VIBE: Can a VLM Read the Room?

Tania Chakraborty, Eylon Caplan, Dan Goldwasser
Purdue University, West Lafayette, IN, USA
{tchakrab,ecaplan,dgoldwas}@purdue.edu

Abstract

Understanding human social behavior such as recognizing emotions and the social dynamics causing them is an important and challenging problem. While LLMs have made remarkable advances, they are limited to the textual domain and cannot account for the major role that non-verbal cues play in understanding social situations. Vision Language Models (VLMs) can potentially account for this gap, however their ability to make correct inferences over such social cues has received little attention. In this paper, we explore the capabilities of VLMs at social reasoning. We identify a previously overlooked limitation in VLMs: the Visual Social-Pragmatic Inference gap. To target this gap, we propose a new task for VLMs: Visual Social-Pragmatic Inference. We construct a high quality dataset to test the abilities of a VLM for this task and benchmark the performance of several VLMs on it.

1 Introduction

"It is only with the heart that one can see rightly; what is essential is invisible to the eye."

— *de Saint-Exupéry,* The Little Prince

Understanding social dynamics—such as recognizing emotions and their cause—is something humans do intuitively (Spunt and Adolphs, 2019), yet *how* we do it remains a challenging question. Even emotion perception, an intuitive ability for most people, poses several levels of complexity (Abbas et al., 2024) arising from a variety of factors such as context, cultural variability and channels of perception (language, vision, etc.). Influential psychology and cognitive neuroscience studies (Ekman and Friesen, 1978; de Gelder, 2006) have shown that significant proportions of socially relevant information is contained in non-verbal cues like facial expression and body language. In certain situations (Mehrabian and Wiener, 1967; Mehrabian, 1972), only 7% of meaning is conveyed through words, with 38% coming from tone of voice and *55% through facial expressions!* Even within the visual modality alone, emotion recognition is not simply a matter of identifying facial expressions (Barrett and Kensinger, 2010), as the same facial expression can convey different emotions depending on body posture (Aviezer et al., 2012). E.g., even an iconic emotion indicator like a smile, may not always indicate joy (Fernández-Dols and Ruiz-Belda, 1995). Despite the intuitive ease with which most people perceive emotions, research has shown that this ability relies on complex mechanisms spanning cognition, perception and understanding.

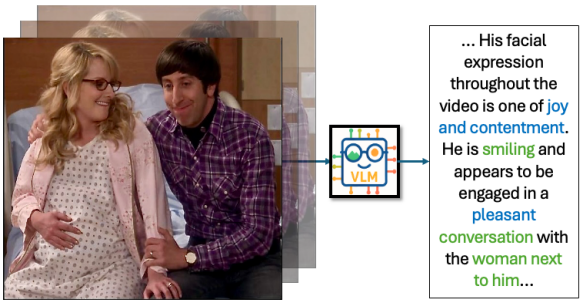

Figure 1: In the video, the man smiles sadly, teary eyed. His partner looks at him with sympathy and pats his leg comfortingly. The VLM (InternVL2 26B) correctly identifies the smile, and the woman next to the man, but is not able to interpret the smile correctly.

In recent years there has been increasing interest in AI systems that can achieve human levels of social intelligence. LLMs in particular have shown remarkable promise in such tasks, though they struggle when presented with complex scenarios (Li et al., 2024; Liu et al., 2024d; Mou et al., 2024; Wu et al., 2025) and are limited to the textual domain. The recent emergence of VLMs (Singh et al., 2022; Bai et al., 2023; Bordes et al., 2024; Zhang et al., 2024a) provides an exciting opportunity to study social scenes by connecting textual and visual information and leveraging the strong reasoning capabilities of LLMs. However, while some works have explored the efficacy of using

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22992–23008
November 4-9, 2025

VLMs for social intelligence tasks (Yang et al., 2024; Sun et al., 2024; Bhattacharyya and Wang, 2025; Liu et al., 2025a), this area still remains relatively underexplored.

In this paper, we conduct a thorough investigation of VLMs' ability to perform social reasoning tasks. We focus on emotion-related inferences drawn from videos that depict a scene's social dynamics. Our goal is to test whether VLMs can *"read the room"*, i.e., assess the alignment between their inferences and human-level understanding of the scene. Our assessment is done at two levels of analysis. First, in terms of the *visual cues* the model identifies as relevant for understanding the social situation, and second in terms of the *pragmatic inferences* made on top of these cues. Most works have focused on the correctness of the first step, i.e., testing whether models incorrectly hallucinate a visual pattern (Li et al., 2023a; Yin et al., 2023; Liu et al., 2024a; Favero et al., 2024; Rawte et al., 2025). In contrast, we focus on the second and as a result expose a previously overlooked limitation in VLMs: even when the VLM correctly identifies a relevant visual cue, it struggles to interpret it correctly. We refer to this as the **Visual Social-Pragmatic (VSP) Inference gap**. To further elucidate what we mean by this, we use Figure 1 (a frame representing a short video clip) as an example: while a man is clearly smiling, it does *not* indicate joy, as evident by his slumped posture and teary eyes. Furthermore, the second person in the scene (a pregnant woman in hospital gown) looks at him in sympathy and comforts him. The smile should be interpreted as sad or bittersweet. However, while the VLM correctly identifies the smile, it misinterprets it as evidence for joy.

Our diagnostic process proceeds in three steps. First we conduct a diagnostic analysis over the emotion recognition task in conversational videos (see Sec. 4) by evaluating the contribution of the visual modality. While offering limited (often negative) contribution when fusing the modalities directly, we note an interesting finding - when visual cues derived from the VLM augment the conversational transcript, they can lead to improved performance. This improvement is not guaranteed and careful tuning is needed, indicating that these cues or inferences made over them are noisy. Second, through human evaluation we separate between noisy cues (i.e., hallucinated) vs. misinterpreted cues (see Sec. 5.3). We curate a dataset, VIBE (VSP Inference of Behavior and Emotion), consisting of 994 unique instances of the VSP Inference task. We benchmark several VLMs on our dataset, and we compare model performances to human performance on the dataset, revealing that humans outperform the best VLM by 17.2% in accuracy. This indicates that VSP Inference–and not only hallucination–is an important limitation that VLMs struggle with. We perform analyses to elucidate exactly which kinds of social visual cues are hallucinated, misinterpreted, or uninformative. Specifically, we see that the most common failure mode on a downstream emotion recognition task is *misinterpretation*, particularly among subtle facial movements like *gaze and eye behavior* and *furrowed brows* (Sec. 5.3). Our contributions are:

1. Expose the VSP inference gap in VLMs. Propose a task and dataset that isolates the gap and provides a tool for measuring it.[1]
2. Analyze how the effects of this gap influence performance on a downstream social science task (emotion recognition).

This paper has two main components, presented in Section 4 and Section 5.

1. Section 4 (Diagnostic Emotion Prediction Task): This section builds an understanding of current VLM capabilities on tasks that require social common sense and cognitive reasoning and serves as motivation for why this is an important problem. Section 3.2 gives a high level definition and motivation for this task.

2. Section 5 (Novel VSPI Task): This section describes in detail the construction of the dataset VIBE, and the subsequent results of evaluating several models on it. Section 3.3 provides a high level introduction of the dataset VIBE and the novel VSPI task.

2 Related Work

Multi-modality in Social Science Tasks: Many methods have been proposed for multi-modal emotion prediction that involve parametric training (Zheng et al., 2023; Yun et al., 2024; Yang et al., 2023a; Huang et al., 2025). Several other approaches combine LLMs with vision models for multi-modal social understanding, e.g. Zhang et al. (2024b); You et al. (2025); Lei et al. (2024) propose prompting strategies to work around the limited reasoning of VLMs. Other works use the reasoning

[1]We will release the dataset under MIT license.

powers of an LLM in conjunction with vision tools (Hyun et al., 2024; Kelly et al., 2024; Etesam et al., 2024). Additionally, some datasets have been proposed for exploring models' capabilities in theory of mind and in emotion interpretation (Lin et al., 2025; Chen et al., 2024). The key difference in our work comes from (1) the use of videos for the temporal dimension and (2) the explicit separation that our dataset makes between the problem of hallucination and VSP inference.

Hallucinations in VLMs: VLMs are known to suffer from hallucinations (Liu et al., 2024b). Many methods attempt to measure and mitigate hallucinations, either by breaking down outputs (Li et al., 2023a; Yin et al., 2023; Petryk et al., 2024), training (Ben-Kish et al., 2024; Xie et al., 2024), or decoding algorithms (Manevich and Tsarfaty, 2024). Visual hallucinations fundamentally differ from VSP in that a description or explanation may be factually correct in isolation (identifying a smile), but provide the wrong pragmatic interpretation, and therefore, incorrect meaning (not realizing it is a sad smile).

Pragmatics: Pragmatics is the study of how context contributes to meaning (Morris, 1938), and has a rich history in linguistics. Recently, works have sought to understand and improve the pragmatics in LLMs via grounding of various forms (Sravanthi et al., 2024; Fried et al., 2023; Mohapatra et al., 2024; White et al., 2024). In the visual domain, VLMs have also been provided with context coming from outside sources, (Luo et al., 2024; Li et al., 2023b; Willemsen et al., 2023), while other works address generating contrastive captions using pragmatic inferences (Ou et al., 2023; Tsvilodub and Franke, 2023). However, in the *visual* domain, the study of pragmatics in *social* contexts remains underexplored.

Multimodal Datasets: There have been excellent datasets that tackle similar problems of social reasoning in multimodal settings. VCR (Zellers et al., 2019) is one such popular image based common-sense dataset. A closely related class of such datasets are image based emotion prediction datasets such as Kosti et al. (2019) and Yang et al. (2023b). VIBE differs from these in 2 main respects: (1) in contrast to images, videos add a layer of complexity due to longer context, subtle temporal effects changing the meaning and (2) VIBE tackles a deeper problem than these tasks by specifically testing the abilities of models to *interpret* a visual artifact.

3 Task Definitions

In this section we briefly define the two tasks that the paper addresses: The *diagnostic task* of Emotion Prediction (in multi-party conversation videos) and the *novel proposed task* of Visual Social-Pragmatic Inference (VSP Inference). We also define vocabulary used in the paper.

3.1 Vocabulary

We define with examples the terms and vocabulary used in this paper.

Video Description: The text description of a video, output by a VLM.

Visual Cue: The text representation of a *directly observable artifact* in a video. E.g. a smile, wave, laughter, etc. Anti-example: happiness (requires inference over what is seen).

VSP Inference: The interpretation of a Visual Cue. Examples:

1. Smile indicating joy: If everything else about the person aligns with happiness then the VSP inference of joy is correct. (visual cue: smile, VSP inference: joy).
2. Smile indicating sadness: If the person is smiling but also has tears in their eyes and a down-turned mouth, then the VSP inference is that it is a sad smile. (visual cue: smile, VSP inference: sadness).

Levels of VSP Inference: Via prompting, we can 'toggle' how much VSP inference a VLM is allowed to insert into its descriptions. We assign a "Level" (1, 3 or 5)[2] as:

1. Level 1: Only Visual Cues. Eg., The woman had raised eyebrows and pursed lips.
2. Level 3: Visual Cues with some inference. Eg., The woman raised her eyebrows in disapproval and pursed her lips angrily.
3. Level 5: Complete inferences. Eg., The woman was furious, with angry eyes and pursed mouth.

3.2 Diagnostic Emotion Prediction Task Overview

The input for this task is the text-transcript of a conversation between multiple people and a video clip for the target utterance. The task is to predict the emotion of the speaker of the target utterance.

[2]We do not list (or use in the paper) Levels 2 and 4. They would fall somewhere between their nearest neighbors.

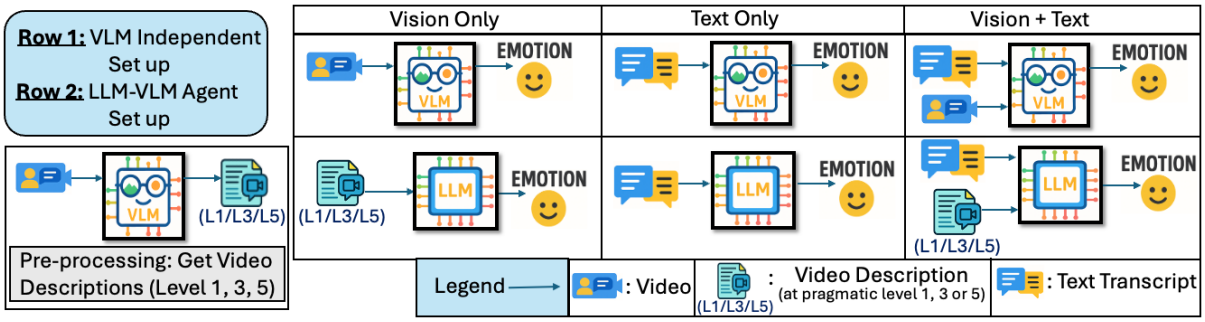

Figure 2: Emotion Prediction Experiment Set-ups. Legend on the bottom, preprocessing on the left.

Several works such as (Aviezer et al., 2012) and (Barrett and Kensinger, 2010) show that emotion recognition is not as simple as identifying facial expressions, but rather highly dependent on context. (Aviezer et al., 2011) showed that the same facial expression is interpreted differently by people when there are changes in body posture. Based on these works, in order for a VLM to do well at emotion prediction, it would need to be good at identifying relevant *Visual Cues* in the video, but just as importantly, would need to excel at making the right *VSP Inference* about the Visual Cues.

To the best of our knowledge, the capabilities of VLMs in this domain remain underexplored, and there are no datasets that isolate VSP Inference as a standalone task. For this reason, we chose emotion prediction as a starting point. This task served as a *diagnostic tool* for us to first determine the capabilities of VLMs in social intelligence.

If a VLM struggles with the emotion prediction task, it could be due to hallucinations, or due to VSP inference mistakes (seeing the cues but misinterpreting them). The goal of the following proposed task was to investigate whether the VLM struggles with one or both types of mistakes. Our intent here was not to train the best multi-modal emotion predictor, but rather to gauge the social intelligence of a VLM.

3.3 Novel Visual Social-Pragmatic Inference Task Overview

Visual Social-Pragmatic Inference is a new task that we propose in this paper. The input to the task is a video, a Visual Cue and two different VSP Inferences about the Visual Cue. The task is to pick the correct inference, based on visual context present in the video.

E.g., in Figure 1, the man is smiling. But in the video, he has tears in his eyes, and a slouched downcast body posture. His partner looks at him and rubs his hand sympathetically. Given the context, the correct interpretation is that the man is smiling sadly, reminiscing about something bittersweet.

With extensive human annotation, we constructed VIBE: a dataset of challenging VSP Inferences. VIBE was carefully curated to isolate the VSP Inference task. The Visual Cue that the VLM is required to interpret is guaranteed to be in the video, which mitigates hallucinations. Additionally, unlike other video datasets, the questions in VIBE are guaranteed to be answerable since we only have videos where the speaker is clearly visible, and the context required to interpret a Visual Cue is contained locally in the video.

Using VIBE, we conducted an analysis of several VLMs and were able to get a better understanding of VLMs abilities at VSP.

4 Diagnostic Task: Emotion Prediction

The conversations and video clips we used were sourced from the MC-EIU dataset (Liu et al., 2024c) which consists of text transcripts of conversations and video clips of the utterances. Since our focus was videos that require VSP Inferences, we used only the videos explicitly labeled with a non neutral emotion. Additionally, it is known that VLMs struggle with long context videos (Qu et al., 2025), so we only used videos that were under 4 seconds. The final dataset included 3,536 Chinese video clips (from three different shows) and 5,589 English clips (from two shows). This formed a robust, diverse set of **over 9000 videos** spanning multiple **languages, cultures, and genres**.

As illustrated in Figure 2, we had two main diagnostic experimental settings: **(1) The VLM acts independently** to do the task, and **(2) The VLM acts as a perception agent** only, and a stronger LLM (GPT-4o-mini) reasons over the visual information. For both settings, we compare performance with the very strong baseline of an LLM (GPT-4o-mini)

operating over the text transcript. The motivation behind having the two settings was to probe two different capabilities of the VLM. In setting (1) the goal was to test the VLM's ability to *directly identify the emotion of the speaker*. It was not obvious how the VLM would perform compared to the stronger LLM; while the LLM has stronger reasoning capabilities, the VLM is aided by rich visual data in addition to the text. In contrast, it was reasonable to expect a boost in performance when the VLM had both modalities, compared to either one alone. In setting (2) the goal was to test the ability of the VLM to *identify, accurately interpret and communicate* what it saw to the LLM (GPT-4o-mini). In this setting, it was reasonable to expect that the additional visual information would boost the performance of the LLM compared to the text only baseline. For both settings, to give the VLMs the best chance, we sampled the maximum possible frames that our compute resources allowed (30 frames). This decision was based on prior research that showed that performance generally increases as more frames are sampled (Hu et al., 2025; Liu et al., 2025b).

4.1 VLM Independent Setting

We benchmarked the performance of 4 VLMs of sizes varying from 3B to 26B, from 2 different families of models: InternVL2 and Qwen (Chen et al., 2025; Qwen et al., 2025). We chose the models based on their performance on vision benchmarks, support for video and the computational cost of running them. As illustrated in Figure 2, we set up the emotion prediction task for each model under 3 different settings: (1) Vision only, (2) Text Only, and (3) Text + Vision. We set the number of frames to be the maximum allowed frames for the smallest-context model (the InternVL2 models), which is 30 frames, and uniformly sampled this many frames for all models to ensure fairness. The 30-frame, 4-second limitations guaranteed at least 7 frames per second, but in most cases was more than 15.

Diagnostic Results: As seen in Table 1, the best performing model was InternVL2 26B with text only, having also very strong performance under the other two settings. Based on these results, we decided to use InternVL2 26B for further experiments. The key take-away here was, contrary to expectations, in general ***combining the modalities did not perform significantly better***, even though ***each modality had strong individual performance***.

Model	Vision	Text	Text + Vision
GPT-4o-mini[3]	–	0.538	–
Qwen 3B	0.387	0.446	0.456
Qwen 7B	0.373	0.449	0.476
InternVL-8B	0.466	0.488	0.422
InternVL-26B	0.457	**0.493**	0.468
CogVLM2-Video	0.47	–	0.387

Table 1: Weighted F1 for VLM Independent Experiments.

4.2 VLM-LLM Agent Setting

In this second diagnostic experiment, we used the VLM as a perception agent. We prompted the VLM to describe the speaker's facial expression and body language at 3 Levels of VSP Inference (see definitions) to get Video Descriptions and fed them into the LLM for classification. For some examples of the 3 Levels of VSP Inference, and an experiment confirming the toggling of levels see App. A.1. In this experiment, we implemented the same three settings as in the previous experiment: V, T, and T + V (Figure 2).

Results: This experiment's results are shown in Figure 3. The takeaways were: (1) There are strong signals in the vision which the VLM was able to communicate, (2) yet combining the modalities did not lead to significant improvement. Both the visual and textual modalities exhibited strong individual performance, but naively combining them did not yield a clear performance gain—when the VLM operated independently and also when paired with an LLM. Works like (Zheng et al., 2023) suggest that directly incorporating vision is difficult because of the noise in the visual domain. To better understand the results, we broke them down by emotion. As shown in Figure 8 (App. A), the relative performance of each modality varies across emotions. We discuss the implications of these results in the next subsection.

4.3 Emotion Prediction Discussion and Implications

From the experiments in sections 4.1 and 4.2, we saw that while VLMs show reasonable strong performance over the visual domain, there was not a clear improvement over the text based performance (Table 1 and Figure 3). This could imply two things: (1) That there is no visual information that would help improve over the text and/or (2) current VLMs do not have the ability to extract and communicate the necessary information accurately. We designed a simple yet effective algorithm to fur-

[3]Strong closed model (skyline), to contextualize results.

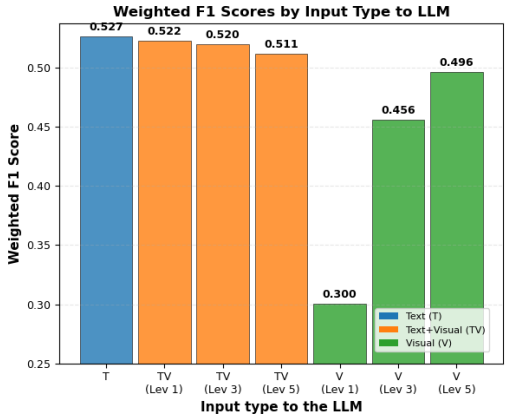

Figure 3: Weighted F1-Score for the VLM-LLM Agent experimental setting. The x-axis represents the various types of input to the LLM. Here Text is the text transcript of the conversation, and Visual means the visual cues generated from the VLM when given the conversation video. The Levels correspond to the various VSP levels of the visual cues (see definitions).

ther analyze this: the Weighted Voting Algorithm. The success of the algorithm depended on two key conditions being true: (1) There is **complementary information** in the Video Descriptions that **cannot be recovered** from the text transcripts, and (2) The VLM's performance has a **consistent pattern**: it struggles with certain emotions and does better at others, even with small sample sizes. We first describe and present results of the algorithm, then discuss its implications:

Weighted Voting Algorithm: As shown in Figure 3, we had seven sources for emotion prediction: text-only (T), visual-descriptions at three pragmatic levels (V1, V3, V5), and combined modalities with the visual description at 3 pragmatic levels (TV1, TV3, TV5). We exclude Level 1 (due to subpar performance) from the weighted voting algorithm, and refer to the remaining 5 sources as *Agents*. Each agent casts an equal vote, which is accepted if there is a clear majority. If no majority is reached, then the votes are weighted as follows:

For each of the five agents, we constructed a simple trust model by estimating their per-emotion precision on a small subset of the data, referred to as the *Calibration Set*. Each agent casts a vote for an emotion, and the influence of their vote is weighted according to the corresponding trust score.

Formally, let the agents be:

$$A = \{A_T, A_{V3}, A_{V5}, A_{TV3}, A_{TV5}\} \quad (1)$$

Let the emotion space be:

$$E = \{\text{all candidate emotion labels}\} \quad (2)$$

Using the calibration set, we compute the precision for each agent $a \in A$ and emotion $e \in E$ and use it to create a Trust function:

$$T : A \times E \to [0, 1] \quad (3)$$

Each agent $a \in A$ votes for an emotion $e_a \in E$. Their vote is weighted by the trust function T. Thus for each $e \in E$ we compute a score, $S(e)$:

$$S(e) = \sum_{a \in A} \mathbf{1}(e_a = e)\, T(a, e) \quad (4)$$

Finally the predicted emotion $\hat{e}$ is computed:

$$\hat{e} = \arg\max_{e \in E} S(e) \quad (5)$$

Overall Performance and Takeaways: Results are in Table 2. For both languages for both calibration settings we showed that our simple algorithm led to much better performance than any of the baselines. This validates our assumptions: *(1) There is **complementary information** in video descriptions **not recoverable** from the text, and (2) The VLM struggles (and succeeds) in a **consistent** manner across emotions.* From these results, we can answer the following questions:

- Is there complementary information in the videos beyond text? → Yes
- Can a VLM extract and communicate this information? → Yes (sometimes)
- Do VLM successes and failures follow systematic patterns? → Yes
- Can these patterns be leveraged to do social and cognitive reasoning tasks better? → Yes

Clearly, vision can play an important role in understanding social scenarios, and VLMs are able to (sometimes) give us valuable information that is not recoverable from the text. This encouraging result motivates us to further study exactly what kinds of things does a VLM struggle with? In the next section we show that hallucinations (well known in the literature) are not the only limitation: VLMs also struggle with VSPI. That is, even when they correctly identify visual elements, there are cases when they cannot interpret the correct VSP *meaning* of what they identified.

	Within lang en	Within lang zh	Cal. zh test en	Cal. en test zh
Vision	0.481	0.526	0.481	0.526
Text	0.498	0.568	0.501	0.567
Vis+Text	0.509	0.543	0.509	0.538
Voting	**0.523**	**0.592**	**0.530**	**0.609**

Table 2: Weighted Voting Algorithm at calibration size 50.

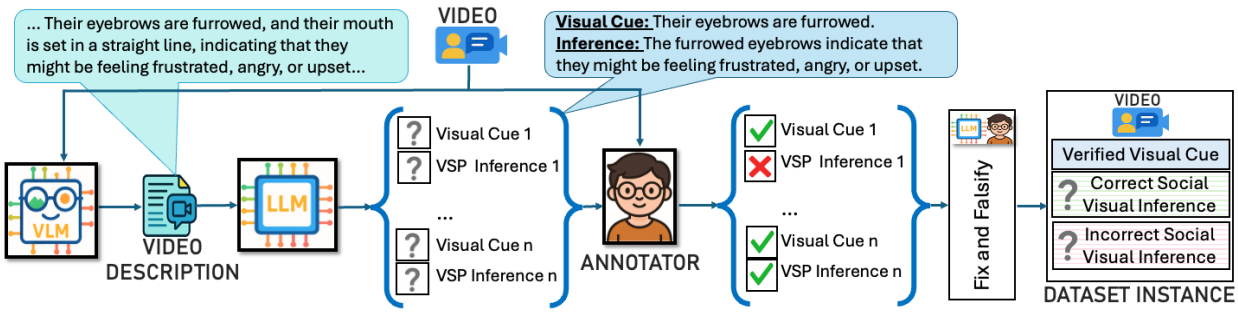

Figure 4: Main steps of the VIBE dataset curation pipeline.

5 Proposed Task: Visual Social-Pragmatic Inference

The VSP Inference task is a new task we propose in this paper. Given a Visual Cue and a video, the task is to correctly interpret the pragmatic meaning of the cue. In the following sections we describe the process of creating the dataset and formalize the inputs and outputs. We benchmark select VLMs on the dataset and present the results and analysis.

5.1 VIBE Dataset Creation

This section details the process of curating the dataset we name VIBE (VSP Inference Based on Evidence). Our vision for VIBE was a dataset of rich video clips that contain Visual Cues which can be interpreted in more than one way. At a high level, the first step we took was to carefully filter for videos that were likely to contain such information. Once we had a pool of candidate videos, we began the process of curating the dataset. The main steps are shown in Figure 4, and detailed below.

Video Selection: To ensure the dataset consisted of videos rich in challenging Visual Cues, we used performance on the emotion prediction task as a heuristic for video selection. We note that the emotion prediction task was originally designed for a multi-modal setting. Whereas for our dataset, we wanted to isolate videos that were informative and interpretable based on visual information *alone*, as is required by the VSP Inference task. To that end, we applied a three-stage filtering process to identify promising videos. Specifically, we selected videos where the emotion was misclassified for three different settings: (1) VLM given only the video, (2) LLM provided with Level 3 visual information, and (3) LLM provided with Level 5 visual cues. Furthermore, since some emotion pairs naturally co-occur very frequently (e.g., joy and surprise), we only retained videos where the predicted and gold emotions are highly unlikely to co-exist, such as joy and anger. The list of excluded and retained emotion pairs is provided in App. E.

Video Description Generation: For each video, in our candidate pool of videos, we extracted Video Descriptions that would be pertinent to the dataset. We generated these descriptions to be at Level 3 and Level 5 of VSP Inference and made sure there was a focus on facial expressions and body language. The prompts can be found in App. D.

Visual Cue and Inference Extraction: We implemented few-shot prompting to GPT-4o to break down the descriptions into candidate[4] Visual Cues and VSP Inferences.

Human Annotation: A human was shown the video clip and asked to predict the main speaker's emotion. The annotator could pick up to 3 Ekman emotions. Next, they verified the presence (or absence) of the candidate Visual Cues in the video. If the Visual Cue was confirmed to be present in the video, they rated the correctness of the Pragmatic Visual Inference drawn from the Visual Cue.

For every Visual Cue-VSP Inference pair, we had 2-4 humans do the annotation. We chose annotators from a variety of backgrounds and included both native English speakers and native Chinese speakers. All annotators had at least a Bachelor's degree and were proficient in English. The annotation involved a total of about 25 hours of high quality human annotation time. Even though this was an inherently subjective task, we saw an agreement of 74.5% for visual cues (random being 50%), and 52.8% for VSP inferences (random being 33%). To ensure the high quality of VIBE, the dataset only includes Visual Cue-VSP Inference pairs for which at least 2 annotators had perfect agreement on both Visual Cue and VSP inference.

Fix and Falsify: Post human annotation, we were left with a set of verified Visual Cues, VSP Inferences that were confirmed to be correct or

[4] "candidate" since Visual Cues are unverified at this stage.

incorrect, and a pool of emotions for the speaker labeled by annotators. For each video, we computed a "max scoring emotion", which was the emotion that was labeled the most times by annotators (ties broken arbitrarily). We envisioned VIBE as a multiple choice dataset, so as the final step, we needed to come up with correct and incorrect counter-parts for all the inferences we had. We used an LLM (GPT-4o-mini) and the human labeled emotions to do this final step of "fixing" and "falsifying" the inferences.

Creating Correct Inference Choices: For all the inferences that were confirmed *incorrect* by annotators, we came up with the correct inference using the following methods:

1. When the max scoring emotion from annotators agreed with the gold label for the video, we used this emotion to correct the wrong inference. We call these *human fixed inferences*.
2. When the agreement was not there, we simply negated the incorrect inference. We call these *fixed by negation inferences*.

Creating Incorrect Inference Choices: For all the inferences that were confirmed *correct* by annotators, we came up with the incorrect inference using the following methods:

1. When the max scoring emotion from annotators agreed with the gold label for the video, we used this emotion to come up with a mutually exclusive emotion (App. E). We then used this emotion to falsify the inference. We call these *distractor inferences*.
2. When an agreement could not be reached, we employed two methods: We either used the emotion that no annotator voted for to falsify the inference (these are also distractor inferences) or we simply negated the correct inference, which we call *false negation inferences*.

Final Dataset Composition: In the final dataset we had 50% negation style choices (which could be correct or incorrect), and 50% distractor style choices (which could also be correct or incorrect). The diverse methods were carefully designed to safeguard the dataset against being hacked by always choosing negations or distractors or even by discerning between VLM and LLM generated text.

Final Dataset: The final dataset contains 433 unique video clips and 994 unique instances of the VSP Inference task. An instance of the task is: Given a video, a Visual Cue, and two candidate VSP Inferences, output the correct inference.

5.2 VSP Inference Experiments

We benchmarked all VLMs used in the emotion prediction task on our VIBE dataset to illustrate its difficulty. We additionally include the expensive OpenAI GPT-4o mini model as a reference to large, closed VLMs. All models are prompted using CoT, though we include results for standard prompting in App. B. Finally, we had two humans do the VIBE dataset on 100 instances each as comparison. The results can be seen in Table 3. These results demonstrate that (1) VLMs struggle with VSPI (which comes intuitively to humans), and (2) The question types in VIBE that VLMs struggle with most (fixed and NT) are consistently hard for all the models we evaluated (but not for humans!). This result underscores the importance of targeting VSPI for VLMs in order to make them socially and emotionally intelligent. In section 5.3 we dive into a deeper analysis of the kinds of mistakes made on VIBE and connect the results back to the emotion prediction task from Section 4

Model	Fixed	Distractor	NF	NT	All
IVL2-26B[5]	29.5	74.5	78.8	18.1	63.4
IVL2-8B	13.1	93.5	**97.3**	5.5	73.8
IVL2-4B	20.5	86.6	87.7	29.1	71.5
Q2.5-3B	25.4	90.3	93.8	22.8	75.1
Q2.5-7B	26.2	93.3	87.7	26.8	74.4
4o-mini	31.1	90.1	94.6	17.3	75.3
CogVLM 2-Video	25.4	78.0	95.4	12.6	69.7
Humans	**88.9**	**94.7**	95.2	**81.8**	**92.5**

Table 3: Accuracy on VIBE by model and question type (CoT). NF and NT are *negation false* and *negation true* respectively.

5.3 VSP Inference Analysis

In this section, we use the VIBE dataset to quantitatively analyze the VLM's ability to describe a scene (i.e. *read the room*). We do this by investigating the impacts of its visual cues on the downstream emotion prediction task. We aim to answer two questions: (1) "Which kinds of visual cues help reach the correct emotion label?" and (2) "When a mistake is made, is it due to hallucination, misinterpretation, or something else (like an uninformative visual cue, necessary textual context, etc.)?".

From our annotations, using only annotator-agreed labels, we had over 1.2 K visual cues. Among these, 27.3% were labeled as hallucinations. In the remaining, non-hallucinated visual

[5]Evaluation on this model is not wholly fair as the dataset was partially created from its mistakes.

cues, 27.5% were labeled as being misinterpreted.

In order to better understand the distribution, we came up with 12 clusters of visual cues (e.g. "Emphatic gestures", "Smile/Laughter", etc.), created keyword representations of them, and used cosine similarity over SBERT embeddings to retrieve the top 200 visual cues for each. Since all visual cues in the dataset were there because they led to an incorrect emotion prediction, we aimed to explain that error using the human annotations. Specifically, we calculated the hallucination and misinterpretation rate within each cluster. If the visual cue was neither hallucinated nor misinterpreted, then it was correct, but for other reasons led to an incorrect emotion prediction (e.g. being uninformative, textual context being necessary, etc.). Figure 5 shows the composition of the kinds of errors among the different visual cue clusters.

We see that the "Smile/Laughter" cluster has both a high misinterpretation and hallucination rate, meaning that the VLM often mistakes non-smiles for smiles, and even when it identifies a real smile, it often cannot interpret its meaning (e.g. a *sad smile*). Likewise, we see that the cluster "Leaning & Body Orientation" visual cues (e.g. "leaning forward", "upright posture", "laying down") are less commonly hallucinated, but often misinterpreted.

In order to further investigate how these errors led to misclassification, we mapped the clusters onto the full, un-annotated space of visual cues used for the emotion prediction task. To do this, we used the same cluster keywords and the same cosine similarity search mechanism to retrieve the top 500 visual cues for each cluster. We then computed the precision and recall for each emotion, for each cluster. The result for 'Joy' can be seen in Figure 6 (other emotions in App. C). We plot $1-P$ and $1-R$ on the y-axis to emphasize bigger errors and visualize their compositions.

We see that for 'Joy', recall failures substantially exceed precision failures across every cluster, indicating that the VLM more often *misses true visual cues* than it hallucinates spurious ones. Particularly, subtle facial movements such as "Brow Furrows" and "Gaze & Eye Behavior" exhibit the highest total failure rates, with nearly all true instances going undetected (very low recall). In sharp contrast, the 'Smile/Laughter" cluster is almost perfectly handled (both precision and recall failures < 0.1), showing the model's strength on overt expressions like smiles. Across the board, misinterpretation comprises the largest slice of the error budget, followed by hallucination, while other error types remain minimal. These results suggest that, to improve emotion-prediction accuracy, future work should focus on bolstering the model's sensitivity to subtle nonverbal signals and reducing its tendency to misinterpret correctly detected cues.

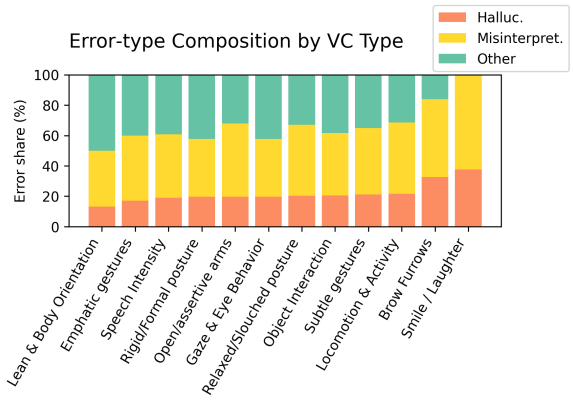

Figure 5: Distribution of error type by type of visual cue.

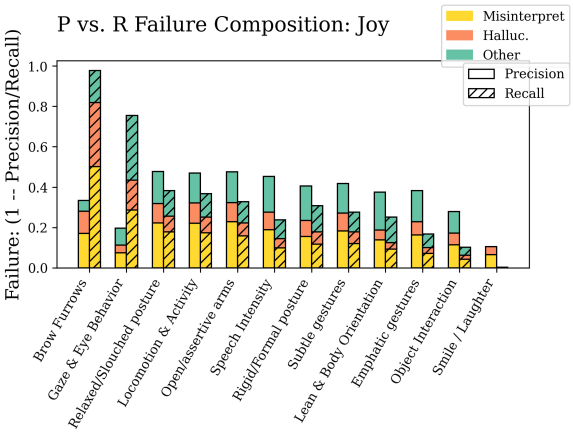

Figure 6: Error analysis for 'Joy'. Bigger bars = bigger errors.

6 Conclusion

We introduced the VSP Inference task, exposing a previously overlooked limitation in VLMs. Through diagnostic analyses, human annotation and the VIBE dataset, we showed that current VLMs struggle with VSP Inference. Our results underscore the need for improved social reasoning in VLMs and provide a benchmark for future research.

7 Limitations and Future Work

Variety in Vision models chosen: In this paper we limited our analyses to 3 families of VLMs up to size 26B. The VIBE dataset was tested on 7 models of varying families and sizes. More extensive analysis including more models and architectures could lead to more insights.

Bias Due to Video Description Source: The Visual Cues in our dataset were selected from failure cases of InternVL2 26B, which may have inherent biases. Anything the model did not identify as relevant would be missed in the dataset. The dataset would benefit from having more models in the video description generation process.

Inherent subjectivity in human annotation: While we saw high inter-annotator agreement, the annotation task was inherently subjective. Cultural and personal biases may have some impact on the dataset, even if it only meant that we were not able to include some hard cases because of disagreement within annotators.

Future Work The limitations mentioned above give rise to some natural directions for future work. It would be worthwhile to study more models on this family of tasks, with an emphasis on architecture and training data to see if there are significant performance differences. Additionally, given the subjective nature of the task, a promising extension of this paper would be to model variations in human judgments as uncertainty estimation of the models. Finally, this paper is a first step into investigating the alignment of multi-modal models and human perceptions of social scenarios; We believe that further investigations and subsequent improvements to these models will lead to socially competent AI systems.

8 Ethics Statement

This paper addresses social reasoning capabilities of VLMs, and proposes a dataset for it. We acknowledge that the inherent subjective nature of the task might affect the dataset. To mitigate any issues we employed a diverse group of annotators that were willing to volunteer their time. We recognize that systems that aim to mimic or understand human social dynamics raise ethical concerns. The purpose of this paper is not to endow VLMs with any malicious capabilities but rather to foster better understanding of such models in the research community. The video clips in our data-set were sourced from an existing research dataset and were obtained with the required permissions. This data is meant to be used for research purposes only.

Acknowledgments

We thank the reviewers for their insightful comments that helped improve the paper. This work was supported by NSF CAREER award IIS2048001 and the DARPA CCU program. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by, DARPA, or the US Government.

References

Rizwan Abbas, Bingnan Ni, Ruhui Ma, Teng Li, Yehao Lu, and Xi Li. 2024. Context-based emotion recognition: A survey. *Neurocomputing*, 618:129073.

Hillel Aviezer, Shlomo Bentin, Veronica Dudarev, and Ran R. Hassin. 2011. The automaticity of emotional face-context integration. *Emotion*, 11(6):1406–1414.

Hillel Aviezer, Yaacov Trope, and Alexander Todorov. 2012. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, 338(6111):1225–1229.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Lisa Feldman Barrett and Elizabeth A. Kensinger. 2010. Context is routinely encoded during emotion perception. *Psychological Science*, 21(4):595–599.

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. 2024. Mitigating open-vocabulary caption hallucinations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22680–22698, Miami, Florida, USA. Association for Computational Linguistics.

Sree Bhattacharyya and James Z. Wang. 2025. Evaluating vision-language models for emotion recognition. In *In Proceedings of NAACL-findings*, volume abs/2502.05660.

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, and 1 others. 2024. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*.

Zhawnen Chen, Tianchun Wang, Yizhou Wang, Michal Kosinski, Xiang Zhang, Yun Fu, and Sheng Li. 2024. Through the theory of mind's eye: Reading minds with multimodal video large language models. *Preprint*, arXiv:2406.13763.

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

Beatrice de Gelder. 2006. Towards the neurobiology of emotional body language. *Nature Reviews Neuroscience*, 7(3):242–249.

Antoine de Saint-Exupéry. 1943. *The Little Prince*. Reynal & Hitchcock, New York. Originally published in French as *Le Petit Prince*.

Paul Ekman and Wallace V. Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA.

Yasaman Etesam, Özge Nilay Yalçın, Chuxuan Zhang, and Angelica Lim. 2024. Contextual emotion recognition using large vision language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 4769–4776. IEEE.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

José M. Fernández-Dols and Marisol A. Ruiz-Belda. 1995. Are smiles a sign of happiness? gold medal winners at the olympic games. *Journal of Personality and Social Psychology*, 69(6):1113–1119.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. *Preprint*, arXiv:2211.08371.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, and Trishul Chilimbi. 2025. M-llm based video frame selection for efficient video understanding. *Preprint*, arXiv:2502.19680.

Dawei Huang, Qing Li, Chuan Yan, Zebang Cheng, Yurong Huang, Xiang Li, Bin Li, Xiaohui Wang, Zheng Lian, and Xiaojiang Peng. 2025. Emotion-qwen: Training hybrid experts for unified emotion and general vision-language understanding. *Preprint*, arXiv:2505.06685.

Lee Hyun, Kim Sung-Bin, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. 2024. Smile: Multimodal dataset for understanding laughter in video with language models. *Preprint*, arXiv:2312.09818.

Chris Kelly, Luhui Hu, Bang Yang, Yu Tian, Deshun Yang, Cindy Yang, Zaoshan Huang, Zihao Li, Jiayin Hu, and Yuexian Zou. 2024. Visiongpt: Vision-language understanding agent using generalized multimodal framework. *Preprint*, arXiv:2403.09027.

Ronak Kosti, Jose Alvarez, Adria Recasens, and Agata Lapedriza. 2019. Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1.

Yuxuan Lei, Dingkang Yang, Zhaoyu Chen, Jiawei Chen, Peng Zhai, and Lihua Zhang. 2024. Large vision-language models as emotion recognizers in context awareness. *Preprint*, arXiv:2407.11300.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023a. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

Yunxin Li, Baotian Hu, Chen Xinyu, Yuxin Ding, Lin Ma, and Min Zhang. 2023b. A multi-modal context reasoning approach for conditional inference on joint textual and visual clues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10757–10770, Toronto, Canada. Association for Computational Linguistics.

Zaijing Li, Gongwei Chen, Rui Shao, Yuquan Xie, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing emotional generation capability of large language models via emotional chain-of-thought. *Preprint*, arXiv:2401.06836.

Yuxiang Lin, Jingdong Sun, Zhi-Qi Cheng, Jue Wang, Haomin Liang, Zebang Cheng, Yifei Dong, Jun-Yan He, Xiaojiang Peng, and Xian-Sheng Hua. 2025. Why we feel: Breaking boundaries in emotional reasoning with multimodal large language models. *Preprint*, arXiv:2504.07521.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024b. A survey on hallucination in large vision-language models. *Preprint*, arXiv:2402.00253.

Rui Liu, Haolin Zuo, Zheng Lian, Xiaofen Xing, Björn W. Schuller, and Haizhou Li. 2024c. Emotion and intent joint understanding in multimodal conversation: A benchmarking dataset. *Preprint*, arXiv:2407.02751.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025a. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv preprint arXiv:2501.01282*.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, and 8 others. 2025b. Nvila: Efficient frontier visual language models. *Preprint*, arXiv:2412.04468.

Ziyi Liu, Abhishek Anand, Pei Zhou, Jen tse Huang, and Jieyu Zhao. 2024d. Interintent: Investigating social intelligence of llms via intention understanding in an interactive game context. *Preprint*, arXiv:2406.12203.

Fuwen Luo, Chi Chen, Zihao Wan, Zhaolu Kang, Qidong Yan, Yingjie Li, Xiaolong Wang, Siyu Wang, Ziyue Wang, Xiaoyue Mi, Peng Li, Ning Ma, Maosong Sun, and Yang Liu. 2024. Codis: Benchmarking context-dependent visual comprehension for multimodal large language models. *Preprint*, arXiv:2402.13607.

Avshalom Manevich and Reut Tsarfaty. 2024. Mitigating hallucinations in large vision-language models (LVLMs) via language-contrastive decoding (LCD). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6008–6022, Bangkok, Thailand. Association for Computational Linguistics.

Albert Mehrabian. 1972. *Nonverbal Communication*. Aldine-Atherton, Chicago.

Albert Mehrabian and Morton Wiener. 1967. Decoding of inconsistent communications. *Journal of Personality and Social Psychology*, 6(1):109–114.

Biswesh Mohapatra, Manav Nitin Kapadnis, Laurent Romary, and Justine Cassell. 2024. Evaluating the effectiveness of large language models in establishing conversational grounding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9767–9781, Miami, Florida, USA. Association for Computational Linguistics.

Charles W. Morris. 1938. Foundations of the theory of signs. In Otto Neurath, editor, *International Encyclopedia of Unified Science*, volume 1, pages 1–59. University of Chicago Press, Chicago.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and Zhongyu Wei. 2024. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios. *Preprint*, arXiv:2410.19346.

Jiefu Ou, Benno Krojer, and Daniel Fried. 2023. Pragmatic inference with a clip listener for contrastive captioning. *Preprint*, arXiv:2306.08818.

Suzanne Petryk, David M. Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph E. Gonzalez, and Trevor Darrell. 2024. Aloha: A new measure for hallucination in captioning models. *Preprint*, arXiv:2404.02904.

Tianyuan Qu, Longxiang Tang, Bohao Peng, Senqiao Yang, Bei Yu, and Jiaya Jia. 2025. Does your vision-language model get lost in the long video sampling dilemma? *Preprint*, arXiv:2503.12496.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Vipula Rawte, Aryan Mishra, Amit Sheth, and Amitava Das. 2025. Defining and quantifying visual hallucinations in vision-language models. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 501–510.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650.

Robert P. Spunt and Ralph Adolphs. 2019. The neuroscience of understanding the emotions of others. *Neuroscience Letters*, 693:44–48. Funding by NIH.

Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. *Preprint*, arXiv:2401.07078.

Haomiao Sun, Mingjie He, Tianheng Lian, Hu Han, and Shiguang Shan. 2024. Face-mllm: A large face perception model. *arXiv preprint arXiv:2410.20717*.

Polina Tsvilodub and Michael Franke. 2023. Evaluating pragmatic abilities of image captioners on a3ds. *Preprint*, arXiv:2305.12777.

Isadora White, Sashrika Pandey, and Michelle Pan. 2024. Communicate to play: Pragmatic reasoning for efficient cross-cultural communication. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12201–12216, Miami, Florida, USA. Association for Computational Linguistics.

Bram Willemsen, Livia Qian, and Gabriel Skantze. 2023. Resolving references in visually-grounded dialogue via text generation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 457–469, Prague, Czechia. Association for Computational Linguistics.

Zhaoqing Wu, Dan Goldwasser, Maria Leonor Pacheco, and Leora Morgenstern. 2025. Identifying power relations in conversations using multi-agent social reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 855–865.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association*

for Computational Linguistics: EMNLP 2024, pages 13258–13273, Miami, Florida, USA. Association for Computational Linguistics.

Haozhe Yang, Xianqiang Gao, Jianlong Wu, Tian Gan, Ning Ding, Feijun Jiang, and Liqiang Nie. 2023a. Self-adaptive context and modal-interaction modeling for multimodal emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6267–6281, Toronto, Canada. Association for Computational Linguistics.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2023b. Emoset: A large-scale visual emotion dataset with rich attributes. *Preprint*, arXiv:2307.07961.

Qu Yang, Mang Ye, and Bo Du. 2024. Emollm: Multimodal emotional understanding meets large language models. *arXiv preprint arXiv:2406.16442*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12).

Haoxuan You, Zhecan Wang, Rui Sun, Long Chen, Gengyu Wang, Hammad A. Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. 2025. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. *Preprint*, arXiv:2305.14985.

Taeyang Yun, Hyunkuk Lim, Jeonghwan Lee, and Min Song. 2024. Telme: Teacher-leading multimodal fusion network for emotion recognition in conversation. *Preprint*, arXiv:2401.12987.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. *Preprint*, arXiv:1811.10830.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024a. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Qixuan Zhang, Zhifeng Wang, Dylan Zhang, Wenjia Niu, Sabrina Caldwell, Tom Gedeon, Yang Liu, and Zhenyue Qin. 2024b. Visual prompting in llms for enhancing emotion recognition. *Preprint*, arXiv:2410.02244.

Wenjie Zheng, Jianfei Yu, Rui Xia, and Shijin Wang. 2023. A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15445–15459, Toronto, Canada. Association for Computational Linguistics.

A Diagnostic Results

A.1 Experiment to support toggling Levels of VSP Inference

In order to generate output at the correct level of VSP Inference, prompts were manually tuned with care. Additionally we did an experiment to confirm that the outputs generally matched the expected VSP Inference Level; We extracted the top 20 n-grams (n=5, 7, 10) from each of the 3 levels. These were manually inspected (and 10 are provided below in Table5. Additionally, the 20 10-grams were provided to GPt o4-mini-high to compare them with shuffled level order using the prompt:

> I'm going to give you top phrases from three different sources. Summarize the main differences between the 3 sources. Output a table where each row is a source. The columns should be "Focus" and "Characteristic Phrases".
>
> Source x: <top n-grams from level 1>
> Source y: <top n-grams from level 5>
> Source z: <top n-grams from level 3>

Figure 7: Prompt to confirm VSP Level.

The summary result (re-ordered for readability):

Level 1 (Source: X)
Focus: Detailed, low-level facial movements **Characteristic Phrases:** – "his eyebrows move up and down" – "her mouth and lips open and close as she speaks"
Level 3 (Source: Z)
Focus: Broader body-language/facial cues **Characteristic Phrases:** – "exhibits a range of body language cues that suggest she is" – "his eyebrows are slightly furrowed indicating concern"
Level 5 (Source: Y)
Focus: Explicit emotion labeling/classification **Characteristic Phrases:** – "emotions such as joy sadness fear disgust surprise or anger" – "he is feeling joy. he is smiling and appears to be"

Table 4: The Focus and Characteristic Phrases confirm that the toggling worked as expected.

A.2 Complete VLM-LLM Agent Experiment Results

Here are the f1-scores of the VLM-LLM agent setting for both languages (and overall). Chinese scores are generally slightly higher than the Egnlish ones, which could be due to quirks of the VLM model or the data itself.

10-gram	Freq	7-gram	Freq	5-gram	Freq
Level 1 (Detailed, low-level facial movements)					
changes in his facial features the video. his eyebrows move	1031	mouth and lips open and close as	2117	eyebrows move up and down	3621
and her mouth and lips open and close as she	1020	maintains a relatively static posture the video.	1975	the the man in the	2690
her mouth and lips open and close as she speaks.	1005	changes in his facial features the video.	1294	mouth and lips open and	2482
changes in her facial features the video. her eyebrows move	998	in his facial features the video. his	1294	and lips open and close	2473
facial features the video. his eyebrows move up and down	975	his facial features the video. his eyebrows	1294	there are no significant changes	2433
in his facial features the video. his eyebrows move up	960	changes in her facial features the video.	1180	are no significant changes in	2433
his facial features the video. his eyebrows move up and	960	in her facial features the video. her	1178	lips open and close as	2128
and his mouth and lips open and close as he	855	her facial features the video. her eyebrows	1178	maintains a relatively static posture	2040
in her facial features the video. her eyebrows move up	854	eyebrows move up and down her forehead	1144	a relatively static posture the	2007
her facial features the video. her eyebrows move up and	854	her mouth and lips open and close	1142	relatively static posture the video.	1975
...		...		...	
Level 3 (Broader body-language/facial cues)					
a range of body language cues that suggest she is	647	exhibits a range of facial expressions the	2150	a range of facial expressions	3760
exhibits a range of body language cues that suggest she	644	a range of facial expressions the video.	1798	exhibits a range of facial	3704
exhibits a range of facial expressions that suggest he is	601	exhibits a range of facial expressions that	1522	is feeling a mix of	3547
that suggest she is feeling a mix of emotions. her	544	exhibits a range of body language cues	1488	the a man in a	2297
exhibits a range of body language cues that suggest he	513	a range of body language cues that	1482	he is engaged in a	2285
a range of body language cues that suggest he is	513	range of body language cues that suggest	1430	range of facial expressions the	2192
that suggest he is feeling a mix of emotions. his	512	a range of facial expressions that suggest	1332	is engaged in a serious	2036
his eyebrows are slightly furrowed indicating concentration or concern. his	501	his body language suggests that he is	1156	body language cues that suggest	1941
eyebrows are slightly furrowed indicating concentration or concern. his mouth	482	shirt exhibits a range of facial expressions	938	be feeling a mix of	1891
are slightly furrowed indicating concentration or concern. his mouth is	482	the slight furrow in his brow and	936	that he might be feeling	1831
...		...		...	
Level 5 (Explicit emotion labeling)					
emotions such as joy sadness fear disgust surprise or anger.	680	the in the appears to be feeling	2188	the in the is the	5982
the in the appears to be expressing a mix of	510	the in the appears to be expressing	1446	in the appears to be	4947
strong emotions such as joy sadness fear disgust surprise or	440	joy sadness fear disgust surprise or anger.	1345	the in the appears to	4808
the in the appears to be expressing a sense of	406	appears to be engaged in a conversation	1335	to be engaged in a	3396
such as joy sadness fear disgust surprise or anger. based	404	in the is the man in the	1247	in the is the man	3015
as joy sadness fear disgust surprise or anger. based on	404	the in the is the man in	1244	the appears to be feeling	2554
that he is feeling joy. he is smiling and appears	383	appears to be expressing a mix of	1226	appears to be engaged in	2534
based on these observations it seems that the is feeling	380	to be engaged in a conversation with	1054	in the is the woman	2104
a positive and happy emotion. the in the is the	349	in the appears to be expressing a	963	appears to be expressing a	2090
is feeling joy. he is smiling and appears to be	329	the in the is the man wearing	946	her body language including her	1685
...		...		...	

Table 5: Representative n-grams (10, 7, and 5) across Levels 1, 3, and 5.

Modality	Overall	Chinese	English
T	0.5266	0.5673	0.5006
TV_1	0.5224	0.5582	0.4991
TV_3	0.5199	0.5376	0.5091
TV_5	0.5114	0.5422	0.4952
V_1	0.3002	0.2326	0.3438
V_3	0.4559	0.4568	0.4583
V_5	0.4961	0.5260	0.4805

Table 6: Weighted F1-scores for LLM operating over various modalities. Visual Cues are generated by a VLM given videos of the conversation.
T: Transcript, TV: Transcript & Visual Cues, V: Visual Cues

A.3 Weighted Voting Algorithm Details and Full Results

We split our dataset into a very small calibration pool (300 chinese, 550 english) which still left a large test set to test on (3236 for chinese, 5039 for english). For various calibration sizes, we randomly sampled from the calibration pool, and repeated the algorithm described above 50 times. As shown in Table 7 even at just calibration size of 50, we were able to do much better than any of the other LLM classifications. We also report standard error and confidence intervals for the calibration sets.

B Benchmarking on VIBE

Table 8 shows results for benchmarking on VIBE using a direct prompt instead of CoT prompting.

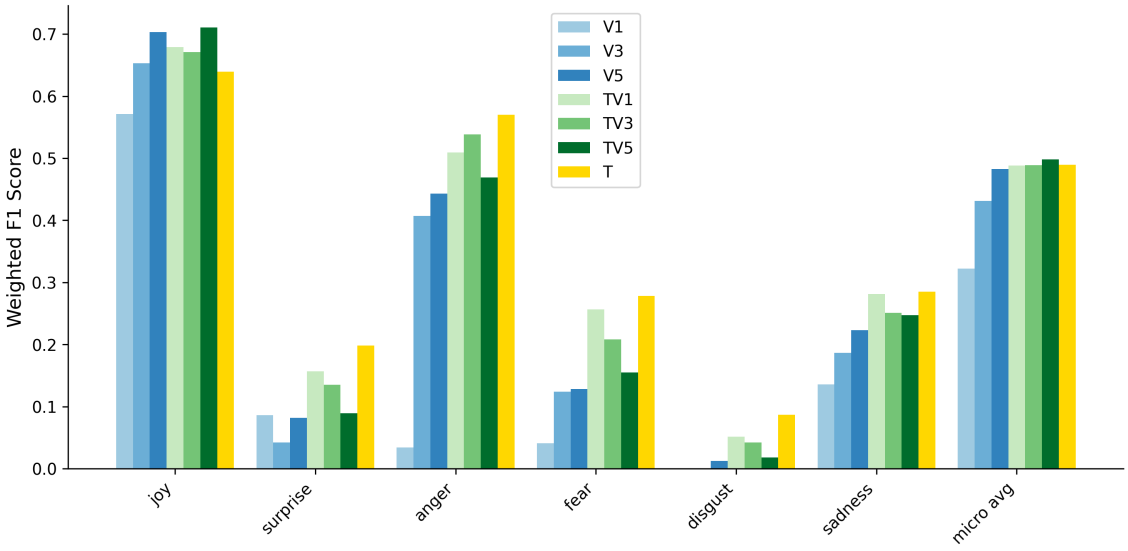

Figure 8: Weighted f1-by emotion.

Cal size	Text	Vision	Vision+Text	Ensemble
Within Language: en (Test: 5039, Calibration Pool: 550)				
0	0.498	0.481	0.509	0.522
50	0.498	0.481	0.509	0.523 (7.49e-04, 1.51e-03)
100	0.498	0.481	0.509	0.523 (6.88e-04, 1.38e-03)
250	0.498	0.481	0.509	0.524 (5.52e-04, 1.11e-03)
500	0.498	0.481	0.509	0.526 (1.45e-04, 2.91e-04)
Within Language: zh (Test: 3236, Calibration Pool: 300)				
0	0.568	0.526	0.543	0.567
50	0.568	0.526	0.543	0.592 (2.42e-04, 4.85e-04)
75	0.568	0.526	0.543	0.592 (1.72e-04, 3.46e-04)
100	0.568	0.526	0.543	0.592 (1.16e-04, 2.33e-04)
250	0.568	0.526	0.543	0.592 (6.17e-05, 1.24e-04)
Cross Language: zh Calibration, en Test (Test: 5589, Cal Pool: 300)				
0	0.501	0.481	0.509	0.542
50	0.501	0.481	0.509	0.530 (5.10e-04, 1.02e-03)
75	0.501	0.481	0.509	0.530 (3.97e-04, 7.98e-04)
100	0.501	0.481	0.509	0.530 (3.42e-04, 6.87e-04)
250	0.501	0.481	0.509	0.531 (1.56e-04, 3.13e-04)
Cross Language: en Calibration, zh Test (Test: 3536, Cal Pool: 550)				
0	0.567	0.526	0.538	0.598
50	0.567	0.526	0.538	0.609 (1.65e-03, 3.33e-03)
100	0.567	0.526	0.538	0.613 (1.16e-03, 2.33e-03)
250	0.567	0.526	0.538	0.614 (5.87e-04, 1.18e-03)
500	0.567	0.526	0.538	0.614 (2.97e-04, 5.97e-04)

Table 7: Results from Weighted Voting Algorithm

C Failure Analysis

Figures 10, 11, and 12 show the error breakdowns for the remaining emotions. Emotions of disgust and surprise had too few data points to contain significant results. Among the figures, clusters with too few points to be significant are also excluded.

D Prompts

Figures 13-22 contain all prompts used during diagnostic experiments and for dataset creation.

E Implementation Details

Model	Fixed	Distractor	NF	NT	All
IVL2-26B	24.6	88.7	93.8	19.7	73.9
IVL2-8B	31.1	87.1	94.4	15.0	73.7
IVL2-4B	27.9	88.2	90.6	29.1	74.1
Q2.5-3B	30.3	86.9	92.3	16.0	72.8
Q2.5-7B	26.2	88.2	93.0	19.7	73.6

Table 8: Accuracy by model and question type (Direct Prompt).

Task Overview
You are a socially intelligent body language expert. Your task is to interpret a person's body language. You will be given a video clip with one main speaker and asked which interpretation of their body language is better. Follow the Task Guidelines and the Response Format.

Task Guidelines
- You are given a video clip with one main speaker.
- You are given one fact about the speaker's body language, and two possible interpretations of that body language.
- Think out loud about which interpretation is better given what you see in the video (2-3 sentences).
- Finally, give your answer according to the Response Format.

Response Format
Thinking out loud: <your thoughts about which interpretation is better (2-3 sentences)>
Answer: <A OR B>

Video Clip
{clip} ### Fact
{ fact_text }

Interpretations
A. { inference_A }
B. { inference_B }

Response

Figure 9: CoT prompt used for all VLMs on the VIBE benchmarking.

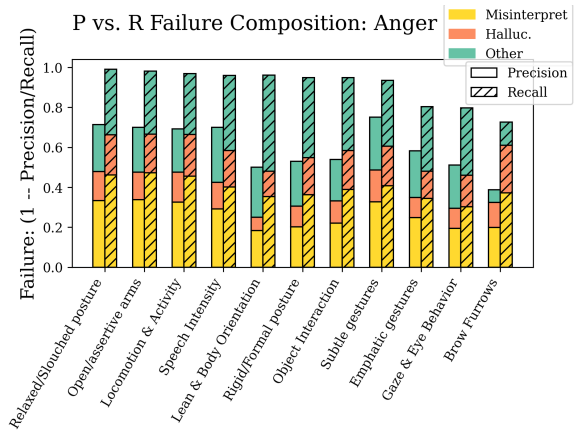

Figure 10: Error analysis for 'Anger'. Bigger bars = bigger errors.

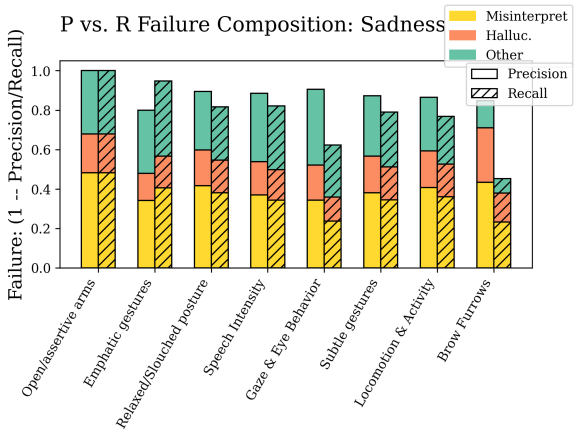

Figure 11: Error analysis for 'Sadness'. Bigger bars = bigger errors.

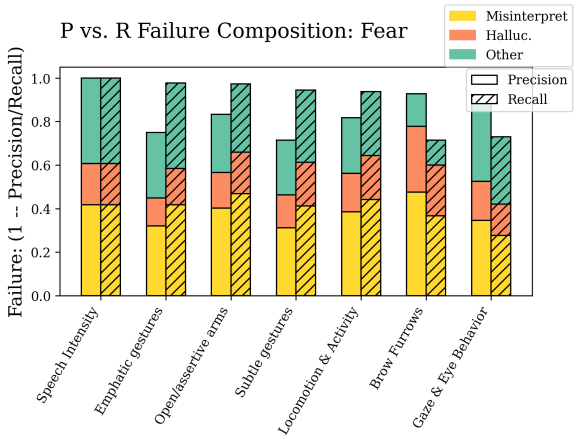

Figure 12: Error analysis for 'Fear'. Bigger bars = bigger errors.

You are given a short video clip from a TV show. There is one main speaker. Throughout the video, think out loud about the mouth movements of the people in the foreground. Based on this, decide who the main speaker is.

Figure 13: VLM Prompt to get speaker.

In 3-4 sentences, without using any adjectives or emotions, describe the changes in the facial features (eyebrows, forehead, eyes, nose, cheeks, mouth and lips.) of the identified speaker. Do not interpret what they mean.

Figure 14: VLM Prompt to get facial expression at Level 1

In 3-4 sentences, describe the facial expressions of the speaker, with an emphasis on the features that hint at the person's emotion. Describe what each of the features indicates about the person's emotional state. What emotion might they be feeling?

Figure 15: VLM Prompt to get facial expression at Level 3

We want to decide the emotion of the speaker. The options are joy, sadness, fear, disgust, surprise and anger. In 3-4 sentences, think about the facial expression of the speaker and what they indicate. Based on this, what emotion is the speaker feeling?

Figure 16: VLM Prompt to get facial expression at Level 3.

In 3-4 sentences, without using any adjectives or emotions, describe the changes in the body language (arms, hands, legs, torso) of the identified speaker. Do not interpret what they mean.

Figure 17: VLM Prompt to get body language at Level 1.

In 3-4 sentences, describe the body language of the speaker, with an emphasis on the features that hint at the person's emotion. Describe what each of the features indicates about the person's emotional state. What emotion might they be feeling?

Figure 18: VLM Prompt to get body language at Level 3.

We want to decide the emotion of the speaker. The options are joy, sadness, fear, disgust, surprise and anger. In 3-4 sentences, think about the body language of the speaker and what they indicate. Based on this, what emotion is the speaker feeling?

Figure 19: VLM Prompt to get body language at Level 5.

Emotion 1	Emotion 2
joy	anger
joy	sadness
joy	fear
joy	disgust
sadness	surprise
sadness	fear
fear	disgust
sadness	disgust
anger	fear
anger	disgust

Table 9: Pairs of emotions that were included in list of "bad" mistakes.

Emotion	Opposite Emotion
anger	joy
fear	joy
joy	sadness
sadness	joy
disgust	joy
surprise	sadness

Table 10: Emotion-opposite pairs.

Task Overview:

You will be given a piece of text. The text will consist of facts (Eg: she has a smile) and inferences (Eg: indicating she is relaxed). Your task is separate the facts and inferences. Try to pair up the facts and inferences. The same fact can be repeated for multiple inferences (explicitly repeat it). Say "No Fact" if there is no fact with an inference. Follow the guidelines to do this.

Strict Guidelines:

- Facts are physical traits like smiles, furrowed brows, other physical traits **without adjectives**.
- The fact may come after the inference in some sentences. Example: Her facial expression is one of happiness and contentment, with a smile on her face. Fact: There is a smile on the speaker's face. Inference: The smile suggests happiness and contentment.
Definition of Inferences:
- Inferences are **what the facts mean or indicate** such as happiness, sadness etc.
- All emotions are inferences.
- ALL adjectives are inferences (in distress, tense, calm, sadly, etc..)
- Any words like "indicates", "suggests", "appears", etc are pointers to inferences.
3 **VERY STRICT RULE**: Do not come up with inferences on your own. Only cluster the information already present in the text.

Examples:

Example text 1:

The speaker, the woman driving the car, has a neutral expression with a slight smile, indicating that she is calm and possibly content. Her eyes are focused forward, suggesting that she is engaged in the conversation. The slight smile on her face hints at a positive emotion, such as happiness or satisfaction. Overall, her facial expression suggests that she is in a relaxed and pleasant emotional state.

Example response 1:

Information breakdown:
1. The speaker is a woman driving the car.
2. The speaker has a neutral expression with a slight smile, indicating that she is calm and possibly content.
3. The speaker's eyes are focused forward, suggesting that she is engaged in the conversation.
...

5. The speaker's facial expression suggests that she is in a relaxed and pleasant emotional state.
- Fact Part: The speaker's facial expression is clearly visible.
- Inference Part: The expression suggests that she is in a relaxed and pleasant emotional state.

Example text 2:

The speaker in the video appears to be feeling joy. Her facial expression is one of ...
3. The speaker seems to be enjoying the conversation and the moment, which indicates a positive and joyful emotion.
- Fact Part: The speaker is participating in a conversation.
- Inference Part: The speaker is enjoying the conversation and the moment, indicating a positive and joyful emotion.

Your text:

text

Your response:

Figure 20: LLM Prompt to break down video descriptions.

Task Overview
You will be given the text of a conversation. Your task is to predict the top emotion of the speaker of the last sentence. The possible emotions are [joy, surprise, anger, fear, disgust, sadness]. Follow the Task Guidelines and the Response Format.

Task Guidelines
- Think out loud about the possible options [joy, surprise, anger, fear, disgust, sadness] using the text. For each emotion, think about whether the last sentence could be an expression of that emotion.
- Finally, output the top emotion according to the Response Format.

Response Format
Thinking out loud: My only allowed emotions are [joy, surprise, anger, fear, disgust, sadness]. Based on the text, I think... <your thoughts>. Therefore... <your choice>.
Emotion: <your top emotion>

Conversation
conversation

Response

Figure 21: LLM Prompt to classify emotion with Text only.

Task Overview
You will be given the text of a conversation and some visual cues about the main speaker of the last sentence. Your task is to predict the top emotion of the speaker of the last sentence. The possible emotions are [joy, surprise, anger, fear, disgust, sadness]. Follow the Task Guidelines and the Response Format.

Task Guidelines
- Think out loud about the possible options [joy, surprise, anger, fear, disgust, sadness] using the text and then the visual cues. Carefully consider if each emotion could apply. - You must pick an emotion.
- Finally, output the top emotion according to the Response Format.

Response Format
Thinking out loud: My only allowed emotions are [joy, surprise, anger, fear, disgust, sadness]. Based on the text I think... <your thoughts>. Now, based on the visual cues, I think... <your thoughts>. Therefore... <your choice>.
Emotion: <your top emotion>

Conversation
conversation
Visual cues
cues

Response

Figure 22: LLM Prompt to classify emotion with Text + Vision.

Task Overview
You will be given some visual cues about a speaker extracted from a video clip. Your task is to predict the top emotion of the speaker. The possible emotions are [joy, surprise, anger, fear, disgust, sadness]. Follow the Task Guidelines and the Response Format.

Task Guidelines
- Think out loud about the possible options [joy, surprise, anger, fear, disgust, sadness] using the visual cues. Carefully consider if each emotion could apply.
- You must pick an emotion.
- Finally, output the top emotion according to the Response Format.

Response Format
Thinking out loud: My only allowed emotions are [joy, surprise, anger, fear, disgust, sadness]. Based on the visual cues, I think... <your thoughts>. Therefore... <your choice>. Emotion: <your top emotion>

Visual cues
cues

Response

Figure 23: LLM Prompt to classify emotion with Vision.