Beyond Accuracy: A Diagnostic Protocol for Fairly Evaluating Multimodal Reasoning

Anonymous Author(s)
Affiliation
Address
email

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, 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

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

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. 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 an old, accuracy-only evaluation versus the proposed multi-faceted report card, which provides a fairer and more actionable assessment of model capabilities.

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 is still good at grounding its reasoning in real evidence.
Plausibility (Set-Recall)	75%	**Insight:** The model's "wrong" answers are often 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. Emo-bias: 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] Anonymous Author(s). When one modality sabotages the others: A diagnostic lens on multimodal reasoning. In *Submitted to the Multimodal Algorithmic Reasoning Workshop at NeurIPS*, 2025. Anonymized for review.

[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.

A Supplementary Material: Full MAR Workshop 2025 Paper

This appendix contains the full, anonymized text of our paper submitted to the Multimodal Algorithmic Reasoning (MAR) Workshop at NeurIPS 2025, from which the empirical results in Section 2 are cited.

The key findings referenced in the main paper can be found in the following locations within this supplementary document:

- **Methodology:** The full methodology for detecting Dominant Modality Override (previously termed 'Modality Sabotage') is detailed in Section 2 of the supplementary paper.
- **Empirical Results:** The specific statistics for the MELD (48.2%) and IEMOCAP (59.7%) datasets are presented in Section 3 and summarized in Table 1 of the supplementary paper.

When One Modality Sabotages the Others: A Diagnostic Lens on Multimodal Reasoning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Despite rapid growth in multimodal foundation models, their reasoning traces remain opaque: it is often unclear which modality drives a prediction, how conflicts are resolved, or when one stream dominates. In this paper, we introduce *modality sabotage*—a failure mode in which a high-confidence unimodal error overrides other evidence and misleads the fused result. To analyze such dynamics, we propose a lightweight, model-agnostic evaluation layer that treats each modality as an agent, producing candidate labels and a brief self-assessment. A simple fusion mechanism aggregates these outputs, exposing *contributors* (modalities supporting correct outcomes) and *saboteurs* (modalities that mislead). Applying our diagnostic layer in a case study on multimodal emotion recognition benchmarks with foundation models revealed systematic reliability profiles, providing insight into whether failures may arise from dataset artifacts or model limitations. More broadly, our framework offers a diagnostic scaffold for multimodal reasoning, supporting principled auditing of fusion dynamics and informing possible interventions.

1 Introduction

Multimodal foundation models have advanced rapidly in tasks that combine vision, language, and audio, from question answering tasks to understanding social signals [1]. Yet in practice, their decisions remain a black box: users cannot tell *which* stream of data the system relied on, *how* conflicting evidence–e.g., when text, audio, and vision suggest different labels–was resolved, if at all, or whether a single sensor dominated the outcome. Prior work has discussed related issues such as *modality collapse*, where vision–language models over-rely on text [2], and *unimodal bias*, where fusion lets one stream dominate across a dataset [3]. In contrast, we highlight a distinct failure mode we call *modality sabotage*: instance-level cases where a high-confidence unimodal error not only fails locally but actively overrides other evidence and pulls the fused prediction off-target. Unlike collapse or bias, which describe systematic trends, sabotage is a diagnostic lens on individual decisions, making visible which modality misled the model and when. Despite strong progress in multimodal fusion [4–14] and impressive results from multimodal large language models (MLLMs) in vision–language understanding, VQA, and video understanding [15–20], current systems mostly emphasize cross-modal feature interaction and modality completion, leaving how cues map to constructs and how conflicts are resolved largely unexplored. Decades of psychology and affective computing show that audio and visual cues carry complementary emotional information [21–23], for example facial expressions correlate with pleasant affect [24] while speech acoustics track arousal [25–27]. Yet these studies typically isolate unimodal contributions rather than addressing how models should integrate, arbitrate, or dominate across modalities in multimodal settings. We address this need with a simple, transparent, model-agnostic framework that treats each modality as an agent, whose outputs constitute a diagnostic layer that records per-modality votes, confidences, and disagreements, enabling systematic analysis of contributions and failure modes before a final fused decision is

Submitted to the Multimodal Algorithmic Reasoning Workshop at NeurIPS (MAR 2025). Do not distribute.

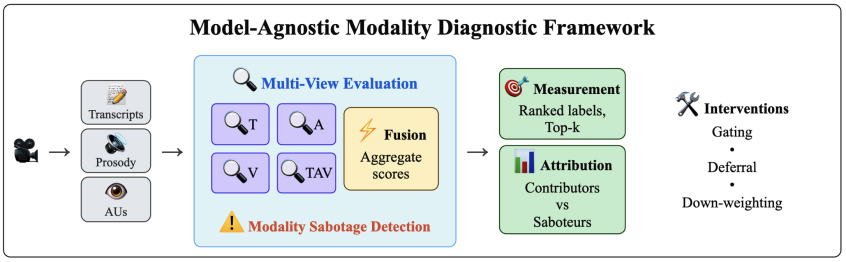

Figure 1: Each modality (T, A, V) and a joint view (TAV) agent outputs classification labels with confidence. A simple fusion aggregates these into a ranked prediction, enabling attribution of *contributors* vs. *saboteurs*. The callout highlights high-confidence unimodal errors that mislead the fused decision (*modality sabotage*); see Section 2 for details.

made. Specifically, we propose a plug-and-play modality-as-agent fusion that queries text (T), audio (A), vision (V), and their joint view (TAV) separately, then aggregates their predictions into a final decision. The design makes attribution explicit at the instance level, surfacing *contributors* (modalities supporting correct answers) and *saboteurs* (modalities that mislead).

Our contributions are threefold: (i) a lightweight framework that yields instance-level attribution without retraining or architectural changes; (ii) a measurable operationalization of modality sabotage for high-confidence but misleading unimodal outputs; and (iii) dataset- and backbone-dependent reliability profiles that clarify whether failures stem from dataset artifacts or model limitations.

2 Methodology

We evaluate the framework across three public and widely-used multimodal emotion datasets and report unimodal and fused performance, top-k coverage, and sabotage diagnostics.

Inputs per modality. For each video segment we derive modality–specific, purely descriptive inputs that avoid direct emotion inference: (i) **Text (T):** Whisper ASR transcripts from the audio track serve as the textual input; (ii) **Audio (A):** Each audio utterance instance is analyzed by *Qwen-Audio* [28] with a structured prompt to elicit non-lexical descriptors—prosody (pitch/intonation, loudness/intensity, tempo/rhythm), voice quality (breathiness/creak/tension), and articulation—while forbidding use of lexical content or emotion labels; (iii) **Vision (V):** we compute facial AUs with *OpenFace* [29], select an AU-peak frame, and ask a VLM (*GPT-4 Vision* [30]) to produce an objective caption of observable cues (e.g. facial expressions, posture, gestures, and context) without mental-state attributions. These modality-specific descriptors feed the corresponding modality agents.

Agents and outputs. We propose a simple, model-agnostic framework (Figure 1) that treats each modality as an *agent* and makes fusion decisions legible. For each sample, T, A, V, and TAV are queried with a structured prompt; each agent returns a sorted set of candidate labels with confidence scores (1–100) together with a *data-quality report* (score, issues, and a short rationale). The confidences are sparse and uncalibrated (an agent may emit only a few labels), and the data-quality field is used to probe whether the LLM can self-diagnose issues such as noisy transcripts or occluded faces. We fuse agents by aggregating their confidences per label and normalizing to obtain a single ranked score vector. Let $S_m(y) \in [0, 100]$ be the confidence assigned by agent $m \in \{\mathrm{T}, \mathrm{A}, \mathrm{V}, \mathrm{TAV}\}$ to label y (zero if y is not proposed), and let $q_m \in [0, 1]$ denote the agent's self-reported quality score (rescaled from 1–100). We compute

$$\tilde{s}(y) \;=\; \sum_m w_m \, S_m(y), \qquad p(y) \;=\; \frac{\tilde{s}(y)}{\sum_{y'} \tilde{s}(y')},$$

where $w_m = 1$ by default and $w_m = q_m$ in a quality-weighted ablation. Across benchmarks, quality weighting did not improve top-1 accuracy (and sometimes reduced it), so we retain the unweighted variant as the main setting and report the weighted variant for completeness. We evaluate using the ranking induced by $p(y)$ and report top-k coverage.

Modality sabotage (diagnostic). Fusion can fail *silently* when an overconfident stream dominates: a wrong modality can pull the final decision off-target, and accuracy alone offers no attribution.

Let $S_m(y) \geq 0$ denote agent m's evidence for label y (we use self-reported confidence 1–100; other signals such as probabilities or logits are interchangeable), $p_m(y) = S_m(y)/\sum_{y'} S_m(y')$, $y_m = \arg\max_y p_m(y)$, $c_m = \max_y p_m(y)$, and $\hat{y} = \arg\max_y \tilde{s}(y)$ with $\tilde{s}(y) = \sum_m w_m S_m(y)$. We distinguish two flavors: **Potential sabotage** for m holds when (i) $c_m \geq \tau$ (high confidence) and (ii) $y_m \neq y^*$ (its own error). **Successful sabotage** strengthens this by requiring (iii) $\hat{y} = y_m$ (the fused model follows m), with $\tau = 0.70$ unless noted. However, due to the nature of fusion, successful sabotage does not establish strict causality—multiple agents may jointly support the same wrong label. For this reason, in Section 3 we focus on **potential sabotage**, which provides a clearer upper bound on each modality's tendency toward overconfident errors. Both definitions nonetheless offer actionable diagnostic signals for gating, down-weighting, or deferral.

Top-k reasoning. Modality sabotage motivates *Acc@k*: in many sabotaged cases the correct label remains in the top-k list even when Top-1 is wrong, indicating recoverable uncertainty rather than pure failure. Concretely, we sort the fused scores $p(y)$ and report coverage at k,

$$\mathrm{Acc@}k \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[y_i^* \in \mathrm{TopK}\big(p_i\big)\big].$$

We report Acc@k alongside Top-1 across datasets and backbones. These statistics do not alter the decision rule; they expose whether the diagnostic layer preserves sufficient signal under modality conflict to support reliable interventions, while being robust to the sparsity and lack of calibration in agent confidences.

3 Case Study Results

3.1 Aggregate accuracy and effect of self-reported quality

Table 1 compares a single-call **TAV** baseline (Top-1) with our **agentic fusion** (reported as "Fus T1–T5" in the table) under confidence-only fusion and reports the ablation when additionally weighting by self-reported data quality. Three patterns emerge. *(i) Top-*1 vs. *Top-k*. The fusion maintains baseline-level Top-1 on **MELD** and **IEMOCAP** and improves markedly on MELD, while Top-k coverage rises steeply across datasets. On **MER**, Top-1 changes from 0.38 (baseline) to 0.33 (fusion, GPT-5-nano), but the correct label appears with high probability in the ranking (Top-5 $= 0.97$). On **MELD**, Top-1 improves by $+0.09$ for GPT-5-nano ($0.27 \rightarrow 0.36$, $+33\%$) and by $+0.15$ for GPT-4-mini ($0.30 \rightarrow 0.45$), with Top-5 $= 0.92/0.90$. On **IEMOCAP**, Top-1 is essentially flat for GPT-5-nano ($0.28 \rightarrow 0.29$) and slightly lower for GPT-4-mini ($0.28 \rightarrow 0.24$), but Top-5 remains substantially higher than Top-1 (GPT-5-nano: 0.76, GPT-4-mini: 0.72). These results indicate that the fusion preserves multiple plausible hypotheses beyond the Top-1 prediction.

Ablation: confidence $\times$ quality weighting. The Δ columns quantify the change when scaling each vote by the product of its confidence and self-reported data-quality. Effects are small and often negative: e.g., on **MELD**/GPT-5-nano, ΔTop-1$= -0.08$ and ΔTop-2$= -0.06$; on **IEMOCAP**/GPT-5-nano, ΔTop-1$= -0.05$ and ΔTop-3$= -0.07$. Occasional mild gains appear (e.g., **MER**/GPT-4-mini: ΔTop-4$= +0.02$, ΔTop-5$= +0.03$), but the overall trend suggests that *self-reported quality* is a noisy proxy that does not consistently align with correctness. Consequently, we report confidence-only fusion as the main setting and treat quality weighting as an ablation.

3.2 Modality behavior and sabotage analysis

We operationalize *modality sabotage* as a measurable, instance-level diagnostic for *high-confidence, misleading* unimodal outputs that dominate the fusion and derail the final decision. This test makes the notion of "pulling the decision away" explicit, yields a *countable event* per example, and supports auditing by answering *who contributed* or *who hurt* each prediction. Figure 2 visualizes unimodal accuracy and sabotage rates per modality for GPT-5-nano under confidence-weighted fusion.

Diagnostic signals revealed. Beyond aggregate rates, the sabotage test surfaces actionable signals at the instance level as reported in Figure 2: (i)Across the columns, we observe a per-modality *calibration gap* (self-reported confidence vs. empirical accuracy), and (ii)Across the rows, we report the *dataset/backbone reliability profiles* ranking modalities by accuracy and sabotage. *Comparing the columns* in Figure 2 (Left), patterns are consistent: audio is the primary saboteur and text most

Table 1: **Top-k coverage and diagnostic effect of quality weighting.** The fusion maintains baseline-level Top-1 accuracy ("Fus T1" vs. "Base T1") while substantially improving Top-k coverage ("Fus T2–T5"). The Δ block reports the change when switching from *confidence-only weighting* to *confidence* $\times$ *data-quality weighting*. Comparisons across datasets and backbones (GPT-5-nano vs. GPT-4o-mini) highlight systematic differences in modality reliability and pipeline robustness.

Dataset / Model	Accuracy						Δ (confidence+quality vs. confidence-only)				
	Base T1	Fus T1	Fus T2	Fus T3	Fus T4	Fus T5	ΔT1	ΔT2	ΔT3	ΔT4	ΔT5
MER / GPT-5-nano	0.38	0.33	0.62	0.85	0.92	0.97	+0.00	+0.01	+0.00	-0.02	+0.01
MER / GPT-4o-mini	0.35	0.23	0.52	0.75	0.83	0.85	-0.03	+0.00	+0.00	+0.02	+0.03
MELD / GPT-5-nano	0.27	0.36	0.58	0.73	0.86	0.92	-0.08	-0.06	-0.03	-0.03	-0.04
MELD / GPT-4o-mini	0.30	0.45	0.64	0.76	0.85	0.90	-0.02	+0.01	-0.02	-0.02	-0.02
IEMOCAP / GPT-5-nano	0.28	0.29	0.47	0.62	0.73	0.76	-0.05	-0.07	-0.07	-0.02	+0.03
IEMOCAP / GPT-4o-mini	0.28	0.24	0.43	0.60	0.70	0.72	+0.01	+0.03	+0.00	-0.02	+0.00

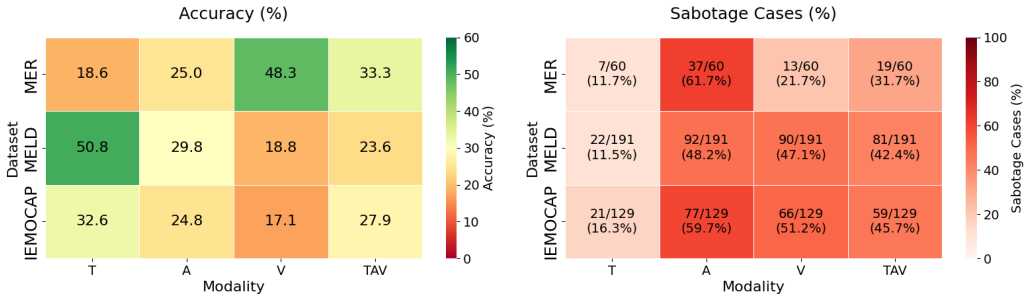

Figure 2: *Left heatmap*: unimodal accuracy for Text (T), Audio (A), Vision (V), and joint view (TAV), highlighting differences across datasets. *Right heatmap*: proportion of cases where a modality *sabotages* the fused decision (high-confidence error flipping Top-1 at threshold 70), where each values show #cases/total (rate%).

contributed. This provides a basis for identifying which components of a model pipeline may require refinement. *Across rows*, we can evaluate which modalities are less reliable within each dataset. This is consistent with each dataset characteristics: **MER** suffers from noisy ASR/translation but benefits from rich video cues; **MELD**'s sitcom-style video with exaggerated cues or multiple actors can mislead vision; **IEMOCAP** features seated dyads, where acted expressions and experimental scenes limit visual reliability.

Proposed Uses of Modality Sabotage Diagnostics Once modality sabotage events are identified, their associated confidence patterns can be leveraged not only for auditing but also for improving reliability. For example, sabotage scores can guide retraining with harder negatives, reweighting modalities that are systematically overconfident, or gating streams that frequently sabotage decisions. Beyond training-time interventions, sabotage scores can be incorporated directly at inference as weights on the final decision layer, adjusting the aggregated ranking to down-weight suspect modalities before producing the output. In this way, the diagnostic layer becomes actionable, informing both model development and real-time decision policies.

4 Conclusions

We introduced a lightweight, model-agnostic modality-as-agent framework that makes multimodal fusion decisions interpretable at the instance level. Each modality produces candidate labels with confidences, and a simple fusion aggregates them into a ranked prediction, making visible both supportive and misleading influences. By operationalizing modality sabotage as a measurable diagnostic, the framework provides concrete signals for auditing fusion dynamics, detecting overconfident errors, and guiding interventions such as gating or down-weighting unreliable streams. In this paper, we demonstrated the feasibility of the approach on multimodal emotion recognition. Looking forward, we position this diagnostic as a scaffold for multimodal algorithmic reasoning and broader agentic AI, offering a lightweight tool to expose shortcomings, handle unreliable sources, and build more interpretable multimodal systems.

References

[1] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56 (10):1–42, 2024.

[2] 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.

[3] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Devi Parikh. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1989–1998, 2019.

[4] Zebang Cheng, Yuxiang Lin, Zhaoru Chen, Xiang Li, Shuyi Mao, Fan Zhang, Daijun Ding, Bowen Zhang, and Xiaojiang Peng. Semi-supervised multimodal emotion recognition with expression mae. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9436–9440, 2023.

[5] Zebang Cheng, Fuqiang Niu, Yuxiang Lin, Zhi-Qi Cheng, Bowen Zhang, and Xiaojiang Peng. Mips at semeval-2024 task 3: Multimodal emotion-cause pair extraction in conversations with multimodal language models. *arXiv preprint arXiv:2404.00511*, 2024.

[6] Zhi-Qi Cheng, Qi Dai, Siyao Li, Teruko Mitamura, and Alexander Hauptmann. Gsrformer: Grounded situation recognition transformer with alternate semantic attention refinement. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3272–3281, 2022.

[7] Zhi-Qi Cheng, Xiang Li, Jun-Yan He, Junyao Chen, Xiaomao Fan, Xiaojiang Peng, and Alexander G Hauptmann. Umetts: A unified framework for emotional text-to-speech synthesis with multimodal prompts. *arXiv preprint arXiv:2404.18398*, 2024.

[8] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6631–6640, 2023.

[9] Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9610–9614, 2023.

[10] Nicolas Richet, Soufiane Belharbi, Haseeb Aslam, Meike Emilie Schadt, Manuela González-González, Gustave Cortal, Alessandro Lameiras Koerich, Marco Pedersoli, Alain Finkel, Simon Bacon, et al. Textualized and feature-based models for compound multimodal emotion recognition in the wild. In *European Conference on Computer Vision*, pages 60–78. Springer, 2024.

[11] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36:17117–17128, 2023.

[12] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[13] Sitao Zhang, Yimu Pan, and James Z Wang. Learning emotion representations from verbal and nonverbal communication. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18993–19004, 2023.

[14] Hengshun Zhou, Debin Meng, Yuanyuan Zhang, Xiaojiang Peng, Jun Du, Kai Wang, and Yu Qiao. Exploring emotion features and fusion strategies for audio-video emotion recognition. In *2019 International conference on multimodal interaction*, pages 562–566, 2019.

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

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[17] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[18] Yuxiang Guo, Faizan Siddiqui, Yang Zhao, Rama Chellappa, and Shao-Yuan Lo. Stimuvar: Spatiotemporal stimuli-aware video affective reasoning with multimodal large language models. *International Journal of Computer Vision*, pages 1–17, 2025.

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

[20] Shuyuan Tu, Qi Dai, Zuxuan Wu, Zhi-Qi Cheng, Han Hu, and Yu-Gang Jiang. Implicit temporal modeling with learnable alignment for video recognition. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 19936–19947, 2023.

[21] Rainer Banse and Klaus R Scherer. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614–636, 1996.

[22] Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. Multimodal affect models: An investigation of relative salience of audio and visual cues for emotion prediction. *Frontiers in Computer Science*, 3:767767, 2021.

[23] Liam Schoneveld, Alice Othmani, and Hazem Abdelkawy. Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognition Letters*, 146:1–7, 2021.

[24] Paul Ekman and Harriet Oster. Facial expressions of emotion. *Annual Review of Psychology*, 30:527–554, 1979.

[25] Jo-Anne Bachorowski. Vocal expression and perception of emotion. *Current Directions in Psychological Science*, 8(2):53–57, 1999.

[26] James A Russell, Jo-Anne Bachorowski, and José-Miguel Fernández-Dols. Facial and vocal expressions of emotion. *Annual Review of Psychology*, 54(1):329–349, 2003.

[27] Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. Dual-constrained dynamical neural odes for ambiguity-aware continuous emotion prediction. In *Proceedings of INTERSPEECH 2024*, pages 3185–3189. ISCA, 2024.

[28] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

[29] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE, 2016.

[30] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper's primary claim, stated in the Abstract and Introduction, is the proposal of a new two-stage protocol for fairer evaluation of multimodal LLMs. The paper's scope is clearly defined as a conceptual and methodological contribution, and it does not overstate its results. The "Intelligence-Accuracy Paradox" is introduced as the motivating problem that the protocol aims to address.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The Conclusion (Section 4, Lines 137-141) explicitly states that the paper's focus is conceptual and that a comprehensive empirical validation is a necessary next step for future work. This acknowledges the primary limitation that the protocol has not yet been applied in a large-scale experiment within this paper.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper is a conceptual and methodological proposal. It does not contain any mathematical theorems, theoretical results, or formal proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The main contribution is a protocol, not a specific experimental result. The protocol itself is described in sufficient detail in Section 3, including the two stages, the logic for partitioning data, and the specific types of metrics to be used in the reasoning evaluation. This provides a clear blueprint for another researcher to implement and apply the protocol. The empirical numbers cited are from a separate, referenced submission.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The paper does not present new experimental results. It is a conceptual work that proposes a new evaluation protocol. Therefore, there is no code or data to release.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: No experiments were conducted for this paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not contain empirical experiments, so measures of statistical significance are not applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: answerNA.

Justification: No experiments were conducted for this paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: This research aims to promote fairer, more transparent, and more robust evaluation of machine learning models, which directly aligns with the principles of responsible research outlined in the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: the primary impact of this work is intended to be positive (leading to less biased and more robust models), we acknowledge a potential negative societal impact in the Conclusion. A fairer evaluation protocol could be misused to "game" benchmarks in new ways if not implemented thoughtfully. For example, the "set of plausible labels" used in Stage 2 could be defined in a biased manner. Proper governance and community consensus on defining these plausible sets will be crucial for the protocol's responsible deployment. This paper aims to start that conversation.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: his paper does not release any models or datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites empirical findings derived from publicly available academic research datasets (IEMOCAP, MELD). These datasets are used in accordance with their terms for research purposes. All cited papers are listed in the References section.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: answerNA

Justification: paper does not release any new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No new research involving human subjects or crowdsourcing was conducted for this paper.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No new human subjects research was conducted for this paper, so IRB approval was not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This paper proposes a protocol for evaluating LLMs. While the protocol involves prompting LLMs to extract reasoning (as described in Section 3.2.1), no LLM was used as a core component in the development of the research methodology itself. The methodology is a human-designed conceptual framework.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.