Beyond Average Contribution: Class-Conditional Modality-Missingness Signatures Across Emotion Datasets, Models, and Humans

Anonymous authors
Paper under double-blind review

Abstract

Average performance differences after modality removal can conceal extensive changes in individual predictions. We propose a paired evaluation framework for multimodal emotion recognition combining the Decision Redistribution Index (DRI), the proportion of predictions that change, with class-conditional transition signatures that track rescue, harm, and wrong-to-wrong transitions. Our study spans six emotion datasets, open- and closed-source models, and 90 human participants. In the MELD subset of EmoS, video removal changes 17.2% of predictions while reducing accuracy by only 0.2 percentage points. Continuous ratings likewise exhibit opposing error changes that cancel in the mean. With variable linguistic content, text forms the most stable basis for judgment, while video removal repeatedly affects transitions between positive or high-arousal emotions and neutral. Audio supports arousal, intensity, and negative-emotion distinctions, but its correction directions vary across datasets, encoders, decoding rules, and observers; cross-dataset variation also appears in human judgments. Paired evaluation reveals the benefits and harms behind average gains and assesses whether modality effects recur across settings.

1 Introduction

Emotion recognition supports empathic feedback in digital companionship, educational interaction, health support, and embodied intelligence (Calvo & D'Mello, 2010; Paiva et al., 2017). Multimodal systems commonly combine text, audio, and video, but black-box predictions and existing fusion mechanisms rarely reveal directly how each modality affects specific samples and classes (Zadeh et al., 2017; Hu et al., 2022b). Existing work therefore still often measures the contribution of each modality with a single average: remove one modality, then compare accuracy, F1, or regression error (Hazarika et al., 2022). A near-zero aggregate effect has two possible sources. The modality may not participate in the judgment, or it may frequently change predictions while corrections and harms cancel after aggregation. In discrete tasks, the latter case may involve large-scale transitions of some emotions toward neutral; in continuous tasks, deactivation and overactivation can also produce similar means. Reporting only an average score cannot reveal which judgments are changed by modality missingness or where the errors move. Figure 1 summarizes two issues: (a) systematic paired full–missing comparisons on the same samples remain limited across datasets, models, and human observers; (b) zero net accuracy change can reflect unchanged predictions or conceal rescue–harm cancellation and wrong-to-wrong transitions amid extensive prediction changes.

Text, video, and audio provide different evidence for emotion judgments. In tasks where linguistic content varies across samples, text conveys events, appraisal targets, and emotion concepts, providing a semantic basis for class judgments. Psychological experiments have found that temporarily reducing access to emotion word meanings slows facial emotion judgments or reduces their accuracy (Lindquist et al., 2006). On IEMOCAP and MELD, MultiEMO's text-only baseline outperforms its audio-only and video-only baselines, with further improvements when other modalities are added (Shi & Huang, 2023). Our text-deletion, permutation, and representation analyses on tasks with variable linguistic content also support the role of text as a backbone for judgment (Appendix G). Video provides facial expressions and their changes, and dynamic presentation aids recognition of subtle expressions (Ambadar et al., 2005); the relative advantages of faces, semantics, and prosody

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

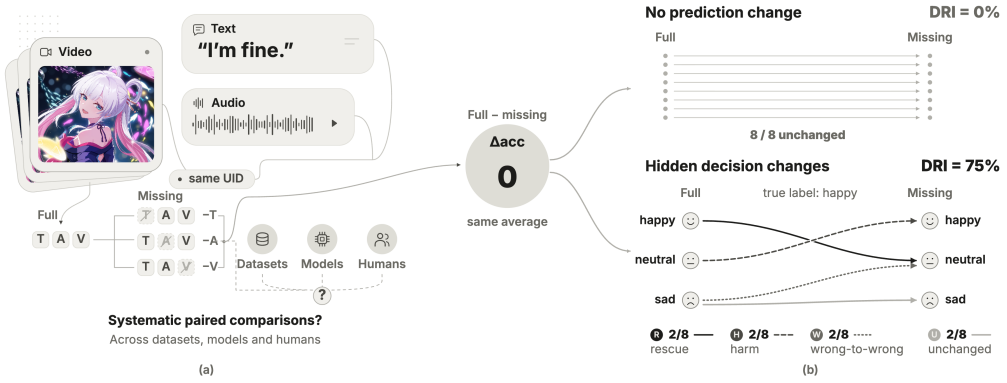


Figure 1: (a) Limited paired comparisons across datasets, models, and humans. (b) Equal net effects: no changes or rescue–harm cancellation and wrong-to-wrong shifts (schematic).

vary by emotion class, with faces being more effective for classes such as happiness (Paulmann & Pell, 2011). In our cross-corpus results, video mainly supplies boundary evidence for overt expression, especially helping maintain distinctions between happy, joy, or excitement and neutral.

Prosody, pitch, energy, speaking rate, and voice quality carry rich emotional cues; audio alone can support emotion recognition (Scherer, 2003), and multimodal systems can also benefit from its inclusion (Lian et al., 2025). In our cross-corpus results, audio mainly reflects arousal and expressive intensity and can also help distinguish some negative emotions; however, similar high-arousal cues can occur in anger, fear, joy, or surprise and cannot be mapped directly to a single class. The dataset, noise conditions, and acoustic representation affect audio-based recognition performance (Lian et al., 2026), while the direction of prediction changes in our study also varies with the decoding method and observer. In our human experiments, the leading transitions after audio removal also vary across datasets, suggesting that this dependence on conditions is not unique to particular models or encoders. Audio can be highly valuable, yet the classwise direction of its corrections is relatively unstable and may introduce error transitions. Average contribution characterizes the net benefit in a particular setting, whereas stability asks whether that benefit persists across conditions.

Prior work has analyzed modality roles through unimodal performance, feature attribution, Shapley contributions, information decomposition, and missing-modality reconstruction, including naturalistic data such as CMU-MOSEI (Bagher Zadeh et al., 2018) and human judgments. These methods estimate different objects: unimodal performance measures independent decodability, conditional gain measures additional predictive utility given the other modalities, input deletion from a fixed model measures sensitivity to missingness at deployment, and model attribution describes internal allocation. Cross-setting stability is rarely tested as an independent object; naturalistic settings still lack systematic paired ablations of text, audio, and video, and humans and models are rarely compared on the same samples under the same missingness conditions. Existing results generally cannot show how much rescue–harm cancellation and wrong-to-wrong redistribution is contained within near-zero or similar aggregate effects, nor whether these changes concentrate on the same class boundaries.

To address the inability of average effects to reveal error destinations and the lack of systematic tests of cross-setting stability, we study paired modality missingness across datasets, models, and human–model comparisons. Specifically, we make the following contributions:

- We propose a paired evaluation framework for multimodal ablation. The Decision Redistribution Index (DRI) and its decomposition measure prediction changes, rescues, and harms; source-level unique identifiers (UIDs) preserve condition pairs. Class-conditional transition signatures trace error destinations, while the correctness-cancellation index and wrong-to-wrong share characterize changes concealed by net effects. Continuous tasks track error changes and prediction direction.

- We conduct a unified comparison across six datasets, multiple open- and closed-source models, and 90 human participants. MER2025, EmoS (incl. MELD/CH-SIMS v2), and CMU-MOSEI (Lian et al., 2025; Guo et al., 2026; Bagher Zadeh et al., 2018) provide the main evidence from naturalistic settings, while IEMOCAP, RAVDESS, and CREMA-D (Busso et al., 2008; Livingstone & Russo, 2018; Cao et al., 2014) test boundary conditions involving dialogue, fixed text, and expressive intensity. Each corpus retains its native task, and cross-corpus interpretation is organized only by valence, arousal, and expressive intensity.
- We examine modality contribution and stability as two dimensions and identify class-conditional patterns across settings: linguistic semantics forms the most stable basis for judgment, video mainly maintains the boundary between positive, high-arousal emotions and neutral, and audio supplements distinctions in arousal, intensity, and negativity. Audio does not necessarily have the lowest average contribution, but the direction of its corrections is relatively unstable. Point estimates of human conditional effects also change direction across datasets.

2 Related Work

2.1 Modality Information and Emotion Judgment

Multimodal emotion research commonly reports unimodal results and modality ablations, yet no stable performance ordering has emerged among text, audio, and video (Baltrušaitis et al., 2019; Zadeh et al., 2017; Tsai et al., 2019; Hazarika et al., 2020; Yu et al., 2021; Zhang et al., 2023). When linguistic content is rich, text usually dominates; the relative advantages of audio and video vary with the dataset, representation, and comparison method, and may reverse when the same model is evaluated in a different domain (Poria et al., 2017; Shi & Huang, 2023). Unimodal performance measures independent decodability and is not equivalent to conditional gain given text. Aggregate rankings therefore do not establish what emotional information each modality provides, how it corrects specific predictions, or whether those effects persist across conditions.

Comparisons of modality function also depend on the experimental protocol: input deletion from a fixed checkpoint measures sensitivity to missingness, whereas retraining modality combinations assesses the ability to relearn from the remaining information. Attention and related analyses describe internal weights or sensitivities and cannot directly substitute for differences in output risk (Jain & Wallace, 2019). Methods for decomposing contributions further examine how information is distributed across modalities and how they interact (Wörtwein et al., 2022; 2024; Parcalabescu & Frank, 2023; Wang et al., 2025). SHAPE uses model accuracy as the utility to compute each modality's marginal contribution and cross-modal cooperation (Hu et al., 2022b). PID distinguishes task-relevant redundant, unique, and synergistic information, and can separately quantify interactions in the data and those captured by a model (Liang et al., 2023). These analyses rarely trace the specific classes toward which predictions move after modality removal.

Imbalanced training can cause modality competition or unimodal collapse, leaving some information underused (Wang et al., 2020; Du et al., 2023). Theoretical analysis of late-fusion networks shows that features learnable under unimodal training may remain unlearned by the corresponding encoders during joint training (Huang et al., 2022). Gradient modulation and modality decoupling can alleviate some of these problems (Peng et al., 2022; Wang et al., 2024). Yet analogous reversals occur in human judgment, so the cross-condition stability of correction directions cannot be attributed entirely to model optimization.

2.2 Prediction Changes Under Modality Missingness

Research on missing modalities primarily seeks to preserve performance under incomplete inputs, using cross-modal translation, meta-learning, representation imagination, reconstruction, distillation, diffusion, and dynamic parameter sharing (Pham et al., 2019; Ma et al., 2021; Zhao et al., 2021; Wang et al., 2023; Li et al., 2024). MultiBench incorporates missingness and noise into cross-task evaluation, and subsequent work further addresses modality dominance and estimation bias induced by shifts in the missingness mechanism (Liang et al., 2021; Wang et al., 2024; Choi et al., 2026).

Beyond aggregate performance, some diagnostic studies have begun to examine prediction behavior: Hazarika et al. (2022) analyze model sensitivity to a single modality, and EMMR detects polarity inconsistency after modality removal (Zeng et al., 2022). These studies still do not distinguish rescue, harm, and wrong-to-wrong changes or trace the specific destinations of discrete predictions. This distinction affects the interpretation of a near-zero net effect, which may reflect either unchanged predictions or positive–negative cancellation after frequent prediction changes.

2.3 Human Multisensory Emotion Perception

Human emotion perception integrates faces, voices, and linguistic semantics; even when a task requires an irrelevant channel to be ignored, faces and voices continue to influence each other (Massaro & Egan, 1996; de Gelder & Vroomen, 2000; Vroomen et al., 2001; Paulmann & Pell, 2010). Multisensory research also shows that people adjust their weights according to cue reliability (Ernst & Banks, 2002). This integration does not produce a fixed dominant modality, and class distributions also affect aggregate comparisons: faces are more helpful for majority classes in the datasets, such as happiness and surprise, whereas voice is more helpful for sadness, placing voice at a disadvantage under such class distributions. Modality advantages also vary with attentional instructions and visual quality (Paulmann & Pell, 2011; Takagi et al., 2015; Collignon et al., 2008).

Related reviews indicate that faces and voices are integrated while retaining distinct specializations (Klasen et al., 2012; Schirmer & Adolphs, 2017). Acoustic cues are generally related more directly to arousal and expressive intensity than to discrete emotion classes (Russell, 1980; Scherer, 2003; Juslin & Laukka, 2003). Prior work has used human informativeness ratings to guide models or compared computational decompositions with human judgments (Wörtwein et al., 2021; 2022; 2024). This paper further asks human observers to respond to the same samples and missingness conditions used for the models, testing whether paired decision redistribution concealed by average effects also occurs beyond models.

3 Paired Evaluation Framework and Experimental Design

We pair predictions for the same sample under full and reduced conditions using source-level unique identifiers (UIDs). Alongside native aggregate metrics, we report the extent of prediction changes and their class destinations to reveal rescues, harms, and wrong-to-wrong transitions concealed by average effects. Classification tasks retain class labels, and continuous tasks retain native rating scales; Figure 2 shows the paired design and analysis workflow.

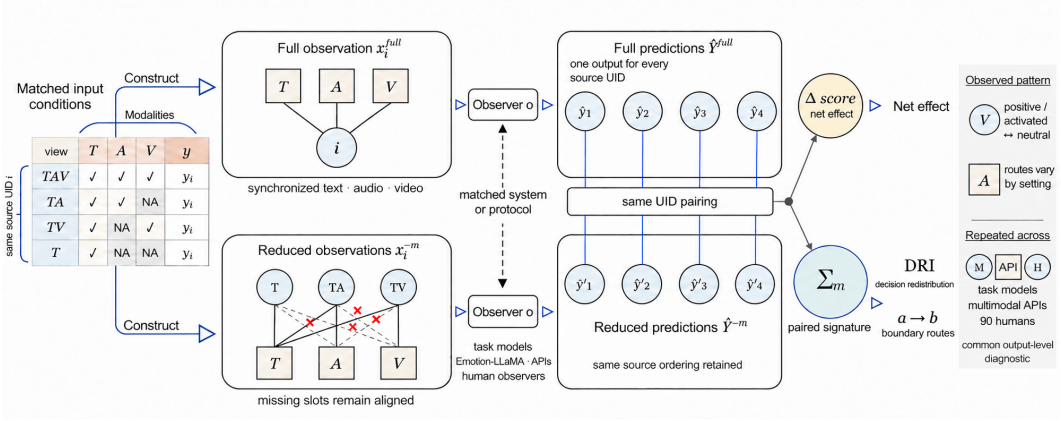


Figure 2: Overview of the paired modality-missingness design. Matched full and reduced observations retain alignment through source-level unique identifiers (UIDs) across task models, multimodal APIs, and human observers. The net score difference summarizes the aggregate effect, while the paired signature Σ_m records decision movement and directed boundary transitions.

3.1 Paired Transition Signature

For observer o, dataset d, and removed modality m, let the ground-truth target be $Y = y$ and predictions under full and reduced conditions be $\hat{Y}_{\text{full}}$ and $\hat{Y}_{\setminus m}$. The paired transition signature is

$$\Sigma_{o,d,m}(y, a, b) = \Pr\Big(\hat{Y}_{\text{full}} = a,\ \hat{Y}_{\setminus m} = b \mid Y = y\Big). \tag{1}$$

The signature preserves the true class, the prediction before removal, and its destination afterward. Diagonal terms $a = b$ indicate unchanged predictions, while off-diagonal terms record directed class-boundary crossings. The two conventional confusion matrices are the marginals obtained by summing Σ over b or a.

Non-identifiability. Two marginal confusion matrices generally cannot uniquely recover Σ and therefore cannot determine the extent or destinations of prediction changes. Two systems can even have identical accuracy, Macro-F1, and before-and-after confusion matrices while satisfying $\text{DRI} = 0$ and $\text{DRI} = 1$, respectively (Appendix A.2). The coupling between conditions preserved by UID-level pairing is thus necessary to identify transitions.

3.2 Decision Redistribution and Changes in Continuous Ratings

All summary quantities are derived from Σ. From the perspective of retaining the modality, a rescue is correct under the full condition but incorrect after removal; a harm is incorrect under the full condition but correct after removal; and a wrong-to-wrong transition is incorrect in both conditions but changes predicted class. Their proportions among all paired samples are R, H, and W, respectively, giving

$$\Delta_{\text{acc}} = \text{Acc}_{\text{full}} - \text{Acc}_{\setminus m} = R - H, \qquad \text{DRI} = R + H + W. \tag{2}$$

Here, Δ_{acc} is the net accuracy effect, and the Decision Redistribution Index $\text{DRI} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big(\hat{y}_{\text{full},i} \neq \hat{y}_{\setminus m,i}\big)$ is the proportion of N paired samples whose predicted class changes. The correctness-cancellation index χ and wrong-to-wrong share ω are

$$\chi = \frac{2\min(R, H)}{R + H}, \qquad \omega = \frac{W}{\text{DRI}}. \tag{3}$$

When $R + H > 0$, $\chi \in [0, 1]$, with larger values indicating stronger rescue–harm cancellation; when $\text{DRI} > 0$, ω is the share of all prediction changes that are wrong-to-wrong. A zero denominator makes the corresponding index NA, and raw counts are reported. By definition,

$$\text{DRI} - |\Delta_{\text{acc}}| = 2\min(R, H) + W, \tag{4}$$

where the right-hand terms represent rescue–harm cancellation and accuracy-invisible wrong-to-wrong transitions. The boundary signature $q_s(y, a, b)$ retains the normalized off-diagonal transition distribution from a to b for true class y, describing the class boundaries on which changes concentrate; its definition and edge-level distance appear in Appendix A.4.

For continuous tasks, let the target be y and predictions under full and reduced conditions be z_{full} and $z_{\setminus m}$. The directed change $z_{\setminus m} - z_{\text{full}}$ describes upward or downward prediction shifts, and the paired error change for sample i is

$$d_i := |y_i - z_{\setminus m,i}| - |y_i - z_{\text{full},i}|. \tag{5}$$

MAE is the mean absolute difference between predictions and ground truth. The net error effect is $\Delta\text{MAE} = \text{MAE}_{\setminus m} - \text{MAE}_{\text{full}} = \frac{1}{N}\sum_i d_i$, and total error migration is $G = \frac{1}{N}\sum_i |d_i|$. Continuous outputs have no discrete class boundaries, so no boundary signature is defined.

3.3 Datasets and Observers

The core evidence comes from MER2025 (Lian et al., 2025), EmoS (Guo et al., 2026), and CMU-MOSEI (Bagher Zadeh et al., 2018). The first two characterize discrete emotion judgments under variable linguistic content, while the third tests net effects and paired transitions on each native continuous rating dimension. IEMOCAP provides a laboratory-data reference, and RAVDESS and CREMA-D provide fixed-text controls; task-model architectures, frozen-feature probes, and Stage 1–5 results appear in Appendix G.

Observers include task models, Emotion-LLaMA, APIs, and humans. Gemini 3.1 Pro preview (Google DeepMind, 2026) is the main API observer; extended results appear in Appendix I. Ninety participants respond under matched TAV, TV, and TA conditions to estimate the conditional contributions of audio and video when text is present and test whether the phenomenon also occurs beyond models.

Input deletion from a fixed checkpoint, separately optimized modality branches, API requests with a single modality removed, and human judgments under matched stimulus conditions correspond to different estimands whose effect sizes cannot be directly equated. Σ provides a unified output-level paired description for these protocols. Full inference specifications appear in Appendix D.

4 Results and Discussion

We first examine sample-level changes concealed by average effects, then compare the affective regions influenced by each modality and their consistency across settings. The human study further tests whether these decision-change patterns and modality differences also occur beyond models.

4.1 Net Effects Conceal Cancellation in Correctness Changes and Redistribution Across Error Classes

For classification we report net accuracy effect Δ_{acc} and DRI; for continuous tasks, net error effect ΔMAE and total error migration G (complete event counts and derived metrics: Table E1). Human classification rows use four-person plurality labels per stimulus–condition pair for paired sensitivity analysis; individual responses define the Primary estimand (Section 4.3; Appendix F).

Table 1: Net effects, DRI, total error migration, and major transitions.

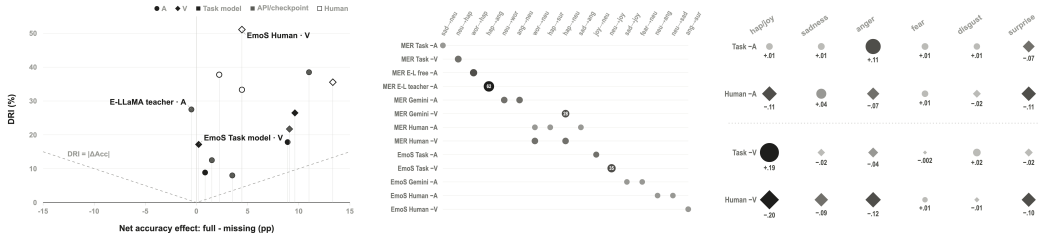
Observer	Removed	Δ_{acc} (pp)	DRI (%)	Major transition or dimension
MER2025				
Task model	Audio	$+8.9$	17.8	sad → neutral (13%)
Task model	Video	$+9.6$	26.5	neutral → happy (21%)
Emotion-LLaMA (free generation)	Audio	$+11.0$	38.5	worried → happy (25%)
Emotion-LLaMA (teacher forcing)	Audio	-0.5	27.5	happy → angry (62%)
Gemini	Audio	$+3.5$	8.0	neutral → worried; angry → neutral (19%)
Gemini	Video	$+9.1$	21.7	happy → neutral (28%)
Human plurality (sensitivity)	Audio	$+4.4$	33.3	worried → neutral / happy → surprise / sad → angry (13% each)
Human plurality (sensitivity)	Video	$+13.3$	35.6	worried → neutral; happy → neutral (19%)
EmoS				
Task model	Audio	$+0.8$	8.9	joy → neutral (15%)
Task model	Video	$+0.2$	17.2	neutral → joy (35%)
Gemini	Audio	$+1.5$	12.5	sadness → joy; fear → neutral (12%)
Human plurality (sensitivity)	Audio	$+2.2$	37.8	neutral → anger; neutral → sadness (12%)
Human plurality (sensitivity)	Video	$+4.4$	51.1	anger → surprise (13%)
CMU-MOSEI				
Observer	Removed	ΔMAE	G	Major transition or dimension
Task model	Audio	$+0.026$	0.075	anger $+0.11$; surprise -0.07
Task model	Video	$+0.023$	0.079	happiness $+0.19$
Human	Audio	$+0.024$	0.300	joy -0.11; surprise -0.11
Human	Video	$+0.030$	0.292	joy -0.20

Δ_{acc}: net accuracy effect; DRI: Decision Redistribution Index; ΔMAE: net error effect; G: total error migration. pp denotes percentage points and MAE mean absolute error. DRI is a percentage; ΔMAE and G use the native rating scale. Positive net effects favor the full condition.

Equation (2) and the nonnegativity of R, H, and W imply DRI $\geq |\Delta_{\text{acc}}|$. The dashed V-shaped line in Figure 3a marks the theoretical lower bound DRI $= |\Delta_{\text{acc}}|$; by Equation (4), the vertical distance from each point to this boundary reflects rescue–harm cancellation and wrong-to-wrong redistribution. Points with near-zero net effects still exhibit high DRI, showing that small accuracy differences can accompany widespread prediction changes.

Across the 13 discrete comparisons in Table 1, the median correctness-cancellation index χ is 83.3%, with a range of 38.0%–98.6%; the median share of wrong-to-wrong transitions among all prediction changes is 25.0%, with a range of 2.3%–65.5%. These settings consistently contain substantial rescue–harm cancellation, accompanied by varying degrees of error-class redistribution.

The MELD stratum of EmoS provides the most direct example. In the video-removal comparison, the counts of rescue, harm, and wrong-to-wrong events are 35, 34, and 12, respectively. DRI reaches 17.2%, whereas the net accuracy effect of the full condition relative to the missing condition is only $+0.2$ pp; the correctness-cancellation index is 98.6%, and wrong-to-wrong transitions account for 14.8% of all prediction changes.



(a) (b) (c)

Figure 3: Paired decision changes induced by modality removal. (a) Net accuracy effect versus DRI, with the V-shaped line indicating the theoretical lower bound; (b) dominant class transitions from the full condition to the removal condition, where the numbers and marker sizes indicate each transition's share of all prediction changes; (c) mean shifts in continuous emotion dimensions. $-$A/$-$V denote audio/video removal, and circles/diamonds in (c) denote upward/downward shifts. pp denotes percentage points.

Teacher-forced decoding in Emotion-LLaMA exhibits stronger error-class redistribution. Audio removal produces 9 rescues, 10 harms, and 36 wrong-to-wrong transitions; the net accuracy effect is -0.5 pp, DRI is 27.5%, the correctness-cancellation index is 94.7%, and wrong-to-wrong transitions account for 65.5% of all prediction changes.

Continuous judgments exhibit the corresponding structure on their native scales. For the CMU-MOSEI task model, the net MAE increase is 0.023–0.026 and total error migration is 0.075–0.079; for mean human ratings, the net increase is 0.024–0.030 and total error migration reaches 0.292–0.300. Sample-level increases and decreases in error likewise cancel in mean MAE.

This cancellation is also visible within a single emotion dimension. After video removal from the CMU-MOSEI task model, the net ΔMAE for sadness is -0.0017, yet 47.5% of samples incur larger errors. Meanwhile, 61.0% of predicted ratings shift downward (Appendix E.2). The small improvement in average error coexists with sample-level harm, while ratings shift more often toward lower sadness intensity. Such shifts may correct overestimation bias or deepen existing underestimation.

A small net effect alone cannot establish modality redundancy, nor can high DRI establish benefit. Distinguishing sample-level benefits and harms can help preserve useful corrections and reduce harm; assessing recurrence across settings also requires comparing the affective regions involved and the directions of change.

4.2 Prediction Transition Patterns after Modality Removal

Text forms the most stable backbone for judgment in tasks where linguistic content varies across samples; paired-deletion, representation, and fusion results from Stages 1–5 provide consistent evidence (Appendix G). In the MER2025 task model, text removal reduces Macro-F1 by 0.3358, more than audio or video removal; frozen unimodal probes also favor text across our three core datasets (Appendix G). Similar advantages appear in prior work: text is MultiEMO's strongest single modality on IEMOCAP and MELD (Shi & Huang, 2023), while sentiment models on MOSI and MOSEI are highly sensitive to missing language representations (Hazarika et al., 2022). Our interventions further show that sample correspondence and word order affect judgments: within-video text mismatching and word-order shuffling that preserves the original words on CMU-MOSEI reduce Pearson correlation by 0.1146 and 0.0489, respectively (Appendix H).

Figure 3b shows that changes after video removal concentrate in relatively similar affective regions. Across the MER2025 task model, Gemini, and human judgments, transitions between happy and neutral recur; for the EmoS task model, the leading transition is neutral → joy; in CMU-MOSEI, the dimensions with the largest changes for the task model and human ratings are happiness and joy, respectively. What recurs across settings is a transition between positive or high-arousal emotions and neutral, consistent with prior observations that faces have an advantage for classes such as happiness (Paulmann & Pell, 2011). The specific arrows preserve differences between label spaces and observers.

Audio is less consistent across settings. The MER2025 task model, Gemini, and human judgments have different leading transitions; on the same Emotion-LLaMA checkpoint, changing only the decoding rule changes the leading transition from worried → happy to happy → angry. In EmoS, the task model mainly exhibits joy → neutral, Gemini mainly exhibits sadness → joy and fear → neutral, and human judgments mainly move from neutral to anger or sadness. Figure 3c shows that the CMU-MOSEI task model emphasizes an upward shift in anger and a downward shift in surprise, while the principal dimensions and directions change again for human ratings.

We distinguish regional recurrence from exact edge-level agreement. To quantify the latter, we use D_{edge}, the Hellinger distance between normalized changed-edge distributions, where 0 indicates identical allocation and 1 indicates disjoint transition edges (Appendix A.4). On the same 200 MER2025 samples and the same Emotion-LLaMA checkpoint, changing only the decoding rule shifts the audio net effect from $+11.0$ to -0.5 pp and yields $D_{\text{edge}} = 0.953$ (95% CI $[0.920, 1.000]$). Thus, audio transition structure can change almost completely without changing the dataset or checkpoint.

Late fusion provides the same evidence at the representation level: the audio weight is close to 0 for EmoS, approximately 0.04 for CMU-MOSEI, and 0.42–0.44 for HuBERT in the MER2025 frozen-feature probe (Appendix G). Audio representation differences also appear in conditional gains: on the same MER2025 samples with FaceMAE, the wav2vec and HuBERT configurations yield audio conditional Macro-F1 increments of 0.0665 and 0.1676, respectively (Appendix G.1). Output transitions, decoding rules, and fusion weights jointly show that the magnitude and direction of audio effects depend strongly on the particular setting, whereas video effects recur more stably in similar affective regions.

This comparison characterizes the cross-setting stability of modality function. The visual results support a summary of the affected regions; the audio results require interpretation in conjunction with the encoder, fusion method, and observer. Systems with similar average gains may still rely on entirely different audio cues, so paired transitions provide information beyond a single performance difference.

4.3 Decision Changes and Modality Differences in Human Judgments

In the individual-response analysis, the point estimate of the accuracy effect is higher for video than audio on both MER2025 and EmoS (Table 2; Appendix F).

Table 2: Human Primary effects and same-UID alignment with Gemini.

A. Individual-response Primary effects		
Dataset	Audio Δ_{acc}	Video Δ_{acc}
MER2025	$+2.8$ pp	$+4.4$ pp
EmoS	-3.9 pp	-1.7 pp
B. Same-UID human–Gemini alignment under video removal		
Metric	Result	
Raw rescue/harm/unchanged-correctness agreement	0.600	
Unchanged correctness for both observers (proportion of all UIDs)	0.556	
Same-direction rescue/harm intersection (excluding joint unchanged correctness)	0.100	
Cohen's κ	-0.031	
Six-class recall-Δ correlation	$\rho = 0.273,\ p = 0.617$	

A: individual-response effects. B: MER45 video-removal plurality alignment. Unchanged correctness includes unchanged labels and wrong-to-wrong transitions; joint unchanged correctness does not require matching predictions.

After aggregating the four responses for each stimulus–condition pair into a plurality label, 33.3%–51.1% of stimuli receive different judgments under modality removal. For example, among the 45 EmoS stimuli, 9 are judged correctly with video but incorrectly without it, 7 show the reverse, and another 7 move between incorrect labels. Retaining video therefore yields a net accuracy gain of only 4.4 pp, although 23 stimuli receive different aggregate judgments; wrong-to-wrong transitions leave accuracy unchanged. On MER2025, audio and video removal change 15 and 16 aggregate judgments, respectively, while the corresponding net accuracy gains from retaining each modality are 4.4 and 13.3 pp (Table E1). Beneficial corrections outweigh harmful changes more strongly for video, whereas audio effects show greater cancellation. Human continuous ratings on CMU-MOSEI exhibit a similar pattern: net ΔMAE is 0.024–0.030, whereas total error migration is 0.292–0.300 (Table 1), with stimulus-level increases and decreases in error likewise canceling in the mean.

Transition destinations also partly resemble those of models. Under video removal, the human happy $\rightarrow$ neutral transition on MER2025 matches Gemini and involves the same class pair as the task model's reverse transition. On CMU-MOSEI, the largest shifts in human and model ratings occur in joy and happiness, respectively, although in opposite directions (Figure 3b,c). The leading human video-removal transition on EmoS is instead anger $\rightarrow$ surprise. Leading audio transitions vary across datasets and observers: on EmoS, human judgments mainly move from neutral to anger or sadness, whereas the task model mainly moves from joy to neutral.

Human and Gemini judgment changes overlap little on the shared MER45 video-removal subset. Their rescue/harm/unchanged-correctness agreement is 60.0%, but joint unchanged correctness accounts for 55.6% of all stimuli and thus most of this raw agreement. Excluding these cases, the same-direction rescue/harm intersection is 0.100; κ is near zero, and the correlation of class-wise recall changes is not significant (Table 2). Human results thus support decision redistribution and cancellation beyond models, alongside partly similar trends in modality effects; the specific stimuli that benefit or suffer still differ across observers.

5 Conclusion

Across datasets, models, APIs, and human observers, small average gains can accompany widespread decision redistribution. Their value depends on the mix and distribution of beneficial corrections, harmful changes, and error destinations, including wrong-to-wrong transitions invisible to accuracy. In tasks with variable linguistic content, text forms a stable backbone for judgment; video removal shows relatively consistent effects on positive or high-arousal emotions and neighboring classes, whereas the classwise directions, continuous dimensions, decoding outcomes, and fusion weights associated with audio removal vary substantially across settings. With DRI and transition destinations, the framework locates differences in specific samples and decision boundaries.

Future work should preserve valuable corrections and reduce avoidable harms, using paired evaluation to clarify the prediction-change costs accompanying aggregate gains. Audio effects should be interpreted alongside encoder replacement, fusion weights, and output rules, while human experiments offer an independent reference for testing whether model conclusions depend on training objectives. Reporting affected samples, classes, and error destinations alongside native task metrics helps test whether valuable corrections persist across datasets, architectures, and observers.

6 Limitations

Our findings depend on the experimental conditions. Models trained only on complete inputs may struggle with missing inputs, so deletion effects cannot be equated with results from retraining modality combinations. API behavior may vary with version and input processing, while human trials retain text. Emotion categories and rating scales also differ across datasets; we therefore compare recurring affective regions and use MOSEI's continuous ratings to supplement directional evidence. Even within a dataset, agreement among three annotators cannot exclude bias: samples may not represent other populations or settings, and annotators may share judgment preferences. Accuracy and the correctness of prediction changes depend on these labels. DRI is label-independent but still depends on sample selection and cannot alone establish more accurate emotion recognition.

AI USE STATEMENT

Generative AI tools were used to assist with the English translation of the manuscript and the organization of supplementary supporting materials. The authors take responsibility for the final content of the manuscript and supplementary materials.

REFERENCES

Zara Ambadar, Jonathan W. Schooler, and Jeffrey F. Cohn. Deciphering the Enigmatic Face: The Importance of Facial Dynamics in Interpreting Subtle Facial Expressions. *Psychological Science*, 16(5):403–410, 2005. doi: 10.1111/j.0956-7976.2005.01548.x. URL https://doi.org/10.1111/j.0956-7976.2005.01548.x.

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

AmirAli Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *ACL 2018*, pp. 2236–2246, 2018. doi: 10.18653/v1/P18-1208. URL https://doi.org/10.18653/v1/P18-1208.

T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. doi: 10.1109/TPAMI.2018.2798607. URL https://doi.org/10.1109/TPAMI.2018.2798607.

C. Busso et al. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language Resources and Evaluation*, 42(4):335–359, 2008. doi: 10.1007/s10579-008-9076-6. URL https://doi.org/10.1007/s10579-008-9076-6.

Rafael A. Calvo and Sidney K. D'Mello. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, 2010. doi: 10.1109/T-AFFC.2010.1. URL https://doi.org/10.1109/T-AFFC.2010.1.

H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244. URL https://doi.org/10.1109/TAFFC.2014.2336244.

Z. Cheng, Z.-Q. Cheng, J.-Y. He, J. Sun, K. Wang, Y. Lin, Z. Lian, X. Peng, and A. Hauptmann. Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. In *NeurIPS 2024*, volume 37, pp. 110805–110853, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c7f43ada17acc234f568dc66da527418-Abstract-Conference.html.

Y. S. Choi, V. Jeanselme, P. Elias, and S. Joshi. ICYM2I: The Illusion of Multimodal Informativeness under Missingness. In *ICLR 2026*, 2026. URL https://openreview.net/forum?id=jC7FK8Rf4s.

O. Collignon, S. Girard, F. Gosselin, S. Roy, D. Saint-Amour, M. Lassonde, and F. Lepore. Audio-Visual Integration of Emotion Expression. *Brain Research*, 1242:126–135, 2008. doi: 10.1016/j.brainres.2008.04.023. URL https://doi.org/10.1016/j.brainres.2008.04.023.

B. de Gelder and J. Vroomen. The Perception of Emotions by Ear and by Eye. *Cognition & Emotion*, 14(3):289–311, 2000. doi: 10.1080/026999300378824. URL https://doi.org/10.1080/026999300378824.

C. Du, J. Teng, T. Li, Y. Liu, T. Yuan, Y. Wang, Y. Yuan, and H. Zhao. On Uni-Modal Feature Learning in Supervised Multi-Modal Learning. In *ICML 2023*, pp. 8632–8656, 2023. URL https://proceedings.mlr.press/v202/du23e.html.

M. O. Ernst and M. S. Banks. Humans Integrate Visual and Haptic Information in a Statistically Optimal Fashion. *Nature*, 415:429–433, 2002. doi: 10.1038/415429a. URL https://doi.org/10.1038/415429a.

Google DeepMind. Gemini 3.1 Pro. Model card, February 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed September 11, 2026.

P. Guo, J. Liang, Z. Xie, Q. Wang, and D. F. Wong. EmoS: A High-Fidelity Multimodal Benchmark for Fine-Grained Streaming Emotional Understanding. In *ACL 2026*, pp. 39074–39089, 2026. doi: 10.18653/v1/2026.acl-long.1813. URL https://doi.org/10.18653/v1/2026.acl-long.1813.

D. Hazarika, R. Zimmermann, and S. Poria. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *ACM Multimedia 2020*, pp. 1122–1131, 2020. doi: 10.1145/3394171.3413678. URL https://doi.org/10.1145/3394171.3413678.

D. Hazarika, Y. Li, B. Cheng, S. Zhao, R. Zimmermann, and S. Poria. Analyzing Modality Robustness in Multimodal Sentiment Analysis. In *NAACL 2022*, pp. 685–696, 2022. doi: 10.18653/v1/2022.naacl-main.50. URL https://doi.org/10.18653/v1/2022.naacl-main.50.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR 2016*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked Autoencoders Are Scalable Vision Learners. In *CVPR 2022*, pp. 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291. URL https://doi.org/10.1109/TASLP.2021.3122291.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR 2022*, 2022a. URL https://openreview.net/forum?id=nZeVKeeFYf9.

P. Hu, X. Li, and Y. Zhou. SHAPE: An Unified Approach to Evaluate the Contribution and Cooperation of Individual Modalities. In *IJCAI 2022*, pp. 3064–3070, 2022b. doi: 10.24963/ijcai.2022/425. URL https://doi.org/10.24963/ijcai.2022/425.

Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang. Modality Competition: What Makes Joint Training of Multi-modal Network Fail in Deep Learning? (Provably). In *ICML 2022*, pp. 9226–9259, 2022. URL https://proceedings.mlr.press/v162/huang22e.html.

S. Jain and B. C. Wallace. Attention is not Explanation. In *NAACL 2019*, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357. URL https://doi.org/10.18653/v1/N19-1357.

P. N. Juslin and P. Laukka. Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code? *Psychological Bulletin*, 129(5):770–814, 2003. doi: 10.1037/0033-2909.129.5.770. URL https://doi.org/10.1037/0033-2909.129.5.770.

M. Klasen, Y.-H. Chen, and K. Mathiak. Multisensory Emotions: Perception, Combination and Underlying Neural Processes. *Reviews in the Neurosciences*, 23(4):381–392, 2012. doi: 10.1515/revneuro-2012-0040. URL https://doi.org/10.1515/revneuro-2012-0040.

M. Li et al. A Unified Self-Distillation Framework for Multimodal Sentiment Analysis with Uncertain Missing Modalities. In *AAAI 2024*, volume 38, pp. 10074–10082, 2024. doi: 10.1609/aaai.v38i9.28871. URL https://doi.org/10.1609/aaai.v38i9.28871.

Z. Lian, L. Sun, Y. Ren, H. Gu, H. Sun, L. Chen, B. Liu, and J. Tao. MERBench: A Unified Evaluation Benchmark for Multimodal Emotion Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(5):5793–5810, 2026. doi: 10.1109/TPAMI.2026.3653457. URL https://doi.org/10.1109/TPAMI.2026.3653457.

Z. Lian et al. MER 2025: When Affective Computing Meets Large Language Models. arXiv:2504.19423, 2025. URL https://arxiv.org/abs/2504.19423.

P. P. Liang et al. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. In *NeurIPS 2021 Datasets and Benchmarks Track*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/37693cfc748049e45d87b8c7d8b9aacd-Abstract-round1.html.

P. P. Liang et al. Quantifying & Modeling Multimodal Interactions: An Information Decomposition Framework. In *NeurIPS 2023*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/575286a73f238b6516ce0467d67eadb2-Abstract-Conference.html.

Kristen A. Lindquist, Lisa Feldman Barrett, Eliza Bliss-Moreau, and James A. Russell. Language and the Perception of Emotion. *Emotion*, 6(1):125–138, 2006. doi: 10.1037/1528-3542.6.1.125. URL https://doi.org/10.1037/1528-3542.6.1.125.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make Acoustic and Visual Cues Matter: CH-SIMS v2.0 Dataset and AV-Mixup Consistent Module. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 247–258, 2022. doi: 10.1145/3536221.3556630. URL https://doi.org/10.1145/3536221.3556630.

S. R. Livingstone and F. A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, 13(5):e0196391, 2018. doi: 10.1371/journal.pone.0196391. URL https://doi.org/10.1371/journal.pone.0196391.

M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng. SMIL: Multimodal Learning with Severely Missing Modality. In *AAAI 2021*, volume 35, pp. 2302–2310, 2021. doi: 10.1609/aaai.v35i3.16330. URL https://doi.org/10.1609/aaai.v35i3.16330.

D. W. Massaro and P. B. Egan. Perceiving Affect from the Voice and the Face. *Psychonomic Bulletin & Review*, 3(2):215–221, 1996. doi: 10.3758/BF03212421. URL https://doi.org/10.3758/BF03212421.

A. Mollahosseini, B. Hasani, and M. H. Mahoor. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. doi: 10.1109/TAFFC.2017.2740923. URL https://doi.org/10.1109/TAFFC.2017.2740923.

Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. Empathy in Virtual Agents and Robots: A Survey. *ACM Transactions on Interactive Intelligent Systems*, 7(3):11:1–11:40, 2017. doi: 10.1145/2912150. URL https://doi.org/10.1145/2912150.

L. Parcalabescu and A. Frank. MM-SHAP: A Performance-Agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks. In *ACL 2023*, pp. 4032–4059, 2023. doi: 10.18653/v1/2023.acl-long.223. URL https://doi.org/10.18653/v1/2023.acl-long.223.

S. Paulmann and M. D. Pell. Contextual Influences of Emotional Speech Prosody on Face Processing: How Much Is Enough? *Cognitive, Affective, & Behavioral Neuroscience*, 10(2):230–242, 2010. doi: 10.3758/CABN.10.2.230. URL https://doi.org/10.3758/CABN.10.2.230.

S. Paulmann and M. D. Pell. Is There an Advantage for Recognizing Multi-Modal Emotional Stimuli? *Motivation and Emotion*, 35(2):192–201, 2011. doi: 10.1007/s11031-011-9206-0. URL https://doi.org/10.1007/s11031-011-9206-0.

X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu. Balanced Multimodal Learning via On-the-Fly Gradient Modulation. In *CVPR 2022*, pp. 8238–8247, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Peng_Balanced_Multimodal_Learning_via_On-the-Fly_Gradient_Modulation_CVPR_2022_paper.html.

H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos. Found in Translation: Learning Robust Joint Representations by Cyclic Translations between Modalities. In *AAAI 2019*, volume 33, pp. 6892–6899, 2019. doi: 10.1609/aaai.v33i01.33016892. URL https://doi.org/10.1609/aaai.v33i01.33016892.

S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency. Context-Dependent Sentiment Analysis in User-Generated Videos. In *ACL 2017*, pp. 873–883, 2017. doi: 10.18653/v1/P17-1081. URL https://doi.org/10.18653/v1/P17-1081.

J. A. Russell. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980. doi: 10.1037/h0077714. URL https://doi.org/10.1037/h0077714.

K. R. Scherer. Vocal Communication of Emotion: A Review of Research Paradigms. *Speech Communication*, 40(1–2):227–256, 2003. doi: 10.1016/S0167-6393(02)00084-5. URL https://doi.org/10.1016/S0167-6393(02)00084-5.

A. Schirmer and R. Adolphs. Emotion Perception from Face, Voice, and Touch: Comparisons and Convergence. *Trends in Cognitive Sciences*, 21(3):216–228, 2017. doi: 10.1016/j.tics.2017.01.001. URL https://doi.org/10.1016/j.tics.2017.01.001.

T. Shi and S.-L. Huang. MultiEMO: An Attention-Based Correlation-Aware Multimodal Fusion Framework for Emotion Recognition in Conversations. In *ACL 2023*, pp. 14752–14766, 2023. doi: 10.18653/v1/2023.acl-long.824. URL https://doi.org/10.18653/v1/2023.acl-long.824.

S. Takagi, S. Hiramatsu, K. Tabei, and A. Tanaka. Multisensory Perception of the Six Basic Emotions Is Modulated by Attentional Instruction and Unattended Modality. *Frontiers in Integrative Neuroscience*, 9:1, 2015. doi: 10.3389/fnint.2015.00001. URL https://doi.org/10.3389/fnint.2015.00001.

Z. Tong, Y. Song, J. Wang, and L. Wang. VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/416f9cb3276121c42eebb86352a4354a-Abstract-Conference.html.

Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *ACL 2019*, pp. 6558–6569, 2019. doi: 10.18653/v1/P19-1656. URL https://doi.org/10.18653/v1/P19-1656.

J. Vroomen, J. Driver, and B. de Gelder. Is Cross-Modal Integration of Emotional Expressions Independent of Attentional Resources? *Cognitive, Affective, & Behavioral Neuroscience*, 1(4):382–387, 2001. doi: 10.3758/CABN.1.4.382. URL https://doi.org/10.3758/CABN.1.4.382.

H. Wang, S. Luo, G. Hu, and J. Zhang. Gradient-Guided Modality Decoupling for Missing-Modality Robustness. In *AAAI 2024*, volume 38, pp. 15483–15491, 2024. doi: 10.1609/aaai.v38i14.29474. URL https://doi.org/10.1609/aaai.v38i14.29474.

J. Wang, F. Cao, Z. Xue, X. Zhao, and J. Liang. CMoB: Modality Valuation via Causal Effect for Balanced Multimodal Learning. In *NeurIPS 2025*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/76a0f4215050c7d4e88c03cb8a11fee6-Abstract-Conference.html.

W. Wang, D. Tran, and M. Feiszli. What Makes Training Multi-Modal Classification Networks Hard? In *CVPR 2020*, pp. 12695–12705, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_What_Makes_Training_Multi-Modal_Classification_Networks_Hard_CVPR_2020_paper.html.

Y. Wang, Y. Li, and Z. Cui. Incomplete Multimodality-Diffused Emotion Recognition. In *NeurIPS 2023*, 2023. URL https://openreview.net/forum?id=BuGFwUS9B3.

T. Wörtwein, L. B. Sheeber, N. Allen, J. F. Cohn, and L.-P. Morency. Human-Guided Modality Informativeness for Affective States. In *ICMI 2021*, pp. 728–734, 2021. doi: 10.1145/3462244.3481004. URL https://doi.org/10.1145/3462244.3481004.

T. Wörtwein, L. Sheeber, N. Allen, J. Cohn, and L.-P. Morency. Beyond Additive Fusion: Learning Non-Additive Multimodal Interactions. In *Findings of EMNLP 2022*, pp. 4681–4696, 2022. doi: 10.18653/v1/2022.findings-emnlp.344. URL https://doi.org/10.18653/v1/2022.findings-emnlp.344.

T. Wörtwein, N. B. Allen, J. F. Cohn, and L.-P. Morency. SMURF: Statistical Modality Uniqueness and Redundancy Factorization. In *ICMI 2024*, pp. 339–349, 2024. doi: 10.1145/3678957.3685716. URL https://doi.org/10.1145/3678957.3685716.

A. Yang et al. Qwen3 Technical Report. arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

W. Yu, H. Xu, Z. Yuan, and J. Wu. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. In *AAAI 2021*, volume 35, pp. 10790–10797, 2021. doi: 10.1609/aaai.v35i12.17289. URL https://doi.org/10.1609/aaai.v35i12.17289.

A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor Fusion Network for Multimodal Sentiment Analysis. In *EMNLP 2017*, pp. 1103–1114, 2017. doi: 10.18653/v1/D17-1115. URL https://doi.org/10.18653/v1/D17-1115.

J. Zeng, J. Zhou, and T. Liu. Mitigating Inconsistencies in Multimodal Sentiment Analysis under Uncertain Missing Modalities. In *EMNLP 2022*, pp. 2924–2934, 2022. doi: 10.18653/v1/2022.emnlp-main.189. URL https://doi.org/10.18653/v1/2022.emnlp-main.189.

H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, and T. Yu. Learning Language-Guided Adaptive Hyper-Modality Representation for Multimodal Sentiment Analysis. In *EMNLP 2023*, pp. 756–767, 2023. doi: 10.18653/v1/2023.emnlp-main.49. URL https://doi.org/10.18653/v1/2023.emnlp-main.49.

J. Zhao, R. Li, and Q. Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. In *ACL-IJCNLP 2021*, pp. 2608–2618, 2021. doi: 10.18653/v1/2021.acl-long.203. URL https://doi.org/10.18653/v1/2021.acl-long.203.

A Notation, Tensor Definitions, and Non-identifiability

A.1 Discrete Signatures and Three Summary Axes

For the same set of samples under removal of modality m, let Y denote the ground-truth label, A the prediction with the complete input, and B the prediction after removal. The class-conditional signature is $\Sigma_m(y, a, b) = \Pr(A = a, B = b \mid Y = y)$. The complete-input confusion matrix is the marginal obtained by summing over b, whereas the removed-input confusion matrix is the marginal obtained by summing over a.

For the same set of N paired UIDs, the proportions of rescue, harm, and wrong-to-wrong events among all samples are defined as

$$R = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{A_i = Y_i,\ B_i \neq Y_i\},$$

$$H = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{A_i \neq Y_i,\ B_i = Y_i\},$$

$$W = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{A_i \neq Y_i,\ B_i \neq Y_i,\ A_i \neq B_i\}.$$

The net accuracy effect and Decision Redistribution Index are, respectively,

$$\Delta_{\text{acc}} = R - H, \qquad \text{DRI} = R + H + W.$$

The cancellation index for changes in correctness and the share of wrong-to-wrong events among all decision changes are, respectively,

$$\chi = \frac{2\min(R, H)}{R + H}, \qquad \omega = \frac{W}{\text{DRI}}.$$

The relationship between these two components and the net effect is

$$\text{DRI} - |\Delta_{\text{acc}}| = 2\min(R, H) + W.$$

When $R + H = 0$, χ is recorded as NA; when $\text{DRI} = 0$, ω is recorded as NA, and the raw event counts are reported as well.

A.2 Why Two Confusion Matrices Are Insufficient

After the marginal confusion matrices have been fixed, the coupling of $(A, B) \mid Y$ remains non-unique. Consider binary classification in which each ground-truth class contains two equally weighted samples, and both the complete and removed conditions classify one sample correctly and one incorrectly. One coupling makes the two conditions correct on the same sample, yielding $A = B$ and $\text{DRI} = 0$; a second coupling swaps the identity of the correctly classified sample in the removed condition, yielding a flip for every sample and $\text{DRI} = 1$. The two systems have identical accuracy, Macro-F1, and before-and-after confusion matrices. Therefore, no statistic that depends only on the two marginals can recover the paired transitions.

When the number of classes is $K \geq 3$, the destination class of an error may change even if a sample is misclassified under both conditions. Wrong-to-wrong mass does not enter the accuracy difference, but it changes the downstream class burden and the distribution of transition edges. This is precisely why W and the full Σ_m must be retained separately.

A.3 Continuous-Task Counterparts

For sample i, define the removal-induced change in error as $d_i = |y_i - b_i| - |y_i - a_i|$. The net ΔMAE is $\bar{d}$, total error migration is $G = N^{-1}\sum_i |d_i|$, and the prediction displacement is $b_i - a_i$. The quantity $\bar{d}$ describes the aggregate balance, G describes sample-level activity, and the proportions of upward and downward shifts describe direction. Reporting only ΔMAE once again compresses the cancellation between increases and decreases in error into a single number.

A.4 Transition-Edge Distance

Normalizing over the edges (y, a, b) on which a decision change occurs gives q_s. The distance between two settings is defined as $D_{\text{edge}}(s, t) = \sqrt{1 - \sum_{y,a,b} \sqrt{q_s(y, a, b) q_t(y, a, b)}}$. Comparisons on shared samples use a common empirical class proportion and jointly resample shared source groups; comparisons across samples fix common class weights in advance. A distance of zero indicates that decision-change mass is allocated along the same edges, whereas a value close to one indicates that the transition edges barely overlap.

B Data, Splits, and Modality-Removal Implementation

B.1 Units of Analysis

Data/evidence	Task	Paired samples	Clustering unit	Primary use
MER C2 video	Discrete classification	2,930	342 source groups	Same-checkpoint video removal
MER C2 audio	Discrete classification	2,930	342 source groups	Independent TAV/TV rule comparison
Gemini MER	Discrete classification	198	97 source groups	Physical media removal
EmoS paired outputs	Fine-grained classification	1,309	231 source groups	TAV/TA paired comparison
EmoS-MELD	Fine-grained classification	472	96 source groups	Net effect without source-group overlap
MOSEI fixed checkpoint	Six-dimensional continuous intensity	4,521	672 source videos	Same-model audio/video removal
MOSEI independent T/A	Six-dimensional continuous intensity	1,842 validation	299 source videos	Independent decoding and conditional gain
Human Primary	MER and EmoS classification; MOSEI continuous	135 stimuli: 45 from each of three datasets	stimulus	Marginal risk across TAV/TV/TA conditions

Table 1 and Table E1 report MER audio results on 2,930 development samples (ARCH-DEV), not the separate 1,063-sample held-out evaluation. Sample sizes are reported separately according to independent units such as people, stimuli, clips, source videos, sessions, speakers, or actors. IEMOCAP, RAVDESS, and CREMA-D retain their respective grouping structures to test results under dialogue, fixed-text, and expression-intensity conditions.

B.2 Input Removal and Missingness States

Paired removal for a frozen model holds the samples, labels, checkpoint, retained modalities, and preprocessing fixed, and removes the target modality at the input layer. API comparisons hold the model ID, prompt template, and output mapping fixed and omit the corresponding audio or video at the request layer; independently trained comparisons optimize prediction rules separately for each modality subset. Generative responses are mapped uniformly to the native labels, with unparsed outputs recorded as $\perp$; sensitivity analyses separately use mappable responses and a neutral fallback mapping.

The TAV and TA outputs of EmoS are paired by UID. The MELD stratum is used to estimate the net effect, the CH-SIMS-v2 stratum (Liu et al., 2022) is used to characterize Joy–Neutral transitions, and the independent T/A branches for CMU-MOSEI are fitted on the training set and evaluated for conditional gain on 1,842 validation clips.

C Protocols, Estimands, and Interpretability Boundaries

Protocol	Comparison	Estimand
Fixed-checkpoint removal	Complete and reduced inputs for the same UID	The specified model's predictive response to modality removal
Independently trained systems	Separately optimized rules with the modality included and with the modality removed	Conditional utility and error distribution under the corresponding training pathways
Independent T/A fusion	T, A, and TA from frozen branches	Independent decodability, conditional gain, and fusion weights
Physical removal through an API	Cross-request outputs under the same ID and prompt	Interface input sensitivity at the time of recording

Protocol	Comparison	Estimand
Human Primary	TAV, TV, and TA conditions on matched stimuli	Stimulus-conditioned marginal differences at the individual-response level
Human consensus	Condition-wise pluralities for the same UID	Transitions and boundary sensitivity among aggregated observers

Our assessment of "modality contribution" follows the estimand rather than superficial terminological similarity. Fixed-checkpoint results are used primarily for deployment sensitivity, independently trained results describe learnable conditional utility, independent branches separate unimodal decoding from conditional gain, and human results support observer-level condition comparisons. Only quantities from the same protocol are compared directly.

D Inference, Sensitivity, and Multiple Comparisons

D.1 Cluster Resampling

The primary intervals resample the independent generative unit of each dataset: MER and Gemini use source groups, EmoS uses source groupings, MOSEI uses source videos, and the human data use stimuli. Every round retains all observations within each sampled group and recomputes the overall metrics, transition tensor, and class effects accordingly. The primary model signatures use 10,000 bootstrap replicates; the independent T/A validation diagnostics use 2,000 source-video bootstrap replicates; and the Human Primary and sensitivity analyses use 50,000 and 20,000 resamples, respectively. Confidence intervals are based on empirical quantiles.

D.2 Paired Tests and Class-Level Control

For classification effects on the same UID, the sample-level difference in correctness is used; for continuous effects, the sample-level difference in absolute error is used. Edge-distribution comparisons jointly sample shared source groups so that the two conditions retain identical sample composition. The primary sparse-edge example for Emotion-LLaMA uses 20,000 paired cluster permutations. Holm correction is applied to the prespecified families of class-wise and destination-edge analyses for humans, whereas class-wise model effects are summarized with cluster intervals.

D.3 Pairing Coverage, Output Mapping, and Zero Inflation

The Gemini MER video comparison contains 198 complete pairs, with a net accuracy effect from $+0.09$ to $+0.10$. The free-form Emotion-LLaMA results compute edge distance separately using mappable responses and the neutral fallback rule. Human–model directional comparisons report the raw rescue/harm/unchanged-correctness agreement rate, the joint unchanged-correctness proportion, the same-direction rescue/harm intersection after excluding those cases, κ, and the class-rank correlation.

E Natural-Scene Results Summary

E.1 Discrete Tasks

Setting	Overall effect	Paired transitions
MER C2 video	Macro-F1 $0.6300 \rightarrow 0.5518$ ΔAcc $+0.0962$	Happy recall $+26.9$ pp; Neutral→Happy 165 (21.3%); Happy→Neutral 156 (20.1%)
EmoS MELD TAV→TA ($N = 472$)	ΔAcc $+0.0021$ 95% CI $[-0.0326, 0.0345]$	DRI 17.16%; rescue/harm/wrong-to-wrong counts of 35/34/12; Neutral→Joy is the dominant transition
EmoS all TAV→TA ($N = 1{,}309$; sensitivity)	ΔAcc $+0.0092$ 95% CI $[-0.0110, 0.0288]$	DRI 21.31%; rescue/harm/wrong-to-wrong counts of 109/97/73; Joy–Neutral transitions

The main text and main figure use the MELD stratum ($N = 472$), while the aggregate over all samples ($N = 1{,}309$) constitutes the sensitivity analysis. Both scopes exhibit a small net change in accuracy and a higher proportion of paired transitions, primarily involving the Joy–Neutral boundary.

Table E1. Event counts and derived metrics for the discrete comparisons.

Setting	Removed	N	#R/#H/#W	ΔAcc (pp)	DRI (%)	χ (%)	ω (%)
MER2025							
Task model	A	2930	341/80/102	+8.91	17.85	38.00	19.50
Task model	V	2930	427/145/204	+9.62	26.48	50.70	26.29
Emotion-LLaMA free-form	A	200	30/8/39	+11.00	38.50	42.11	50.65
Emotion-LLaMA teacher-forced	A	200	9/10/36	−0.50	27.50	94.74	65.45
Gemini	A	200	10/3/3	+3.50	8.00	46.15	18.75
Gemini	V	198	30/12/1	+9.09	21.72	57.14	2.33
Human plurality (sensitivity)	A	45	7/5/3	+4.44	33.33	83.33	20.00
Human plurality (sensitivity)	V	45	9/3/4	+13.33	35.56	50.00	25.00
EmoS							
Task model	A	1309	48/37/31	+0.84	8.86	87.06	26.72
Task model (MELD)	V	472	35/34/12	+0.21	17.16	98.55	14.81
Task model (all; sensitivity)	V	1309	109/97/73	+0.92	21.31	94.17	26.16
Gemini	A	200	12/9/4	+1.50	12.50	85.71	16.00
Human plurality (sensitivity)	A	45	5/4/8	+2.22	37.78	88.89	47.06
Human plurality (sensitivity)	V	45	9/7/7	+4.44	51.11	87.50	30.43

The main-text summary uses 13 primary analysis units; the all-sample EmoS video result is included as a sensitivity row. #R/#H/#W are the raw event counts; χ and $\omega = W/\text{DRI}$ are calculated according to the ratio definitions in Section 3.2 of the main text.

E.2 Continuous CMU-MOSEI Task

Removed	Dimension	Net ΔMAE	Error movement	Interpretation
Video	sadness	−0.00169 CI $[-0.00516, 0.00191]$	$G = 0.06138$ 47.5% worsened	The net value is near zero, but 61.0% of predictions shift downward, indicating strong cancellation
Video	happiness	+0.11170	Other five dimensions: −0.00169 to +0.01372	Substantially larger than the range of −0.00169 to +0.01372 for the other five dimensions
Audio	anger	+0.06002	MAE increases on all six dimensions	Removing audio increases MAE on all six dimensions, with the largest increase for anger

The continuous task retains all six native intensity dimensions. The largest change under video removal is concentrated in happiness, whereas audio removal increases MAE across all six dimensions, with the largest increase in anger; the two modalities therefore exhibit different regional distributions.

F Human Study and Same-UID Human–Model Comparison

F.1 Individual Responses and Stimulus-Level Plurality

The Primary analysis includes 90 participants selected from 108 completers according to the prespecified quality criteria. The 135 stimuli comprise 45 each from MER2025, EmoS, and CMU-MOSEI, under the TAV, TV, and TA conditions, with four responses obtained for each stimulus–condition pair. The classification subset contains 90 stimuli and 1,080 responses, while the continuous CMU-MOSEI subset contains 45 stimuli and 540 responses. The Primary analysis estimates between-condition differences using individual responses and treats the stimulus as the clustering unit.

The same-UID comparison between models and humans uses the plurality label from the four responses for each stimulus–condition pair. This aggregation constitutes a paired sensitivity analysis

at the stimulus level, used to test whether effects with similar average directions arise from the same stimuli; individual responses remain the Primary level of analysis for the human experiment.

F.2 Human Effects and Overlap

Data	Audio accuracy effect	Video accuracy effect	Conclusion
MER2025	$+0.028$	$+0.044$	Point estimates are positive, and both intervals cover zero
EmoS	-0.039	-0.017	Point estimates are negative, and both intervals cover zero

The shared MER45 set covers 36 source groups. The raw rescue/harm/unchanged-correctness agreement rate between the human plurality and Gemini under video removal is 0.600; joint unchanged correctness accounts for 0.556 of all UIDs. After excluding these cases, the same-direction rescue/harm intersection is 0.100, with Cohen's $\kappa = -0.031$. The Spearman correlation of the six class-wise recall deltas is $\rho = 0.273$, with exact label-permutation $p = 0.617$. Similar directions in the average effects arise primarily from different stimuli. The continuous CMU-MOSEI condition results are summarized by stimulus-level differences in absolute error. Unchanged correctness includes unchanged labels and wrong-to-wrong transitions; joint unchanged correctness does not require matching predictions.

F.3 Participant Background and Compensation

All participants were students in psychology or medicine and received compensation for taking part in the study.

F.4 Ethical Considerations

Participants were informed of the study procedures before taking part. The study did not involve participants' private information.

G Task-Model Architecture, Stages, and Fusion Analysis

G.1 Task-Model Architecture

The task model uses Qwen3-4B as the shared text backbone (Yang et al., 2025), with LoRA applied to the attention projections and feed-forward layers (Hu et al., 2022a). The audio branch receives 16-kHz waveforms and uses wav2vec 2.0 small to produce frame-level representations (Baevski et al., 2020), followed by a temporal Transformer, attentive-statistics pooling, and a 512-dimensional adapter that form two audio-prefix tokens. The video branch extracts frozen AffectNet ResNet18 representations from selected face frames (He et al., 2016; Mollahosseini et al., 2019) and passes them through a projection, temporal convolution, a single-layer Transformer, and dual-query statistical pooling to form two video-prefix tokens.

The fusion layer uses independent sigmoid gates to modulate the audio and video prefixes and generates two text-conditioned cross-attention tokens. The final sequence comprises the cross tokens, audio prefixes, video prefixes, and text tokens, with the last valid position connected to the dataset-native task head. Training updates the Qwen LoRA parameters, audio pooling and adapter, visual temporal encoder, prefix projections, gates, cross-attention module, and task head; the AffectNet image backbone remains frozen, while the final four wav2vec layers are progressively unfrozen according to $0 \to 1 \to 2 \to 4$.

We conduct a frozen-feature probe on 200 MER2025 validation samples (97 source groups), comparing wav2vec 2.0 small and TencentGameMate/chinese-hubert-large (Hsu et al., 2021) with their respective pooling settings and feature dimensions. Evaluation uses nested group cross-validation with five outer folds and four inner folds. Visual representations are the FaceMAE local facial features and VideoMAE temporal features provided by Emotion-LLaMA (He et al., 2022; Tong et al.,

2022; Cheng et al., 2024). Audio-only Macro-F1 is 0.4898 for wav2vec and 0.6176 for HuBERT, with higher audio conditional increments for HuBERT under both visual representations.

Audio encoder	Visual encoder	Audio-only F1	TAV F1	TV F1	Audio increment
wav2vec	FaceMAE	0.4898	0.5998	0.5333	+0.0665
wav2vec	VideoMAE	0.4898	0.6175	0.5701	+0.0474
HuBERT	FaceMAE	0.6176	0.7009	0.5333	+0.1676
HuBERT	VideoMAE	0.6176	0.6776	0.5701	+0.1075

G.2 Stage 1–5 Analysis Chain

Stage	Analysis objective	Implementation	Reported result
Stage 1	Target-domain TAV adaptation	Joint training of the shared multimodal modules and dataset-native task head	Complete-input checkpoints and native metrics for six datasets
Stage 2	Paired input dependence	Comparison of complete, removed, muted, shuffled, and donor conditions using the same checkpoint	Net effects, actual changes, and class- or dimension-level transitions
Stage 3	Independent decodability	Frozen T/A/V representations and grouped-null references	Task-relevant information in each modality representation
Stage 4	Conditional increments and fusion	Calibrated late fusion of T, A, V, TA, TV, AV, and TAV	Combination performance, weights across seeds, and conditional increments
Stage 5	Conditional mechanisms	Stratification by quality, source, class, time reversal and shifting, and donor replacement	Local structure of how modality effects vary with input conditions

Across Stages 1–5, the foundational role of text is most stable in data with variable linguistic content. Table G1 summarizes native task performance, paired ablations, and frozen-representation analyses for each dataset, while Table G2 summarizes fusion and structural-intervention results; classification tasks retain their native class metrics, and CMU-MOSEI retains its six-dimensional continuous metrics.

Table G1. Native task performance, paired ablations, and frozen probes in Stages 1–3.

Dataset	Stage 1 complete input	Stage 2 paired ablation	Stage 3: T/A/V probes
EmoS	W-F1 / Macro-F1 / UAR 0.5807 / 0.4903 / 0.5184	Real A versus silence: +0.0062 Macro-F1; versus source+language mismatch: +0.0079	Macro-F1 0.4290 / 0.1954 / 0.2035
MER2025	Macro-F1 / UAR / Acc 0.5890 / 0.5846 / 0.6471	Remove T / A / V: −0.3358 / −0.0125 / −0.1952 Macro-F1	Macro-F1 0.5070 / 0.3295 / 0.3194
CMU-MOSEI	MAE 0.1695; Pearson 0.3820; mean corr 0.3436; AP 0.4406	Remove T: ΔPearson −0.2222; remove A / V: ΔMAE +0.0257 / +0.0229	Pearson 0.2954 / 0.1911 / 0.1729
IEMOCAP	Macro-F1 / UAR / Acc 0.2880 / 0.2823 / 0.4516	Remove T / A / V: −0.1271 / −0.0869 / −0.0747 Macro-F1	Macro-F1 0.2541 / 0.2200 / 0.1094
RAVDESS	Macro-F1 / UAR / Acc 0.8287 / 0.8242 / 0.8292	Remove A / V: −0.6152 / −0.2244 Macro-F1	Macro-F1 0.0480 / 0.7197 / 0.4857
CREMA-D	Macro-F1 / UAR / Acc 0.6182 / 0.6188 / 0.6199	Remove A / V: −0.2330 / −0.2268; matched shuffle: −0.3140 / −0.3064	Macro-F1 0.0809 / 0.5243 / 0.4568

Table G2. Fusion and structural interventions in Stages 4–5.

Dataset	Stage 4 fusion	Stage 5 conditional mechanisms
EmoS	TAV−TV Macro-F1 $+0.0030$; T / A / V weights 0.8462 / 0.000009 / 0.1538	Quality Q4 / Q3: $+0.0122$ / $+0.0158$; Q2 / Q1: -0.0040 / -0.0152; lexical–semantic coupling 0.1346
MER2025	T / TA / TV / TAV Macro-F1 0.5070 / 0.5361 / 0.5691 / 0.5829; weights 0.5579 / 0.0819 / 0.3603	Text donor -0.2360; visual identity/background proxy -0.1280; A reverse -0.0005
CMU-MOSEI	TV is the best bimodal condition for all five seeds; TAV−TV ΔPearson $+0.0021$; weights 0.3177 / 0.0443 / 0.6380	Text mismatch: ΔPearson -0.1146, ΔAP -0.0786; word-order shuffle: -0.0489 / -0.0401
IEMOCAP	Late-TAV Macro-F1 0.1943, Joint-TAV 0.2880; weights 0.970 / 0.012 / 0.018	Text matched shuffle -0.1328; word order -0.0513; A/V matched shuffle -0.0978
RAVDESS	Late AV=TAV Macro-F1 0.7512; T / A / V weights approximately 0 / 0.4858 / 0.5142	Audio reverse -0.3340; audio half-shift -0.0040; visual reverse -0.0004
CREMA-D	Late AV=TAV Macro-F1 0.6533; T / A / V weights approximately 0 / 0.4574 / 0.5426	Audio reverse -0.1166; the effect varies with class and audio RMS

These original results establish three layers of consistency. When linguistic content varies, text removal or permutation produces the largest structural loss, and frozen text representations are also the strongest; video retains a relatively stable conditional increment in MER2025, CMU-MOSEI, and IEMOCAP; audio is clearly decodable, while its fusion weight and structural-intervention effect vary across datasets. The two fixed-text datasets further delineate the role of semantic variation: when the text weight is near zero, audio and video still support strong native task performance.

Stages 1–5 form a unified analysis chain across RAVDESS, CREMA-D, IEMOCAP, CMU-MOSEI, EmoS, and MER2025. The main text uses the paired results from MER2025, EmoS, and CMU-MOSEI as the core evidence, while the remaining datasets provide references under fixed text, laboratory speech, and native class spaces.

G.3 Audio Late-Fusion Weights

Setting	Audio weight or share	Corresponding structure
RAVDESS	0.4858	Fixed text, native eight classes
CREMA-D	0.4574	Fixed text, native six classes
EmoS	0.000009	Global-simplex fusion
MER2025	0.0819	wav2vec baseline
CMU-MOSEI	0.0443	Mean across five-seed trimodal fusion
IEMOCAP	0.0120	Native ten classes, five folds
MER2025 HuBERT	0.42–0.44	Frozen-feature probes with two visual encoders (G.1)

The audio weight ranges from nearly 0 to approximately 0.49 and changes markedly with the dataset and representation pipeline. EmoS, CMU-MOSEI, and IEMOCAP show strong compression, whereas the HuBERT combinations in the MER2025 frozen-feature probe yield 0.42–0.44; this range is consistent with the instability of audio transitions observed in the main text.

G.4 Independent Text/Audio Branches and Conditional Increment

Diagnostic	Point estimate	95% CI	Meaning
Audio−Text $Q = -$bMSE	-0.0104207	$[-0.0148969, -0.0061890]$	Audio alone is weaker than Text
signed-ridge TA bMSE gain relative to calibrated T	$+0.0046612$	$[0.0024969, 0.0070681]$	A stable audio conditional increment is present after fusion
global-simplex OOF optimal audio weight	0.29	$[0.20, 0.38]$	The outer-video-group audio weight is stably greater than 0
partial A|T macro Pearson	0.0981074	$[0.0686118, 0.1249767]$	After controlling for text, the audio residual remains associated with the target residual

The balanced MSE of signed-ridge TA is 0.0834462, and its macro Pearson correlation is 0.3701163. The intervals in the table are obtained by cluster resampling over 299 source videos.

G.5 Fusion of Independent T/A and Video Branches

Input	MAE	Pearson	Interpretation
T	0.172966	0.364523	Independent text baseline
TA	0.175095	0.379281	Pearson improves while MAE worsens
TAV	0.175653	0.387182	Joint use of the video branch and T/A branches

Under mixed fusion, Pearson increases from T to TA and TAV, while MAE moves in a different direction, showing that correlational structure and absolute error characterize different fusion outcomes. Adding a modality can improve the correlational structure while changing the calibration scale of the predictions.

G.6 Comparison of Decoding Rules at the Same Checkpoint

Emotion-LLaMA rule	DRI	ΔAccuracy	95% CI	Edge distance from the other rule
Free-form generation	38.5%	$+0.110$	$[0.0485, 0.1690]$	$D_{\text{edge}} = 0.953$ CI $[0.920, 1.000]$
Teacher-forced candidate scoring	27.5%	-0.005	$[-0.0472, 0.0388]$	

The 184 responses mappable under both rules yield $D_{\text{edge}} = 0.949$, compared with 0.937 under neutral-fallback mapping. Across the four audio–video combinations on the 200-sample MER2025 subset, the audio Macro-F1 increment ranges from 0.0474 to 0.1676, a 3.54-fold difference; joint changes in the encoder, pooling, and representation dimensionality further demonstrate the sensitivity of audio results to implementation choices.

H Text and Boundary Datasets

Data	Intervention	Intervention−baseline (metric)	Boundary implication
MER joint checkpoint	Remove text	-0.3358 (Macro-F1)	Variable natural language provides the primary semantic reference
CMU-MOSEI	Same-video text mismatch	-0.1146 (Pearson)	Semantic correspondence remains critical after controlling for visual content
CMU-MOSEI	Text word-order perturbation	-0.0489 (Pearson)	Word-order structure contributes beyond the set of lexical items
IEMOCAP	Matched text shuffle	-0.1328 (Macro-F1)	Correspondence in dialogue text still has a measurable effect
RAVDESS, CREMA-D	Fixed/nearly fixed scripts	T/A/V Macro-F1: RAVDESS 0.0480/0.7197/0.4857; CREMA-D 0.0809/0.5243/0.4568	Text varies too little across samples, so audio and visual expression carry the distinctions

Fixed-text data delineate the identification conditions for semantic effects: when lexical content varies little across samples, class differences are carried mainly by vocal and visual expression. The text effects in natural-language data, together with the weak role of text in fixed-script data, show that the semantic reference varies with task content.

I Extended API Results

I.1 Gemini Paired Comparisons

Gemini's MER video-removal comparison covers 198 TAV/TA pairs and 97 source groups, with an accuracy effect of $+9$ to $+10$ percentage points. The requests keep the model ID and prompt fixed and physically remove the video at the input.

Gemini's overall Macro-F1 point effects in the MER and EmoS audio comparisons are $+0.0303$ and $+0.0212$, respectively. MER covers both audio and video removal, while EmoS adds class-level transitions across datasets.

I.2 Extended API Panel

The extended panel includes Gemini 2.5 Pro, Muse Spark 1.1/1.2, and Doubao Seed 2.0 Lite, with Gemini 3.1 Pro preview retained as a same-batch reference. CMU-MOSEI compares T, TA, TV, and TAV on the same 20 source videos; MER2025 and EmoS provide TV-condition results on 20 samples each.

Table I1. Paired-condition MSE on CMU-MOSEI for the extended API panel.

Model	T	TA	TV	TAV	TV−TAV
Gemini 3.1 Pro preview	0.3157	0.2520	0.2387	0.2230	$+0.0157$
Gemini 2.5 Pro	0.3406	0.2852	0.3346	0.3134	$+0.0212$
Muse Spark 1.1	0.3014	0.3405	0.2657	0.2442	$+0.0215$
Muse Spark 1.2	0.3404	0.2951	0.2179	0.2764	-0.0584
Doubao Seed 2.0 Lite	0.3303	0.2967	0.2475	0.2631	-0.0156

Table I2. Macro-F1 under the TV condition on the discrete tasks for the extended API panel.

Model	MER2025	EmoS
Gemini 3.1 Pro preview	0.8135	0.6810
Gemini 2.5 Pro	0.6984	0.7050
Muse Spark 1.1	0.8040	0.6515
Muse Spark 1.2	0.7974	0.5476
Doubao Seed 2.0 Lite	0.8694	0.4807

In CMU-MOSEI, adding audio reduces MSE for Gemini 3.1, Gemini 2.5, and Muse Spark 1.1, but increases it for Muse Spark 1.2 and Doubao; adjacent Muse versions also exhibit opposite directions. The ranking under the TV condition likewise changes between MER2025 and EmoS. The extended panel broadens the system dependence of audio effects to different API families and model versions.