Summary of Changes

Response to Reviews July 28th, 2025

We sincerely thank our reviewers for providing feedback that improved our paper. In this version, we address key concerns raised by the meta-reviewer and reviewers by adding additional experiments and elaborations.

Model Performances

We have added the mean ± standard deviation computed over five runs and explicitly noted this in the table caption.

Utterance Length & Context (Appendix F)

As suggested by Reviewer Xg3C, we include annotator disagreement results specifically for longer utterances. In Appendix F, we report Cohen's κ for clips containing full clauses and for clips ≥ 6 tokens (the median) and provide examples that showcase the differences between shorter and longer utterances. This filter reproduces the same results, showing that regardless of clip length, annotators still struggle specifically on examples that yielded modality disagreement.

Gating Behavior (Appendix A.2)

In response to Reviewer Xg3C's concern that independent gating might bias the fusion model toward a single modality, we report per-modality gate-weight distributions (mean ± std) showing that the model instead achieves balanced, sample-adaptive integration across text, audio, and video.

Broader Context (Appendix B.2)

In response to Reviewer L1W1 and the meta-reviewer's suggestion to better situate our findings, we have added a detailed discussion in Appendix B contrasting listener-centric empathy datasets (e.g., EmpatheticDialogues, OMG-Empathy) with speaker-centric settings, and explaining why our chosen corpus uniquely isolates expressed empathy signals. We also outline how our modality-disagreement diagnostic could be applied to listener-centric tasks to flag high-variance clips, thereby guiding more focused annotation and model development.

Video and Transcript Examples (Table 12)

As per Reviewer L1W1's suggestion, we added a table showcasing representative clips from each disagreement region (Red, Blue, Yellow, and Green) complete with still frames, transcripts, true labels, and annotator judgments.

This visualization highlights real instances where unimodal and multimodal predictions diverge.

General Changes
We streamlined the prose throughout, removed redundant background references, and tightened legends and callouts to improve overall clarity. In response to Reviewer 8Jj1's feedback, we also enhanced our figures to improve readability and clarity.

We believe these revisions substantially strengthen our empirical grounding, clarify our claims, and better situate our work. Once again, we thank all our reviewers for their time and thoughtful, detailed feedback.