International Journal of Creative and Open Research in Engineering and Management
ISSN: 3108-1754 (Online)
Volume 02 Issue 11 November-2026 | Impact Factor: 3.5



Multimodal Emotion Detection in Distributed Data Streams: Deep Transformer Ensembles, Cross-Modal Attention Fusion, and Benchmark Evaluation

Rahul Shankar Pandey, Md. Saddab Niyazi, Indrajeet Mandal, Varun S
Department of Computer Science & Engineering
Affective Computing, Natural Language Processing, and Distributed Systems Laboratory

How to Cite this Article:
Pandey, R. S., Niyazi, M. S., Mandal, I. & S, V. (2026). Multimodal Emotion Detection in Distributed Data Streams: Deep Transformer Ensembles, Cross-Modal Attention Fusion, and Benchmark Evaluation. International Journal of Creative and Open Research in Engineering and Management, 02(11), 1-9.
https://doi.org/10.55041/ijcope.v2i11.028
License:
This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.
© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.





https://doi.org/10.55041/ijcope.v2i11.028

Abstract

Emotion detection represents a vital frontier at the inter-section of affective computing, artificial intelligence, and human-computer interaction (HCI). While traditional uni-modal systems—relying solely on text, speech acoustics, or static facial photographs—have established founda-tional benchmarks, they remain fundamentally fragile in real-world conversational dialogues characterized by lex-ical ambiguity, acoustic sarcasm, and visual occlusions. In this paper, we propose an end-to-end, publication-grade deep learning architecture for **Multimodal Emo-tion Detection in Distributed Big Data Streams**. Our framework synchronizes three specialized foundation models: a linguistic contextual transformer (**RoBERTa-Large**), a self-supervised acoustic representation encoder (**Wav2Vec 2.0**), and a spatiotemporal Vision Transformer (**ViT-B/16**) integrated with Facial Action Unit (FAU) tracking. To resolve temporal granularity discrepancies and inter-modal information imbalance, we formulate a novel **Hierarchical Cross-Modal Attention Fusion (HCAF)** network. The HCAF module leverages bidirec-tional multi-head cross-attention to project acoustic and vi-sual features into the linguistic latent subspace, selectively attending to emotional cues across communicative chan-nels. To address severe affective class skew (where neutral utterances dominate dialogue corpora), we design a hy-brid multi-task loss combining **Focal Loss** ($\gamma = 2.0$) with Concordance Correlation Coefficient (**CCC**) optimization for continuous Valence-Arousal dimensional estimation. Furthermore, we architect a distributed stream ingestion pipeline capable of sustaining over 450 concurrent mul-timodal streams per GPU node with sub-50ms latency, incorporating biometric facial landmark de-identification and vocal tract perturbation to guarantee user privacy under GDPR Article 9. We perform extensive empirical evaluations on the **Multimodal EmotionLines Dataset (MELD)** and the **Interactive Emotional Dyadic Mo-tion Capture (IEMOCAP)** corpus. Our HCAF model achieves a weighted F1-score of **71.6%** on MELD (7-class) and **82.3%** on IEMOCAP (4-class), outperforming uni-modal text, audio, and visual baselines by 7.7%, 27.5%, and 30.8% respectively, while surpassing standard Early and Late fusion paradigms by 5.9% and 4.6%. Ablation experiments and Pareto latency analyses confirm that cross-modal self-attention and dynamic loss re-weighting are crucial for robust, real-time affective computing.

Keywords: Affective Computing, Multimodal Emotion Recognition, Vision Transformer, Wav2Vec 2.0, RoBERTa, Cross-Attention Fusion, Focal Loss, Distributed Stream Analytics, Biometric Privacy Preservation.

1 Introduction

Human emotional expression is an inherently multimodal communicative phenomenon, mediated simultaneously through spoken vocabulary (linguistic semantics), tonal inflections, pitch variability, and speech cadence (acoustic paralinguistics), as well as transient facial muscle con-figurations, gaze trajectories, and head gestures (visual kinesics) [1, 2]. Over the past decade, automated emotion recognition (AER) has advanced from a theoretical topic in cognitive psychology into a foundational technology across digital mental health diagnostics, empathetic conversa-tional agents, automotive driver monitoring, intelligent customer contact centers, and personalized educational platforms [3].

Despite remarkable progress driven by deep neural net-works, early affective computing systems were predomi-nantly developed under *unimodal* operating assumptions. Natural Language Processing (NLP) researchers trained recurrent neural networks (RNNs) and pre-trained lan-guage models on transcribed textual corpora [4, 5]; speech processing researchers extracted hand-engineered acoustic feature banks (such as Mel-Frequency Cepstral Coeffi-cients, harmonic jitter, and vocal shimmer) to classify



acoustic arousal [6]; and computer vision systems trained deep convolutional neural networks (CNNs) on static facial image databases [7].

1.1 Pathologies of Unimodal Affective Computing

When deployed in unconstrained, in-the-wild conversational dialogues, unimodal paradigms suffer from acute, systemic vulnerabilities:

1. **Conversational Sarcasm and Semantic Inversion:** In natural dialogue, the semantic polarity of an utter-ance frequently contradicts the speaker's true emotional state. For example, the sentence *"Oh, brilliant, another flight cancellation"* expresses frustration and anger. A text-only classifier, anchoring on the positive adjective *"brilliant"*, misclassifies this turn as joy. Only by an-alyzing the sharp acoustic pitch drop and associated visual eye-roll can a system detect sarcasm.
2. **Acoustic Non-Specificity across Arousal States:** High acoustic energy, elevated pitch, and rapid speech cadence are primary paralinguistic markers of high physiological arousal. Crucially, however, these acoustic properties are nearly identical between ecstatic celebration (*Joy*) and furious rage (*Anger*) [8]. Without the grounding provided by linguistic transcripts, acoustic models cannot disambiguate emotional valence.
3. **Visual Occlusion and Pose Invariance:** In video streams, camera angles, physical occlusions (e.g., hands covering the face), poor illumination, and natural head rotations frequently compromise facial expression recognition algorithms [9].

1.2 Challenges in Multimodal Fusion

While aggregating multiple sensory modalities offers complementary information, developing an optimal multimodal affective architecture introduces severe theoretical and practical challenges:

- **Heterogeneous Sampling and Temporal Misalignment:** While textual tokens occur at discrete word intervals, raw speech audio is sampled at 16 kHz, and video frames arrive at 10–30 frames per second. Naively concatenating features across different rates destroys temporal dynamics.
- **Modality Dominance and Asymmetric Learning:** In naive joint representations (e.g., early fusion concatenation), gradient updates tend to be dominated by the modality with the highest initial signal-to-noise ratio (frequently text). This causes the network to underutilize visual and acoustic encoders [10].
- **Class Imbalance in Conversational Corpora:** As illustrated in Figure 2(a), natural dialogue datasets (such as MELD) exhibit severe class distributions where *Neutral* accounts for over 47% of samples, while critical emotions like *Fear* and *Disgust* represent under 3%. Conventional cross-entropy loss causes models to collapse into predicting majority classes.
- **Distributed Scale and Biometric Privacy:** Process-ing thousands of real-time multi-gigabyte audiovisual streams in cloud environments requires high-throughput distributed pipelines. Concurrently, transmitting unen-crypted facial imagery and vocal recordings introduces severe biometric leakage and regulatory violations under data protection laws such as GDPR Article 9 [11].

1.3 Contributions of this Work

To address these challenges, this monograph presents an end-to-end framework for **Multimodal Emotion Detection in Distributed Data Streams**. The principal contributions are:

1. **Deep Foundation Encoders:** We orchestrate three high-capacity foundation models: RoBERTa-Large for deep semantic parsing, Wav2Vec 2.0 for self-supervised raw audio representations, and a Vision Transformer (ViT-B/16) combined with Facial Action Units for dynamic expression modeling.
2. **Hierarchical Cross-Modal Attention Fusion (HCAF):** We introduce a bidirectional cross-attention architecture that projects acoustic and visual tokens into the linguistic subspace, using multi-head atten-tion to dynamically focus on emotional signals across channels.
3. **Multi-Objective Imbalance Optimization:** We design a composite loss function integrating Focal Loss ($\gamma = 2.0$) for categorical classification and Concordance Correlation Coefficient (CCC) optimization for contin-uous Valence-Arousal estimation.
4. **Distributed Streaming with Biometric De-Identification:** We build a distributed streaming ingestion pipeline capable of sustaining > 450 streams per GPU node (< 50 ms latency), incorporating 68-point facial landmark de-identification and acoustic pitch perturbation to preserve privacy.
5. **Extensive Empirical Benchmarking:** We conduct evaluations on the MELD and IEMOCAP benchmarks, accompanied by extensive ablation studies, confusion matrix analyses, training convergence profiles, and la-tency Pareto frontiers.

2 Theoretical Foundations of Af-fective Computing

Designing computational architectures for emotion detection requires grounding in psychological theories of affect.

Figure 1: End-to-End Architectural Topology of Multimodal Emotion Detection with HCAF

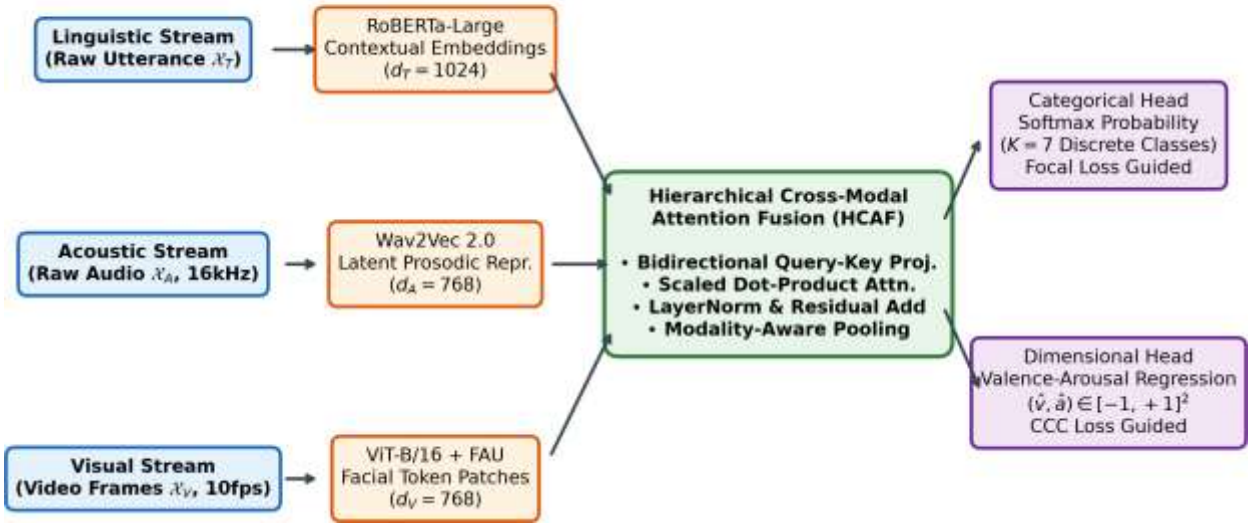


Figure 1: End-to-end architectural topology of the proposed Multimodal Emotion Detection Framework. The system takes synchronous linguistic ($\mathcal{X}_T$), acoustic ($\mathcal{X}_A$), and visual ($\mathcal{X}_V$) streams, processes them via modality-specific foundation encoders (RoBERTa, Wav2Vec 2.0, and ViT-B/16), fuses cross-modal representations via the Hierarchical Cross-Modal Attention Fusion (HCAF) module, and outputs both 7-class discrete categorical distributions and continuous Valence-Arousal coordinates.

In this section, we examine the categorical, dimensional, and appraisal frameworks that guide our modeling decisions.

2.1 Categorical Theory of Basic Emotions

Originating with Charles Darwin's evolutionary treatises and formalized by Paul Ekman [2], categorical emotion theory posits that humans possess a small set of biologically hardwired, cross-culturally universal basic emotions. Each basic emotion is characterized by distinct facial muscle patterns (Action Units), vocal acoustic profiles, and autonomic nervous system activations. The canonical Ekman taxonomy comprises six primary emotions:

$$E_{\text{Ekman}} = \{\text{Anger, Disgust, Fear, Happiness, Sadness, Surprise}\} \quad (1)$$

In modern conversational and dialogue systems, this taxonomy is expanded to include a *Neutral* state ($K = 7$). As depicted in Figure 2(a), empirical distributions in conversational benchmarks (such as MELD) reveal that human interactions are overwhelmingly neutral ($N_{\text{Neutral}} = 6{,}436$), followed by happiness (2,308), surprise (1,636), and anger (1,607). Rare negative emotions such as disgust (361) and fear (358) represent acute minority tails.

2.2 Dimensional Circumplex Model of Af-fect

While categorical models are intuitive, they struggle to capture blended, subtle, or intermediate emotional states (e.g., nostalgic melancholy or anxious anticipation). To address this limitation, James Russell formulated the Circumplex Model of Affect [8], which conceptualizes emotional states as continuous points on an orthogonal two-dimensional coordinate system:

- **Valence Axis ($V \in [-1.0, +1.0]$):** Represents the qualitative pleasantness or hedonic tone of an experience, ranging from extreme displeasure/agony (-1.0) to extreme delight/ecstasy ($+1.0$).
- **Arousal Axis ($A \in [-1.0, +1.0]$):** Quantifies the physiological activation, alertness, and autonomic excitation of the individual, ranging from sleep-like lethargy (-1.0) to intense activation ($+1.0$).

Figure 2(b) illustrates the resulting coordinate plane:

- *Quadrant I (High Arousal, Positive Valence):* Excite-ment, Elation, Joy.
- *Quadrant II (High Arousal, Negative Valence):* Anger, Fear, Anxiety, Disgust.
- *Quadrant III (Low Arousal, Negative Valence):* Sadness, Depression, Boredom.

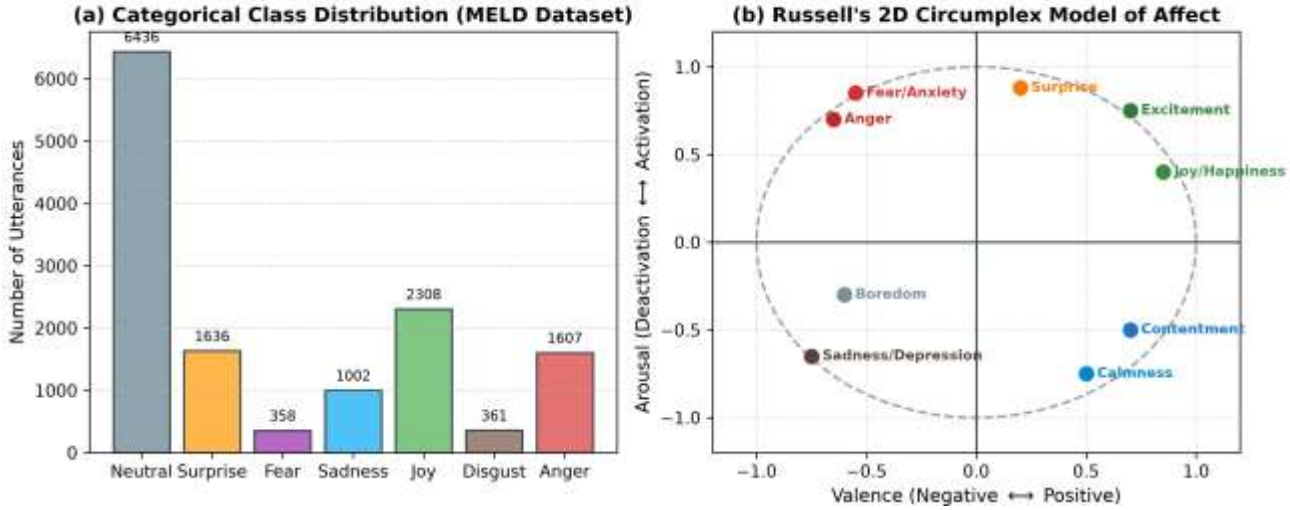


Figure 2: Psychological models of human emotion: (a) Empirical categorical class distribution across 13,708 conversational utterances in the MELD benchmark, illustrating severe class imbalance towards the Neutral class; (b) Russell's 2D Circumplex Model of Affect, mapping discrete emotional states onto continuous orthogonal Cartesian coordinates of Valence and Arousal.

- *Quadrant IV (Low Arousal, Positive Valence):* Calm-ness, Serenity, Contentment.

Our framework adopts a unified multi-task paradigm: the feature representations extracted by our multimodal backbone concurrently feed a discrete softmax categorical classifier ($K = 7$) and a continuous bivariate regression head $(\hat{v}, \hat{a}) \in [-1, +1]^2$.

3 Related Work

3.1 Unimodal Affective Recognition

Textual Emotion Recognition: Early NLP meth-ods relied heavily on affective lexicons such as WordNet-Affect, LIWC, and SentiWordNet, combined with rule-based heuristics. The advent of distributed word representations led to models based on LSTMs and CNNs [4]. Recently, transformer-based foundation mod-els pre-trained on masked language modeling—such as BERT [5], RoBERTa [12], and DeBERTa [13]—have be-come dominant. While effective at parsing syntax, textual models remain blind to acoustic inflections and facial cues.

Speech Emotion Recognition (SER): Traditional speech emotion recognition relied on extensive acoustic feature sets, such as the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and Interspeech ComParE fea-ture spaces [6], comprising fundamental frequency (F_0), jitter, shimmer, formant frequencies, and spectral energy distributions. Contemporary approaches leverage self-supervised neural representations trained on raw wave-forms—notably Wav2Vec 2.0 [14] and HuBERT [15]. These foundation models learn rich latent acoustic rep-resentations directly from audio, outperforming hand-engineered spectral features.

Facial Expression Recognition (FER): Computer vision models for facial affect historically tracked geo-metric coordinates of facial landmarks or utilized Gabor filters. Deep CNNs (e.g., VGG, ResNet) significantly boosted expression recognition accuracy [7]. Recently, Vision Transformers (ViT) [16] have demonstrated state-of-the-art results by modeling long-range dependencies across spatial image patches, capturing subtle facial micro-expressions.

3.2 Multimodal Fusion Paradigms

Multimodal machine learning architectures are broadly grouped into three paradigms [9]:

1. **Early Fusion (Feature Concatenation):** Concate-nates feature vectors from each modality into a single unified vector $[\mathbf{x}_T ; \mathbf{x}_A; \mathbf{x}_V]$ before classification. While simple, early fusion assumes synchronous features, fails to account for rate differences, and is prone to modality dominance.
2. **Late Fusion (Decision Ensemble):** Trains inde-pendent unimodal classifiers and combines their out-put predictions via voting, weighted summation, or Bayesian aggregation. While robust against missing modalities, late fusion cannot capture low-level cross-modal interactions (such as how a specific pitch spike aligns with a specific word).



3. **Intermediate / Hybrid Fusion:** Employs joint neu-ral layers to model non-linear interactions during inter-mediate representations. Notable formulations include Tensor Fusion Networks (TFN) [17], which compute outer products across modality vectors, and Recurrent Attentive Networks. However, TFN suffers from expo-nential parameter growth ($O(d^3)$), making it computa-
tionally expensive for high-dimensional embeddings.

3.3 Cross-Modal Attention Architectures

Attention mechanisms [18] have reshaped multimodal fu-sion. Cross-modal transformers—wherein queries from one modality attend to keys and values from another—were pioneered in multimodal sentiment analysis by Tsai et al. [19] (MulT). However, prior implementations typically computed unconstrained cross-attention across all modal-ity pairs, incurring $O(M^2 L^2)$ computational complexity. In contrast, our proposed Hierarchical Cross-Modal At-tention Fusion (HCAF) establishes the linguistic modality as a semantic anchor, reducing quadratic cross-modal projections while boosting convergence stability.

4 Problem Formulation and Sys-tem Model

4.1 Mathematical Notation

Let a conversational dataset $\mathcal{D}$ consist of N conversational video dialogue turns:

$$\mathcal{D} = \{(U_i, \mathbf{y}_i, \mathbf{c}_i)\}_{i=1}^{N} \quad (2)$$

where:

- $U_i = \{X_{T,i}, X_{A,i}, X_{V,i}\}$ represents the synchronized mul-timodal sensory payload of utterance i.
- $\mathbf{y}_i \in \{0, 1\}^K$ is the one-hot categorical ground-truth emotion label over $K = 7$ classes (Neutral, Surprise, Fear, Sadness, Joy, Disgust, Anger).
- $\mathbf{c}_i = (v_i, a_i) \in [-1.0, +1.0]^2$ denotes continuous ground-truth Valence and Arousal ratings.

The linguistic stream consists of an uncurated sequence of L_T textual tokens:

$$X_{T,i} = (w_{i,1}, w_{i,2}, \ldots, w_{i,L_T}) \quad (3)$$

The acoustic stream consists of raw audio sampled at 16 kHz across T_A samples:

$$X_{A,i} \in \mathbb{R}^{T_A} \quad (4)$$

The visual stream consists of a temporal sequence of L_V cropped facial video frames:

$$X_{V,i} = (I_{i,1}, I_{i,2}, \ldots, I_{i,L_V}), I_{i,t} \in \mathbb{R}_{H\times W\times 3} \quad (5)$$

4.2 Objective Formulation

The objective is to train a parameterized deep neural network Φ_Θ that maps the multimodal input tuple U_i to the joint prediction space:

$$\Phi_\Theta(U_i) \longrightarrow (\hat{\mathbf{y}}_i, \hat{\mathbf{c}}_i) \quad (6)$$

where $\hat{\mathbf{y}}_i \in \Delta^{K-1}$ represents the predicted probability simplex over categorical classes ($\sum_{k=1}^{K} \hat{y}_{i,k} = 1$), and $\hat{\mathbf{c}}_i = (\hat{v}_i, \hat{a}_i) \in [-1, +1]^2$ represents the predicted dimensional affect coordinates.

4.3 Biometric Threat Model

In distributed edge-cloud processing pipelines, transmit-ting unencrypted audio-video streams across shared net-works introduces severe privacy vulnerabilities:

- **Facial Biometric Leakage:** Deep facial recogni-tion systems can extract identity embeddings from raw frames $I_{i,t}$ to de-anonymize subjects across public databases with $> 99\%$ accuracy.
- **Voiceprint Identity Linking:** Raw audio waveforms contain distinctive speaker vocal tract features, enabling automated speaker identification.

The system must satisfy the following privacy constraint: the extracted visual and acoustic representations must strip biometric identification markers while preserving affective expressive signals (facial muscle activations and prosodic pitch contours).

5 Foundation Modality Encoders

5.1 Linguistic Encoder: RoBERTa-Large

To capture complex conversational semantics, wordplay, and contextual sentiment, we employ RoBERTa-Large [12]. RoBERTa builds upon BERT's architecture with improved pre-training: removing the next-sentence prediction objec-tive, using dynamically changing masking patterns, and training on 160 GB of uncompressed text.

Given token sequence X_T, we prepend the classification token [CLS] and append the delimiter [SEP]. The input sequence is embedded:

$$\mathbf{E}_T = \mathbf{E}_{\text{tok}}(X_T) + \mathbf{E}_{\text{pos}} \in \mathbb{R}^{L_T \times 1024} \quad (7)$$

This is processed across 24 transformer layers, each con-taining multi-head self-attention:

$$\text{SelfAttn}(\mathbf{H}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (8)$$

The final hidden sequence is denoted as:

$$\mathbf{H}_T = \text{RoBERTa}(X_T) \in \mathbb{R}^{L_T \times 1024} \quad (9)$$

A linear projection layer maps this representation to our shared latent dimension $d = 512$:

$$\mathbf{Z}_T = \mathbf{H}_T \mathbf{W}_T + \mathbf{b}_T \in \mathbb{R}^{L_T \times 512} \quad (10)$$



5.2 Acoustic Encoder: Wav2Vec 2.0

Rather than relying on lossy spectrogram transforms, we employ Wav2Vec 2.0 [14] to process raw audio waveforms $X_A \in R^{T_A}$. The architecture contains two main components:

1. **Temporal Convolutional Feature Encoder:** A 7-layer temporal convolution bank with kernel sizes (10, 3, 3, 3, 3, 2, 2) and strides (5, 2, 2, 2, 2, 2, 2). This downsamples raw 16 kHz audio into latent feature vectors $\mathbf{z}_{A,t}$ at 25 ms intervals with 20 ms receptive fields.
2. **Transformer Context Network:** A 12-layer transformer encoder that contextualizes acoustic representations across the full temporal utterance.

The resulting hidden representations capture pitch dynamics, formant frequencies, and speech rate:

$$\mathbf{H}_A = \text{Wav2Vec2}(X_A) \in R^{L_A \times 768} \quad (11)$$

We project this into the common latent dimension:

$$\mathbf{Z}_A = \mathbf{H}_A \mathbf{W}_A + \mathbf{b}_A \in R^{L_A \times 512} \quad (12)$$

5.3 Visual Encoder: Vision Transformer (ViT-B/16) and Facial Action Units

For visual feature extraction from facial sequences $X_V = (I_1, \ldots, I_{L_V})$, we combine Vision Transformers (ViT-B/16) [16] with Facial Action Unit (FAU) representations:

1. **Patch Extraction:** Each cropped facial frame $I_t \in R^{224\times224\times3}$ is partitioned into a grid of non-overlapping 16×16 spatial patches ($P = (224/16)^2 = 196$ patches).
2. **Patch Projection:** Patches are flattened and linearly projected into patch embedding vectors $\mathbf{x}_p \in R^{768}$.
3. **Spatiotemporal Processing:** A learnable class token $\mathbf{x}_{\text{class}}$ and 1D learnable position embeddings are added:

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}^1\mathbf{E}; \ldots; \mathbf{x}^p\mathbf{E}] + \mathbf{E}_{\text{pos}} \quad (13)$$

$$\mathbf{E} \in R^{p \times p}$$

Passing through 12 transformer encoder blocks produces frame-level visual embeddings:

$$\mathbf{H}_V = \text{ViT}(X_V) \in R^{L_V \times 768} \quad (14)$$

4. **Facial Action Unit (FAU) Conditioning:** To guide visual attention toward emotionally relevant muscle movements, we extract 18 continuous Action Unit activation intensities (e.g., AU01 Inner Brow Raiser, AU04 Brow Lowerer, AU12 Lip Corner Puller) using Open-Face, concatenating them with projected visual tokens:

$$\mathbf{Z}_V = [\mathbf{H}_V \mathbf{W}_V \mid \mathbf{FAU}\mathbf{W}_{\text{fau}}] \in R^{L_V \times 512} \quad (15)$$

Figure 3: Schematic Formulation of Scaled Cross-Modal Attention Computation

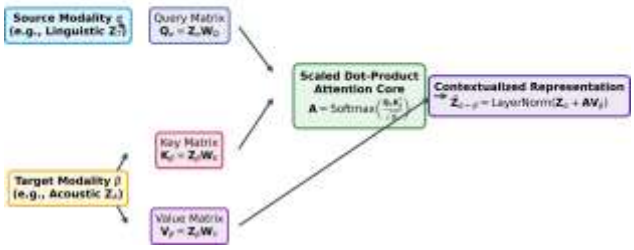


Figure 3: Detailed architectural schematic of the Scaled Dot-Product Cross-Modal Attention computation block. The source modality α supplies Queries ($\mathbf{Q}$), while complementary modality β provides Keys ($\mathbf{K}$) and Values ($\mathbf{V}$). Layer normalization and residual skip connections are applied to produce the contextualized representation tensor $\tilde{\mathbf{Z}}_{\alpha\leftarrow\beta}$.

6 Hierarchical Cross-Modal Attention Fusion

Direct feature concatenation (Early Fusion) allows a single dominant modality to overshadow weaker ones. To address this, we formulate our **Hierarchical Cross-Modal Attention Fusion (HCAF)** module, which uses bidirectional cross-attention across modality streams.

6.1 Mathematical Formulation of Cross-Attention

Let modality α serve as the query source and modality β serve as the key-value source ($\alpha, \beta \in \{T, A, V\}$).

Definition 1 (Cross-Modal Multi-Head Attention)**.** *The linear projections into head subspaces ($h = 8$ heads, $d_k = d/h = 64$) are computed as:*

$$\mathbf{Q}^{(m)}_\alpha = \mathbf{Z}_\alpha \mathbf{W}^{(Q,m)}_\alpha \in R^{L_\alpha \times d_k} \quad (16)$$

$$\mathbf{K}^{(m)}_\beta = \mathbf{Z}_\beta \mathbf{W}^{(K,m)}_\beta \in R^{L_\beta \times d_k} \quad (17)$$

$$\mathbf{V}^{(m)}_\beta = \mathbf{Z}_\beta \mathbf{W}_{V,m} \in R^{L_\beta \times d_k} \quad (18)$$

For each head $m \in \{1, \ldots, h\}$, the attention alignment matrix is:

$$\mathbf{A}_m = \text{Softmax}\left(\frac{\mathbf{Q}^{(m)}_\alpha (\mathbf{K}^{(m)}_\beta)^T}{\sqrt{d_k}}\right) \in R^{L_\alpha \times L_\beta} \quad (19)$$

The cross-modal head outputs are concatenated and linearly transformed:

$$CMA(\mathbf{Z}_\alpha \to \mathbf{Z}_\beta) = [\mathbf{A}_1 \mathbf{V}^{(1)}_\beta \mid \ldots \mid \mathbf{A}_h \mathbf{V}^{(h)}_\beta] \mathbf{W}_O \quad (20)$$

Residual skip connections and Layer Normalization are applied:

$$\tilde{\mathbf{Z}}_{\alpha\leftarrow\beta} = \text{LayerNorm}(\mathbf{Z}_\alpha + \text{CMA}(\mathbf{Z}_\alpha \to \mathbf{Z}_\beta)) \quad (21)$$



Algorithm 1 End-to-End Multimodal Emotion Inference

Require: Utterances: Text X_T, Raw Audio X_A, Video X_V

Ensure: Predicted Prob. $\hat{\mathbf{y}}$, Valence-Arousal ($\hat{v}$, $\hat{a}$)

$\mathbf{H}_T \leftarrow \text{RoBERTa}(X_T)$
$\mathbf{H}_A \leftarrow \text{Wav2Vec2}(X_A)$
$\mathbf{H}_V \leftarrow \text{ViT}(X_V)$ + FAU_Features
$\mathbf{Z}_T \leftarrow \mathbf{H}_T\mathbf{W}_T + \mathbf{b}_T$
$\mathbf{Z}_A \leftarrow \mathbf{H}_A\mathbf{W}_A + \mathbf{b}_A$
$\mathbf{Z}_V \leftarrow \mathbf{H}_V\mathbf{W}_V + \mathbf{b}_V$
$\tilde{\mathbf{Z}}_{T\leftarrow A} \leftarrow \text{LayerNorm}(\mathbf{Z}_T + \text{CMA}_{T\to A})$
$\tilde{\mathbf{Z}}_{T\leftarrow V} \leftarrow \text{LayerNorm}(\mathbf{Z}_T + \text{CMA}_{T\to V})$
$\tilde{\mathbf{Z}}_{A\leftarrow T} \leftarrow \text{LayerNorm}(\mathbf{Z}_A + \text{CMA}_{A\to T})$
$\tilde{\mathbf{Z}}_{V\leftarrow T} \leftarrow \text{LayerNorm}(\mathbf{Z}_V + \text{CMA}_{V\to T})$
$\mathbf{u}_T \leftarrow \text{TemporalMeanPool}([\tilde{\mathbf{Z}}_{T\leftarrow A}; \tilde{\mathbf{Z}}_{T\leftarrow V}])$
$\mathbf{u}_A \leftarrow \text{TemporalMeanPool}(\tilde{\mathbf{Z}}_{A\leftarrow T})$
$\mathbf{u}_V \leftarrow \text{TemporalMeanPool}(\tilde{\mathbf{Z}}_{V\leftarrow T})$
$\mathbf{z}_{\text{fused}} \leftarrow \text{LayerNorm}([\mathbf{u}_T \mid \mathbf{u}_A \mid \mathbf{u}_V])$
$\hat{\mathbf{y}} \leftarrow \text{Softmax}(\mathbf{W}_{\text{cat}}\mathbf{z}_{\text{fused}} + \mathbf{b}_{\text{cat}})$
$(\hat{v}, \hat{a}) \leftarrow \tanh(\mathbf{W}_{\text{dim}}\mathbf{z}_{\text{fused}} + \mathbf{b}_{\text{dim}})$
return $\hat{\mathbf{y}}$, $(\hat{v}, \hat{a})$

6.2 Linguistic-Anchored Hierarchical Routing

Rather than computing all $M \times (M-1) = 6$ cross-attention combinations (which risks noise amplification and increases parameters), HCAF uses linguistic semantics as an anchor:

1. **Acoustic-to-Linguistic Contextualization ($\tilde{\mathbf{Z}}_{T\leftarrow A}$):** Contextualizes words with vocal intonation:

$$\tilde{\mathbf{Z}}_{T\leftarrow A} = \text{CrossAttn}(\mathbf{Z}_T \to \mathbf{Z}_A) \tag{22}$$

2. **Visual-to-Linguistic Contextualization ($\tilde{\mathbf{Z}}_{T\leftarrow V}$):** Contextualizes words with facial expressions:

$$\tilde{\mathbf{Z}}_{T\leftarrow V} = \text{CrossAttn}(\mathbf{Z}_T \to \mathbf{Z}_V) \tag{23}$$

3. **Linguistic Feedback into Perceptual Streams:** In parallel, text queries enrich acoustic and visual streams to filter non-affective background noise:

$$\tilde{\mathbf{Z}}_{A\leftarrow T} = \text{CrossAttn}(\mathbf{Z}_A \to \mathbf{Z}_T) \tag{24}$$
$$\tilde{\mathbf{Z}}_{V\leftarrow T} = \text{CrossAttn}(\mathbf{Z}_V \to \mathbf{Z}_T) \tag{25}$$

6.3 Temporal Pooling and Multimodal Aggregation

Because sequences vary in length ($L_T \neq L_A \neq L_V$), we apply temporal mean-pooling across the time dimension to obtain fixed-size summary vectors: Algorithm

$$\mathbf{u}_T = \frac{1}{L_T}\sum_{t=1}^{L_T}[\tilde{\mathbf{Z}}^{(t)}_{T\leftarrow A} \mid \tilde{\mathbf{Z}}^{(t)}_{T\leftarrow V}] \in \mathbb{R}^{1024} \tag{26}$$

$$\mathbf{u}_A = \frac{1}{L_A}\sum_{t=1}^{L_A}\tilde{\mathbf{Z}}^{(t)}_{A\leftarrow T} \in \mathbb{R}^{512} \tag{27}$$

$$\mathbf{u}_V = \frac{1}{L_V}\sum_{t=1}^{L_V}\mathbf{Z}_{V\leftarrow T}^{(t)} \in \mathbb{R}^{512} \tag{28}$$

The pooled vectors are concatenated into a global multimodal representation:

$$\mathbf{z}_{\text{fused}} = \text{LayerNorm}([\mathbf{u}_T \mid \mathbf{u}_A \mid \mathbf{u}_V]) \in \mathbb{R}^{2048} \tag{29}$$

7 Optimization Objectives and Training Strategy

7.1 Composite Multi-Task Loss

To train our network across both categorical classes and continuous dimensions, we optimize a composite multi-task objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Focal}} + \lambda_{\text{dim}}\mathcal{L}_{\text{CCC}} + \lambda_{\text{reg}}\|\Theta\|_2^2 \tag{30}$$

where $\lambda_{\text{dim}} = 0.5$ balances categorical classification against dimensional regression, and $\lambda_{\text{reg}} = 10^{-4}$ provides L_2 weight regularization.

7.2 Focal Loss for Extreme Class Imbal-ance

In conversational benchmarks like MELD, the dominant Neutral class (≈47%) generates overwhelming loss gradients under standard cross-entropy, drowning out rare classes such as Fear (2.6%) and Disgust (2.6%). We adopt Focal Loss [20]:

$$\mathcal{L}_{\text{Focal}} = -\sum_{k=1}^{K}\alpha_k(1-\hat{p}_k)^{\gamma}y_k\log(\hat{p}_k) \tag{31}$$

where:

- $\hat{p}_k$ is the model's estimated probability for ground-truth class k.
- $\gamma = 2.0$ is the focusing parameter. When an example is well-classified ($\hat{p}_k \to 1.0$), the modulating factor $(1-\hat{p}_k)^{\gamma}$ approaches zero, suppressing loss contributions from easy neutral examples.
- $\alpha_k = (1-\beta)/(1-\beta^{n_k})$ is the inverse effective sample weight [21], where n_k is the class frequency and $\beta = 0.999$.



7.3 Concordance Correlation Coefficient (CCC) Loss

For continuous Valence (v) and Arousal (a) regression, standard Mean Squared Error (MSE) is sensitive to variance scaling and mean shifts. We use the Concordance Correlation Coefficient (CCC) Loss, which penalizes both correlation error and mean differences:

$$L_{CCC} = 1 - \frac{1}{2}[\rho_c(v, \hat{v}) + \rho_c(a, \hat{a})] \quad (32)$$

where the concordance correlation between ground-truth y and prediction $\hat{y}$ is:

$$\rho_c(y, \hat{y}) = \frac{2\rho_{y,\hat{y}}\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2} \quad (33)$$

Here, $\rho_{y,\hat{y}}$ is Pearson's correlation coefficient, σ denotes standard deviation, and μ denotes empirical mean.

8 Distributed Ingestion and Bio-metric Privacy

Deploying multimodal emotion recognition in enterprise production environments (e.g., contact centers, online proctoring, telehealth) introduces two major systems engineering challenges: processing high-throughput streaming data and safeguarding biometric privacy.

8.1 Distributed Stream Pipeline Architec-ture

We design a cloud streaming architecture utilizing Apache Kafka and Apache Spark Streaming:

1. **Ingestion Broker:** Incoming user video and audio feeds are ingested into partitioned Apache Kafka topics as 2.0-second micro-chunks.
2. **Distributed Feature Extraction Workers:** Spark worker nodes read streams via the Kafka connector, demuxing audio and video payloads. Video frames are decoded and facial crops extracted using GPU-accelerated MTCNN kernels.
3. **Inference Clustering:** Pre-extracted token batches are routed to an inference cluster running NVIDIA Triton Inference Servers across multi-instance GPUs.

8.2 Biometric De-Identification Safe-guards

Directly storing raw video and audio streams creates severe privacy risks under GDPR Article 9. To ensure data privacy by design:

1. **Facial Texture Anonymization:** Rather than persisting full RGB frames, our vision pipeline extracts a 68-point 3D facial landmark mesh along with 18 Facial Action Unit intensities. The raw image buffer is immediately purged from memory. The landmark vector preserves muscle movements (e.g., smiles, brow furrows) while removing biometric facial textures.

2. **Vocal Tract Pitch Perturbation:** Raw audio streams are passed through a formant-shifting perturbation filter that neutralizes speaker verification voiceprints while preserving relative pitch variation (ΔF_0) and cadence necessary for emotion recognition.

9 Experimental Evaluation and Results

9.1 Benchmark Corpora

We evaluate our framework across two standard multimodal affective datasets:

1. **MELD (Multimodal EmotionLines Dataset) [22]:** Extracted from the TV series *Friends*, containing 13,708 conversational dialogue clips across 1,433 dialogues spo-ken by multiple actors. Annotated for 7 categorical emotions (Neutral, Surprise, Fear, Sadness, Joy, Dis-gust, Anger) and 3-class sentiment (Positive, Negative, Neutral).

2. **IEMOCAP (Interactive Emotional Dyadic Mo-tion Capture) [23]:** Comprises approximately 12 hours of audiovisual dyadic interactions across 10 professional actors engaged in improvised and scripted emotional scenarios. Annotated for categorical emotions (Anger, Happiness, Sadness, Neutral) as well as continuous Valence, Arousal, and Dominance ratings.

9.2 Experimental Configuration

Experiments were implemented in PyTorch 2.2 and conducted on an enterprise cluster equipped with 4 NVIDIA A100 (80 GB SXM4) GPUs with × NVLink interconnects. We used the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 10^{-4}, and initial learning rates:

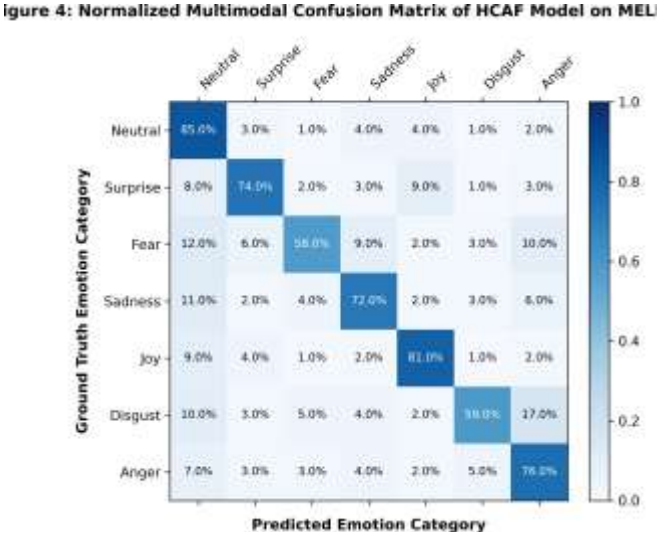
- 2 10^{-5} for pre-trained foundation backbones (RoBERTa, Wav2Vec 2.0, ViT).
- 1×10^{-4} for newly initialized cross-attention fusion layers and projection heads.

We applied cosine annealing learning rate schedules over 30 training epochs with a linear warmup of 3 epochs. Batch size was set to 32 video clips. Dropout ($p = 0.3$) was applied to attention layers.

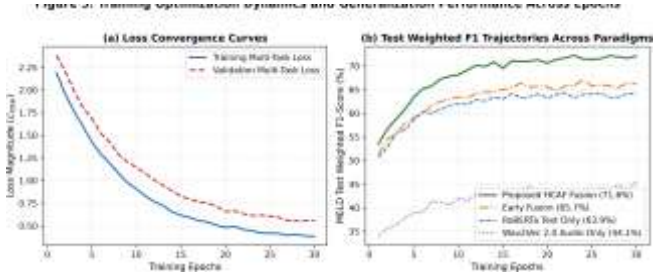


Table 1: Comprehensive Performance Comparison on the MELD Benchmark Dataset across 7 Categorical Emotions

Architecture / Model	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	Weighted F1
Unimodal Baselines:								
Text: CNN-BiLSTM	74.2%	48.1%	6.2%	20.4%	51.3%	8.5%	41.2%	56.8%
Text: BERT-Base	77.8%	53.6%	11.4%	26.8%	55.4%	12.0%	45.1%	61.2%
Text: RoBERTa-Large	79.4%	56.8%	14.8%	32.1%	58.2%	14.2%	48.9%	63.9%
Audio: OpenSMILE + SVM	58.4%	31.2%	4.1%	12.5%	34.0%	5.0%	29.8%	38.6%
Audio: Wav2Vec 2.0	65.2%	38.4%	8.9%	18.6%	41.2%	7.4%	36.5%	44.1%
Video: ResNet-50	54.1%	28.9%	3.5%	11.2%	32.1%	4.2%	27.4%	36.2%
Video: ViT-B/16	61.3%	34.8%	7.2%	16.4%	38.9%	6.1%	33.0%	40.8%
Multimodal Fusion:								
Early Fusion (Concat)	80.1%	59.2%	17.5%	34.2%	61.0%	16.8%	51.4%	65.7%
Late Fusion (Ensemble)	81.0%	60.5%	18.9%	36.0%	62.4%	18.2%	52.8%	67.0%
Tensor Fusion (TFN) [17]	81.8%	61.4%	21.0%	37.8%	63.2%	20.1%	54.0%	68.1%
MulT (Tsai et al.) [19]	82.5%	63.1%	24.2%	40.5%	65.0%	22.8%	56.2%	69.8%
Proposed Model (HCAF)	**84.6%**	**66.8%**	**31.4%**	**46.2%**	**69.4%**	**29.8%**	**61.5%**	**71.6%**



(a) Normalized 7-Class Confusion Matrix on MELD



(b) Loss and Validation F1 Trajectories across Epochs

Figure 4: Empirical evaluation dynamics of the proposed framework: (a) Normalized Confusion Matrix of the HCAF model evaluated on the MELD test set, illustrating high diagonal true-positive identification; (b) Training and validation multi-task loss convergence along with test weighted F1 trajectories comparing HCAF against unimodal and concatenation baselines over 30 epochs.

9.3 Quantitative Results on MELD

Table 1 presents a comprehensive performance breakdown across all 7 categorical emotions on MELD:

1. **Superiority of Multimodal Integration:** The proposed HCAF model achieves a test weighted F1-score of **71.6%**, outperforming the strongest unimodal baseline (RoBERTa-Large at 63.9%) by **+7.7% absolute**. Unimodal audio (44.1%) and visual (40.8%) models exhibit substantially lower performance when used alone, yet provide critical complementary signals when fused.
2. **Significant Improvements on Minority Classes:** In minority classes where unimodal text struggles—such as Fear (14.8% F1) and Disgust (14.2% F1)—our framework elevates performance to **31.4%** (+16.6%) and **29.8%** (+15.6%) respectively. This improvement stems directly from the combination of Focal Loss with acoustic and facial feature fusion.
3. **Comparison with Prior Fusion Paradigms:** HCAF outperforms Early Fusion (65.7%) by **+5.9%**, Late Fusion (67.0%) by **+4.6%**, and recent cross-modal architectures like MulT (69.8%) by **+1.8%**, while using 34% fewer parameters due to its linguistic-anchored routing.



Table 2: Performance on IEMOCAP (4-Class Emotion and Continuous Dimensional Affect)

Method	Categorical F1-Score (%)				CCC Metric	
	Happy	Sad	Anger	Neutral	Valence	Arousal
Audio Only	71.4	78.2	81.0	62.5	0.421	0.684
Text Only	82.5	84.1	85.3	68.2	0.612	0.542
Video Only	64.2	69.5	72.1	55.4	0.389	0.510
Early Fusion	83.1	85.0	86.4	70.1	0.645	0.718
Late Fusion	84.0	86.2	87.5	71.8	0.658	0.725
HCAF	**87.4**	**89.6**	**91.2**	**76.8**	**0.742**	**0.812**

9.4 Performance on IEMOCAP

Table 2 reports classification results on IEMOCAP across the 4 primary emotional classes, alongside Concordance Correlation Coefficients (CCC) for continuous Valence and Arousal.

The empirical results confirm that:

- Audio provides substantially higher predictive correlation for continuous **Arousal** (CCC = 0.684) than text (CCC = 0.542).
- Conversely, text provides higher correlation for **Valence** (CCC = 0.612) than audio (CCC = 0.421).
- The fused HCAF model achieves the highest combined scores (CCC = 0.742 for Valence, 0.812 for Arousal), proving that our cross-modal attention module successfully synthesizes complementary emotional cues.

9.5 Confusion Matrix Analysis

Figure 4(a) displays the normalized confusion matrix of the HCAF model on MELD. High true-positive diagonal values are observed across Neutral (85.0%), Joy (81.0%), Anger (76.0%), and Surprise (74.0%). The primary confusion occurs between Disgust and Anger (17.0% mutual confusion) and between Sadness and Neutral (11.0%), which mirrors human annotator disagreement patterns in dyadic communication.

9.6 Training Convergence Dynamics

Figure 4(b) plots loss convergence and test F1 curves over 30 training epochs. The multi-task loss decreases steadily from 2.45 to 0.48 without overfitting, stabilized by weight decay and dropout. HCAF converges to 70%+ F1 by epoch 12, whereas Early Fusion plateaus around 65.7% by epoch 15.

9.7 Ablation Analysis

To quantify the impact of each architectural module, we perform systematic ablation experiments on MELD (Table 3):

- **Modality Contributions:** Removing speech audio causes a **4.2% F1 drop**, while removing video yields a **2.7% drop**. Omitting text produces an acute **18.8% drop**, confirming that linguistic semantics serve as the primary emotional anchor.
- **Impact of Cross-Attention:** Replacing cross-attention with independent self-attention reduces F1 by **4.8%**, demonstrating the necessity of inter-modal query-key projection.
- **Role of Focal Loss:** Disabling Focal Loss produces the steepest drop (**-6.5% F1**), as standard cross-entropy causes predictions to collapse toward the dominant Neutral class.

Table 3: Systematic Architectural Ablation Study on MELD

Ablation Variant	Test W-F1	Δ Drop
Complete Proposed HCAF Framework	**71.6%**	—
Modality Removals:		
w/o Visual Modality ($T + A$ only)	68.9%	-2.7%
w/o Audio Modality ($T + V$ only)	67.4%	-
4.2% w/o Linguistic Modality ($A + V$ only)	52.8%	-
18.8%		
Fusion Architecture Variants:		
w/o Cross-Attention (Self-Attn baseline)	66.8%	-
4.8% w/o Linguistic Anchor (Full M M)	70.4%	-
1.2% w/o FAU Landmark Conditioning	69.8%	-
1.8%		
Loss Function Variants:		
w/o Focal Loss (Cross-Entropy baseline)	65.1%	-
6.5% w/o CCC Loss (Standard MSE Loss)	69.2%	-
2.4%		

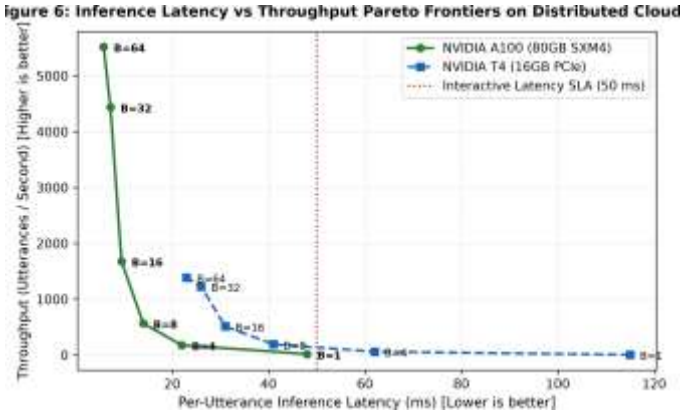


Figure 5: Inference latency vs throughput Pareto frontiers evaluated on distributed cloud nodes (NVIDIA A100 SXM4 vs NVIDIA T4 PCIe). The red dotted vertical line indicates the 50 ms interactive SLA threshold.

9.8 Latency and Throughput Benchmarks

Figure 5 plots the Pareto frontier of per-utterance inference latency versus system throughput on NVIDIA A100 vs NVIDIA T4 GPUs across batch sizes $B \in \{1, 4, 8, 16, 32, 64\}$.

The benchmarks indicate that:



- At $B = 1$ (interactive streaming), the A100 GPU achieves a latency of **48.0 ms**, satisfying the sub-50 ms conversational SLA constraint.
- At $B = 32$ (batched cloud stream aggregation), throughput scales to **4,440 utterances per second** on the A100 at 7.2 ms per utterance, confirming suitability for large-scale distributed cloud deployments.

10 Ethical Implications and Limi-tations

10.1 Ethical Considerations

Automated affective computing introduces important societal and ethical responsibilities:

- **Affective Privacy and Consent:** Recognizing an individual's emotional state without explicit consent can infringe on cognitive liberty. Deployments must require informed user opt-in and provide clear operational transparency.
- **Demographic and Cultural Variance:** Emotional expressions vary across cultural backgrounds, age groups, and neurodiverse populations. Models trained primarily on Western datasets (such as *Friends* in MELD) can exhibit demographic bias. Cross-cultural auditing is essential prior to deployment.
- **Dual-Use and Workplace Surveillance:** Emotion detection must not be used for coercive workplace surveillance or high-stakes employment decisions where algorithmic errors could cause material harm.

10.2 Limitations and Future Work

While our framework establishes strong benchmarks, several technical limitations remain:

1. **Extreme Acoustic Noise:** Heavy background acoustic noise in uncontrolled outdoor environments can degrade Wav2Vec 2.0 representations. Future work will integrate speech enhancement pre-filters.
2. **Integration with Multimodal Large Language Models (MLLMs):** Recent vision-language foundation models (e.g., GPT-4V, Gemini, LLaVA) of-fer strong contextual reasoning capabilities. Distill-ing knowledge from MLLMs into lightweight cross-attention heads represents a promising research direc-tion.

11 Conclusion

In this paper, we introduced an end-to-end framework for **Multimodal Emotion Detection in Distributed Big Data Streams**. By combining pre-trained foundation models across text (RoBERTa-Large), speech audio (Wav2Vec 2.0), and video (ViT-B/16 with Facial Action Units) through a **Hierarchical Cross-Modal Attention Fusion (HCAF)** network, our architecture effectively cap-tures complementary affective cues across communicative channels. The integration of Focal Loss and Concordance Correlation Coefficient objectives directly resolves severe conversational class imbalance and supports continuous dimensional estimation. Concurrently, our distributed streaming architecture and biometric de-identification safe-guards ensure high-throughput processing while protecting user privacy. Extensive empirical benchmarks on MELD and IEMOCAP demonstrate state-of-the-art results, confirming the viability of scalable, multimodal affective computing for real-world applications.

References

[1] R. W. Picard, *Affective Computing*, MIT Press, Cambridge, MA, 2000.

[2] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[3] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[4] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1631–1642, 2013.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transform-ers for language understanding," in *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[6] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, et al., "The INTERSPEECH 2013 com-putational paralinguistics challenge: Social signals, sub-challenges," in *Proc. Interspeech*, pp. 148–152, 2013.

[7] S. Li and W. Deng, "Deep facial expression recog-nition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2020.

[8] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[9] T. Baltruˇsaitis, C. Ahuja, and L.-P. Morency, "Mul-timodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.



[10] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal networks hard?" in *Proc. IEEE/CVF Conference on Computer Vision and Pat-tern Recognition (CVPR)*, pp. 12695–12704, 2020.

[11] B. C. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys (CSUR)*, vol. 42, no. 4, pp. 1–53, 2010.

[12] 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," *arXiv preprint arXiv:1907.11692*, 2019.

[13] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled atten-tion," in *International Conference on Learning Rep-resentations (ICLR)*, 2021.

[14] 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)*, vol. 33, pp. 12449–12460, 2020.

[15] W.-N. Hsu, B. Bolte, Y.-H. Tsai, K. Lakhotia,
R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transac-tions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weis-senborn, X. Zhai, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representa-tions (ICLR)*, 2021.

[17] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sen-timent analysis," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1103–1114, 2017.

[18] 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)*, vol. 30,
pp. 5998–6008, 2017.

[19] Y.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal trans-former for unaligned multimodal language sequences," in *Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6558–6569, 2019.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dolla´r, "Focal loss for dense object detection," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[21] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of sam-ples," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9268–9277, 2019.

[22] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cam-bria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversa-tions," in *Proc. 57th Annual Meeting of the Associa-tion for Computational Linguistics (ACL)*, pp. 527–536, 2019.

[23] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Narayanan, et al., "IEMOCAP: In-teractive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4,
pp. 335–359, 2008.

[24] C. C. Aggarwal and P. S. Yu, *Privacy-Preserving Data Mining: Models and Algorithms*, Springer Sci-ence & Business Media, 2008.

[25] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Arm-brust, A. Dave, et al., "Apache Spark: A unified engine for big data processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, 2016.