International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(11s) 2026 pp. 629-648
https://doi.org/10.70917/ijcisim-2026-3777
Article

Precision-Weighted Fusion of Fine-Tuned Text and Speech Encoders for Calibrated and Trustworthy Multimodal Emotion Recognition

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

[1, 2, 3] Department of Computer Science & Engineering, Jain (Deemed –to-be-university), Bengaluru, India
[1]Email: btrajashree@gmail.com, [2]Email: das.sunanda2012@gmail.com,
[3]Email: phdkumaresan@gmail.com
*Corresponding Mail: btrajashree@gmail.com

Abstract: Multimodal speech emotion recognition (SER) systems are typically optimized and reported solely in terms of accuracy, yet deployment in affect sensitive applications additionally requires calibrated confidence, the ability to abstain on unreliable inputs and uncertainty estimates that reflect the genuine ambiguity of emotional expression, how decision level fusion architectures differ along these trustworthiness dimensions on the IEMOCAP benchmark remains largely unquantified. We propose PWF-FT, a Precision Weighted Fusion framework in which a RoBERTa text encoder and a HuBERT speech encoder, adapted with low rank adaptation (LoRA) and learned attention pooling, each emit class logits together with an utterance level aleatoric log variance, and the two modality decisions are fused by inverse variance (precision) weights computed per utterance, training combines a heteroscedastic unimodal objective with a soft label distillation term derived from crowd annotation distributions and the framework is benchmarked against text only, audio only, feature concatenation, fixed average and attention gated fusion under leave one session out (LOSO) evaluation with temperature scaling, expected calibration error (ECE), area under the risk coverage curve (AURC) and selective prediction analyses. Fusion lifts weighted accuracy from 66.45% (text) and 58.02% (audio) to 71.13-73.14% across the four fusion variants, with attention gated fusion best at 73.14±1.79% WA (72.74±1.88% UA, 72.97±1.77% wF1) and PWF at 71.13±2.22% WA, after temperature scaling all fusion systems calibrate to ECE 0.037-0.052 and PWF is the only variant whose predictive entropy correlates positively and consistently with human soft label entropy (Spearman $\rho = +0.286 \pm 0.029$ across all five folds). We conclude that attention, concatenation and fixed fusion are statistically indistinguishable in accuracy once encoders are fine tuned (paired fold level tests, $p \geq 0.15$), that precision weighting costs a small, quantified two WA points ($p = 0.003$) and that its practical value lies in interpretable, human aligned uncertainty that supports selective prediction and reliable deferral in human computer interaction, affective monitoring, and clinical screening pipelines.
Keywords: Multimodal emotion recognition, uncertainty quantification, model calibration, precision-weighted fusion, IEMOCAP, LoRA, selective prediction, soft labels.

1. INTRODUCTION

Speech emotion recognition (SER) underpins a growing range of human computer interaction systems, from conversational agents and call center analytics to driver monitoring and mental health screening. Because emotion is conveyed jointly through what is said and how it is said, bimodal systems that combine lexical content with acoustic prosody consistently outperform their unimodal counterparts on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) benchmark [1], and recent transformer-based fusion systems report weighted accuracies well above 80% under speaker dependent or random split protocols [2], [3]. This accuracy centric progress, however, conceals a dimension of model quality that matters at least as much in deployment: trustworthiness. Modern neural classifiers are known to be systematically overconfident [4], and a fusion system that assigns 95% confidence to a wrong emotion label is arguably more dangerous than a less accurate system that knows when it does not know [5].

The trustworthiness problem is particularly acute in emotion recognition because the labels themselves are uncertain. Emotional expression is inherently ambiguous: human annotators frequently disagree and the standard practice of majority voting their judgments into a single hard label discards information about which utterances are genuinely equivocal [6], [7]. An emotion recognizer that is well calibrated with respect to hard labels but oblivious to human ambiguity will be confidently wrong precisely on the utterances where humans themselves hesitate. Recent work has therefore begun to model annotator distributions directly [8], yet the interaction between multimodal fusion architecture and uncertainty quality whether the way two modalities are combined changes how well the model's confidence tracks both correctness and human ambiguity remains largely unexplored on standard SER benchmarks.

This paper addresses that gap with PWF-FT, an end to end fine tuned Precision Weighted Fusion framework for four class emotion recognition on IEMOCAP. A RoBERTa text encoder [9] and a HuBERT speech encoder [10] are adapted with low rank adaptation (LoRA) [11] a parameter efficient strategy in the same family as selective layer adaptation methods recently shown to preserve near full fine tuning quality at a fraction of the cost [12] and each modality branch emits both class logits and a learned utterance level log-variance representing aleatoric uncertainty [13]. The fused decision is a convex combination of the two unimodal logit vectors, weighted per utterance by each modality's inverse variance (precision), so that the more certain modality dominates on every individual input. We evaluate PWF-FT against five carefully matched systems text only, audio only, feature concatenation, fixed averaging, and a learned attention gate under leave one session out (LOSO) cross validation, and we report not only accuracy but a full trustworthiness profile: expected calibration error before and after temperature scaling, risk coverage behaviour, selective prediction accuracy and the alignment between model uncertainty and human soft label entropy.

The main contributions of this work are as follows:

1. A precision weighted decision fusion architecture (PWF-FT) in which fine tuned RoBERTa and HuBERT branches each predict logits and heteroscedastic log-variance, and perutterance inverse variance weights produce an interpretable, uncertainty driven fusion of the two modalities.
2. A trustworthiness centred evaluation protocol for multimodal SER that jointly reports WA/UA/wF1, ECE (raw and temperature scaled), AURC, coverage conditioned selective accuracy, and Spearman correlation between model predictive entropy and human annotation entropy, all under LOSO with matched training budgets.
3. Empirical evidence that, once encoders are fine tuned end to end, the accuracy differences among concatenation, fixed, attention gated, and precision fusion show a clear structure under paired fold level tests attention, concatenation and fixed fusion statistically indistinguishable ($p \geq 0.15$), precision fusion about two WA points below ($p = 0.003$) and that temperature scaling largely equalizes calibration across all fusion heads (ECE 0.037-0.052), findings that reframe the value proposition of uncertainty aware fusion.
4. A demonstration that PWF is the only evaluated fusion mechanism producing an intrinsic uncertainty signal that consistently and positively tracks human ambiguity ($\rho = +0.286 \pm 0.029$, positive in every fold, range +0.248 to +0.335), enabling human aligned deferral without any auxiliary uncertainty module.

The remainder of the paper is organized as follows. Section II reviews related work. Section III describes the dataset. Section IV presents the PWF-FT methodology, its mathematical formulation, and the training algorithm. Section V details the experimental setup. Section VI reports and discusses all results, and Section VII concludes.

2. RELATED WORK

A. Multimodal Emotion Recognition on IEMOCAP

Decision and feature level fusion of pretrained text and speech encoders dominates recent IEMOCAP research. Zhao et al. [3] fused wav2vec 2.0 and BERT representations at multiple granularities, and their follow-up knowledge aware Bayesian coattention model reached 75.5% WA / 77.0% UA under five fold cross validation [16]. Zou et al. [14] combined MFCC, spectrogram, and wav2vec 2.0 streams through a co-attention mechanism, reporting 69.80% WA / 71.05% UA under the stricter leave one session out protocol that we also adopt. Chen et al. [15] introduced a key sparse transformer that filters emotion irrelevant tokens during audio text interaction, and Sun et al. [17] showed that auxiliary tasks further regularize wav2vec 2.0 + BERT fusion. On the audio only side, TIM Net [18] models multi-scale temporal dynamics (68.29% WA / 69.00% UA), while speaker-aware wav2vec 2.0 pipelines such as Park et al. [19] reach 72.14% WA / 72.97% UA by disentangling speaker identity from emotion. Cross modal transformer systems such as MemoCMT [2] report above 81% accuracy, but under random utterance splits in which speaker identity leaks between training and testing; Antoniou et al. [20] demonstrated that such protocol choices inflate IEMOCAP results

substantially, motivating our exclusive use of LOSO. Critically, none of these systems reports calibration, risk–coverage, or uncertainty-alignment metrics, which is precisely the axis this paper targets.

B. Uncertainty and Calibration in Emotion Recognition

Guo et al. [4] established that modern networks are miscalibrated and that temperature scaling is a strong post hoc remedy, our protocol applies it to every system. Within affective computing, the closest work to ours is COLD fusion by Tellamekala et al. [21], which learns calibrated and ordinally ranked latent distribution variances for audiovisual emotion recognition and demonstrates that uncertainty aware fusion improves robustness under modality corruption. Evidential approaches form a second family: Sensoy et al. [22] replaced softmax with Dirichlet evidence, Han et al. [5] extended this to trusted multiview classification with dynamic evidential fusion at the opinion level and Wu et al. [8], [23] modeled utterance specific Dirichlet priors and deep evidential regression to capture annotator uncertainty in emotion labels. Schrüfer et al. [24] benchmarked uncertainty quantification methods for real world SER and found that even simple entropy based signals carry useful reliability information. Zhang et al. [25] provided theoretical grounding for dynamic multimodal weighting under low quality inputs. Our precision weighting mechanism can be viewed as the simplest member of this family a heteroscedastic Gaussian assumption on each modality's decision [13] with the advantage of adding only two scalar heads to a standard late fusion model.

C. Ambiguity, Soft Labels, and Annotator Disagreement

Han et al. [6] argued early that SER should move from hard to soft targets to reflect perception uncertainty, and Uma et al. [7] surveyed learning from disagreement methods across NLP and vision, concluding that annotator distributions carry systematic signal rather than noise. The Hugging Face release of IEMOCAP used in this work exposes per utterance crowd score distributions over emotion categories, which we exploit both as a KL based soft training target and as an independent ground truth measure of human ambiguity: the entropy of the annotation distribution. To our knowledge, correlating an end to end fused model’s predictive entropy against this human entropy under LOSO has not previously been reported for IEMOCAP fusion systems.

D. Parameter-Efficient Fine-Tuning of Pretrained Encoders

Full fine tuning of two transformer encoders is memory prohibitive on commodity GPUs. LoRA [11] freezes pretrained weights and learns low rank updates to attention projections, Feng and Narayanan [26] showed in PEFT-SER that LoRA is the strongest parameter efficient strategy for adapting speech models (wav2vec 2.0 [27], HuBERT [10], and related encoders benchmarked in SUPERB [28]) to emotion recognition. Basahel et al. [12] recently proposed SCALE, which profiles transformer layers by activation magnitude and attention entropy and adapts only the most influential ones with lightweight adapters, achieving up to 99% of full fine tuning performance on BERT-base sentiment analysis with roughly 40% of the parameters, conceptually, SCALE and this work share the premise that internal signal statistics layer activations there, per utterance decision variance here should govern where adaptation capacity and decision weight are allocated. We adopt plain LoRA on the query/key/value projections of both encoders as a well understood, reproducible baseline instantiation of this principle, extending resource efficient adaptation from unimodal sentiment classification [12] to bimodal emotion recognition.

3. DATASET

All experiments use the IEMOCAP corpus [1] as distributed in the AbstractTTS/IEMOCAP dataset on the Hugging Face. This release packages the corpus at the utterance level, with each row containing the 16 kHz audio waveform, the reference transcription, the majority vote categorical label (major_emotion), the source file name from which the recording session can be parsed and crucially for this work per utterance crowd annotation scores across emotion categories (angry, happy, excited, sad, neutral, frustrated, and others), from which a soft label distribution can be constructed. Following the dominant four class IEMOCAP protocol, we merge excited into happy and retain the classes {angry, happy, neutral, sad}. Utterances whose majority label falls outside these categories are discarded, leaving 6,877 utterances spanning all five recording sessions (Ses01-Ses05). Table I summarizes the statistics.

For each retained utterance, the soft target vector is formed by accumulating the crowd scores of the contributing categories onto the four target classes (excited scores are added to happy), followed by L1 normalization, utterances with no valid soft scores fall back to a one hot vector on the majority label. The Shannon entropy of this normalized annotation distribution serves as the human ambiguity reference used in Section VI-D. Audio is truncated to a maximum of 6 s (96,000 samples) to bound memory, and transcripts are tokenized to at most 64 sub word tokens.

TABLE I. DATASET STATISTICS AFTER FOUR-CLASS FILTERING (ABSTRACTTTS/IEMOCAP)

Property	Value
Retained utterances	6,877
Classes (4)	angry / happy / neutral / sad
Class counts	1,269 / 2,632 / 1,726 / 1,250
Sessions	Ses01-Ses05 (LOSO folds)
Per-fold test size	1,243-1,507 utterances
Audio	16 kHz, truncated to 6 s
Text	reference transcripts, max 64 tokens
Soft labels	normalized crowd score distributions

4. PROPOSED METHODOLOGY

A. Overview

PWF-FT is a two branch late fusion network. The text branch encodes the transcript with RoBERTa-base [9], the audio branch encodes the raw waveform with HuBERT-base-ls960 [10], whose convolutional feature extractor is kept frozen for stability. Both encoders are adapted with LoRA modules (rank r = 8, scaling α = 16, dropout 0.05) on the query, key, and value projections [11], with gradient checkpointing enabled so that the full pipeline trains within 8-12 GB of GPU memory. Each branch applies learned attention pooling over its token or frame sequence, projects the pooled vector to a shared 256 dimensional space, and feeds a modality specific classification head plus, in the precision configuration, a scalar log variance head. The fusion stage combines the two unimodal logit vectors with per utterance precision weights. Five reference systems share every component except the fusion rule: text_only and audio_only single branches, concat, which classifies the concatenated projections, fixed, which averages unimodal logits with equal weights and attention, which predicts the two weights with a learned gate.

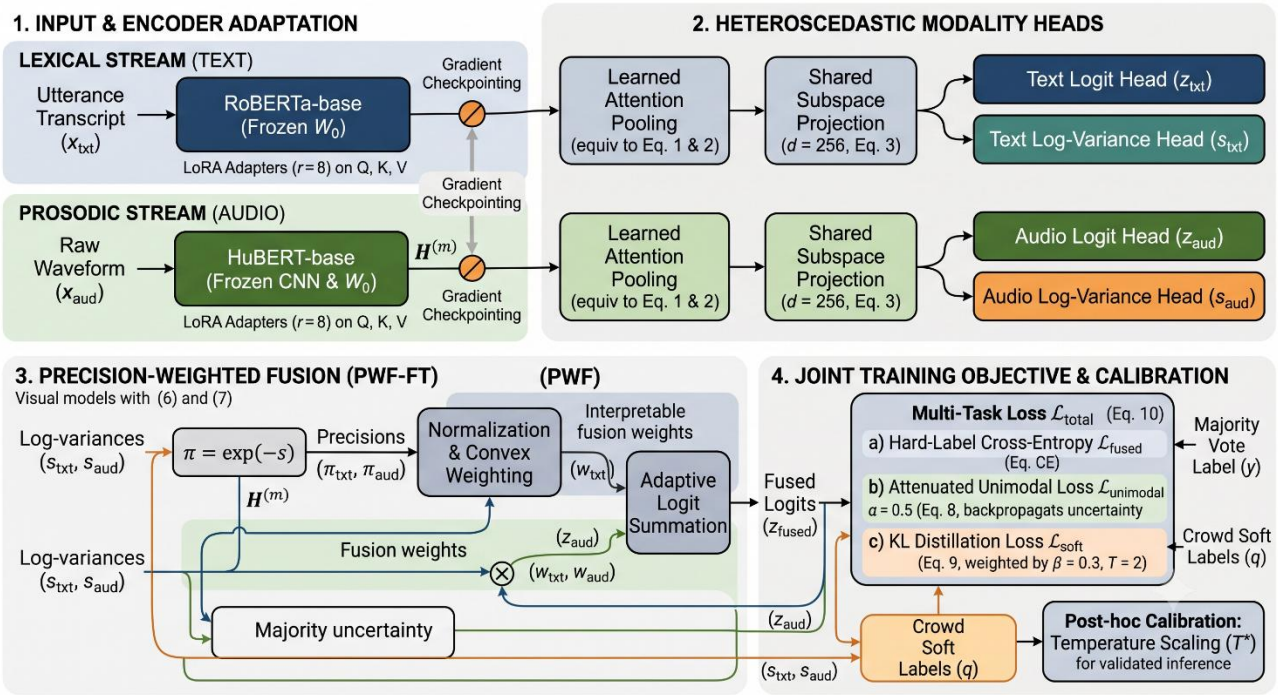

Fig. 1. End to End Architecture of the Proposed PWF-FT Framework

Fig. 1. Architectural schematic of the proposed PWF-FT framework, illustrating the end to end processing pipeline from LoRA adapted unimodal encoders and attention pooling to precision weighted decision level fusion under a joint heteroscedastic and soft label distillation objective.

B. Attention Pooling and Modality Encoding

Let $\mathcal{X}_{\text{txt}}$ and $\mathcal{X}_{\text{aud}}$ denote the input lexical transcript sequence and raw speech waveform respectively. We utilize two pretrained deep transformer architectures as our backbone encoders: a RoBERTa-base model $f_{\theta_{\text{txt}}}(\cdot)$ for text representation and a HuBERT-base-ls 960 model $f_{\theta_{\text{aud}}}(\cdot)$ for audio representation. Both backbones are parameterized by θ_{txt} and θ_{aud} which encompass their frozen pretrained weights augmented with active Low Rank Adaptation (LoRA) projection matrices.

Let $H^{(m)} \in \mathbb{R}^{T_m \times D_m}$ denote the sequence of hidden states extracted from the final layer of the encoder for modality $m \in \{\text{txt,aud}\}$:

$$H^{(m)} = f_{\theta_m}(\mathcal{X}_m)$$

where T_m is the sequence length (the number of sub word tokens for text and down sampled acoustic frames for audio), and D_m is the hidden state dimensionality ($D_{\text{txt}} = 768$, $D_{\text{aud}} = 768$). To handle variable length inputs within batched operations, we define a binary validity mask vector $m^{(m)} \in \{0,1\}^{T_m}$, where $m_t^{(m)} = 1$ indicates a valid sequence element and $m_t^{(m)} = 0$ indicates a padded placeholder at index $t \in \{1, \dots, T_m\}$.

Rather than performing simple temporal or token wise averaging, we introduce a content adaptive attention-pooling mechanism to compress each sequence into an utterance level representation. For a given modality m, let $h_t^{(m)} \in \mathbb{R}^{D_m}$ be the hidden vector at time step t. A two layer neural scorer computes an unnormalized scalar salience weight $s_t^{(m)}$:

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

where $W_1^{(m)} \in \mathbb{R}^{d_{\text{attn}} \times D_m}$ is a projection matrix, $b_1^{(m)} \in \mathbb{R}^{d_{\text{attn}}}$ is a bias vector, $w_2^{(m)} \in \mathbb{R}^{d_{\text{attn}}}$ is a linear readout weight vector, and d_{attn} represents the projection dimension (set to 128). The normalized pooling weight $a_t^{(m)}$ is obtained via a masked softmax operation:

$$a_t^{(m)} = \frac{m_t^{(m)} \exp\left(s_t^{(m)}\right)}{\sum_{k=1}^{T_m} m_k^{(m)} \exp\left(s_k^{(m)}\right)}$$

Using these adaptive weights, the pooled representation $\bar{h}^{(m)} \in \mathbb{R}^{D_m}$ is calculated as the convex combination:

$$\bar{h}^{(m)} = \sum_{t=1}^{T_m} a_t^{(m)} h_t^{(m)}$$

To bridge the dimensional and representational gap between the two distinct processing streams, we project each pooled vector $\bar{h}^{(m)}$ into a shared, unified multimodal subspace of dimension $d_{\text{shared}} = 256$:

$$u^{(m)} = \text{Dropout}\left(\text{GELU}\left(\text{LayerNorm}\left(W_p^{(m)} \bar{h}^{(m)} + b_p^{(m)}\right)\right)\right)$$

where $W_p^{(m)} \in \mathbb{R}^{d_{\text{shared}} \times D_m}$is a linear mapping matrix, $b_p^{(m)} \in \mathbb{R}^{d_{\text{shared}}}$ is the bias vector, and LayerNorm$(\cdot)$ and GELU$(\cdot)$ represent Layer Normalization and Gaussian Error Linear Unit activation functions, respectively. This sequence yields the aligned, low dimensional modality embeddings $u^{(\text{txt})}, u^{(\text{aud})} \in \mathbb{R}^{d_{\text{shared}}}$.

C. Precision Weighted Fusion

Under the precision weighted fusion (PWF-FT) paradigm, we formulate late decision level fusion under a heteroscedastic uncertainty framework. For each modality $m \in \{\text{txt,aud}\}$, we establish two distinct linear heads operating on the embedding $u^{(m)}$: a classification logit head and a scalar log variance regression head.

The category wise raw prediction logits $z_m \in \mathbb{R}^C$ (where $C = 4$ represents the emotion categories) are computed as:

$$z_m = W_z^{(m)} u^{(m)} + b_z^{(m)}$$

where $W_z^{(m)} \in \mathbb{R}^{C \times d_{\text{shared}}}$ and $b_z^{(m)} \in \mathbb{R}^C$ are projection weights and biases. Simultaneously, the aleatoric log variance $s_m \in \mathbb{R}$ is estimated as:

$$s_m = w_s^{(m)\top} u^{(m)} + b_s^{(m)}$$

where $w_s^{(m)} \in \mathbb{R}^{d_{\text{shared}}}$ and $b_s^{(m)} \in \mathbb{R}$ are learnable parameters. The scalar output s_m is interpreted as the logarithm of the input dependent aleatoric variance σ_m^2, such that:

$$s_m = \ln(\sigma_m^2) \Longrightarrow \sigma_m^2 = \exp(s_m)$$

The predictive precision π_m is defined as the inverse of the aleatoric variance:

$$\pi_m = \frac{1}{\sigma_m^2} = \exp(-s_m)$$

For each individual utterance, we dynamically construct a convex set of fusion weights $(w_{\text{txt}}, w_{\text{aud}})$ by normalizing the modality precisions through a softmax transformation:

$$w_{\text{txt}} = \frac{\pi_{\text{txt}}}{\pi_{\text{txt}} + \pi_{\text{aud}}} = \frac{\exp(-s_{\text{txt}})}{\exp(-s_{\text{txt}}) + \exp(-s_{\text{aud}})}$$

$$w_{\text{aud}} = \frac{\pi_{\text{aud}}}{\pi_{\text{txt}} + \pi_{\text{aud}}} = \frac{\exp(-s_{\text{aud}})}{\exp(-s_{\text{txt}}) + \exp(-s_{\text{aud}})}$$

The fused representation logits $z_{\text{fused}} \in \mathbb{R}^C$ are obtained by a precision weighted convex linear combination of the unimodal logits:

$$z_{\text{fused}} = w_{\text{txt}} z_{\text{txt}} + w_{\text{aud}} z_{\text{aud}}$$

This formulation enforces that on any arbitrary utterance, the modality exhibiting lower predicted observation uncertainty (higher precision) is assigned a proportionally higher weight in determining the final multimodal logit vector.

Reference Fusion Baselines

To evaluate the efficacy of the proposed precision weighting mechanism, we benchmark it against the following architectural variants:

1. **Fixed Average Fusion (fixed)**: The fusion weights are static and nonadaptive:

$$w_{\text{txt}} = w_{\text{aud}} = 0.5 \Longrightarrow z_{\text{fused}} = 0.5 z_{\text{txt}} + 0.5 z_{\text{aud}}$$

2. **Attention Gated Fusion (attention)**: A two-layer gating network G estimates the dynamic weights from the concatenated modality projections:

$$[w_{\text{txt}}, w_{\text{aud}}] = \text{softmax}\left(W_g^{(2)} \text{GELU}\left(W_g^{(1)}\left[u^{(\text{txt})}; u^{(\text{aud})}\right] + b_g^{(1)}\right) + b_g^{(2)}\right)$$

where $\left[u^{(\text{txt})}; u^{(\text{aud})}\right] \in \mathbb{R}^{2d_{\text{shared}}}$ is the concatenated embedding vector, $W_g^{(1)} \in \mathbb{R}^{d_{\text{shared}} \times 2d_{\text{shared}}}$, $b_g^{(1)} \in \mathbb{R}^{d_{\text{shared}}}$, $W_g^{(2)} \in \mathbb{R}^{2 \times d_{\text{shared}}}$, and $b_g^{(2)} \in \mathbb{R}^2$.

3. **Feature-Concatenation Fusion (concat)**: Unimodal embeddings are concatenated directly and fed to a single classification head:

$$z_{\text{fused}} = W_c\left[u^{(\text{txt})}; u^{(\text{aud})}\right] + b_c$$

where $W_c \in \mathbb{R}^{C \times 2d_{\text{shared}}}$ and $b_c \in \mathbb{R}^C$.

D. Training Objective

The optimization of our end to end framework is guided by a joint multitask loss function defined over a mini batch of size N. For any sample $i \in \{1, \dots, N\}$, let $y_i \in \{1, \dots, C\}$ represent the ground truth categorical label (derived from majority voting), and let $q_i = [q_{i,1}, \dots, q_{i,C}]^\top \in \mathbb{R}^C$ represent the normalized crowd annotation distribution satisfying $\sum_{c=1}^{C} q_{i,c} = 1$ with $q_{i,c} \geq 0$.

The total objective function is a linear combination of three distinct losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{fused}} + \alpha \mathcal{L}_{\text{unimodal}} + \beta \mathcal{L}_{\text{soft\_distill}}$$

1. Primary Fused Loss ($\mathcal{L}_{\text{fused}}$)

This term represents the standard empirical cross entropy loss applied directly to the fused logit output z_{fused} to enforce predictive accuracy over the hard targets:

$$\mathcal{L}_{\text{fused}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{H}_{\text{CE}} \left(\text{softmax}\left(z_{\text{fused},i}\right), y_i \right) = -\frac{1}{N} \sum_{i=1}^{N} \ln \left(\frac{\exp\left(z_{\text{fused},i,y_i}\right)}{\sum_{j=1}^{C} \exp\left(z_{\text{fused},i,j}\right)} \right)$$

where $\mathcal{H}_{\text{CE}}(\cdot)$ denotes the categorical cross entropy function.

2. Heteroscedastic Unimodal Loss ($\mathcal{L}_{\text{unimodal}}$)

To ensure the learned log variance parameters s_m correctly approximate the sample dependent aleatoric uncertainty, we implement an attenuated loss formulation for both modalities:

$$\mathcal{L}_{\text{unimodal}} = \sum_{m \in \{\text{txt,aud}\}} \left[\frac{1}{N} \sum_{i=1}^{N} \left(\exp\left(-s_{m,i}\right) \mathcal{H}_{\text{CE}}\left(\text{softmax}\left(z_{m,i}\right), y_i\right) + \frac{1}{2} s_{m,i} \right) \right]$$

The exponential factor $\exp\left(-s_{m,i}\right) = \frac{1}{\sigma_{m,i}^2}$ serves to scale down the classification loss for highly challenging or noisy inputs, allowing the model to decrease the gradient magnitude of samples it identifies as highly uncertain. The regularize term $\frac{1}{2} s_{m,i}$ acts as a logarithmic barrier that penalizes the model for trivially predicting infinite uncertainty. The loss reaches its analytical minimum when the estimated variance $\sigma_{m,i}^2 = \exp\left(s_{m,i}\right)$aligns with the actual mean squared error scale of that modality's prediction.

3. Soft-Target Distillation Loss ($\mathcal{L}_{\text{soft\_distill}}$)

To regularize training and capture the human perceptual uncertainty encoded within the annotator distributions, we add a Kullback Leibler (KL) divergence term between the crowd score distribution q_i and the temperature-softened probability distribution computed from the fused logits:

$$\mathcal{L}_{\text{soft\_distill}} = \tau^2 \cdot \left[\frac{1}{N} \sum_{i=1}^{N} D_{\text{KL}} \left(q_i \parallel \mathbf{p}_i^{\tau} \right) \right]$$

where $\mathbf{p}_i^{\tau} \in \mathbb{R}^C$ is the softened predictive probability vector defined by:

$$p_{i,c}^{\tau} = \frac{\exp\left(z_{\text{fused},i,c}/\tau\right)}{\sum_{j=1}^{C} \exp\left(z_{\text{fused},i,j}/\tau\right)}$$

The KL divergence is explicitly formulated as:

$$D_{\text{KL}}(q_i \parallel \mathbf{p}_i^{\tau}) = \sum_{c=1}^{C} q_{i,c} \ln \left(\frac{q_{i,c}}{p_{i,c}^{\tau}} \right)$$

The scaling coefficient τ^2preserves the relative gradient magnitudes when backpropagating through the temperature scaled activations. We set the hyperparameter values to $\alpha = 0.5$, $\beta = 0.3$, and $\tau = 2.0$.

E. Calibration and Trustworthiness Metrics

To validate the deployment viability of our systems in safety critical human machine pipelines, we evaluate performance along several critical trustworthiness dimensions.

1. Post-hoc Calibration (Temperature Scaling)

To guarantee that the predicted probabilities map precisely to the empirical frequency of correct predictions, we apply temperature scaling post training. Let $\mathcal{D}_{\text{val}} = \{(z_i, y_i)\}_{i=1}^{N_{\text{val}}}$ be the validation subset. We optimize a global scalar temperature parameter $T > 0$ via L-BFGS to minimize validation cross-entropy:

$$T^* = \arg\min_{T>0} \left[-\frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} \sum_{c=1}^{C} \mathbb{I}\,(y_i = c) \ln \left(\frac{\exp(z_{i,c}/T)}{\sum_{j=1}^{C} \exp(z_{i,j}/T)} \right) \right]$$

where $\mathbb{I}(\cdot)$ is the standard indicator function. The calibrated test probabilities are then computed as $P(y_i = c \mid \mathcal{X}) = \text{softmax}\left(z_{\text{fused},i}/T^*\right)_c$.

2. Expected Calibration Error (ECE)

To measure calibration performance, we partition the calibrated test set predictions of size N_{test}into $M = 15$ equally spaced, mutually exclusive confidence intervals (bins). Let B_b denote the set of indices of predictions falling into the interval of the b-th bin:

$$B_b = \left\{ i \in \{1, \dots, N_{\text{test}}\} \;\middle|\; \hat{p}_i \in \left(\frac{b-1}{M}, \frac{b}{M} \right] \right\}, \quad b \in \{1, \dots, M\}$$

where $\hat{p}_i = \max_c P(y_i = c \mid \mathcal{X})$ represents the model's confidence for sample i, and $\hat{y}_i = \arg\max_c z_{\text{fused},i,c}$ represents the predicted class label. The empirical accuracy $\text{acc}(B_b)$ and confidence $\text{conf}(B_b)$ of each bin are computed as:

$$\text{acc}(B_b) = \frac{1}{|B_b|} \sum_{i \in B_b} \mathbb{I}\,(\hat{y}_i = y_i)$$

$$\text{conf}(B_b) = \frac{1}{|B_b|} \sum_{i \in B_b} \hat{p}_i$$

The Expected Calibration Error (ECE) is calculated as the weighted average of the absolute discrepancies:

$$\text{ECE} = \sum_{b=1}^{M} \frac{|B_b|}{N_{\text{test}}} |\text{acc}(B_b) - \text{conf}(B_b)|$$

3. Selective Prediction & Area Under the Risk Coverage Curve (AURC)

In selective prediction, the model is equipped with a confidence based selection function $g_i \in \{0,1\}$ to either output the prediction ($g_i = 1$) or abstain ($g_i = 0$). We evaluate this capability by sorting the test set in descending order of predictive confidence:

$$\hat{p}_{\pi(1)} \geq \hat{p}_{\pi(2)} \geq \cdots \geq \hat{p}_{\pi(N_{\text{test}})}$$

where π is the sorting permutation. At any given sample index coverage $k \in \{1, \dots, N_{\text{test}}\}$, the corresponding selective risk (error rate) is:

$$\mathcal{R}(k) = \frac{1}{k} \sum_{j=1}^{k} \mathbb{I}\left(\hat{y}_{\pi(j)} \neq y_{\pi(j)}\right)$$

The Area Under the Risk Coverage Curve (AURC) represents the cumulative sum of these risks across all possible coverage boundaries:

$$\text{AURC} = \frac{1}{N_{\text{test}}} \sum_{k=1}^{N_{\text{test}}} \mathcal{R}\,(k) = \frac{1}{N_{\text{test}}} \sum_{k=1}^{N_{\text{test}}} \left(\frac{1}{k} \sum_{j=1}^{k} \left(1 - \mathbb{I}\left(\hat{y}_{\pi(j)} = y_{\pi(j)}\right)\right) \right)$$

A lower AURC score demonstrates that the model correctly assigns low confidence to its errors, yielding an effective signal for selective abstention.

4. Human-Ambiguity Alignment (ρ)

To measure the alignment between the model's internal uncertainty and actual human perceptual ambiguity, we evaluate the Spearman rank correlation between the Shannon entropy of the model's prediction and the Shannon entropy of the crowd annotator distribution. For each sample i, these entropies are computed as:

$$\mathcal{H}(\mathbf{p}_i) = -\sum_{c=1}^{C} p_{i,c} \ln(p_{i,c})$$

$$\mathcal{H}(\mathbf{q}_i) = -\sum_{c=1}^{C} q_{i,c} \ln(q_{i,c})$$

where $p_{i,c} = \text{softmax}(z_{\text{fused},i})_c$ is the fused predictive probability of class c. Let $R(\mathcal{H}(\mathbf{p}))_i$ and $R(\mathcal{H}(\mathbf{q}))_i$ denote the rank-order positions of these entropies within the test set. The Spearman coefficient ρ is computed as:

$$\rho = 1 - \frac{6\sum_{i=1}^{N_{\text{test}}} d_i^2}{N_{\text{test}}(N_{\text{test}}^2 - 1)}$$

where $d_i = R(\mathcal{H}(\mathbf{p}))_i - R(\mathcal{H}(\mathbf{q}))_i$ is the rank difference for the i-th utterance. A positive and statistically significant ρ indicates that the model's confidence is physically aligned with human perceptual disagreement.

To stabilize memory consumption on consumer hardware, gradient checkpointing is enabled on both backbones, and the convolutional feature extractor of the HuBERT backbone is kept completely frozen. The complete optimization execution, cyclic hyperparameter selection, post-hoc calibration, and validation gated inference sequence are mathematically formalized in Algorithm 1.

Algorithm 1: PWF-FT Training and Speaker-Independent LOSO Evaluation

Input: Dataset of complete dyadic utterances $\mathcal{D} = \{(\mathcal{X}_{\text{aud},i}, \mathcal{X}_{\text{txt},i}, y_i, q_i, \mathcal{S}_i)\}_{i=1}^{N_{\text{total}}}$

- Pre-trained backbone encoders $f_{\theta_{\text{txt}}}$ (RoBERTa-base) and $f_{\theta_{\text{aud}}}$ (HuBERT-base)
- Hyperparameters: LoRA rank $r = 8$, scaling $\alpha_{\text{LoRA}} = 16$, max epochs $E_{\text{max}} = 8$, validation patience $P = 3$, loss weights $\alpha = 0.5$ and $\beta = 0.3$, distillation temperature $\tau = 2.0$, batch size $B_s = 12$, gradient accumulation steps $K_{\text{accum}} = 2$ (effective batch size $N_{\text{eff}} = 24$).

Output: Out-of-sample speaker-independent metrics: {WA,UA,wF1,ECE, ECE_{ts},AURC, $\rho\}_f$ evaluated across all five session folds.

1. **Initialize** global out-of-sample metrics log $\mathcal{M} \leftarrow \emptyset$.
2. **For each** cross-validation fold $f \in \{1, \ldots, 5\}$ **do**:
    a. Partition corpus $\mathcal{D}$ into isolated speaker sets $S_{\text{test}} \leftarrow \{i \mid \mathcal{S}_i = \mathcal{S}_f\}$, $S_{\text{val}} \leftarrow \{i \mid \mathcal{S}_i = \mathcal{S}_{(f \bmod 5)+1}\}$, and $S_{\text{train}} \leftarrow \mathcal{D} \setminus (S_{\text{test}} \cup S_{\text{val}})$.
    b. Initialize framework parameters $\Theta^{(f)} = \{A_m, B_m \text{ for } m \in \{\text{txt,aud}\}\} \cup \{W_{\text{pool}}, W_p, W_z, w_s, b\}^{(f)}$, keeping backbone weights W_0 and the HuBERT CNN front end completely frozen.
    c. Instantiate AdamW optimizer with discriminative learning rates ($\eta_{\text{LoRA}} = 3 \times 10^{-4}$, $\eta_{\text{heads}} = 10^{-3}$) and learning rate scheduler $\mathcal{S}_{\text{sched}}$ featuring 10% linear warmup.
3. **For each** epoch $e \in \{1, \ldots, E_{\text{max}}\}$ **do**:
4. **For each** mini-batch step t, $\mathcal{B}_t = \{(\mathcal{X}_{\text{aud}}, \mathcal{X}_{\text{txt}}, y, q)_n\}_{n=1}^{B_s}$ partitioned from S_{train} **do**:

a. Execute forward pass under AMP: Eq. (1)-(7) → $z_{\text{fused},n}$; compute loss scaled by accumulation slices: $\mathcal{L}_{\text{scaled}} \leftarrow \mathcal{L}_{\text{total}}/K_{\text{accum}}$ via Eq. (10); backpropagate, apply gradient norm clipping at 5.0, update parameters $\Theta^{(f)}$ via AdamW every K_{accum} steps, and step $\mathcal{S}_{\text{sched}}$.
b. Evaluate Unweighted Accuracy (UA) on validation set S_{val}; update historical tracker $\Theta_{\text{best}}^{(f)} \leftarrow \Theta^{(f)}$ upon tracking higher val_UA, and execute early stopping if validation improvement stalls for $P = 3$ consecutive epochs.

5. Solve optimal validation calibration temperature T^* on validation logits $z_{\text{fused},S_{\text{val}}}$ via validation loss minimization using the L-BFGS optimizer.
6. Perform out of sample model forward inference on held out test session S_{test}using parameters $\Theta_{\text{best}}^{(f)}$ to obtain raw logits $z_{\text{fused,test}}$ and compute calibrated probabilities $\text{softmax}\left(z_{\text{fused,test}}/T^*\right)$.
7. Compute out-of-sample metrics for fold f: recognition accuracy ($\text{WA}_f, \text{UA}_f, \text{wF1}_f$), Expected Calibration Error ($\text{ECE}_f, \text{ECE}_{\text{ts},f}$ via Eq. 12), selective risk coverage (AURC_f via Eq. 13), and human perceptual rank-order alignment coefficient (ρ_f via Eq. 14), then append metrics tuple $\mathcal{M}_f$ to tracker $\mathcal{M}$.
8. **Return** final out of sample speaker-independent evaluation matrices: $\text{Mean}(\mathcal{M}) \pm \text{Std}(\mathcal{M})$.

5. EXPERIMENTAL SETUP

All six systems share identical data pipelines, encoders, pooling, projection width (256), dropout (0.3), optimizer, schedule, and stopping criterion, only the fusion rule and the applicable loss terms differ, so any observed gap is attributable to the fusion mechanism. Training uses mixed precision AMP on a single 8-12 GB GPU with batch size 12 and gradient accumulation of 2 (effective batch 24), for at most 8 epochs with patience 3 on validation UA. Under LOSO, each of the five folds holds out one full session (both speakers) for testing, uses the cyclically next session for validation and temperature fitting, and trains on the remaining three, guaranteeing speaker independent evaluation [20]. Table II lists the complete configuration; all values are taken directly from the released training script for reproducibility.

TABLE II. COMPLETE TRAINING CONFIGURATION

Component	Setting
Text encoder	Roberta base, max 64 tokens
Audio encoder	Hubert base-ls960, conv front-end frozen
PEFT	LoRA r=8, α=16, dropout 0.05 on Q/K/V of both encoders
Pooling / projection	attention pooling, Linear→LayerNorm→GELU→Dropout(0.3), 256-d
Loss weights	α (unimodal) = 0.5, β (soft KL) = 0.3, τ = 2.0
Optimizer	AdamW, LR 3×10^{-4} (LoRA) / 10^{-3} (heads); wd 10^{-4}
Schedule	10% linear warmup, linear decay, grad clip 5.0
Batching	batch 12 × accum 2, AMP, grad checkpointing
Epochs / patience	8 / 3 (monitor validation UA)
Evaluation	LOSO over 5 sessions, temperature scaling on validation fold
Seed	42 (identical across systems)

Metrics. Weighted accuracy (WA) is overall accuracy, unweighted accuracy (UA) is macro averaged recall, wF1 is the support weighted F1. ECE uses 15 bins, Eq. (12), the _ts suffix denotes post temperature scaling values. AURC follows Eq. (13) (lower is better), and selective accuracy is reported at coverage levels 1.0 down to 0.5. The

human alignment coefficient ρ of Eq. (14) is defined only for the precision system, which is the only one equipped with intrinsic perutterance uncertainty heads.

6. RESULTS AND DISCUSSION

A. Main Recognition Results

Table III reports LOSO recognition performance, mean ± standard deviation over the five session folds. Every fusion system outperforms both unimodal branches by a wide margin: the best fusion (attention) exceeds text-only by +6.7 WA points and audio only by +15.1 points, confirming the complementarity of lexical and prosodic evidence on IEMOCAP [3], [14]. Among the four fusion variants, attention is nominally best (73.14% WA), followed by concat (72.99%), fixed (72.01%), and precision (71.13%). Because the five LOSO folds are shared across systems, paired significance tests over the perfold values of Table IV are applicable. The fusion over unimodal gain is highly reliable (attention vs. text_only: paired t-test $p = 0.002$ on WA). Among fusion heads, attention, concat and fixed are statistically indistinguishable from one another (paired t-test $p \geq 0.15$, Wilcoxon signed rank $p \geq 0.19$ for every pair, on both WA and UA). Precision fusion, by contrast, sits consistently about two points below attention in every fold (mean $\Delta WA = -2.01$, paired t-test $p = 0.003$, $\Delta UA = -1.64$, $p = 0.012$, Wilcoxon $p = 0.062$, the smallest value attainable at $n = 5$). The honest reading, consistent with the reality check analysis of Antoniou et al. [20], is therefore twofold: the encoders, not the fusion rule, carry the bulk of the accuracy, and precision weighting carries a small but real accuracy cost of roughly two WA points, the explicit price of the trustworthiness properties quantified in the remainder of this section.

TABLE III. MAIN LOSO RESULTS ON IEMOCAP 4-CLASS (MEAN ± STD OVER 5 SESSION FOLDS, %)

System	WA	UA	wF1
text_only	66.45 ± 2.66	65.93 ± 3.48	66.56 ± 2.69
audio_only	58.02 ± 3.09	59.53 ± 3.22	56.89 ± 3.53
concat	72.99 ± 1.77	73.12 ± 1.73	73.09 ± 2.03
fixed	72.01 ± 1.89	71.54 ± 2.37	71.90 ± 1.87
attention	**73.14 ± 1.79**	**72.74 ± 1.88**	**72.97 ± 1.77**
precision (PWF, ours)	**71.13 ± 2.22**	**71.10 ± 2.58**	**71.07 ± 2.16**

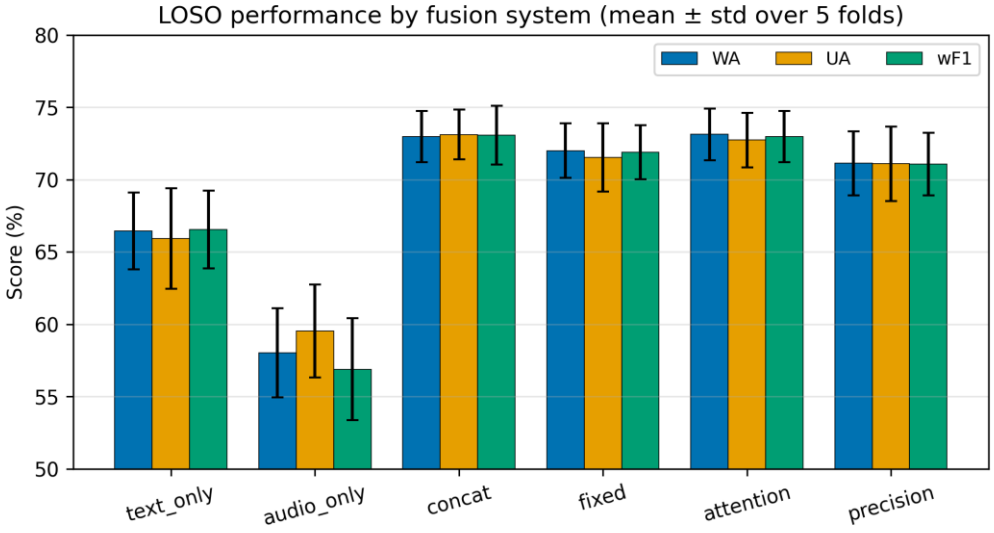

Fig. 2. LOSO recognition performance (WA, UA, wF1; mean ± std over the five session folds) for all six systems, visualizing Table III. The four fusion systems form a statistically overlapping band 5–15 points above the unimodal baselines.

Fig. 2 renders the same comparison graphically: the error bars of the four fusion systems overlap almost completely, while both unimodal systems sit clearly below the band. Table IV decomposes the aggregate into its five constituent folds. The per fold view exposes two effects hidden by the means. First, fold difficulty varies substantially Ses02 is the easiest test session for every system (up to 75.62% WA) and Ses04 the hardest for audio (53.08%) so cross paper comparisons that use a single fixed test session [20] should be interpreted with care. Second, the fusion ranking is not stable across folds (attention wins folds 1-3, concat wins folds 4-5), reinforcing the conclusion that no fusion head dominates. The final column of Table IV shows that the human alignment correlation of the precision system is positive in every single fold, ranging from +0.248 to +0.335; unlike accuracy, this property is fold invariant.

TABLE IV. PER-FOLD TEST WEIGHTED ACCURACY (%) BY HELD-OUT SESSION, WITH THE HUMAN-ALIGNMENT P OF THE PRECISION SYSTEM

Test session	**text_only**	**audio_only**	**concat**	**fixed**	**attention**	**precision**	**ρ (precision)**
Ses01	63.50	61.41	70.16	70.38	72.25	69.86	+0.299
Ses02	71.20	58.65	75.54	75.46	75.62	74.50	+0.335
Ses03	65.09	60.84	72.22	72.57	74.28	72.77	+0.248
Ses04	65.17	53.08	73.35	71.17	70.35	68.24	+0.276
Ses05	67.29	56.14	73.66	70.47	73.19	70.27	+0.272
Mean ± std	**66.45 ± 2.66**	**58.02 ± 3.09**	**72.99 ± 1.77**	**72.01 ± 1.89**	**73.14 ± 1.79**	**71.13 ± 2.22**	**+0.286 ± 0.029**

B. Calibration

Table V presents the calibration and trustworthiness profile. Before temperature scaling, the unimodal systems are the most miscalibrated (ECE 0.147-0.150), consistent with the overconfidence phenomenon of [4], among fusion systems, fixed averaging is intrinsically best calibrated (ECE 0.060), which is expected since averaging two logit vectors implicitly tempers the fused distribution, while concat is worst (0.113). PWF sits between them (0.098). After per fold temperature scaling, however, every fusion system converges to a narrow ECE band of 0.037-0.052 a single scalar fitted on one held out session essentially erases the architectural calibration differences. This is an important negative result for uncertainty aware fusion research: raw ECE advantages reported without a temperature scaled control [21] may overstate the architectural contribution. Figure 3 shows the PWF reliability diagram, the curve tracks the diagonal closely across the confidence range with a mild residual overconfidence in mid confidence bins, matching its 0.052 post scaling ECE.

TABLE V. CALIBRATION AND TRUSTWORTHINESS (MEAN ± STD; ECE/AURC LOWER IS BETTER; _TS = AFTER TEMPERATURE SCALING)

System	**ECE**	**ECE_ts**	**AURC**	**AURC_ts**	**ρ (vs. human)**
text_only	0.147 ± 0.022	0.046 ± 0.007	0.171 ± 0.019	0.169 ± 0.018	–
audio_only	0.150 ± 0.052	0.065 ± 0.031	0.263 ± 0.043	0.263 ± 0.044	–
concat	0.113 ± 0.023	0.042 ± 0.015	0.123 ± 0.016	0.122 ± 0.016	–
fixed	0.060 ± 0.022	0.037 ± 0.013	0.127 ± 0.019	0.126 ± 0.018	–
attention	0.091 ± 0.009	0.039 ± 0.011	0.126 ± 0.014	0.125 ± 0.015	–
precision (PWF, ours)	**0.098 ± 0.015**	**0.052 ± 0.016**	**0.140 ± 0.021**	**0.140 ± 0.021**	**+0.286 ± 0.029**

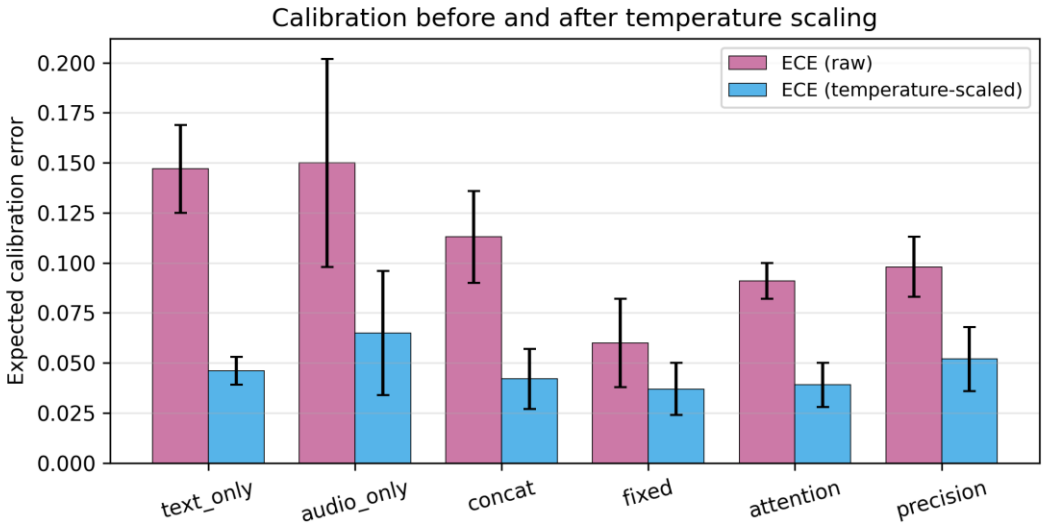

Fig. 3. Expected calibration error before (raw) and after temperature scaling for all six systems (mean ± std over folds), visualizing the ECE columns of Table V. The large architectural spread in raw ECE collapses to a narrow 0.037-0.052 band after a single validation-fitted temperature.

Fig. 3 makes the temperature-scaling effect of Table V immediately visible: raw ECE spans a 2.5× range across architectures, yet the scaled bars are nearly level. The one exception to the equalization is audio_only (0.065 after scaling), whose validation and test sessions differ enough in difficulty that a single temperature transfers imperfectly another symptom of the modality's weaker generalization already seen in Tables III and IV.

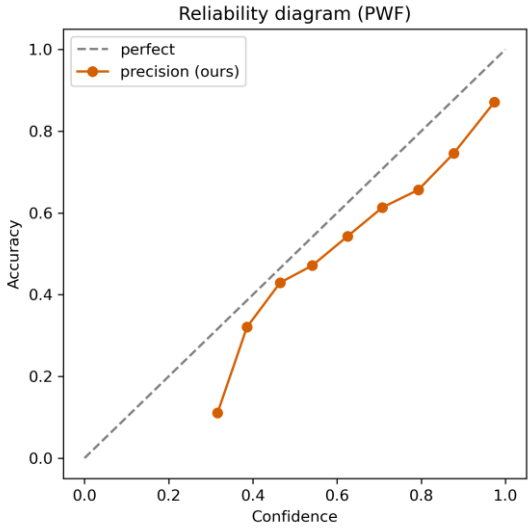

Fig. 4. Reliability diagram of the proposed PWF system (pooled LOSO test predictions). The curve lies close to the diagonal across the confidence range, with mild residual overconfidence in the mid confidence bins, consistent with the 0.052 post scaling ECE in Table V.

C. Selective Prediction and Risk–Coverage

Table VI, Fig. 4, and Fig. 5 evaluate whether each system's confidence supports abstention. All four fusion systems produce near parallel risk coverage curves that dominate both unimodal systems at every coverage level: at 50% coverage, fusion accuracies reach 86.8-88.2% versus 83.1% (text) and 71.8% (audio). PWF attains 86.82% at half coverage and improves monotonically as coverage decreases, confirming that its confidence ranks errors effectively, its AURC (0.140) is slightly higher than concat/fixed/attention (0.123-0.127), which mirrors its slightly

lower base accuracy rather than a defect of the uncertainty signal, since AURC lower bounds at the base error rate. The audio only curve in Fig. 4 remains far above all others across the full range, showing that prosody alone yields both lower accuracy and less useful confidence, while the text curve occupies an intermediate band. Practically, any of the fusion systems can trade 30% coverage for roughly a 9-10 point accuracy gain an operating point directly relevant to human in the loop affective applications where uncertain utterances are deferred to a human rater [24], [31].

TABLE VI. SELECTIVE PREDICTION: ACCURACY (%) AT EACH COVERAGE LEVEL (LOSO MEAN)

System	1.00	0.90	0.80	0.70	0.60	0.50
text_only	66.45	70.04	73.22	76.56	79.73	83.12
audio_only	58.02	60.43	62.50	65.52	68.47	71.81
concat	72.99	76.28	78.97	82.36	85.33	88.11
fixed	72.01	75.13	78.31	81.44	84.69	88.15
attention	73.14	76.33	78.94	82.10	85.04	87.73
precision (PWF, ours)	**71.13**	**73.95**	**76.94**	**80.21**	**83.31**	**86.82**

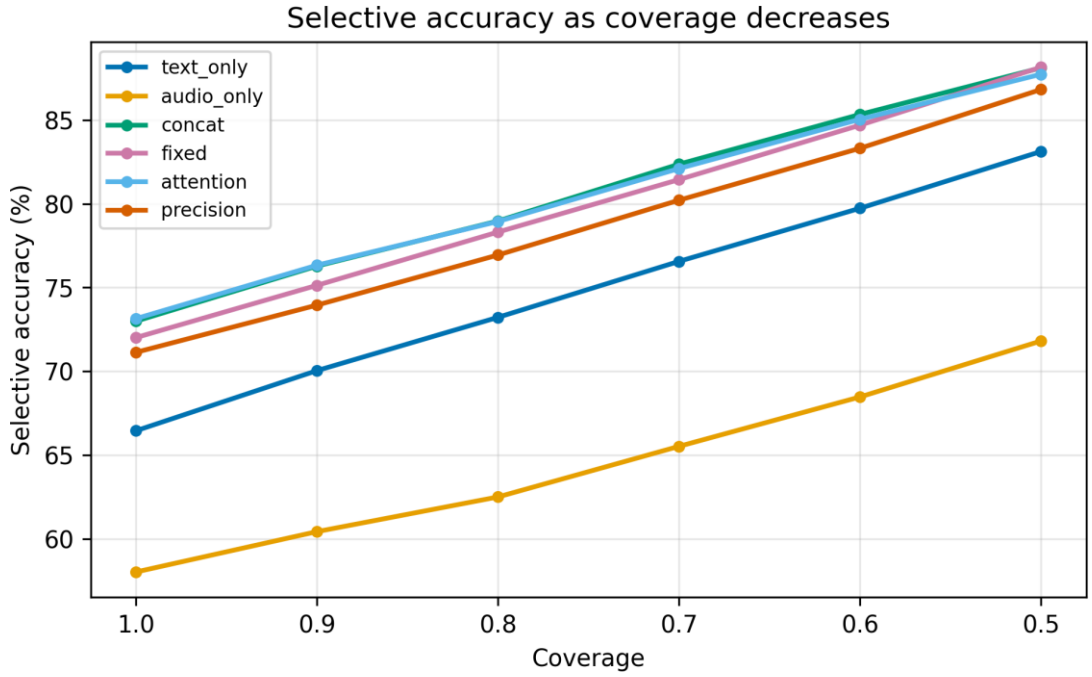

Fig. 5. Selective accuracy as coverage decreases from 1.0 to 0.5, plotting the rows of Table VI. All four fusion curves rise in near parallel, gaining roughly 15 points at half coverage; the unimodal curves remain strictly below across the entire range.

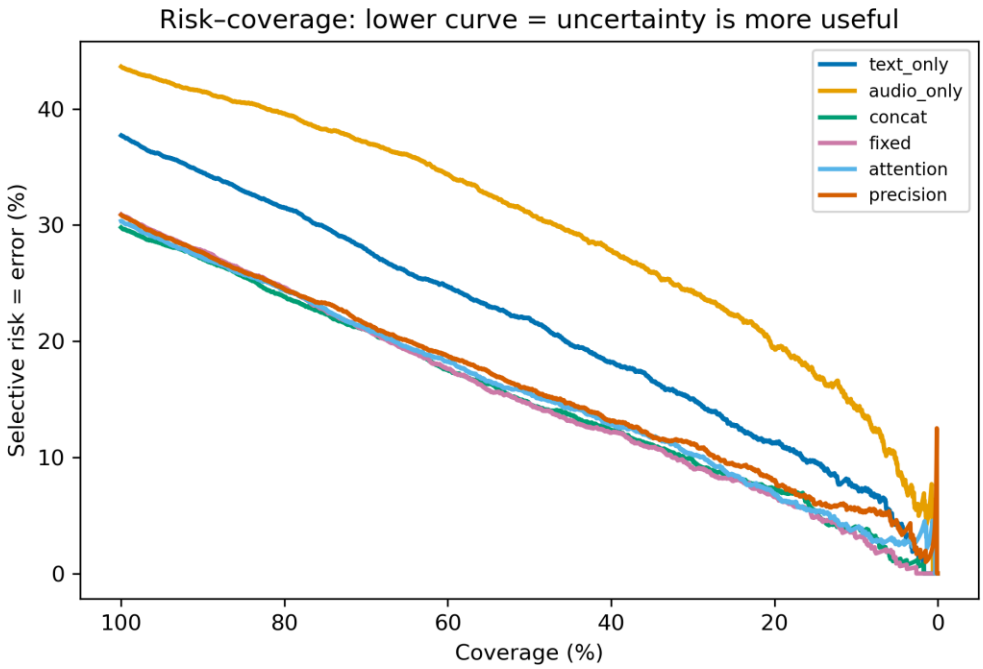

Fig. 6. Risk-coverage curves (pooled LOSO test predictions, lower is better). All four fusion systems form a tight dominant band below the unimodal curves, audio only is uniformly worst. The near vertical spikes at the extreme right correspond to the trivially small retained sets at <2% coverage.

D. Human Ambiguity Alignment

The distinguishing property of PWF is measured by Eq. (14). Across the five LOSO folds, the Spearman correlation between the fused predictive entropy and the human soft-label entropy is +0.299, +0.335, +0.248, +0.276, and +0.272, giving the aggregate $\rho = +0.286 \pm 0.029$ in Table V positive in every fold. With test sets of 1,243-1,507 utterances, these correlations are individually decisive: the t-approximation to the Spearman null distribution yields $p < 10^{-21}$ in every fold (e.g., $\rho = +0.248$ with $n = 1{,}458$ gives $t \approx 9.8$, $p \approx 7\times10^{-22}$), so the alignment cannot be attributed to sampling noise. Fig 6 visualizes the pooled relationship: the human entropy axis is quantized into bands because a small number of annotators produces a discrete set of possible disagreement levels, and within the high ambiguity bands (annotation entropy above roughly 0.6) the mass of model entropy shifts visibly upward relative to the low ambiguity bands on the left. The alignment is moderate rather than strong, which is the expected ceiling: annotation entropy reflects perceiver disagreement that includes information (e.g., facial expression, conversational context) unavailable to an audio text model. Nevertheless, no other system in the comparison possesses any intrinsic mechanism for this alignment softmax entropy of a hard label trained fusion head is optimized to be confident everywhere whereas PWF's heteroscedastic heads are explicitly trained by Eq. (8) to expand variance on inputs where the evidence is genuinely conflicting. This property connects PWF to the annotator distribution modelling line of Wu et al. [8], [23] while requiring only two scalar heads on top of a standard fusion network.

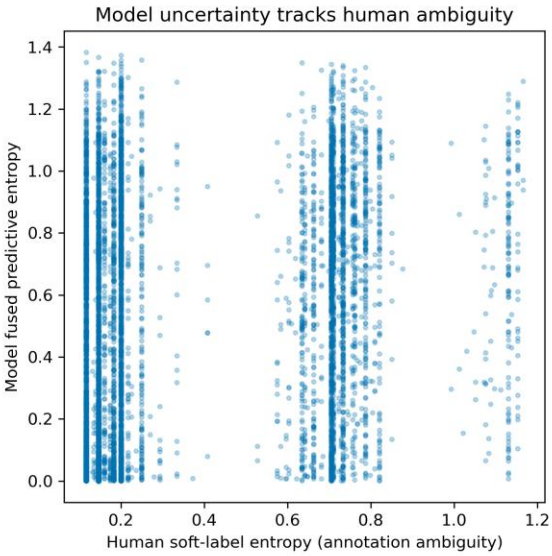

Fig. 6. Model fused predictive entropy versus human soft label entropy (pooled LOSO test set, PWF system). Vertical banding reflects the discrete disagreement levels attainable with a small annotator pool. Model uncertainty shifts upward in the high ambiguity bands, yielding Spearman $\rho = +0.286 \pm 0.029$ across folds.

E. Behavior Under Modality Disagreement

Fusion should matter most on utterances where the two modalities disagree. Fig. 7 isolates the test subset on which the text only and audio only systems predict different labels and reports each system's accuracy there. All four fusion systems land in a narrow band around 60-62%, whereas the full text only system achieves roughly 51% and the audio-only system roughly 42%, the frozen unimodal heads extracted from inside the fusion models (text_only_head, audio_only_head) are lower still, at roughly 43% and 36%. Three observations follow. First, the roughly 10 point margin of every fusion system over the best unimodal system on this subset is where the overall fusion gain of Table III is concentrated on modality agreement utterances all systems perform similarly. Second, PWF is statistically on par with concat, fixed, and attention on this subset, showing that per utterance precision weighting resolves conflicts as effectively as a learned gate. Third, PWF alone reports which modality it trusted through its explicit weights w_txt and w_aud of Eq. (6), turning conflict resolution from an opaque decision into an auditable one, an interpretability property that black-box gates or feature concatenation cannot provide and that evidential fusion frameworks obtain only with heavier machinery [5], [25].

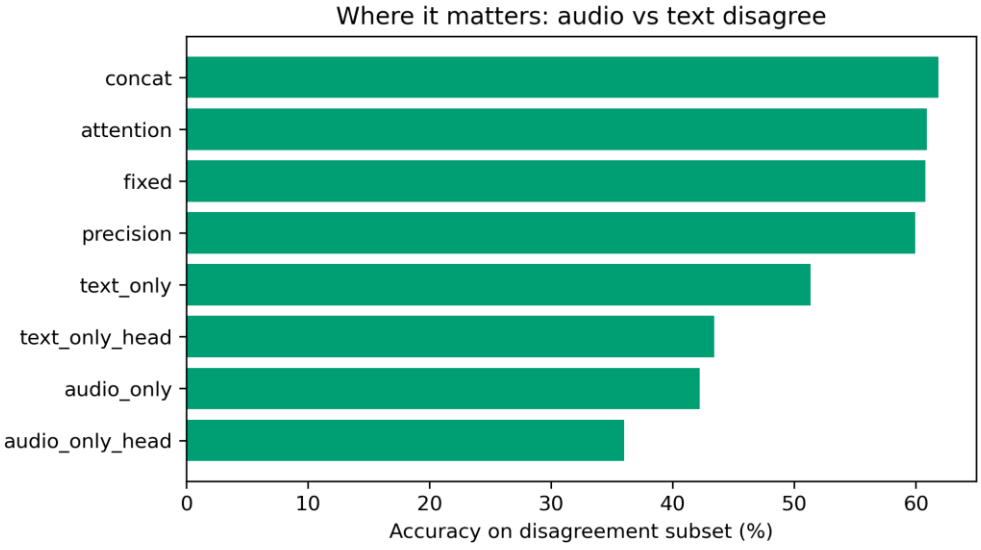

Fig. 7. Accuracy on the audio text disagreement subset (utterances where the two unimodal systems predict different labels). Fusion systems cluster near 60-62%, roughly 10 points above the best standalone unimodal system; the _head variants are the unimodal branches read out from inside the fusion models.

F. Comparison With Published Methods

Table VII and Fig. 8 situate PWF-FT against published IEMOCAP four-class systems that report speaker independent results. Under the matched LOSO protocol, our attention and precision configurations exceed the coattention multilevel system of Zou et al. [14] by up to +3.3 WA points, and both surpass the temporal modelling audio only TIM-Net [18] and the frozen HuBERT Large probe from the SUPERB benchmark [28], the speaker specific wav2vec 2.0 system of Park et al. [19] is comparable to our attention configuration in accuracy (72.14 vs. 73.14 WA). Systems evaluated under random utterance splits or five fold CV with speaker overlap, such as MemoCMT (81.85% WA) [2] and knowledge aware Bayesian coattention (75.5% WA) [16], report higher numbers that are not directly comparable, as protocol leakage on IEMOCAP is documented to inflate accuracy substantially [20]. More importantly, none of the listed methods reports calibration, risk coverage or human alignment measurements; the final columns of Table VII are populated only for our systems. The comparison therefore supports the central claim: PWF-FT delivers accuracy competitive with the published speaker independent state of practice while additionally providing a calibrated, abstention ready, and human aligned uncertainty profile.

TABLE VII. COMPARISON WITH PUBLISHED IEMOCAP 4 CLASS METHODS (VALUES AS REPORTED BY THE ORIGINAL AUTHORS; '–' = NOT REPORTED)

Method	Year	Protocol	WA (%)	UA (%)	ECE_ts	AURC	ρ
PWF-FT, attention gate (ours)	**2026**	**LOSO**	**73.14**	**72.74**	**0.039**	**0.126**	**–**
PWF-FT, precision fusion (ours)	**2026**	**LOSO**	**71.13**	**71.10**	**0.052**	**0.140**	**+0.286**
Park et al., speaker-specific wav2vec 2.0 [19]	2023	Spk-indep.	72.14	72.97	–	–	–
Zou et al., coattention multilevel [14]	2022	LOSO	69.80	71.05	–	–	–
Ye et al., TIM-Net (audio only) [18]	2023	10-fold CV	68.29	69.00	–	–	–
HuBERT-Large frozen probe, SUPERB [28]	2021	Spk-indep.	67.62	–	–	–	–

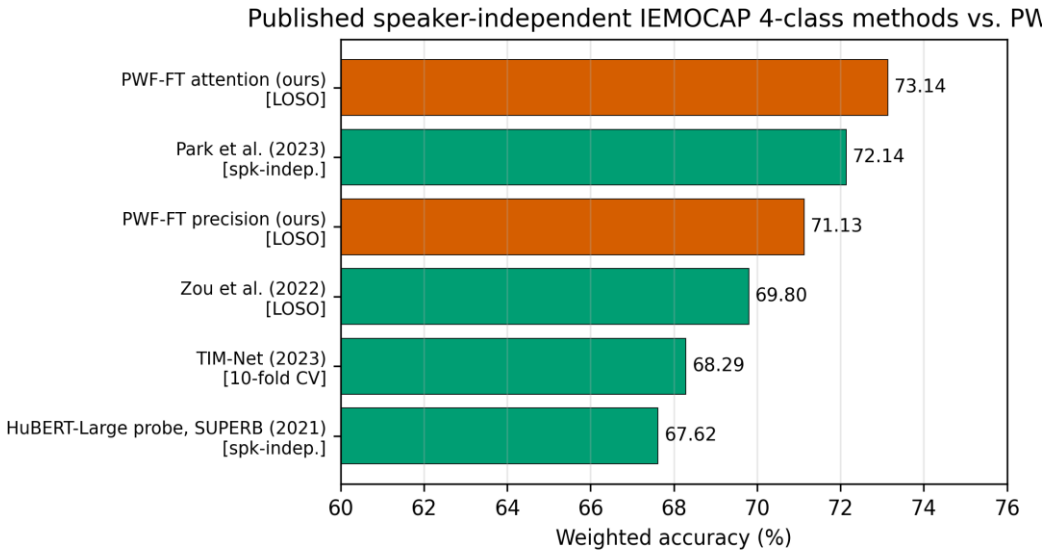

Fig. 8. Weighted accuracy comparison of the speaker independent published methods in Table VII with the two proposed configurations (highlighted); the evaluation protocol of each method is annotated in brackets. PWF-FT attention leads the group; PWF-FT precision remains competitive while uniquely adding the calibrated, human aligned uncertainty profile of Tables V and VI.

G. Discussion

Why do the fusion systems outperform the unimodal baselines? Section VI-E localizes the gain to modality conflict utterances, where a second evidence stream corrects errors that neither branch could fix alone. Why does precision fusion not additionally beat the learned gate in accuracy? Both mechanisms ultimately produce a two way convex combination of the same unimodal logits, and with fine tuned encoders the unimodal logits already saturate most of the learnable signal, leaving the weighting rule little accuracy headroom, the paired tests of Section VI-A quantify the residual cost of the heteroscedastic constraint at roughly two WA points the price of forcing the fusion weights to be meaningful variances rather than free parameters. The genuine differentiators are elsewhere: PWF's weights are supervised by a principled heteroscedastic objective, Eq. (8), which is what produces the fold consistent human-ambiguity alignment of Section VI-D and the auditable per utterance modality attribution of Section VI-E. In deployment terms, the recommendation that follows from Tables V-VII is concrete: if only raw accuracy matters, any fusion head suffices and attention gating is marginally preferable; if the system must abstain, defer, or explain in clinical triage, safety monitoring, or annotation assist pipelines precision fusion supplies calibrated confidence (ECE 0.052 after scaling), competitive selective accuracy (86.8% at half coverage), human aligned uncertainty ($\rho = +0.286$), and explicit modality trust weights at a cost of two scalar heads.

7. CONCLUSION

This paper presented PWF-FT, a precision weighted fusion framework that couples LoRA adapted RoBERTa and HuBERT encoders with heteroscedastic per modality uncertainty heads and inverse variance decision fusion, and evaluated it under a trustworthiness centered LOSO protocol on the crowd annotated Hugging Face release of IEMOCAP. Three findings stand out. First, fusion delivers large, consistent accuracy gains over unimodal systems (up to 73.14% WA versus 66.45% text only and 58.02% audio-only), concentrated on modality disagreement utterances. Second, once encoders are fine tuned, attention gated, concatenation, and fixed fusion are statistically indistinguishable in accuracy (paired tests over shared LOSO folds, $p \geq 0.15$), post temperature scaling calibration is nearly uniform across all fusion heads, and precision fusion carries a small, quantified accuracy cost of about two WA points ($p = 0.003$) an honest accounting that cautions against accuracy only claims for uncertainty aware fusion architectures. Third, precision fusion uniquely yields intrinsic uncertainty that aligns with human annotation ambiguity in every fold ($\rho = +0.286 \pm 0.029$) and exposes interpretable per utterance modality weights, capabilities directly usable for selective prediction and human in the loop deferral.

Limitations and future work. Results derive from a single seed per fold on one corpus; the paired fold-level tests of Section VI-A control for fold difficulty but not for initialization variance, so multiseed replication and cross-corpus transfer (e.g., MSP-IMPROV, MELD) are natural next steps. The loss weights $\alpha = 0.5$ and $\beta = 0.3$ follow the conventions of heteroscedastic multitask training [13] and standard distillation temperature practice rather than a dedicated ablation, and per class error structure (e.g., the well known happy neutral confusion on IEMOCAP) is not broken out here, component wise ablation of Eq. (10) in particular, isolating how much of the human alignment ρ stems from the precision heads versus the soft label term and per-class confusion analysis are the first items of planned follow up work. The human alignment ceiling is bounded by the small IEMOCAP annotator pool, which quantizes the ambiguity axis (Fig. 6), corpora with richer annotation distributions would sharpen the measurement. The precision heads model a single scalar variance per modality, class wise or evidential extensions [5], [22] and conformal wrappers offering coverage guarantees are promising refinements. Finally, integrating the per utterance modality weights into user facing explanations, and testing whether they anticipate sarcasm and other text prosody conflicts, would directly exercise the interpretability that motivates the design.

References

1. C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008.
2. M. Khan, P.-N. Tran, N. T. Pham, A. El Saddik, and A. Othmani, “MemoCMT: Multimodal emotion recognition using cross-modal transformer-based feature fusion,” Scientific Reports, vol. 15, art. no. 5473, 2025, doi: 10.1038/s41598-025-89202-x.
3. Z. Zhao, Y. Wang, and Y. Wang, “Multi-level fusion of wav2vec 2.0 and BERT for multimodal emotion recognition,” in Proc. Interspeech, Incheon, South Korea, 2022, pp. 4725–4729.
4. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. 34th Int. Conf. Machine Learning (ICML), Sydney, Australia, 2017, pp. 1321–1330.

5. 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, 2023, doi: 10.1109/TPAMI.2022.3171983.
6. J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, “From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty,” in Proc. 25th ACM Int. Conf. Multimedia, Mountain View, CA, USA, 2017, pp. 890–897.
7. 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, 2021.
8. W. Wu, C. Zhang, X. Wu, and P. C. Woodland, “Estimating the uncertainty in emotion class labels with utterance-specific Dirichlet priors,” IEEE Transactions on Affective Computing, 2022, doi: 10.1109/TAFFC.2022.3221801.
9. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A robustly optimized BERT pretraining approach,” Facebook AI technical report, arXiv:1907.11692, 2019.
10. 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.
11. 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 Proc. 10th Int. Conf. Learning Representations (ICLR), 2022.
12. 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, art. no. 159, 2026, doi: 10.3390/computers15030159.
13. A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?” in Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 5574–5584.
14. H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, “Speech emotion recognition with co-attention based multi-level acoustic information,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Singapore, 2022, pp. 7367–7371.
15. W. Chen, X. Xing, X. Xu, J. Yang, and J. Pang, “Key-sparse transformer for multimodal speech emotion recognition,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Singapore, 2022, pp. 6897–6901.
16. Z. Zhao, Y. Wang, and Y. Wang, “Knowledge-aware Bayesian co-attention for multimodal emotion recognition,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1–5.
17. D. Sun, Y. He, and J. Han, “Using auxiliary tasks in multimodal fusion of wav2vec 2.0 and BERT for multimodal emotion recognition,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1–5.
18. J. Ye, X.-C. Wen, Y. Wei, Y. Xu, K. Liu, and H. Shan, “Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1–5.
19. 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.
20. N. Antoniou, A. Katsamanis, T. Giannakopoulos, and S. Narayanan, “Designing and evaluating speech emotion recognition systems: A reality check case study with IEMOCAP,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1–5.
21. 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, 2024, doi: 10.1109/TPAMI.2023.3325770.
22. M. Sensoy, L. Kaplan, and M. Kandemir, “Evidential deep learning to quantify classification uncertainty,” in Advances in Neural Information Processing Systems (NeurIPS), Montréal, Canada, 2018, pp. 3179–3189.
23. W. Wu, C. Zhang, and P. C. Woodland, “Estimating the uncertainty in emotion attributes using deep evidential regression,” in Proc. 61st Annu. Meeting of the Association for Computational Linguistics (ACL), Toronto, Canada, 2023, pp. 15681–15695.
24. O. Schrüfer, M. Milling, F. Burkhardt, F. Eyben, and B. Schuller, “Are you sure? Analysing uncertainty quantification approaches for real-world speech emotion recognition,” in Proc. Interspeech, Kos, Greece, 2024, pp. 3210–3214, doi: 10.21437/Interspeech.2024-977.
25. Q. Zhang, H. Wu, C. Zhang, Q. Hu, H. Fu, J. T. Zhou, and X. Peng, “Provable dynamic fusion for low-quality multimodal data,” in Proc. 40th Int. Conf. Machine Learning (ICML), Honolulu, HI, USA, 2023, pp. 41753–41769.
26. 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,” in Proc. 11th Int. Conf. Affective Computing and Intelligent Interaction (ACII), Cambridge, MA, USA, 2023, pp. 1–8, doi: 10.1109/ACII59096.2023.10388152.
27. 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 (NeurIPS), 2020, pp. 12449–12460.
28. S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee,

"SUPERB: Speech processing universal performance benchmark," in Proc. Interspeech, Brno, Czech Republic, 2021, pp. 1194–1198.
29. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 5998–6008.
30. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. 33rd Int. Conf. Machine Learning (ICML), New York, NY, USA, 2016, pp. 1050–1059.
31. Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 4878–4887.
32. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.