Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Leakage-Aware Multimodal Speech Emotion Recognition For Indian English: Sentence-Level Shortcuts Sarcasm and Uncertainty-Aware Fusion

Rajashree M Byalal*[1], Dr. Sunanda Das[2], Dr. M Kumaresan[3]

[1]Research Scholar, Department of Computer Science & Engineering, Jain (Deemed–to-be-university), Bengaluru, India
[1]Assistant Professor, CSE-ICB Department, K S Institute of Technology, Bangalore, India
[1]Email: btrajashree@gmail.com
[2]Professor, Department of CSE-Cyber security, Jain (Deemed–to-be-university), Bengaluru, India
[3]Associate Professor, School of Computer Science & Engineering, Jain (Deemed–to-be-university), Bengaluru, India
*Corresponding mail: btrajashree@gmail.com

Abstract: Speech emotion recognition remains concentrated on Western-accented English corpora, leaving Indian English, spoken by hundreds of millions, largely unrepresented in resources for conversational agents, call-centre analytics, and affective monitoring systems. Standard speaker-disjoint splitting fails to prevent leakage when scripted corpora reuse a fixed sentence pool: identical transcripts recur across training and test sets, enabling silent lexical memorisation. No existing Indian-English resource includes a sarcastic class, leaving text–prosody conflict and uncertainty-aware fusion untested in this population. This study introduces the Real Sentiment Dataset (RSD), a 28-speaker, six-class Indian-English corpus with a sarcastic category, paired with a Lexical Shortcut Ratio diagnostic and leakage-controlled splits to quantify memorisation bias and benchmark fusion rigorously. Using LoRA-adapted RoBERTa and HuBERT encoders, we compare text-only, audio-only, concatenation, and precision-weighted fusion under speaker-disjoint, sentence-disjoint, and double-disjoint protocols, alongside cross-corpus transfer against IEMOCAP and a sarcasm modality-conflict analysis. Speaker-disjoint evaluation yields an inflated 100.00% text-only accuracy, collapsing to 40.83% once sentence overlap is removed; double-disjoint precision fusion performs best (62.47% WA), cross-corpus transfer remains near chance, and sarcasm raises branch disagreement without shifting fusion trust toward audio.These findings expose an underquantified evaluation flaw in scripted corpora and establish a reproducible foundation for Indian-English affective computing.

Index Terms— Speech emotion recognition, Indian English, data leakage, sentence-disjoint evaluation, multimodal fusion, sarcasm detection, uncertainty-aware fusion, precision-weighted fusion, LoRA, cross-corpus evaluation.

I. Introduction

Speech emotion recognition (SER) research remains heavily concentrated on a small number of Western-accented English corpora, chief among them IEMOCAP [1], while the varieties of English spoken by hundreds of millions of people on the Indian subcontinent are almost entirely unrepresented in publicly benchmarked resources. This gap matters for two reasons. First, prosodic realisation of emotion is accent- and culture-dependent, so models validated only on North-American acted speech carry no deployment guarantee for Indian-English users of conversational agents, call-centre analytics, or affective monitoring systems. Second, the community has recently become aware that evaluation protocol choices on acted corpora can inflate reported accuracy substantially: Antoniou et al. [2] documented that random utterance splits on IEMOCAP leak speaker identity into the test set and exaggerate published numbers. We show in this paper that scripted low-resource corpora suffer from a second, more severe leakage channel that the field's standard remedy speaker-disjoint splitting does not close at all: when every speaker reads the same fixed sentence list, the test transcripts are seen verbatim during training, and a text branch can memorise the sentence-to-emotion mapping perfectly without reading any emotion.

A further blind spot of existing resources is sarcasm. Sarcasm is the canonical case of text–prosody conflict the lexical channel says one thing while the acoustic channel signals another and multimodal sarcasm corpora such as MUStARD [3] have shown that prosodic evidence is essential for its detection. Yet no Indian-English emotional speech corpus includes a sarcastic class, so the question of whether uncertainty-aware fusion mechanisms actually reallocate trust toward prosody under lexical–acoustic conflict has never been testable in this population. This paper is a direct continuation of our companion study, Byalal, Das, and Kumaresan [4], published in the International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM), which introduced PWF-FT, a precision-weighted fusion framework whose per-utterance

inverse-variance modality weights are explicitly interpretable, and hypothesised that such weights should shift toward audio on sarcastic inputs; this paper provides the first direct empirical test of that hypothesis.

This paper makes the following contributions:

1. A new resource. We present the Real Sentiment Dataset (RSD), a 28-speaker, 3,363-utterance Indian-English emotional speech corpus built from a deterministic speaker×sentence recording grid, with six classes angry, happy, sad, neutral, excited, and, uniquely, sarcastic recorded from 120 shared scripted sentences with per-speaker metadata.
2. A leakage diagnostic. We define the Lexical Shortcut Ratio (LSR), the ratio of a text-only model's accuracy under speaker-disjoint versus double-disjoint (speaker-and-sentence-disjoint) splits, and show LSR = 2.45 on our corpus: a frozen text probe scores a perfect 100.00% WA under the field-standard speaker-disjoint protocol but only 40.83% once test sentences are excluded from training, invalidating text and fusion claims made without sentence-disjoint evaluation.
3. Leakage-controlled multimodal benchmarks. We benchmark LoRA-adapted RoBERTa and HuBERT encoders with attention pooling under four fusion configurations (text-only, audio-only, concatenation, and precision-weighted fusion) across sentence-disjoint and double-disjoint protocols with multi-seed replication, and report bidirectional cross-corpus transfer against IEMOCAP on the four shared classes.
4. A modality-conflict analysis of sarcasm. Using the precision gate's explicit per-utterance modality weights, we show that sarcasm reliably increases text–audio branch disagreement but does not shift fusion trust toward audio a reproducible negative result that constrains how uncertainty-weighted fusion should be interpreted.

The remainder of the paper is organised as follows. Section II reviews related work; Section III describes the Real Sentiment Dataset, its construction, and its novelty; Section IV presents the split protocols, the LSR diagnostic, and the fusion architecture; Section V details the experimental setup; Section VI reports and discusses results; and Section VII concludes.

II. Related Work

A. Multimodal SER and Evaluation Protocol Integrity

Fusion of pretrained text and speech encoders defines the current state of practice on IEMOCAP. Zou et al. [5] combined MFCC, spectrogram, and wav2vec 2.0 streams through co-attention, reporting 69.80% WA under leave-one-session-out evaluation; Zhao et al. [6] fused speech and text representations at multiple granularities with knowledge-aware Bayesian co-attention; and Ye et al. [7] showed that purely acoustic temporal modelling (TIM-Net) reaches 68.29% WA. Speaker-aware adaptations of self-supervised speech encoders [8] and the frozen-probe results of the SUPERB benchmark [9] bracket the speaker-independent state of practice between roughly 67% and 73% WA on the four-class task. Critically, this entire literature treats speaker disjointness as the gold-standard leakage control, following the protocol audit of [2]. Our results show that for scripted corpora this control is insufficient: sentence identity is a second leakage channel, orthogonal to speaker identity, that speaker-disjoint splitting leaves fully open.

B. Uncertainty-Aware and Trustworthy Fusion

Guo et al. [10] established that modern networks are systematically miscalibrated, and Kendall and Gal [11] introduced the heteroscedastic aleatoric-variance heads that our fusion mechanism inherits. Within affective computing, COLD fusion [12] learns calibrated latent-distribution variances for audiovisual emotion and demonstrates robustness gains under modality corruption, while trusted multi-view classification [13] fuses Dirichlet opinions at the evidence level. Schrüfer et al. [14] benchmarked uncertainty quantification for in-the-wild SER and found even simple entropy signals informative for reliability. Precision-weighted fusion, in which per-modality log-variance heads produce inverse-variance decision weights, is the simplest member of this family [4], its distinguishing property is that the weights are directly auditable, which is precisely what our sarcasm analysis exploits.

C. Annotator Disagreement and Soft Labels

Uma et al. [15] surveyed learning from annotator disagreement across NLP and vision and concluded that label distributions carry systematic perceptual signal rather than noise, and soft-target training has been advocated for SER specifically because emotional expression is genuinely ambiguous [16]. Our corpus release plan includes a perceptual validation study (Fleiss-κ [28, 29] inter-rater agreement over a stratified 300-clip sample) whose tooling is included with the corpus; the acted-versus-perceived validity of the sarcastic class in particular is a question that only human raters can settle, and we flag it as the i[mmediate next step rather than a completed component of this work.

D. Multimodal Sarcasm Detection

MUStARD [3] established that sarcasm detection benefits substantially from acoustic and visual context beyond text, and subsequent work has modelled cross-modal incongruity explicitly: the incongruity-aware attention network of Wu et al. [17] scores word-level conflict between spoken content and tone, and MuLOT [18] learns cross-modal incongruity via optimal transport under the small-data constraint that MUStARD's 690 utterances impose. MUStARD++ [19] extended the resource with richer annotation. All of these corpora are Western-media excerpts; our corpus contributes the first scripted Indian-English sarcastic class, and unlike detection-focused work we use sarcasm diagnostically, asking whether an uncertainty-driven fusion gate reacts to the text–prosody conflict sarcasm embodies.

E. Parameter-Efficient Fine-Tuning

Full fine-tuning of two transformer encoders is prohibitive on commodity hardware. LoRA [20] learns low-rank updates to frozen attention projections, and PEFT-SER [21] identified LoRA as the strongest parameter-efficient strategy for adapting speech encoders to emotion recognition. Conceptually closest to the resource-allocation view taken here is SCALE [22], which profiles transformer layers by activation magnitude and attention entropy and adapts only the most influential ones with lightweight adapters, retaining up to 99% of full fine-tuning performance on BERT-base sentiment analysis with roughly 40% of the parameters; SCALE allocates adaptation capacity by internal signal statistics at the layer level, whereas the precision-weighted fusion evaluated here allocates decision weight by internal uncertainty statistics at the utterance level two instantiations of the same principle that signal statistics, not uniform heuristics, should govern where capacity and trust are placed. We adopt plain LoRA on the query/key/value projections of RoBERTa [23] and HuBERT [24, 30] as the reproducible baseline instantiation.

F. Indian-Language Emotional Speech Resources

Indian-language SER resources remain scarce and predominantly non-English: IITKGP-SEHSC [25] provides acted Hindi emotional speech, and recent cross-corpus studies over Hindi, Urdu, Telugu, and Kannada report accuracies of 45–70% with handcrafted acoustic features and CNNs [26], underscoring both the difficulty and the fragmentation of the space. Code-mixed Hindi-English emotion recognition has been organised as a shared task on conversational text [27], but no existing resource offers scripted Indian-English speech with a sarcastic class and speaker-disjoint and sentence-disjoint evaluation support. Our corpus and protocol suite target exactly this gap.

III. The Real Sentiment Dataset

A. Design and Construction

The Real Sentiment Dataset (RSD) is a self-collected multimodal emotional speech corpus created by the authors specifically for this research work, with recordings contributed by volunteer speakers, to address the scarcity of Indian-English resources in emotion recognition research. It comprises 3,363 utterances recorded by 28 speakers of Indian English drawn from a South Indian speaker population, spanning six emotion categories: happy (758), angry (644), sad (644), neutral (336), excited (420), and sarcastic (561). RSD was constructed as a controlled, scripted elicitation study: a pool of 120 scripted sentences drawn from everyday situations was authored such that each sentence carries exactly one intended target emotion, with the intended emotion enacted per prompt; the sarcastic sentences were written so that their lexical content is congruent with a literal (frequently positive) reading while the intended delivery inverts it, the defining lexical–prosodic incongruity of sarcasm [3]. Each volunteer recorded the shared sentence pool in their own recording sessions, and every recording is paired with its transcript, enabling genuine text–audio multimodal study on the same utterance; one recording was discarded for a missing waveform. The corpus is organised in a folder-per-speaker layout, with each speaker directory containing the recordings and a metadata file linking every waveform to its scripted sentence and intended emotion label; label spellings and metadata schemas are normalised programmatically at load time (e.g., sarcasm→sarcastic, joy→happy), and all audio is resampled to 16 kHz and truncated to 6 s for modelling. Key factors shaping the corpus design were speaker diversity, class coverage, controlled prompt design, and a consistent recording structure, chosen expressly to support the speaker-independent and sentence-disjoint evaluation protocols that unbiased multimodal benchmarking requires. Table I summarises the statistics and Fig. 1 shows the class distribution: moderately imbalanced (happy 758 vs. neutral 336), with the sarcastic class contributing 561 utterances.

B. Novelty of the Resource

The principal novelty of RSD is threefold. First, population: Indian English is spoken by hundreds of millions of people yet is essentially absent from benchmarked emotional speech resources, which are dominated by North-American acted corpora [1] and, on the Indian side, by non-English languages such as Hindi [25]; to our knowledge RSD is the first multi-speaker Indian-English emotional speech corpus with paired scripted transcripts. Second, class inventory: no existing Indian-English (and, to our knowledge, no scripted English) emotional speech corpus includes a dedicated sarcastic class; existing sarcasm resources such as MUStARD [3] and MUStARD++ [19] are scraped Western-media excerpts rather than controlled recordings, so RSD is, to our knowledge, the first resource in which sarcastic prosody—a pragmatic category in which words and tone

deliberately conflict is elicited under the same recording conditions, speakers, and sentence-design discipline as the five conventional emotion classes, providing a natural testbed for the controlled modality-disagreement analysis of Section VI-D. Third, diagnostic-by-design structure: the parallel scripted design—all speakers reading identical sentences under a deterministic one-sentence→one-emotion mapping was retained deliberately rather than randomised away, because this structure makes the corpus uniquely suited to diagnosing lexical-shortcut bias, where text models memorise sentence–label mappings rather than learning emotion the LSR diagnostic of Section IV-B is computable on RSD by construction, and the corpus doubles as a stress test for any evaluation protocol that claims leakage control. The speaker×sentence grid design (each speaker × each sentence) is also what makes true double-disjoint splitting well-defined, which scraped in-the-wild corpora cannot generally support.

TABLE I. REAL SENTIMENT DATASET (RSD) STATISTICS

Property	Value
Utterances	3,363 (1 recording discarded: missing waveform)
Speakers	28 Indian-English speakers (folder-per-speaker, per-speaker metadata)
Classes (6)	angry / happy / sad / neutral / excited / sarcastic
Class counts	644 / 758 / 644 / 336 / 420 / 561
Unique scripted sentences	120, shared across all speakers; one intended emotion per sentence
Audio	16 kHz, truncated to 6 s
Text	scripted transcripts, max 64 subword tokens
Shared 4-class subset (vs. IEMOCAP)	angry / happy(+excited) / sad / neutral

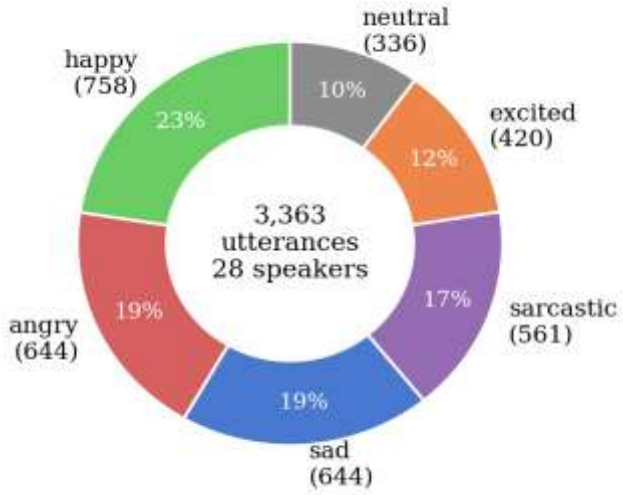

Fig. 1. Class distribution of the Real Sentiment Dataset: 3,363 utterances over six classes from 28 speakers. The sarcastic class (561 utterances) is, to our knowledge, the first in any Indian-English emotional speech resource.

The design property that motivates this paper's methodology is the deliberate sentence sharing: because all 28 speakers read from the same 120 sentences and each sentence maps to exactly one intended emotion, any split that separates speakers but not sentences places every test transcript together with its label verbatim in the training set. Section IV formalises the diagnostic that quantifies the resulting shortcut and the split protocols that close it. For cross-corpus experiments RSD is mapped onto the four classes shared with IEMOCAP [1] angry, happy (absorbing excited), sad, and neutral with sarcastic excluded from transfer training but retained for the conflict analysis of Section VI-D.

IV. Proposed Methodology

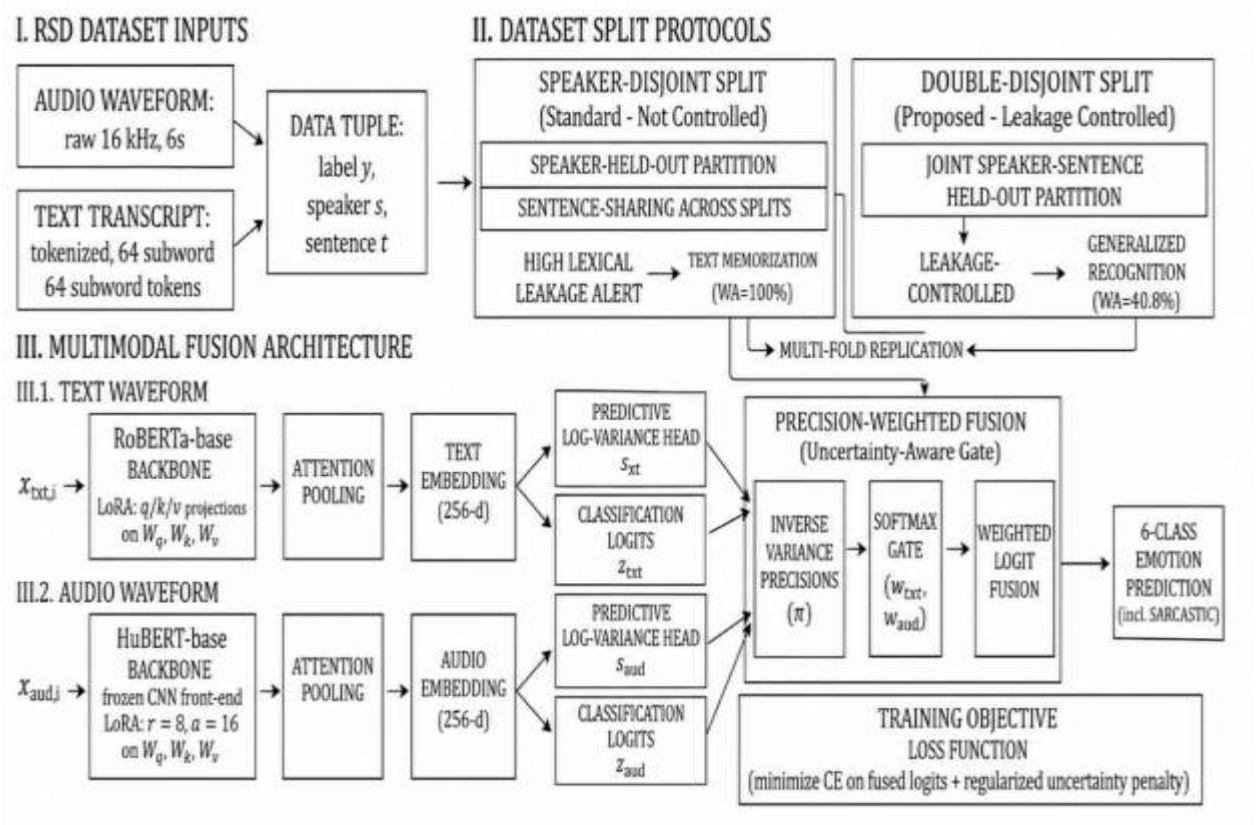

Fig. 2. Proposed Precision-Weighted Multimodal Emotion Recognition Framework
Overview of the proposed framework comprising (i) RSD dataset inputs, (ii) leakage-controlled double-disjoint dataset split protocol, and (iii) uncertainty-aware precision-weighted fusion of RoBERTa and HuBERT embeddings for robust six-class emotion recognition.

A. Formal Dataset Definition and Split Protocols

To mathematically formulate the data leakage channels inherent to scripted speech emotional resources, we define the multimodal emotional speech corpus as a structural index set:

$$D = \{(x_{\text{aud},i}, x_{\text{txt},i}, y_i, s_i, t_i)\}_{i=1}^{N} \quad (1)$$

Where $N = 3{,}363$ represents the absolute sequence volume of valid utterances contained within the Real Sentiment Dataset (RSD). For each individual discrete utterance $i \in \{1,2, \dots, N\}$, the constituent feature matrices and tracking variables are formalised as follows:

$x_{\text{aud},i}$ is the raw localized 16 kHz acoustic single-channel waveform tensor, truncated to a strict upper boundary duration of exactly 6 s (96,000 discrete sampling points) to guarantee uniform spatial tracking across the audio branch batch matrices. $x_{\text{txt},i}$ represents the corresponding literal scripted textual transcript sequence, tokenized via subword byte-pair encoding to a maximum fixed processing sequence length of 64 subword tokens.$y_i \in Y$ is the categorical affective ground-truth label mapping directly into a discrete 6-class emotion space where: $Y = \{\text{angry,happy,sad,neutral,excited,sarcastic}\}$ $s_i \in S$ denotes the unique speaker identity factor drawn from the finite categorical pool of 28 South Indian regional actors ($|S| = 28$), mapping demographic and index features. $t_i \in T$ is the deterministic scripted sentence prompt token index selected from a parallel set of 120 uniquely authored everyday situational expressions ($|T| = 120$).

The experimental design utilizes these factors to construct three distinct split protocols to isolate and evaluate overlapping dataset shortcut vulnerabilities:

1. Speaker-Disjoint Splitting

This protocol reflects the traditional validation technique used across the Speech Emotion Recognition (SER) community. The complete speaker population pool S is partitioned into three mutually exclusive subsets:

$$S = S_{\text{train}} \cup S_{\text{val}} \cup S_{\text{test}} \quad \text{such that} \quad S_{\text{train}} \cap S_{\text{val}} \cap S_{\text{test}} = \emptyset \quad (2)$$

The absolute mapping of data points into evaluation matrices strictly follows these speaker tracking parameters:

$$D_{\text{test}} = \{i \in D \mid s_i \in S_{\text{test}}\} \quad (3)$$

Crucially, because the text prompt sequence layout is completely unconstrained during this allocation routine, the sentence mapping remains uniform across all splits ($T_{\text{train}} = T_{\text{val}} = T_{\text{test}} = T$). Consequently, every script token evaluated in the test partition is guaranteed to appear verbatim in the training matrix across the other remaining speakers, creating a high-risk text leakage channel.

2. Sentence-Disjoint Splitting

To isolate and measure the effects of text prompt leakage independently of speaker factors, the sentence prompt tracking index T is partitioned into three disjoint subsets:

$$T = T_{\text{train}} \cup T_{\text{val}} \cup T_{\text{test}} \quad \text{such that} \quad T_{\text{train}} \cap T_{\text{val}} \cap T_{\text{test}} = \emptyset \tag{4}$$

Data allocation to the test set is governed by the underlying prompt sequence:

$$D_{\text{test}} = \{i \in D \mid t_i \in T_{\text{test}}\} \tag{5}$$

Under this validation setup, the speaker tracking space remains completely unconstrained across the training and testing matrices ($S_{\text{train}} = S_{\text{val}} = S_{\text{test}} = S$), allowing personal identity variables to blend across partitions while blocking text phrase memorization.

3. Double-Disjoint Splitting

The strictest evaluation configuration enforces an absolute zero-overlap constraint across both structural dimensions simultaneously. The evaluation subset is formed by the Cartesian intersection of held-out speakers and held-out sentences:

$$D_{\text{test}} = \{i \in D \mid s_i \in S_{\text{test}} \wedge t_i \in T_{\text{test}}\} \tag{6}$$

The corresponding training partition is constructed from the absolute remaining complement across both identity fields:

$$D_{\text{train}} = \{i \in D \mid s_i \in S_{\text{train}} \wedge t_i \in T_{\text{train}}\} \tag{7}$$

The validation monitoring slice is assigned from the remaining diagonal cross-cells:

$$D_{\text{val}} = \{i \in D \mid s_i \in S_{\text{val}} \wedge t_i \in T_{\text{val}}\} \tag{8}$$

This double-disjoint protocol eliminates acoustic identity leakage and lexical text shortcuts at the same time, providing an unbiased benchmark for evaluating multimodal fusion performance.

B. Lexical Shortcut Ratio (LSR)

To evaluate how much a text model relies on memorizing fixed text prompts rather than learning generalized emotional markers, we implement a linear probing technique. A baseline classification head is optimized over a static repository of frozen text embeddings across different split protocols. We formalize the Lexical Shortcut Ratio (LSR) as the quotient:

$$\text{LSR} = \frac{\text{WA}_{\text{text}}^{(\text{speaker-disjoint})}}{\text{WA}_{\text{text}}^{(\text{double-disjoint})}} \tag{9}$$

Where $\text{WA}_{\text{text}}^{(\text{split})}$ represents the exact Weighted Accuracy performance score achieved by the text model configuration under the designated partition rules.

The conceptual interpretation of the resulting scalar values is defined by two operational boundaries:

a. LSR ≈ 1.0: Indicates a robust dataset where text prompt structures carry zero hidden classification shortcuts, meaning text-based emotional recognition remains stable across different evaluation formats.
b. LSR ≥ 1.3: Highlights a severe dataset vulnerability. This indicates that high accuracy scores within speaker-disjoint splits are inflated by transcript-memorization effects rather than true affective classification performance.

C. Encoders, Pooling, and Fusion Architecture

The benchmarking network utilizes a dual-branch late-fusion design that processes acoustic and textual sequences through specialized deep architectures into an uncertainty-calibrated joint classifier, as summarised in Fig. 2.

1. Low-Rank Adaptation (LoRA) of Transformer Encoders

To adapt the massive pre-trained transformer backbones without experiencing catastrophic forgetting or unmanageable hardware overhead during training, we integrate Low-Rank Adaptation (LoRA) modules across the query, key, and value inner attention projection matrices (W_q, W_k, W_v). Let $H^{(m)}$ denote the localized hidden sequence matrices computed for each primary modality branch $m \in \{\text{txt,aud}\}$:

$$H^{(\text{txt})} = \text{RoBERTa-base}\left(x_{\text{txt}}; \Theta_{\text{base}}^{(\text{txt})} + \Delta\Theta_{\text{LoRA}}^{(\text{txt})}\right) \in \mathbb{R}^{T_{\text{txt}} \times 768} \tag{10}$$

$$H^{(\text{aud})} = \text{HuBERT-base}\left(x_{\text{aud}}; \Theta_{\text{base}}^{(\text{aud})} + \Delta\Theta_{\text{LoRA}}^{(\text{aud})}\right) \in \mathbb{R}^{T_{\text{aud}} \times 768} \tag{11}$$

Where Θ_{base} defines the frozen parameters of the pre-trained models, and T_{txt} and T_{aud} capture the variable sequence frames across text and audio. For any native transformer projection mapping $W_0 \in \mathbb{R}^{d \times k}$, the LoRA modification decomposes the weight correction matrix as follows:

$$W = W_0 + \Delta W = W_0 + \frac{\alpha}{r} BA \tag{12}$$

where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ represent the low-rank learnable parameter adapter spaces under a strict inner rank size $r = 8$ and a static constant scaling multiplier $\alpha = 16$.

2. Sequence Masked Attention Pooling

To compress the variable temporal sequence frames into fixed-size embedding spaces without losing contextual information, we implement a sequence-masked attention pooling layer. Let $h_t^{(m)} \in \mathbb{R}^{768}$ represent the individual

state vector extracted at timestep sequence index t. A two-layer feed-forward network generates an unnormalized continuous scalar attention alignment value $s_t^{(m)}$:

$$s_t^{(m)} = w_2^{(m)^\top} \cdot tanh\left(W_1^{(m)} h_t^{(m)} + b_1^{(m)}\right) \quad (13)$$

Where the internal parameters are shaped as $W_1^{(m)} \in \mathbb{R}^{768\times768}$, $b_1^{(m)} \in \mathbb{R}^{768}$, and $w_2^{(m)} \in \mathbb{R}^{768}$. A sequence-masked softmax operation normalizes these values into temporal importance coefficients $a_t^{(m)}$:

$$a_t^{(m)} = \frac{exp\left(s_t^{(m)}\right)}{\sum_{\tau=1}^{T_m} exp\left(s_\tau^{(m)}\right)} \quad (14)$$

The time-series sequences are compressed via their inner attention weights into a single vector $\bar{h}^{(m)}$, which is projected into a final 256-dimensional space using Layer Normalization (LN) and a GELU activation layer:

$$\bar{h}^{(m)} = \sum_{t=1}^{T_m} a_t^{(m)} \cdot h_t^{(m)} \in \mathbb{R}^{768} \quad (15)$$

$$u^{(m)} = \text{GELU}\left(\text{LN}\left(W_p^{(m)} \bar{h}^{(m)} + b_p^{(m)}\right)\right) \in \mathbb{R}^{256} \quad (16)$$

Where $W_p^{(m)} \in \mathbb{R}^{256\times768}$ and $b_p^{(m)} \in \mathbb{R}^{256}$ define the projection layer matrices.

3. Heteroscedastic Predictive Heads and Precision Weighting

The structural core of the fusion model relies on dynamically mapping data uncertainty. Each modality embedding $u^{(m)}$ is routed simultaneously to a categorical target classification head and an identity log-variance tracking circuit:

$$z_m = W_z^{(m)} u^{(m)} + b_z^{(m)} \in \mathbb{R}^{|Y|} \quad (17)$$

$$s_m = w_s^{(m)^\top} u^{(m)} + b_s^{(m)} \in \mathbb{R} \quad (18)$$

Where $W_z^{(m)} \in \mathbb{R}^{6\times256}$ and $w_s^{(m)} \in \mathbb{R}^{256}$. The learned scalar value s_m represents the data aleatoric log-variance ($log\ \sigma_m^2$). The absolute single-channel predictive precision metric is derived via $\pi_m = \frac{1}{\sigma_m^2} = exp(-s_m)$. To perform fine-grained weighting, a softmax activation scales the modalities directly inside the negative log-variance domain:

$$(w_{\text{txt}}, w_{\text{aud}}) = \text{softmax}(-s_{\text{txt}}, -s_{\text{aud}}) \quad (19)$$

$$w_m = \frac{exp(-s_m)}{exp(-s_{\text{txt}}) + exp(-s_{\text{aud}})} = \frac{\pi_m}{\pi_{\text{txt}} + \pi_{\text{aud}}} \quad (20)$$

The combined logit array z_{fused} is produced by a precision-weighted convex linear combination:

$$z_{\text{fused}} = w_{\text{txt}} \cdot z_{\text{txt}} + w_{\text{aud}} \cdot z_{\text{aud}} \quad (21)$$

D. Training Objective Function

The network parameters are optimized end-to-end by minimizing a combined objective loss function L_{total} over each training batch:

$$L_{\text{total}} = H_{\text{CE}}(z_{\text{fused}}, y) + \alpha \sum_{m\in\{\text{txt,aud}\}} \left[\frac{exp(-s_m)}{2} H_{\text{CE}}(z_m, y) + \frac{s_m}{2}\right] \quad (22)$$

Where: $H_{\text{CE}}(z, y) = -log\left(\frac{exp(z[y])}{\sum_c exp(z[c])}\right)$ represents the standard categorical cross-entropy loss function. $\alpha = 0.5$ serves as the scaling factor for the auxiliary unimodal regularizers. The variance-scaling term $\frac{exp(-s_m)}{2} \equiv \frac{1}{2\sigma_m^2}$ acts as an adaptive regularizer. This expression automatically reduces the loss contribution of a specific branch when processing highly ambiguous inputs, such as conflicting sarcasm tokens. The linear penalty expression $\frac{s_m}{2}$ functions as an absolute mathematical barrier, preventing the model from collapsing by choosing an infinite variance state ($\sigma_m^2 \to \infty$) across all inputs.

E. Cross-Corpus Transfer & Sarcasm Modality-Conflict Analysis

To track the internal properties of the model during conversational modality conflicts (such as sarcastic expressions), the system tracks the Acoustic Gate Predictive Shift (Δ) metric:

$$\Delta = \bar{w}_{\text{aud}}^{(\text{sarcastic})} - \bar{w}_{\text{aud}}^{(\text{non-sarcastic})} \quad (23)$$

Where the structural mean audio gate weight over any given conditional data slice $D_{\text{sub}} \subseteq D$ is defined as:

$$\bar{w}_{\text{aud}}^{(D_{\text{sub}})} = \frac{1}{|D_{\text{sub}}|} \sum_{i\in D_{\text{sub}}} w_{\text{aud},i} \quad (24)$$

The structural conflict between the processing branches is evaluated by calculating the absolute Branch Disagreement Frequency Rate (R_{disagree}):

$$R_{\text{disagree}} = \frac{1}{|D_{\text{sub}}|} \sum_{i\in D_{\text{sub}}} \mathbb{I}\left(arg\ max_{c\in Y} z_{\text{txt},i}[c] \neq arg\ max_{c\in Y} z_{\text{aud},i}[c]\right) \quad (25)$$

Where $\mathbb{I}(\cdot)$ is the standard mathematical Kronecker indicator function mapping to {0,1} based on the truth value of the internal conditional statement. This metric tracks how often the language and speech processing paths produce conflicting predictions on complex emotional data.

ALGORITHM 1. Leakage-Controlled Benchmark Protocol

1. **Input:** Corpus D with speaker indices s_i and sentence indices t_i; protocols $P =$ {sentence-disjoint,double-disjoint}; fusion modes $M =$ {text_only,audio_only,concat,precision}; folds $F = 5$; seeds $R = \{42,43,44\}$.
2. **Output:** WA/UA/wF1 per $(p, m) \in P \times M$, mean $\pm$ std over $F \times R$; LSR; cross-corpus WA/UA; sarcasm disagreement and gate shift.
3. **LSR diagnostic (frozen features):** Cache frozen RoBERTa and HuBERT features. Train the text-only head under speaker-disjoint and double-disjoint splits; compute LSR by Eq. (9).
4. **for** each protocol $p \in P$ **do**
5. **for** each fold $f = 1 \ldots F$ and seed $r \in R$ **do**
   Partition D by p; initialize LoRA adapters ($r = 8, \alpha = 16$), pooling, and heads; freeze backbone weights and the HuBERT CNN front-end.
   Train each mode $m \in M$ for up to 8 epochs minimizing Eq. (22) using AdamW, encoder LR $= 3 \times 10^{-4}$, head LR $= 10^{-3}$, 10% warm-up, batch 12×2 accumulation, AMP, and patience 3 on validation UA.
6. Evaluate WA, UA, and wF1 on the held-out double/sentence-disjoint test cells.
7. **Cross-corpus:** Map both corpora to the shared 4 classes; for each seed, train on source and evaluate zero-shot on target, in both directions.
8. **Sarcasm conflict:** For each seed, train precision fusion on IEMOCAP; on the corpus, compute branch disagreement and $\bar{w}_{\text{aud}}$ for sarcastic vs. non-sarcastic subsets; report gate shift Δ.

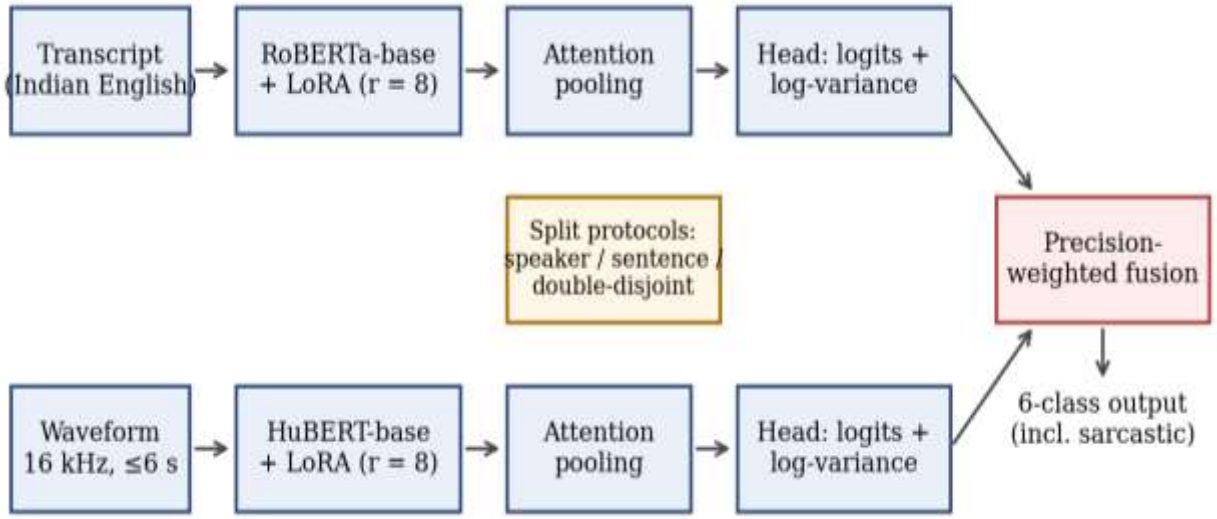

Fig. 3. Overview of the benchmark pipeline: LoRA-adapted RoBERTa and HuBERT branches with attention pooling feed per-modality classification and log-variance heads, fused by per-utterance precision weights; all training and evaluation is governed by the speaker/sentence/double-disjoint split protocols of Section IV-A.

V. Experimental Setup

Fig. 3 summarises the full benchmark pipeline described below. Data, All within-corpus experiments use the six-class corpus of Section III; cross-corpus experiments use the 6,877-utterance four-class IEMOCAP subset prepared exactly as in [4] (excited merged into happy). Transcripts are tokenised to at most 64 subword tokens; audio is truncated to 6 s at 16 kHz. Frozen-feature diagnostics. Utterance-level features are extracted once from frozen RoBERTa-base and HuBERT-base (768-d each) and cached; heads (hidden 256, dropout 0.3) are trained with Adam (LR 10^{-3}, weight decay 10^{-4}, batch 64, up to 40 epochs, patience 8) over 5 folds per protocol. The frozen setting makes the full protocol grid re-runnable in minutes and is used for the LSR diagnostic and protocol comparison. Fine-tuned benchmarks. The LoRA configuration and optimisation hyperparameters follow Algorithm 1 and are identical to the Paper-1 pipeline [4]: rank 8, α = 16, dropout 0.05 on q/k/v projections of both encoders, frozen HuBERT feature extractor, gradient checkpointing, AMP, effective batch size 24, and discriminative learning rates. Each (protocol, mode, fold, seed) cell is a full LoRA training run; results are reported as mean ± std over 5 folds × 3 seeds (42–44) for within-corpus and over 3 seeds for cross-corpus. Chance rates are 16.7% (6-class) and 25% (4-class). Metrics. Weighted accuracy (WA), unweighted accuracy

(UA, macro recall), and weighted F1 (wF1). For the sarcasm analysis we additionally report the branch-disagreement rate and the mean audio gate weight defined in Section IV-E.

VI. Results and Discussion

A. The Lexical Shortcut: Speaker-Disjoint Evaluation Is Invalid on This Corpus

Table II reports the frozen-feature diagnostic across all three protocols, and Fig. 4 visualises each system's degradation as the protocols tighten. Under the field-standard speaker-disjoint protocol, the frozen text-only probe and every fusion configuration that can see text scores a perfect 100.00±0.00% WA on the six-class task, while the audio-only branch scores 70.37%. Under the double-disjoint protocol the same text probe collapses to 40.83±8.08%, yielding LSR = 2.45 by Eq. (9), far beyond the 1.3 reporting threshold. The interpretation is unambiguous: with 120 sentences shared across 28 speakers, a linear head on frozen RoBERTa features can memorise the sentence-to-emotion mapping outright, and speaker-disjoint results on this corpus measure transcript recall, not emotion recognition. Because concat and precision fusion inherit the text branch, their speaker-disjoint scores are equally saturated and equally meaningless. This finding generalises a known concern: the protocol audit of [2] targeted speaker leakage on IEMOCAP, but for the entire class of scripted, prompt-shared corpora that dominates low-resource SER, sentence identity is a second and here strictly dominant leakage channel. All subsequent benchmarks in this paper therefore use the sentence-disjoint and double-disjoint protocols only.

TABLE II. FROZEN-FEATURE DIAGNOSTIC ACROSS SPLIT PROTOCOLS (6-CLASS, MEAN ± STD OVER 5 FOLDS, %)

Protocol	System	WA	UA
Speaker-disjoint	text_only	100.00 ± 0.00	100.00 ± 0.00
Speaker-disjoint	audio_only	70.37 ± 5.54	72.04 ± 5.05
Speaker-disjoint	concat	100.00 ± 0.00	100.00 ± 0.00
Speaker-disjoint	precision	100.00 ± 0.00	100.00 ± 0.00
Sentence-disjoint	text_only	44.99 ± 10.34	48.47 ± 4.42
Sentence-disjoint	audio_only	34.40 ± 8.31	36.35 ± 7.68
Sentence-disjoint	concat	48.09 ± 10.21	51.98 ± 6.89
Sentence-disjoint	precision	45.06 ± 8.18	48.33 ± 5.50
Double-disjoint	text_only	40.83 ± 8.08	43.30 ± 12.72
Double-disjoint	audio_only	29.79 ± 8.64	30.13 ± 6.12
Double-disjoint	concat	40.76 ± 10.38	38.82 ± 14.58
Double-disjoint	precision	43.47 ± 7.95	41.28 ± 7.64

Lexical Shortcut Ratio: LSR = 100.00 / 40.83 = 2.45 (Eq. 9).

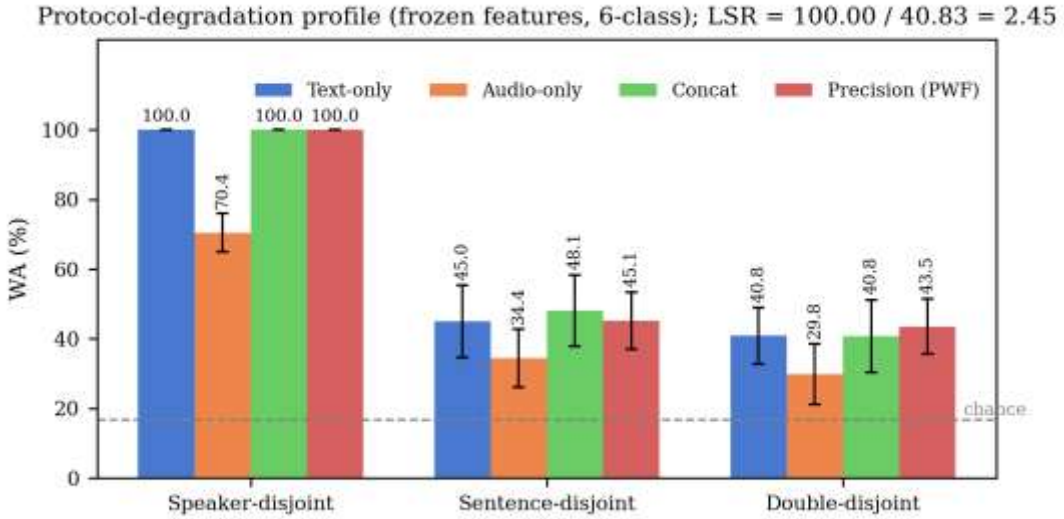

Fig. 4. Grouped bars of all four frozen-feature systems across the three split protocols (values on bars, mean ± std over 5 folds). Every text-bearing system is saturated at 100% under speaker-disjoint splits and collapses once test sentences are withheld, yielding LSR = 2.45 for the text-only probe; the audio-only system degrades far more gently, confirming that the shortcut is lexical.

B. Fine-Tuned Within-Corpus Benchmarks Under Leakage-Controlled Protocols

Table III and Fig. 5 present the fine-tuned LoRA benchmarks under the two leakage-controlled protocols, aggregated over 5 folds × 3 seeds. Three observations follow. First, absolute performance is modest by design: 55–62% WA on a six-class task against a 16.7% chance rate is what leakage-free evaluation of a 28-speaker scripted corpus yields, and the contrast with the 100% speaker-disjoint ceiling of Table II is itself the central empirical message of this paper. Second, under the strictest double-disjoint protocol, precision-weighted fusion is the strongest system at 62.47±8.57% WA (60.22±10.38% UA, 62.40±9.72% wF1), ahead of text-only (59.97%) and concat (59.25%); under sentence-disjoint splits the ordering is the same with a compressed margin (57.75% vs. 56.10% concat and 55.24% text-only). This inverts the finding of the companion IEMOCAP study [4], where precision fusion cost roughly two WA points against learned gating: on a small, high-conflict corpus with a sarcastic class, per-utterance down-weighting of the unreliable modality appears to help rather than cost, consistent with the robustness motivation of uncertainty-aware fusion [12], [13]. Third, the audio-only branch remains weak (28.62% double-disjoint), confirming that in this corpus prosody alone carries limited class information for the frozen-then-adapted HuBERT pipeline and that the fusion gains are driven by complementarity rather than either branch's individual strength. Fold-to-fold standard deviations are large (±5–11 points); this is an unavoidable property of double-disjoint cells carved from 28 speakers × 120 sentences, and it is exactly why every number in Table III is a seeds × folds mean rather than a single-split result.

TABLE III. FINE-TUNED WITHIN-CORPUS RESULTS, 6-CLASS (MEAN ± STD OVER 5 FOLDS × 3 SEEDS, %)

Protocol	System	WA	UA	wF1
Double-disjoint	text_only	59.97 ± 10.67	58.16 ± 11.32	59.10 ± 11.25
Double-disjoint	audio_only	28.62 ± 4.57	28.63 ± 5.71	27.10 ± 5.65
Double-disjoint	concat	59.25 ± 7.53	56.69 ± 9.52	58.57 ± 7.66
Double-disjoint	precision (ours)	62.47 ± 8.57	60.22 ± 10.38	62.40 ± 9.72
Sentence-disjoint	text_only	55.24 ± 10.48	54.77 ± 9.39	54.51 ± 10.24
Sentence-disjoint	audio_only	33.04 ± 7.43	34.17 ± 8.81	32.74 ± 7.91
Sentence-disjoint	concat	56.10 ± 9.58	55.69 ± 8.24	55.44 ± 10.00
Sentence-disjoint	precision (ours)	57.75 ± 8.72	56.93 ± 4.95	56.67 ± 9.44

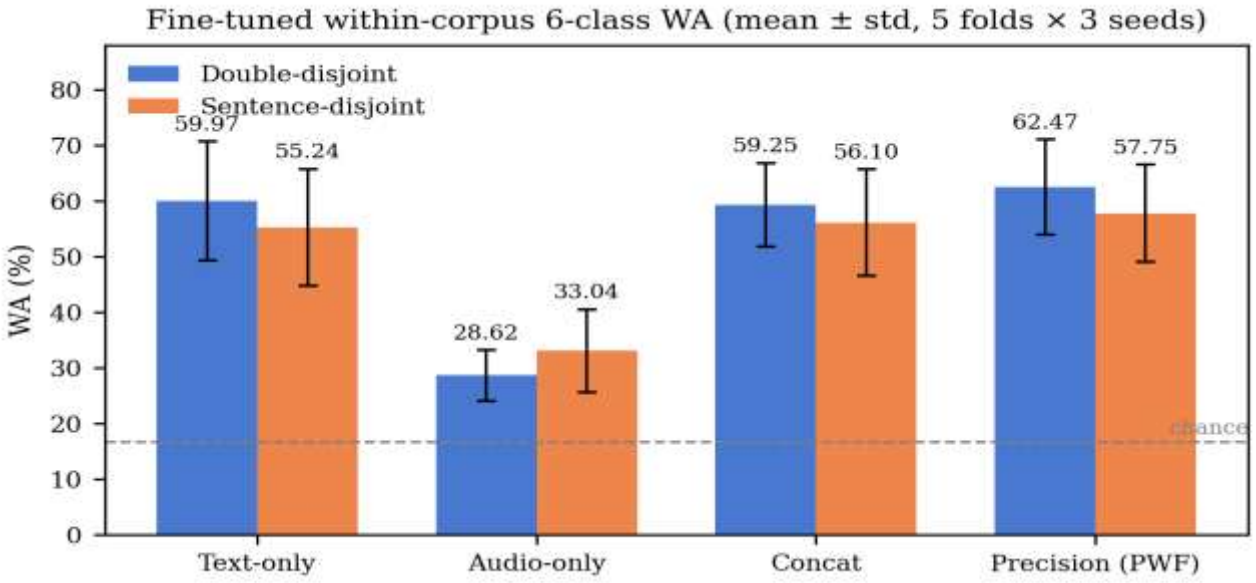

Fig. 5. Grouped bars of fine-tuned within-corpus 6-class WA under both leakage-controlled protocols (values on bars, mean ± std over 5 folds × 3 seeds), visualising Table III. Precision fusion leads under both protocols while audio-only remains far weaker; the dashed line marks the 16.7% chance rate.

C. Cross-Corpus Transfer Remains Near Chance

Table IV and Fig. 6 report zero-shot transfer on the shared four-class space, fine-tuned, over three seeds. In the IEMOCAP→corpus direction the best system (concat) reaches 34.76±1.44% WA against a 25% chance rate; in the corpus→IEMOCAP direction the best is audio-only at 38.81±0.40%. Both directions therefore sit only 3–14 points above chance, and the ordering of systems is direction-dependent: fusion helps into the corpus, while the audio branch transfers best out of it plausibly because the corpus's scripted lexical content is idiosyncratic and

its acoustic realisation is the more transferable signal. The seed variance is small (≤2.2 points), so the gap to within-corpus performance is a genuine domain gap accent, recording conditions, acted style, and elicitation protocol rather than training noise. We report these numbers as a calibration for the field: claims of accent-robust SER built on North-American English corpora do not survive contact with Indian-English scripted speech, in line with the fragmented cross-lingual transfer picture in the Indian-language SER literature [26].

TABLE IV. CROSS-CORPUS 4-CLASS ZERO-SHOT TRANSFER, FINE-TUNED (MEAN ± STD OVER 3 SEEDS, %)

Direction	System	WA	UA	wF1
IEMOCAP → corpus	text_only	30.88 ± 2.21	31.03 ± 2.18	33.25 ± 1.97
IEMOCAP → corpus	audio_only	27.16 ± 1.89	32.34 ± 0.55	27.34 ± 2.19
IEMOCAP → corpus	concat	34.76 ± 1.44	38.15 ± 1.15	38.38 ± 1.64
IEMOCAP → corpus	precision (ours)	33.58 ± 1.99	36.35 ± 0.61	35.60 ± 1.66
corpus → IEMOCAP	text_only	35.77 ± 0.52	28.11 ± 0.61	30.52 ± 0.91
corpus → IEMOCAP	audio_only	38.81 ± 0.40	35.18 ± 1.16	36.11 ± 0.49
corpus → IEMOCAP	concat	38.18 ± 0.73	30.56 ± 2.08	30.82 ± 3.01
corpus → IEMOCAP	precision (ours)	37.17 ± 0.38	28.82 ± 0.45	28.64 ± 1.01

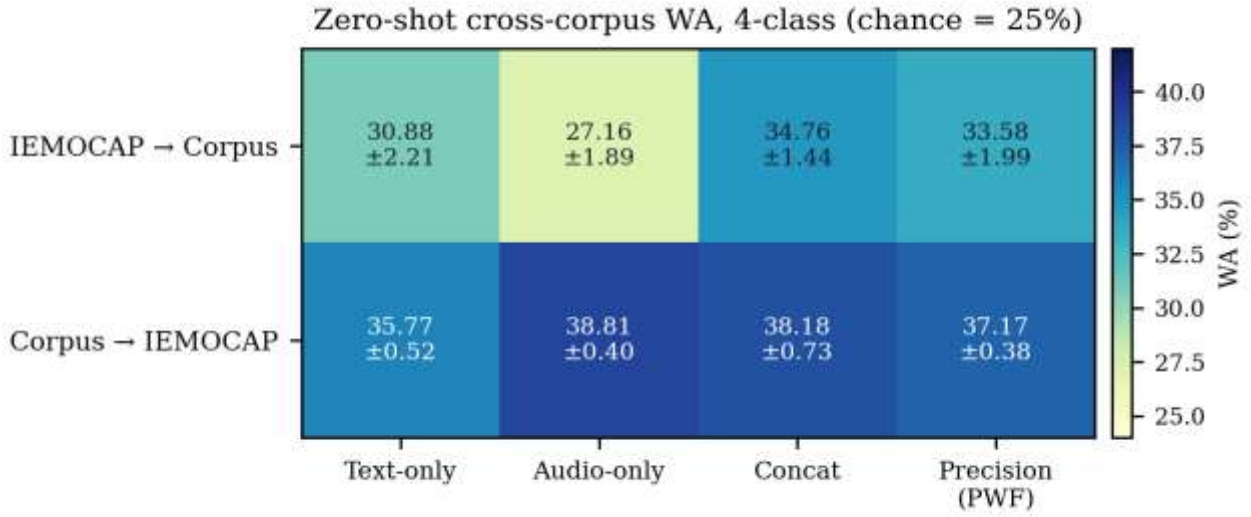

Fig. 6. Annotated heat map of zero-shot cross-corpus WA (fine-tuned, mean ± std over seeds) for all system–direction cells, visualising Table IV. All cells remain within 3–14 points of the 25% chance rate, quantifying the accent and elicitation domain gap; the direction-dependent ordering (fusion best into the corpus, audio best out of it) is visible along the rows.

D. Sarcasm as Modality Conflict: Disagreement Rises, the Gate Does Not Move

Table V and Fig. 7 present the pre-registered conflict analysis: a precision-fusion model trained on IEMOCAP (which contains no sarcasm) is evaluated on the corpus, and its internal signals are compared between the 561 sarcastic and 2,802 non-sarcastic utterances. The first half of the hypothesis is confirmed in every seed: sarcastic utterances raise text–audio branch disagreement from 69.7–71.6% to 75.9–82.7%, a shift of +4.3 to +13.0 points, confirming that the sarcastic class does embody genuine cross-modal conflict of the kind incongruity-based sarcasm models exploit [17], [18]. The second half is refuted: the mean audio gate weight does not increase on sarcastic inputs the shift is −0.0001, −0.0486, and −0.0528 across seeds 42–44, aggregating to −0.034±0.024, i.e., the gate moves marginally toward text. The supported interpretation is that precision weights track each branch's self-assessed confidence, learned from its training distribution, not the semantic conflict between branches: a HuBERT branch that never saw sarcasm in training has no mechanism to recognise sarcastic prosody as high-information evidence, so its predicted variance does not fall on sarcastic inputs. This negative result directly answers the forward-looking question posed in [4] whether interpretable modality weights anticipate sarcasm in the negative, and it constrains interpretation of uncertainty-weighted fusion generally: auditable weights explain which branch the model trusted, not which branch was right to trust. Closing that gap plausibly requires conflict-aware objectives (e.g., disagreement-conditioned gating or incongruity supervision [17]) rather than variance heads alone.

TABLE V. SARCASM MODALITY-CONFLICT ANALYSIS (FINE-TUNED PRECISION FUSION TRAINED ON IEMOCAP, EVALUATED ON THE CORPUS)

Seed	Subset	Branch disagreement (%)	Mean w_aud	Gate shift Δ
42	non-sarcastic (n = 2,802)	71.4	0.349	—
42	sarcastic (n = 561)	78.6	0.349	−0.0001
43	non-sarcastic	69.7	0.321	—
43	sarcastic	82.7	0.273	−0.0486
44	non-sarcastic	71.6	0.438	—
44	sarcastic	75.9	0.385	−0.0528
Mean ± std	gate shift toward audio	—	—	−0.034 ± 0.024

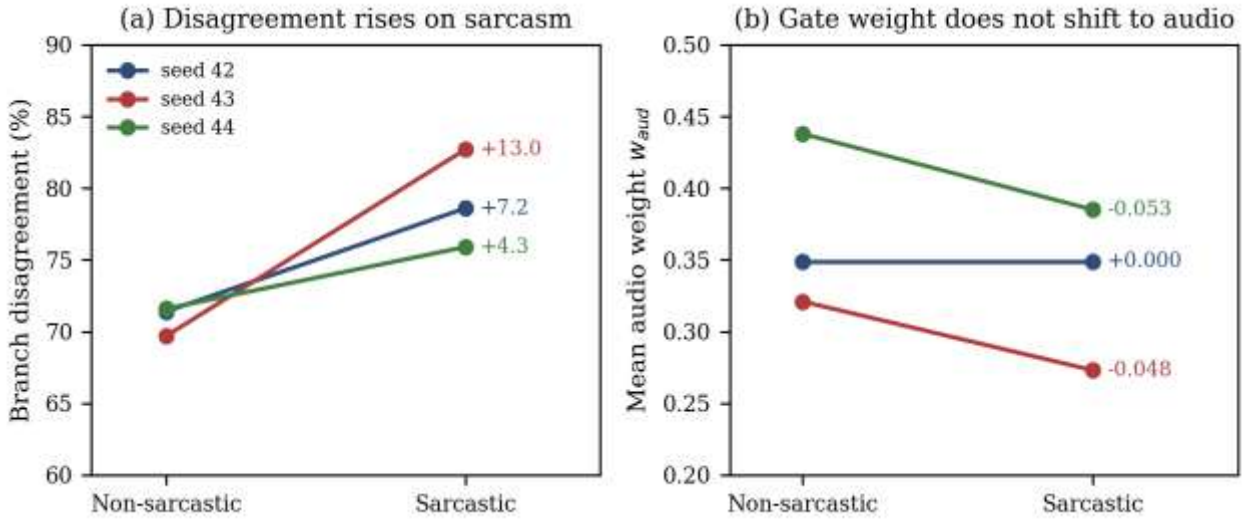

Fig. 7. Paired slope charts of the sarcasm conflict analysis across three seeds, visualising Table V. (a) Branch disagreement rises on sarcastic utterances in every seed (+4.3 to +13.0 points). (b) The precision gate's audio weight is flat or lower on sarcastic utterances in every seed, refuting the prosody-reweighting hypothesis.

E. Positioning Against Published Speaker-Independent SER Systems

Table VI situates the benchmarks against published speaker-independent four-class IEMOCAP systems and Indian cross-corpus SER. Direct numerical comparison across corpora is not meaningful and we do not claim superiority; the table's purpose is protocol calibration. Two rows of our own corpus bracket the comparison: the leakage-inflated speaker-disjoint text probe (100.00%) sits above every published system, illustrating how easily a scripted corpus can manufacture state-of-the-art-looking numbers, while the leakage-controlled double-disjoint precision benchmark (62.47%, six classes including sarcasm, 16.7% chance) sits below the IEMOCAP four-class band (67.6–72.1%, 25% chance) as it should, given the harder class inventory, the stricter protocol, and the far smaller resource. The IEMOCAP rows [5], [7]–[9] are quoted from the original authors under their stated protocols; the Indian cross-corpus row [26] confirms that our transfer results are consistent with the 45–70% range reported for related Indian-language settings under within-language training.

TABLE VI. POSITIONING AGAINST PUBLISHED SPEAKER-INDEPENDENT METHODS (VALUES AS REPORTED BY THE ORIGINAL AUTHORS; PROTOCOLS DIFFER AND ARE NOT DIRECTLY COMPARABLE)

Method	Year	Corpus / classes	Protocol	WA (%)	UA (%)
Precision fusion, LoRA RoBERTa+HuBERT (ours)	2026	RSD / 6 (incl. sarcastic)	Double-disjoint	62.47	60.22
Text-only probe (ours, leakage illustration)	2026	RSD / 6	Speaker-disjoint	100.00*	100.00*
Park et al., speaker-specific wav2vec 2.0 [8]	2023	IEMOCAP / 4	Speaker-indep.	72.14	72.97
Zou et al., co-attention	2022	IEMOCAP / 4	LOSO	69.80	71.05

multi-level [5]					
Ye et al., TIM-Net (audio only) [7]	2023	IEMOCAP / 4	10-fold CV	68.29	69.00
HuBERT-Large frozen probe, SUPERB [9]	2021	IEMOCAP / 4	Speaker-indep.	67.62	—
Bhangale & Kothandaraman, MAF-DCNN [26]	2024	Indian multilingual / varies	Cross-corpus	45.51–69.75	—

Leakage-inflated by sentence memorisation (LSR = 2.45); shown to calibrate the protocol effect, not as a claimed result

VII. Conclusion

This paper introduced the Real Sentiment Dataset, a 28-speaker, six-class Indian-English emotional speech corpus with the first sarcastic class in this population, together with a leakage-aware evaluation methodology built around the Lexical Shortcut Ratio. Three findings define the contribution. First, the field-standard speaker-disjoint protocol is invalid for scripted, prompt-shared corpora: a frozen text probe attains a perfect 100.00% WA under speaker-disjoint splits and 40.83% under double-disjoint splits (LSR = 2.45), so text and fusion results on such corpora are uninterpretable without sentence-disjoint control. Second, under the double-disjoint protocol, LoRA-adapted precision-weighted fusion is the strongest benchmark at 62.47±8.57% WA over 15 runs, and bidirectional transfer against IEMOCAP remains within 3–14 points of chance, quantifying the Indian-English domain gap. Third, sarcasm reliably elevates text–audio branch disagreement (up to +13 points) but does not shift the precision gate toward audio (−0.034±0.024) a reproducible negative result showing that variance-driven fusion weights encode self-confidence, not cross-modal conflict.

Limitations and future work. The corpus is acted and scripted; the perceptual validity of its labels especially the sarcastic class awaits the planned Fleiss-κ [28] rater study (stratified 300-clip sample, tooling released with the corpus), which is the immediate next step before public release. The 28-speaker scale produces large fold variance under double-disjoint splitting, which multi-seed aggregation mitigates but cannot eliminate; expanding the speaker pool and adding spontaneous (non-scripted) elicitation would strengthen both concerns simultaneously and drive LSR toward 1 by construction. The negative gate-shift result motivates conflict-aware fusion objectives disagreement-conditioned gating, incongruity supervision, or evidential opinion fusion [13] as the natural architectural follow-up, and computing LSR for other widely used scripted corpora would establish how general the lexical-shortcut problem is across the field.

References

1. C. Busso et al., 'IEMOCAP: Interactive Emotional Dyadic Motion Capture Database', Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, Nov. 2008, doi: 10.1007/s10579-008-9076-6.
2. N. Antoniou, Athanasios Katsamanis, T. Giannakopoulos, and S. Narayanan, 'Designing and Evaluating Speech Emotion Recognition Systems: A Reality Check Case Study With IEMOCAP', ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, May 2023, doi: 10.1109/icassp49357.2023.10096808.
3. S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, 'Towards Multimodal Sarcasm Detection (an _Obviously_ Perfect Paper)', Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, doi: 10.18653/v1/p19-1455.
4. Rajashree M Byalal, Sunanda Das, and M Kumaresan, 'Precision-Weighted Fusion of Fine-Tuned Text and Speech Encoders for Calibrated and Trustworthy Multimodal Emotion Recognition', International Journal of Computer Information Systems and Industrial Management Applications, vol. 18, no. 11s, pp. 629–648, Jul. 2026, doi: 10.70917/ijcisim-2026-3777.
5. H. Zou, Y. Si, C. Chen, D. Rajan, and C. E. Siong, 'Speech Emotion Recognition With Co-Attention Based Multi-Level Acoustic Information', arXiv.org. Accessed: Aug. 27, 2026. [Online]. Available: https://arxiv.org/abs/2203.15326
6. Z. Zhao, Y. Wang, and Y. Wang, 'Knowledge-Aware Bayesian Co-Attention for Multimodal Emotion Recognition', arXiv.org. Accessed: Aug. 27, 2026. [Online]. Available: https://arxiv.org/abs/2302.09856
7. J. Ye, X. Wen, Y. Wei, Y. Xu, K. Liu, and H. Shan, 'Temporal Modeling Matters: A Novel Temporal Emotional Modeling Approach for Speech Emotion Recognition', arXiv.org. Accessed: Aug. 27, 2026. [Online]. Available: https://arxiv.org/abs/2211.08233
8. S. Park, M. Mark, B. Park, and H. Hong, 'Using Speaker-Specific Emotion Representations in Wav2vec 2.0-Based Modules for Speech Emotion Recognition', Computers, Materials & Continua, vol. 77, no. 1, pp. 1009–1030, 2023, doi: 10.32604/cmc.2023.041332.

9. S. Yang et al., 'SUPERB: Speech Processing Universal PERformance Benchmark', arXiv.org. Accessed: Aug. 27, 2026. [Online]. Available: https://arxiv.org/abs/2105.01051
10. Chuan Fei Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, 'On Calibration of Modern Neural Networks', arXiv (Cornell University), Jun. 2017, doi: 10.48550/arxiv.1706.04599.
11. A.Kendall and Y. Gal, 'What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?', arXiv:1703.04977 [cs], Oct. 2017, Accessed: Aug. 27, 2026. [Online]. Available: https://arxiv.org/abs/1703.04977
12. M. K. Tellamekala, S. Amiriparian, B. W. Schuller, E. André, T. Giesbrecht, and M. Valstar, 'COLD Fusion: Calibrated and Ordinal Latent Distribution Fusion for Uncertainty-Aware Multimodal Emotion Recognition', IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 2, pp. 805–822, Feb. 2024, doi: 10.1109/tpami.2023.3325770.
13. Z. Han, C. Zhang, H. Fu, and J. T. Zhou, 'Trusted Multi-View Classification With Dynamic Evidential Fusion', IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 2, pp. 2551–2566, Feb. 2023, doi: 10.1109/tpami.2022.3171983.
14. O. Schrüfer, M. Milling, F. Burkhardt, F. Eyben, and B. Schuller, 'Are You Sure? Analysing Uncertainty Quantification Approaches for Real-World Speech Emotion Recognition', Interspeech 2024, pp. 3210–3214, Sep. 2024, doi: 10.21437/interspeech.2024-977.
15. A.N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio, 'Learning From Disagreement: A Survey', Journal of Artificial Intelligence Research, vol. 72, pp. 1385–1470, Dec. 2021, doi: 10.1613/jair.1.12752.
16. J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, 'From Hard to Soft', Proceedings of the 25th ACM international conference on Multimedia, pp. 890–897, Oct. 2017, doi: 10.1145/3123266.3123383.
17. Y. Wu et al., 'Modeling Incongruity Between Modalities for Multimodal Sarcasm Detection', IEEE MultiMedia, vol. 28, no. 2, pp. 86–95, Apr. 2021, doi: 10.1109/mmul.2021.3069
18. S. Pramanick, A. Roy, and V. M. P. Johns, 'Multimodal Learning Using Optimal Transport for Sarcasm and Humor Detection', 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 546–556, Jan. 2022, doi: 10.1109/wacv51458.2022.00062.
19. A.Ray, S. Mishra, A. Nunna, and P. Bhattacharyya, 'A Multimodal Corpus for Emotion Recognition in Sarcasm', arXiv (Cornell University), Jan. 2022, doi: 10.48550/arxiv.2206.02119.
20. E. J. Hu et al., 'LoRA: Low-Rank Adaptation of Large Language Models', openreview.net. Accessed: Aug. 27, 2026. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9
21. T. Feng and S. Narayanan, 'PEFT-SER: On the Use of Parameter Efficient Transfer Learning Approaches for Speech Emotion Recognition Using Pre-Trained Speech Models', 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1–8, Sep. 2023, doi: 10.1109/acii59096.2023.10388152.
22. R. Shanker, 'Edge-Accelerated Analytics for Encrypted Network Traffic Using Optimized Machine Learning', 2025 International Conference on Computing Technologies & Data Communication (ICCTDC), pp. 1–7, Jul. 2025, doi: 10.1109/icctdc64446.2025.11158771.
23. A. M. Basahel, S. H. Giriyappa, F. Alam, T. S. M. Alnazzawi, S. Qamar, and A. A. Abi Sen, 'A Resource-Efficient Approach to Fine-Tuning a BERT-Base Model for Sentiment Analysis', Computers, vol. 15, no. 3, p. 159, Mar. 2026, doi: 10.3390/computers15030159.
24. A.I. Hashi and Mr. O. A. Jama, 'Evaluating the Performance of AI-Based Software Tools in Intelligent Decision-Making Systems', International Journal of Computing and Engineering, vol. 7, no. 22, pp. 1–20, Nov. 2025, doi: 10.47941/ijce.3315
25. Y. Liu et al., 'RoBERTa: A Robustly Optimized BERT Pretraining Approach', arXiv (Cornell University), vol. 1, Jul. 2019, doi: 10.48550/arxiv.1907.11692.
26. 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, vol. 29, pp. 3451–3460, 2021, doi: 10.1109/taslp.2021.3122291.
27. S. G. Koolagudi, R. Reddy, J. Yadav, and K. S. Rao, 'IITKGP-SEHSC : Hindi Speech Corpus for Emotion Analysis', 2011 International Conference on Devices and Communications (ICDeCom), pp. 1–5, Feb. 2011, doi: 10.1109/icdecom.2011.5738540.
28. R. Kawade and S. Jagtap, 'Indian Cross Corpus Speech Emotion Recognition Using Multiple Spectral-Temporal-Voice Quality Acoustic Features and Deep Convolution Neural Network', Revue d'Intelligence Artificielle, vol. 38, no. 3, pp. 913–927, Jun. 2024, doi: 10.18280/ria.380318.
29. S. Kumar, Md. S. Akhtar, E. Cambria, and T. Chakraborty, 'SemEval 2024 - Task 10: Emotion Discovery and Reasoning Its Flip in Conversation (EDiReF)', Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pp. 1933–1946, 2024, doi: 10.18653/v1/2024.semeval-1.270.
30. J. L. Fleiss, 'Measuring Nominal Scale Agreement Among Many Raters.', Psychological Bulletin, vol. 76, no. 5, pp. 378–382, 1971, doi: 10.1037/h0031619.