Beyond Accuracy: A Diagnostic Protocol for Fairly Evaluating Multimodal Reasoning

Shohreh Ghorbani[1*] **Chenyu Zhang**[2] **Minsol Kim**[1] **Jingyao Wu**[1]
Rosalind Picard[1] **Patricia Maes**[1] **Paul Pu Liang**[1]
[1]MIT Media Lab [2]Harvard University

Abstract

Current benchmarks for Multimodal Large Language Models (MLLMs) rely on single-accuracy scores, a metric that is fundamentally flawed for subjective tasks like emotion recognition. This paradigm creates an "Intelligence-Accuracy Paradox," where models with sophisticated reasoning about ambiguous human communication are penalized for not conforming to a single, oversimplified ground-truth label, while less intelligent models that exploit dataset biases can achieve higher scores. This paper argues that high accuracy often masks a "hidden failure" on complex, ambiguous instances. To address this, we propose a new, two-stage protocol that is both diagnostic and evaluative. Stage 1 acts as a diagnostic, using **a phenomenon we term Dominant Modality Override (DMO)**, where one modality's high-confidence signal hijacks the final decision to automatically partition a dataset into unambiguous and conflict-rich samples. This diagnosis enables Stage 2, a fairer evaluation where these partitions are assessed differently: unambiguous samples are scored on accuracy, while conflict-rich samples are evaluated on the quality of their reasoning using metrics like clue-based faithfulness and set-based plausibility. This protocol provides a fairer, more faithful "report card" of a model's true capabilities, rewarding intelligent reasoning over brittle pattern matching.

1 Introduction

Human perception and decision-making are often intrinsically ambiguous [1–3]. In tasks such as emotion recognition, intent inference, or multimodal understanding, there is rarely a single definitive ground truth. Conventional evaluation practices, which collapse multiple perspectives into a single accuracy score, therefore fail to reflect this complexity.

Recent work has begun to address this gap by explicitly modeling ambiguity [4–7]. For example, distributional representations are adopted to treat annotator disagreement not as noise but as informative signal, capturing the range of plausible interpretations rather than reducing them to a single label [8–11]. Alongside these advances, new metrics have been proposed to evaluate distributional predictions, moving beyond accuracy to quantify calibration, reliability, and alignment with human variability [12, 13].

Yet a critical dimension remains underexplored: reasoning under ambiguity[14]. While distributional models and metrics acknowledge the existence of multiple plausible outcomes, they do not capture how models reason through conflicting or incomplete evidence to arrive at those outcomes. The challenge of resolving conflicting signals is a foundational problem in multimodal learning, often addressed through advanced fusion mechanisms or modality-weighting strategies [15, 16]. However,

*Corresponding author: ghorbani@mit.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Evaluating the Evolving LLM Lifecycle.

this omission is especially consequential for multimodal large language models (LLMs). Current benchmarks often reward shallow pattern matching, favoring models that ignore conflicts and align with biased labels, while penalizing models that attempt nuanced reasoning and generate responses that diverge from oversimplified ground truth.

This leads to a more profound problem: the **Intelligence-Accuracy Paradox**. A model may appear robust by correctly handling simple, unambiguous cases while failing systematically on complex ones. More capable models that detect subtle contradictions across modalities (e.g., sarcastic prosody versus positive text) risk being judged "wrong" against reductive benchmarks, while simpler models relying on biased heuristics may score higher.

In this paper, we address this gap and propose a novel two-stage protocol that uses a diagnostic signal to enable a fairer, reasoning-focused evaluation of LLMs in ambiguous multimodal contexts.

2 A Diagnostic for Uncovering Benchmark-Model Misalignment

Current evaluation practices in multimodal learning suffer from a critical flaw: they conflate a model's performance on simple, unambiguous tasks with its ability to handle complex, ambiguous ones into a single accuracy score. This practice allows a model to achieve a high score by excelling at easy examples while systematically failing on nuanced cases where inter-modal conflict is present a "hidden failure" that masks a lack of true reasoning capability. To enable a fairer evaluation, we must first disentangle these distinct problem types. We require a diagnostic tool that can automatically identify and partition samples based on the level of conflict between their modalities. To this end, we introduce a diagnostic flag for a specific type of signal divergence we term Dominant Modality Override (DMO). We define this as the event where a high-confidence prediction from a single modality (e.g., audio) is incongruent with the consensus ground-truth label and successfully drives the final fused prediction. This concept builds on our preliminary investigation of such events, which we previously termed 'Modality Sabotage' in our work submitted to the MAR 2025 workshop [17]. The DMO flag is not a model flaw, but a signal for a sample where the benchmark's single label is in high tension with a powerful unimodal signal.DMO is also a characteristic of the dataset's complexity. Our analysis on emotion recognition benchmarks reveals that these misalignments are not rare edge cases, but a common and systemic feature of multimodal communication. For example, in the MELD dataset, a dominant audio signal was the primary driver of a benchmark-incongruent prediction in **48.2%**of high-confidence error cases [17]. On IEMOCAP, this figure rose to **59.7%** of such cases [17].

The commonness of these events reveals why a single accuracy score is so misleading. It forces a critical question: how do high-accuracy models succeed on datasets where nearly half the challenging cases contain strong inter-modal conflict? They can do so in one of two ways. The **Brittle Pattern-Matcher** succeeds by learning a biased heuristic (e.g., "always trust the text"), ignoring the conflict to align with the benchmark's label. In contrast, the **Robust Reasoner** succeeds by correctly interpreting the conflict itself[18, 19]. A single accuracy score cannot distinguish between these two, it rewards the right answer, regardless of whether it was reached for the wrong or right reason. The purpose of our diagnostic is to isolate these common, conflict-rich cases, not to penalize a modality, but to subject the model to a more rigorous test that can finally separate shallow mimicry from intelligent reasoning. The following protocol leverages this diagnostic to achieve that goal.

3 A Two-Stage Protocol for Fairer Evaluation

We propose a two-stage evaluation protocol that leverages the DMO diagnostic to move beyond single-number accuracy and toward a richer, reasoning-sensitive assessment.

3.1 Stage 1: Detect and Partition

For a given test set, we apply the DMO diagnostic to each sample, thereby partitioning the dataset into two subsets. The first consists of *Unambiguous samples*, where modalities are largely consistent and accuracy remains an informative measure. The second consists of *conflict-rich samples*, where DMO flag is detected and a single ground-truth label is insufficient. Accuracy loses validity in this latter subset, and a deeper evaluation of reasoning becomes essential. We explicitly acknowledge

that using single-label accuracy for the unambiguous subset may seem contradictory to our initial critique. However, we argue that this metric's primary flaw, its inability to handle ambiguity, is largely irrelevant for these specific cases where inter-modal signals are coherent and align with a high-agreement consensus label. Indeed, recent work highlights that aligning with human annotator consensus, rather than a single label, is a more robust way to evaluate subjective tasks [20]. For these "easy questions," accuracy remains a pragmatic and efficient proxy for a model's performance on straightforward tasks. Our protocol's core contribution is isolating the conflict-rich cases where this metric's validity completely breaks down, and replacing it with the more appropriate, reasoning-focused evaluation described in Stage 2.

3.2 Stage 2: Interpret Reasoning in Conflict-Rich Cases

Within the conflict-rich subset, the focus of evaluation shifts from outcomes to processes. Rather than asking whether the model matched a single benchmark label, we ask whether its reasoning is faithful, plausible, and nuanced. This stage is operationalized through three components:

For the conflict-rich partition, we shift the evaluation from the *outcome* (the label) to the *process* (the reasoning). We assess the model's reasoning quality using a suite of advanced metrics inspired by recent literature.

3.2.1 Reasoning Extraction

First, similar to approaches in speech emotion captioning [21], we prompt the LLM to externalize its reasoning process. The goal is to make the model articulate its understanding of unimodal signals and explain how it handles any disagreement between them. For instance, we can use a prompt that encourages this synthesis: "Analyze the following multimodal inputs. First, describe the emotional cue suggested by the text modality. Second, describe the emotional cue suggested by the audio/visual modalities. Third, state whether these cues are in conflict. Finally, provide a synthesized final decision and a rationale that explains how you weighed the evidence."

3.2.2 Evaluating Reasoning Faithfulness (Clue/Label Overlap)

We cannot trust a model's reasoning if it is not grounded in evidence. Drawing from foundational work in explainable emotion recognition [22],We evaluate whether the clues cited in the model's rationale are **factually faithful** to the multimodal input. The metric asks two questions: first, does the model provide explanatory clues (e.g., 'tense prosody'), and second, are those clues actually present in the source audio and video? This distinguishes genuine reasoning from plausible-sounding hallucination, ensuring that explanations have a 'certain basis'.

3.2.3 Evaluating Plausibility and Nuance (Set-Based Metrics)

Recognizing that ambiguity invites multiple valid interpretations, we discard the single ground-truth label for conflict-rich cases. Instead, evaluation is performed against a pre-defined **set of plausible labels**. This approach, pioneered in open-vocabulary multimodal emotion recognition [23], uses set-based precision and recall, uses set-based precision and recall. For a sarcastic utterance, the plausible set might be 'Anger, Sarcasm, Neutral'. A model predicting any label in this set receives credit, rewarding it for identifying a legitimate, alternative human interpretation that the original single label ignored.

3.3 The Output: A New Evaluation Report Card

The final output of this our two-stage protocol is not a single, misleading number, but a rich, diagnostic report card that reveals a model's true character. Table 1 illustrates this transformation by comparing the old evaluation paradigm with the output of our proposed protocol. The scores shown are hypothetical but realistic examples designed to demonstrate the insights our method provides. In a real experiment, these metrics would be calculated as follows: **Unambiguous and Conflict-Rich Accuracy**: Standard accuracy calculated separately on the two partitions created in Stage 1. **Reasoning Faithfulness**: For the conflict-rich partition, this is the percentage of samples where the model's rationale correctly cites observable evidence (clues) from the multimodal input, as evaluated in Stage 2.2. **Plausibility (Set-Recall)**: For the conflict-rich partition, this is the set-recall

score calculated against a pre-defined set of plausible labels, as described in Stage 2.3. This metric quantifies how often the model's "wrong" answers are, in fact, legitimate interpretations.

Table 1: Comparison of legacy accuracy-only evaluation versus the proposed multi-faceted report card.

Old Evaluation		
Metric	**Score**	**Interpretation**
Model Accuracy	85%	Looks good, but hides a critical flaw.
New, Fairer Evaluation Report Card		
Metric	**Score**	**Interpretation**
Overall Accuracy	85%	(Legacy score for comparison.)
Unambiguous Accuracy	98%	**Strength:** The model excels at identifying clear emotional cues.
Conflict-Rich Accuracy	15%	**Weakness:** The model fails when faced with conflicting signals.
Reasoning Faithfulness	82%	**Insight:** When it fails, the model still grounds its reasoning in real evidence.
Plausibility (Set-Recall)	75%	**Insight:** Many "wrong" answers are plausible alternative interpretations.

4 Conclusion

The pursuit of higher accuracy scores has led us to a paradox where we may inadvertently favor less intelligent models. A single number cannot capture the complexity of multimodal reasoning. The protocol proposed here offers a concrete path forward. By first using **Modality Sabotage** to diagnose and isolate ambiguity, and then shifting our evaluation to assess the **quality of reasoning** on these complex cases, we can build and deploy models that are not only accurate on the easy problems but also robust, plausible, and trustworthy when faced with the ambiguity of the real world. While this paper has focused on proposing and justifying the evaluation conceptually, our immediate future work involves a comprehensive empirical validation. By applying our method to standard benchmarks like IEMOCAP, MELD etc. we plan to quantify the "hidden failure" of high-accuracy models and demonstrate the utility of our report card in identifying truly robust reasoners.

References

[1] Maria Parmley and Joseph G Cunningham. She looks sad, but he looks mad: The effects of age, gender, and ambiguity on emotion perception. *The Journal of social psychology*, 154(4): 323–338, 2014.

[2] Andrew Mathews. Effects of modifying the interpretation of emotional ambiguity. *Journal of Cognitive Psychology*, 24(1):92–105, 2012.

[3] Emily Mower, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Interpreting ambiguous emotional expressions. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–8. IEEE, 2009.

[4] Ya-Tse Wu, Jingyao Wu, Vidhyasaharan Sethu, and Chi-Chun Lee. Can Modelling Inter-Rater Ambiguity Lead To Noise-Robust Continuous Emotion Predictions? In *Interspeech 2024*, pages 3714–3718, 2024. doi: 10.21437/Interspeech.2024-482.

[5] Yi-Cheng Lin, Haibin Wu, Huang-Cheng Chou, Chi-Chun Lee, and Hung-yi Lee. Emobias: A large scale evaluation of social bias on speech emotion recognition. *arXiv preprint arXiv:2406.05065*, 2024.

[6] Huang-Cheng Chou and Chi-Chun Lee. Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5886–5890. IEEE, 2019.

[7] Wen Wu, Bo Li, Chao Zhang, Chung-Cheng Chiu, Qiujia Li, Junwen Bai, Tara N Sainath, and Philip C Woodland. Handling ambiguity in emotion: From out-of-domain detection to distribution estimation. *arXiv preprint arXiv:2402.12862*, 2024.

[8] Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. Dual-Constrained Dynamical Neural ODEs for Ambiguity-aware Continuous Emotion Prediction. In *Interspeech 2024*, pages 3185–3189, 2024. doi: 10.21437/Interspeech.2024-119.

[9] Deboshree Bose, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. Parametric distributions to model numerical emotion labels. In *Interspeech*, pages 4498–4502, 2021.

[10] T Mani Kumar, Enrique Sanchez, Georgios Tzimiropoulos, Timo Giesbrecht, and Michel Valstar. Stochastic process regression for cross-cultural speech emotion recognition. *Proc. Interspeech 2021*, pages 3390–3394, 2021.

[11] Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. Emotion recognition systems must embrace ambiguity. In *2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 166–170. IEEE, 2024.

[12] Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. Belief mismatch coefficient (bmc): A novel interpretable measure of prediction accuracy for ambiguous emotion states. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2023.

[13] Tejas Siledar, Haw-Shiuan He, Ananya Bhowmick, and Kathleen McKeown. One Prompt to Rule Them All: LLMs for Opinion Summary Evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

[14] Jingwen Liu, Kan Jen Cheng, Jiachen Lian, Akshay Anand, Rishi Jain, Faith Qiao, and Robin Netzorg. Emo-reasoning: Benchmarking emotional reasoning capabilities in spoken dialogue systems. *arXiv preprint arXiv:2405.12345*, 2024.

[15] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Makhzani, Yacine Mroueh, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, 2019.

[16] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913, 2017.

[17] Chenyu Zhang, Minsol Kim, Shohreh Ghorbani, Jingyao Wu, Rosalind Picard, Patricia Maes, and Paul Pu Liang. When one modality sabotages the others: A diagnostic lens on multimodal reasoning, 2025. URL `https://arxiv.org/abs/2511.02794`.

[18] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. *arXiv preprint arXiv:2405.08379*, 2024.

[19] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models. *arXiv preprint arXiv:2402.15248*, 2024.

[20] Ljubiša Bojić, Olga Zagovora, Asta Zelenkauskaite, Vuk Vuković, Milan Čabarkapa, Selma Veseljević Jerković, and Ana Jovančević. Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific Reports*, 2025.

[21] Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shixiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. SECap: Speech Emotion Captioning with Large Language Model. *arXiv preprint arXiv:2309.01257*, 2023.

[22] Zheng Lian, Haiyang Sun, Licai Sun, Hao Gu, Zhuofan Wen, Siyuan Zhang, Shun Chen, Mingyu Xu, Ke Xu, Kang Chen, et al. Explainable multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6814–6823, 2023.

[23] Zheng Lian, Haiyang Sun, Licai Sun, Lan Chen, Haoyu Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, et al. Open-vocabulary multimodal emotion recognition: Dataset, metric, and benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15104–15114, 2023.