanding, where their performances are nearly identical.

F Algorithm

G Criterion of Taxonomy Construction

In this section, we explain why we choose the six dimensions from psychology and social science perspectives. Choosing these six dimensions to form the taxonomy is a deliberate effort to extend and apply the core principles of Daniel Goleman's theory on emotional intelligence (EI) to the development and evaluation of social capabilities in Large Language Models (LLMs). Goleman's model, which includes self-awareness, self-regulation, motivation, empathy, and social skills, provides a robust framework for understanding human emotions and interactions. Here's how these dimensions align and expand upon Goleman's EI components:

Emotion Understanding: Directly correlates with Goleman's self-awareness and empathy components, which involves recognizing and understanding others' feelings (Mayer et al., 2004). In the context of LLMs, this dimension ensures that AI can identify and appropriately respond to a range of human emotions, thus facilitating more empathetic and meaningful interactions.

Social Relationship Reasoning: Builds upon the social skills aspect of Goleman's model. It emphasizes the importance of understanding the nuances of different types of relationships (personal, professional, etc.) and adjusting communications accordingly. This dimension ensures LLMs can navigate the complex web of human relationships in various contexts.

Social Norm Understanding: This dimension aligns with self-awareness, self-regulation and social skills. It focuses on the ability of LLMs to grasp and adhere to the unspoken rules and expec-

<VIDEO>

<CONTEXT>

<QUESTION>

Let’s think step by step:

<ANSWER>

Figure 14: Chain-of-thought prompt for visual social commonsense reasoning.

<VIDEO>

<CONTEXT>

<QUESTION>

1. Emotion Understanding: The ability of LLMs to recognize, interpret, and respond to human emotions expressed through text. This involves not just identifying explicit statements of emotion but also inferring underlying feelings from context, tone, and choice of words.
2. Social Relationship Reasoning: The capacity of LLMs to understand the dynamics and nuances of different social relationships (e.g., familial, professional, casual) and how these relationships influence communication and expectations.
3. Social Norm Understanding: The ability to grasp and apply unwritten rules and behaviors that are considered acceptable within a society or group. This includes understanding politeness, formality levels, and cultural norms.
4. Conflict Resolution: The skill of navigating disagreements or conflicts in a way that seeks to find a resolution that is acceptable to all parties involved. It involves understanding the perspectives of each side, mediating the discussion, and suggesting compromises.
5. Persuasion and Influence: The ability to construct arguments or narratives that can change someone’s beliefs, attitudes, or behaviors. This involves understanding the audience’s values and leveraging rhetorical techniques effectively.
6. Sense of Humor: The capability to understand, generate, and appropriately use humor. This includes recognizing different types of humor (e.g., sarcasm, wit, puns) and knowing when humor is appropriate in social interactions.

Based on the above definition of social commonsense, please first think about the social network between different users, and then think about how their interactions and dialogues affect their emotions and attitudes towards others, and finally answer this question.:

<ANSWER>

Figure 15: Social-of-thought prompt for visual social commonsense reasoning.

Algorithm 1 Self-Training Algorithm of V-AlphaSocial

```
Notation: $\pi$: policy model, $M_D$: multimodal discriminator, $U$: dataset with unlabeled reasoning, $N$: number of iterations, $K$: number of samples, $\theta$: threshold
Initialize negative dataset $D^- \leftarrow \emptyset$ and positive dataset $D^+ \leftarrow \emptyset$
for $i = 1$ to $N$ do                                              ▷ Main training loop
    $D_i^- \leftarrow \emptyset$, $D_i^+ \leftarrow \emptyset$                                  ▷ Reset data
    for each $(v, c, q, gt) \in U$ do
        for $j = 1$ to $K$ do                          ▷ Perform K samplings to generate predictions
            $(r, a) \leftarrow \pi(\text{tool}(v, c, q))$       ▷ Generate reasoning and answer using the policy model
            if $a \neq gt$ then
                $D_i^- \leftarrow D_i^- \cup \{(v, c, q, r, a)\}$    ▷ Add to negative dataset if prediction is incorrect
            else
                if $i = 1$ then
                    $D_i^+ \leftarrow D_i^+ \cup \{(v, c, q, r, a)\}$  ▷ Add all correct predictions to positive dataset in first iteration
                else
                    $rw \leftarrow M_D(v, c, q, r, a)$             ▷ Compute reward from the discriminator
                    if $rw > \theta$ then
                        $D_i^+ \leftarrow D_i^+ \cup \{(v, c, q, r, a)\}$  ▷ Add to positive dataset if exceeds threshold
                    end if
                end if
            end if
        end for
    end for
    Train $M_D$ with $D_i^+ \cup D_i^-$                          ▷ Update the discriminator model
    Train $\pi$ with $D_i^+$                     ▷ Update the policy model using positive samples
end for
return $\pi$                                              ▷ Return the trained policy model
```

tations that govern behavior in different cultures and social groups, ensuring respectful and contextually appropriate interactions. (Fan et al., 2022) describes contextual adaptation as a human-like skill involving reasoning and adjusting to contexts like culture and shared experiences.

Conflict Resolution: This dimension can be seen as an application of several of Goleman's EI components, including self-regulation, empathy, and social skills. It involves understanding different perspectives, mediating discussions, and fostering compromise, enabling LLMs to assist in resolving disputes and reducing tensions. This often involves theory of mind, as (Fan et al., 2022) describes as attributing mental states to others, stepping away from egocentrism to consider other perspectives unlike ones' own.

Persuasion and Influence: This dimension extends from Goleman's social skills component but also involves aspects of self-regulation and motivation. It's about the ability to craft messages that motivate, persuade, and influence others, which is essential for LLMs involved in areas such as marketing, advocacy, or any domain where changing thoughts or behaviors is the goal. As (Fan et al., 2022) delineates, cooperation requires a psychological infrastructure of shared intentionality to create common conceptual ground with others, which is necessary to persuade others.

Sense of Humor: While not directly mentioned in Goleman's theory, a sense of humor is inherently linked to social skills and empathy. It requires understanding context, cultural nuances, and emotional states, making interactions more engaging and human-like. As (Farkas et al., 2021) describes, complex humor activates supramodal areas of the brain strongly associated with emotional processes.

By choosing these six dimensions, the taxonomy not only covers the breadth of Goleman's emotional intelligence components but also adapts and expands them to address specific challenges and opportunities in human-AI interaction. This ensures that LLMs developed under this framework can engage in more nuanced, empathetic, and effective communication, mirroring the complexity of human emotional and social intelligence.

H The Influence of Task Complexity on the Performance of V-ALPHASOCIAL

We display the the influence of task complexity on the performance of V-ALPHASOCIAL. As shown in Figure 16, the accuracy shrinks when increasing the difficulties of the task on all six dimensions.

I Failure Modes in Error Analysis

The failure modes are manually summarized by authors from observations in models' inference results. From superficial errors to profound flaws, these modes could be empirically defined as follows:

1. **Instruction Following Error**: models fail in giving valid answers by choosing from candidate choices.

2. **Visual Perception Error**: models ignore the nuances of visual cues, such as facial expressions.

3. **Misunderstanding of Social Context**: models misunderstand the social context, including emotions, utterances, metaphors, social relationships, etc.

4. **Inconsistent Reasoning Error**: models output inconsistent reasoning chains in which some intermediate claim is incorrect.

5. **Social Expression or Action Error**: models choose the incorrect answer although it gives reasonable deductions. For instance, the chosen candidate expresses the totally different meaning compared with the previous reasoning chain although they could share some same words.

J Data Examplars for different dimensions

1. Emotion Understanding

Definition: The ability of LLMs to recognize, interpret, and respond to human emotions expressed through text. This involves not just identifying explicit statements of emotion but also inferring underlying feelings from context, tone, and choice of words.

Task Construction: Tasks can include identifying emotions in written narratives, generating empathetic responses in conversations, or altering the tone of a response to match the emotional context.

Examplar: Figure 17

2. Social Relationship Reasoning

Definition: The capacity of LLMs to understand the dynamics and nuances of different social relationships (e.g., familial, professional, casual) and how these relationships influence communication and expectations.

Task Construction: Tasks may involve interpreting interactions within specific social contexts, predicting the nature of relationships based on conversation excerpts, or generating responses appropriate to the relationship dynamic.

Examplar: Figure 18

3. Social Norm Understanding

Definition: The ability to grasp and apply unwritten rules and behaviors that are considered acceptable within a society or group. This includes understanding politeness, formality levels, and cultural norms.

Task Construction: Design scenarios where LLMs must choose actions or responses that adhere to social norms, recognize breaches of norms, or adapt communication to fit different cultural or social settings.

Examplar: Figure 19

4. Conflict Resolution

Definition: The skill of navigating disagreements or conflicts in a way that seeks to find a resolution that is acceptable to all parties involved. It involves understanding the perspectives of each side, mediating the discussion, and suggesting compromises.

Task Construction: Create tasks where LLMs mediate simulated conflicts, propose solutions to interpersonal problems, or guide conversations towards de-escalation.

Examplar: Figure 20

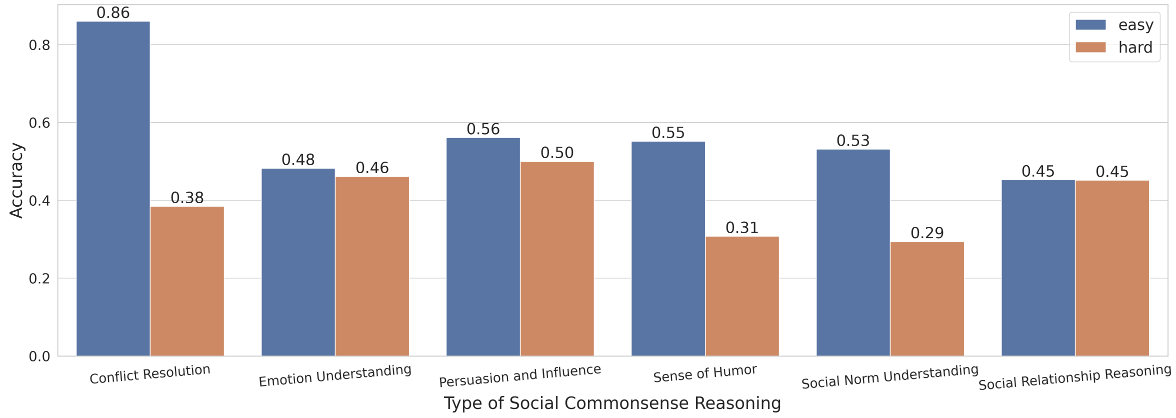


Figure 16: Performance on V-SOCIAL across different difficulties.

5. Persuasion and Influence

Definition: The ability to construct arguments or narratives that can change someone's beliefs, attitudes, or behaviors. This involves understanding the audience's values and leveraging rhetorical techniques effectively.

Task Construction: Tasks could include writing persuasive texts on various topics, crafting messages intended to motivate action, or simulating negotiations where LLMs must achieve specific objectives.

Examplar: Figure **??**

6. Sense of Humor

Definition: The capability to understand, generate, and appropriately use humor. This includes recognizing different types of humor (e.g., sarcasm, wit, puns) and knowing when humor is appropriate in social interactions.

Task Construction: Develop tasks that involve creating humorous responses in context, identifying the humor in texts, or explaining why a particular joke is funny.

Examplar: Figure 21

<sampled 10 frames or the whole video>

Speaker Information: Speaker 1: Father
Speaker 2: Mother

Context: Speaker 1: you guys ready so here's the deal Ellie and Bonnie have no idea that we are coming to the airport they drove to the airport and they're thinking that they're just going to be coming back all by themselves well Joel the kids and I are going down to pick them up and greet them at the airport and we're going to drive back together so we can talk and have more time together and it's going to be a lot of fun are you guys ready let's go get mom let's go get mom necklace those are not for penny no and she certainly want some though doesn't see hey hey now different mommy okay I am ready to rock who is ready to go see mom I'm not very enthusiastic I think Joel and I are much more excited I got a feeling tiny Bonnie is going to be the biggest best hug ever a long day of travel
Speaker 2: we are back home in Salt Lake City and the moment right the moment we walked out of the door we all were like whoa Utah we got a brave of four getting outside our cold clothes cars it was so fun girls I loved our trip we will do it again good again loved it yeah and all the girls like you should talk to her husband's about going somewhere for like the rest of winter for six weeks that would be really fun it would be really fun such an adventure but it's cold but I am so excited to go home and see my babies they're asleep they better believe I'm going to pick them up and kiss them because the week is too long to be gone usually if I get home all just like p.m. and the more I've got to go keep them [Music]
Speaker 1: [Music] you want to treat this is your tree buddy oh I was waiting I'm literally like all on you think I was kind of hoping maybe the boys of touch thinking well
Speaker 2: no I never we saw you she's my co-chair for sure don't tell my didn't you have to scan that Mimi's guys thank you we brought the whole gang with us robbery I've got him right here
Speaker 1: nice surprise Jackson totally ditched your sign he saw you and just dropped it and ran yeah see redcoat hear me hahahahaha ran he follows through the kids will not let me hold them they only want mommy to hold them thanks now I tell or something yeah they haven't been feeling too well the past day

Question: What reaction do the kids have when they see the mom?
A) They cry.
B) They scream.
C) They run to her and hug her.
D) They run away from her and hug dad.

Let's think step by step:

LLM Response:

To determine the reaction of the kids when they see their mom, let's analyze the sequence of frames you provided:

1. Frame 1-4: Shows the father speaking, likely explaining the plan to surprise the mom.
2. Frame 5-7: Shows a child holding a bouquet of flowers, possibly getting ready to greet the mom.
3. Frame 8-9: Shows a younger child, looking towards the camera, likely reacting to something.

Based on these frames and the context provided, it seems the children are excited and preparing to greet their mom. Given this setup and the excitement shown in the initial frames, the most likely reaction of the kids when they see their mom would be:

C) They run to her and hug her.

This conclusion fits with the narrative of a surprise welcome and the excitement visible in the frames.

Figure 17: Example of using chain-of-thought for visual social commonsense reasoning. (Emotion Understanding)

<sampled 10 frames or the whole video>

Speaker Information: Speaker 1: Male news anchor at TV station
Speaker 2: Female news anchor reporting from Korea

Context: Speaker 1: Judging by previous reunions we know just how emotionally charged these events can be. Of course for the families involved but also for the country as a whole. Soa you reported on Monday about the families going through a kind of pre-session to get them prepared what's next on the schedule?
Speaker 2: Right now they are eating breakfast the bus will take them to Goseong where they have to go through the Customs Immigration Quarantine office to head to the Mount Kumgang resort in the North. There they will have their first meeting altogether in a big hall On the second day they will have private meetings. In total the families will meet for three days,but their actual time together will only be 12 hours, but that is an extension from the usual 11, due to a request by the South Korean Unification Ministry. It's seen as a good sign that the North agreed to the request. So many families who haven't been able to see their loved ones yet are desperately waiting for these kinds of reunions to become a regular event

Question: How does speaker 2 feel about the amount of time the families have to meet each other?
A) Happy, the time is just right
B) Angry, it's not enough time
C) Disappointed, wishes there was more
D) Depressed, there's never enough time

Let's think step by step:

LLM Response:

To determine how Speaker 2 feels about the amount of time the families have to meet each other, let's analyze the provided text:

Speaker 2 mentions that families will meet for three days, with their actual time together being only 12 hours, which is an extension from the usual 11 hours due to a request by the South Korean Unification Ministry. They also mention that it's seen as a good sign that the North agreed to this request. However, they also note that many families who haven't been able to see their loved ones yet are desperately waiting for these kinds of reunions to become a regular event.

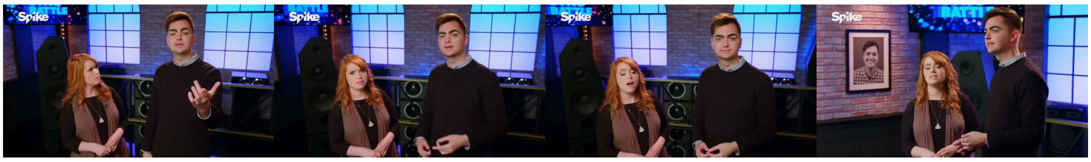
Given this information, it seems Speaker 2 acknowledges the improvement (extension from 11 to 12 hours) but also highlights the desperate wait of many families for these reunions to become regular.

Based on this, the most appropriate choice is:

C) Disappointed, wishes there was more

Speaker 2 likely feels that while the extension is positive, it's still not enough for the families who are longing for more time together and for these reunions to happen more regularly.

Figure 18: Example of using chain-of-thought for visual social commonsense reasoning. (Social Relationship Reasoning)



<sampled 10 frames or the whole video>

Speaker Information: Speaker 1: male interviewer
Speaker 2: Zendaya, actor, dancer, singer
Speaker 3: female interviewer

Context: Speaker 1: hey guys I'm here with zendaya zendaya how are you feeling going into your very first lip sync battle what's going through your
Speaker 2: I feel very good cuz I was really like nervous before I started rehearsing then I got to rehearsals the dance was awesome everything was coming together and I was like you know what this is going to be great
Speaker 1: now you're you're a very successful musician are you aware that there is no actual singing in the show involved
Speaker 2: yes and actually that that gets a a lot of stress is gone because of that I don't have to worry about I out I just get to go out there and do what I would do you know in my car
Speaker 1: y you're going up against Spider-Man yeah um how do you feel about that is that making you a little nervous he's a superhero
Speaker 2: no not because he's not superhero life oh burn Tom is very talented what a lot of people don't know is he is like a trained dancer and like did Alien whatever but you know what that happened then okay this is now
Speaker 1: yeah right what we can do now what do you think would be like the superpowers that somebody would get if they were bitten by a radioactive Zena
· · · <more dialogue transcript>

Question: What is the tone of speaker 2's message to Tom Holland?
A) lighthearted joking insult
B) serious insult
C) cruel joke
D) lovely compliment

Let's think step by step:

LLM Response:

Based on the provided images, the context appears to be an interview scenario with Zendaya and a male interviewer. From the provided text, we see Zendaya's responses and a subsequent humorous exchange between the interviewers.

Let's evaluate the tone of Speaker 2's (Zendaya's) message to Tom Holland step by step:

1. Context of the Message: Zendaya is asked to send a message to Tom Holland, her competitor in a lip sync battle. 2. Content of the Message:

- "Try your best is all I can say."
- "I mean I would be nervous if I were you but you don't have to be nervous if you don't want to be."
- "I wish you the best cuz you're going to need it."

3. Tone Indicators:

- The phrase "Try your best" is encouraging but can be seen as a subtle challenge.
- "I would be nervous if I were you" implies confidence in her abilities, suggesting that Tom should be concerned.
- "I wish you the best cuz you're going to need it" is playful and competitive, implying that she is confident in her own performance while still wishing him well.

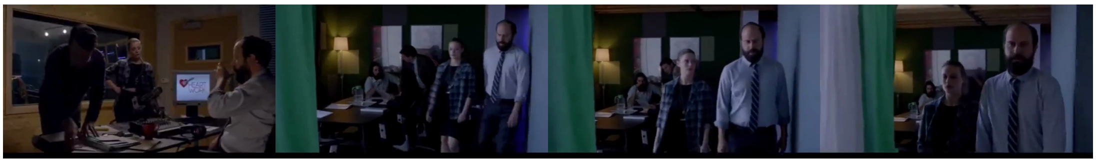
4. Overall Impression: The message combines encouragement with competitive banter, common in friendly rivalries, especially in entertainment contexts like a lip sync battle.

Given these points, the tone of Zendaya's message to Tom Holland is best characterized as:

A) lighthearted joking insult

This option captures the playful nature of her message, which is intended to tease and challenge Tom in a friendly manner.

Figure 19: Example of using chain-of-thought for visual social commonsense reasoning. (Social Norm Understanding)



<sampled 10 frames or the whole video>

Speaker Information: Speaker 1: man in tie, host
Speaker 2: woman, program manager

Context: Speaker 1: these archetypical people in our lives they hurt us and they don't fulfill the archetype archetype is a well I don't really have the exact definition I do know what it means obviously because I just used it well thank you so much for calling I am dr. Greg Coulter and you are listening to the love and living channel on gravity satellite radio you listen we hear
Speaker 2: great great show oh thanks so much Rob and really appreciate it
Speaker 1: good morning Nikki
Speaker 2: hey you said you wanted to talk
Speaker 1: yeah we need to fire rod Rob's gonna go does that make me evil no please this is a business you know what can you fire them for me or do program managers not do that what I mean okay you know what I'll do it if you do it Rob will think on the caller yeah I think that's probably the right move
Speaker 2: no I you know what you do it
Speaker 1: I'm sorry I'll stop being annoying
Speaker 2: so what's the problem what did he do
Speaker 1: look Rob does a great job but the guy creeps me out he's got this look in his eyes you never meet a person and they smile and you feel like they're screaming at ya
Speaker 2: what Rob wraps like the nicest guy in the whole world he's nicer than both of us
Speaker 1: Wow he's just got this bad energy this aura around him
Speaker 2: has he ever done anything to you
Speaker 1: yes me do I like him that's weird right that's not a thing he asked anybody let alone in a workplace look I would fire him but I can't take that on right now this might be a difficult thing for you to understand but I am very vulnerable in this job I have to be I make myself vulnerable to the world I can't be vulnerable around somebody like Rahab so he's got to go
Speaker 2: fine I'll do it just to end this conversation
Speaker 1: I'm a aren't I I'm sorry I'm disgusting okay you know what I'll do it, Oh question what is a good band to work out to something new
Speaker 2: I don't know the monkeys
Speaker 1: that's great thank you the best, now you fire him

Question: What should speaker 2 do to resolve this issue?
A) Agree to whatever speaker 1 says
B) Call HR and speak her opinion to defend Rob
C) Ignore it
D) Fire Rob Immediately

Let's think step by step:

LLM Response:

To resolve the issue in the dialogue between Speaker 1 (the host) and Speaker 2 (the program manager), let's think step by step:

1. Evaluate the Situation: Speaker 1 feels uncomfortable around Rob due to a perceived negative aura and an incident where Rob asked if Speaker 1 liked him. Speaker 1 is in a vulnerable position and cannot work with someone who makes him feel uneasy.
2. Consider Speaker 2's Perspective: Speaker 2 is surprised by Speaker 1's negative feelings towards Rob, who is described as the nicest person. Speaker 2 initially defends Rob but eventually agrees to fire him just to end the conversation.
3. HR Involvement: Instead of making a hasty decision to fire Rob based on subjective feelings, it is more appropriate to involve HR to ensure a fair process. HR can conduct a proper investigation and mediate between the parties.
4. Professional Approach: Speaker 2 should maintain professionalism and address the issue through proper channels. It's important to consider both the host's feelings and Rob's rights as an employee.

Considering these points, the best course of action for Speaker 2 is:

B) Call HR and speak her opinion to defend Rob.

This approach ensures that the issue is handled fairly and professionally, respecting both Speaker 1's concerns and Rob's rights.

Figure 20: Example of using chain-of-thought for visual social commonsense reasoning. (Conflict Resolution)

<sampled 10 frames or the whole video>

Speaker Information: Speaker 1: Rex, lawyer on husband side
Speaker 2: Freddie, lawyer on wife side
Speaker 3: husband, client of speaker 1

Context: Speaker 1: ready ready Bender this is Rex Rex Roth and you are the lovely Maryland please miss Rex Ross and you must be Mr Massie please miles sit sit down oh Freddy I was sorry to hear about the goldberger award, pastry?
Speaker 2: we did very well we did very well,
Speaker 1: not to worry Mrs rexroth you're ably represented I'm sure Freddy's just too modest to tell you he used to clerk for Clarence Thomas pyri go
Speaker 2: and beging don't don't try to bait me miles now if you have a proposal to make let's hear it
Speaker 1: well at this point my client is still prepared to consider reconciliation
Speaker 2: my clients rule that out
Speaker 1: my client is prepared to entertain an amicable disillusion of the marriage without prejudice
Speaker 2: that's a fart stiff of wind
Speaker 1: my client proposes a 30-day cooling off period
Speaker 2: my client feels sufficiently dispassionate
Speaker 1: my client that you not initiate proceedings pending his setting certain Affairs in order
Speaker 3: well what's so goddamn funny
Speaker 1: please let me handle this all right so much for the ice breakers what are you after Freddy
Speaker 2: my client is prepared to settle for 50% of the marital assets
Speaker 1: why only 50 Freddy why not 100 While We're Dreaming why not 150 are you familiar with kersner
Speaker 2: kersner does not apply
· · · <more dialogue transcript>

Question: Why does speaker 1 say we'll eat the pastry?
A) He's hungry
B) He loves desserts
C) Cookies will go bad if no one eats them
D) Ironic, humorous in response to other sides' anager

Let's think step by step:

LLM Response:

Let's analyze the provided images and dialogue to understand why Speaker 1 says "we'll eat the pastry."

From the context and dialogue, we can see that the conversation is tense and involves negotiations between two sides, likely regarding a divorce settlement. The interaction between the lawyers (Speaker 1 and Speaker 2) and their clients is filled with legal jargon and a bit of sarcasm.

Now, examining the context and dialogue:

1. Speaker 1 offers a pastry to Speaker 2 early in the conversation.
2. Speaker 2 responds by dismissing the offer and moving straight to negotiations.
3. The negotiation goes back and forth, with neither side willing to yield easily.
4. After a particularly heated exchange, Speaker 1 makes the comment, "we'll eat the pastry."

Given the situation and tone of the conversation, Speaker 1's remark appears to be a way to diffuse tension with humor. It suggests that the negotiation was tough, and as a conciliatory or ironic gesture, Speaker 1 decides that they will eat the pastry themselves, implying a sense of resolution or acceptance of the situation.

Therefore, the best answer is:

D) Ironic, humorous in response to other sides' anger.

This interpretation fits the tone and context of the dialogue, where legal negotiations are intense, and humor is used to lighten the mood or as a form of irony.

Figure 21: Example of using chain-of-thought for visual social commonsense reasoning. (Sense of Humor)