Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Audio--Lyrics Cross-Attention for Improved Music Emotion Recognition in an Adaptive Dual-View Architecture

Mahesh Bhoopathi C S
Mepco Schlenk Engineering College
Yashwanth D
Mepco Schlenk Engineering College
Dharani M
dharanim@mepcoeng.ac.in

Mepco Schlenk Engineering College

Research Article

Keywords: music emotion recognition, multimodal deep learning, dual-view acoustic features, cochleagram, cross-attention, progressive confidence labelling, style-anchored memory learning, circumplex model

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11565875/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Audio–Lyrics Cross-Attention for Improved Music Emotion Recognition in an Adaptive Dual-View Architecture

Mahesh Bhoopathi C S[1], Yashwanth D[1], Dharani M[1*]

[1*]Department of Artificial Intelligence and Data Science, Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India.

*Corresponding author(s). E-mail(s): dharanim@mepcoeng.ac.in;
Contributing authors: achinnadurai965_bai27@mepcoeng.ac.in;
yashwanthd_bai27@mepcoeng.ac.in;

Abstract

Music emotion recognition (MER) estimates the emotion a song conveys to its listeners. Acoustic signals describe arousal reliably, whereas lyrics say more about valence, yet most bimodal systems simply concatenate the two feature vectors. We present a dual-view architecture in which a Dual-Stream Attention Fusion (DSAF) encoder couples a 128-band log-mel spectrogram with an 84-band ERB-spaced gammatone cochleagram. Bidirectional cross-attention then allows the acoustic tokens and the embeddings of a frozen RoBERTa-base lyric encoder to query each other. Two training mechanisms keep joint quadrant classification and valence–arousal regression stable: Progressive Confidence Labelling (PCL), which weights pseudo-labels by the Jensen–Shannon divergence between the two acoustic views under a threshold rising from 0.80 to 0.90, and Style-Anchored Memory Learning (SAML), which anchors the embedding space with a momentum-updated contrastive queue of 512 entries. On a balanced 2,000-song subset of the MERGE corpus, the best of three seeds reaches 73.33% accuracy and 73.50% Macro-F1; the three-seed mean Macro-F1 is 72.97% and the soft-voting ensemble scores 72.90%. The concordance correlation coefficient is 0.762 for valence and 0.676 for arousal, and the mean one-vs-rest AUC is 0.915. Ablations and an axis-wise error analysis show where the gains come from and that calm and sad music remain the hardest pair to separate.

Keywords: music emotion recognition, multimodal deep learning, dual-view acoustic features, cochleagram, cross-attention, progressive confidence labelling, style-anchored memory learning, circumplex model

1 Introduction

Music can lift, unsettle or calm a listener, and the same song rarely produces the same reaction in two people. Automatic systems that recognise the emotion expressed by a piece of music are useful for streaming platforms, recommendation engines, soundtrack selection for video, and affective-computing applications such as music-based therapy (Yang and Chen 2012; Panda et al. 2023). They are also hard to build. Annotators disagree about the emotion of a given track, artists often leave their intent ambiguous, and the sound of a song interacts with the meaning of its words in ways that a simple model does not capture.

The literature describes musical emotion either with discrete categories or along continuous dimensions (Eerola and Vuoskoski 2011). This work takes the dimensional route and uses the circumplex model of Russell (1980). Along one

axis valence runs from unpleasant to pleasant, and along the other arousal runs from subdued to energetic. Cutting the plane at the origin yields four regions, which we call quadrants:

- Q1 (positive valence, high arousal): happy or excited music;
- Q2 (negative valence, high arousal): angry or tense music;
- Q3 (negative valence, low arousal): sad or depressed music;
- Q4 (positive valence, low arousal): calm or relaxed music.

Audio and lyrics are complementary sources of evidence. Energy, harmonic motion, rhythmic regularity and spectral flux follow arousal closely, but they leave valence ambiguous (Panda et al. 2023; Juslin and Laukka 2004). Lyrics behave the other way round: sentiment is frequently stated outright, while the text carries no information about tempo or loudness (Hu and Downie 2010; Malheiro et al. 2018). Consider a slow, minor-key song with muted drums whose final lines turn hopeful. An audio-only model will tend to assign it low valence, and a text-only model will miss how subdued the performance is.

We see three weaknesses in current bimodal MER systems. First, audio and text are usually joined by early or late concatenation of feature vectors, so neither modality can query the other at the token level. Second, the acoustic front end rarely goes beyond the mel spectrogram, although the cochleagram describes basilar-membrane filtering more directly and has recently been shown to help when used next to the mel view (Li et al. 2026a). Third, training quadrant classification jointly with coordinate regression can produce conflicting gradients, and self-training on the model's own predictions makes the problem worse when early errors are reinforced.

In response, we design a dual-view cross-attention network that is trained end to end and tested on the MERGE benchmark (Louro et al. 2026). Our main contributions are listed below.

1. **Two-stage fusion.** DSAF first combines log-mel and cochleagram features by two rounds of bidirectional cross-attention. A second stage applies cross-attention between the acoustic tokens and frozen RoBERTa-base lyric embeddings in both directions, so the lyrics can guide which acoustic frames receive weight and the audio can do the same for the text.
2. **Progressive Confidence Labelling.** A curriculum turns the Jensen–Shannon divergence between the predictions of the two acoustic views into a per-sample reliability score, and uses that score together with a rising threshold to decide which pseudo-labels are accepted.
3. **Style-Anchored Memory Learning.** A momentum queue trained with a supervised InfoNCE objective keeps intermediate representations close to stable emotion prototypes and limits drift during training.
4. **Thorough evaluation.** We report per-quadrant precision and recall, ROC and precision–recall curves, regression diagnostics, component ablations, an error decomposition along the valence and arousal axes, and a confidence analysis on 300 held-out songs, and we describe a working deployment.

2 Background and Related Studies

2.1 How emotion in music is modelled

Psychological work on musical emotion is organised around two families of models. Categorical models assign a piece to one of a small set of labels such as happy, sad or angry, whereas dimensional models place it in a continuous space. Russell's circumplex model (Russell 1980) is the most common dimensional choice in MER because it needs only two axes and because its quadrants map naturally onto familiar mood words. Eerola and Vuoskoski (2011) compared the two families on music excerpts and found that dimensional ratings were reliable and that categorical judgements could be recovered from them reasonably well, which supports the quadrant-plus-coordinate formulation used here. Like most MER corpora, MERGE is annotated for perceived rather than felt emotion (Juslin and Laukka 2004), and so is this study.

2.2 Acoustic representations and unimodal MER

Early MER systems relied on hand-crafted descriptors such as MFCCs, spectral centroid,

zero-crossing rate and chroma. The survey of Panda et al. (2023) shows that such features support coarse arousal classification fairly well but describe temporal structure poorly.

The mel scale approximates perceived pitch spacing, but it is not a model of the auditory periphery. The gammatone filterbank, which spaces filters on the equivalent rectangular bandwidth (ERB) scale of Glasberg and Moore (1990), imitates cochlear frequency selectivity more closely. The recent Memo2496 work (Li et al. 2026a) reported on its own expert-annotated dataset that pairing log-mel spectrograms with gammatone cochleagrams separates emotions more clearly than using either view alone, and the dual-view idea in our model is inspired by that result. That framework is audio-only, so the question of how a dual-view acoustic encoder should interact with lyrics has not been examined.

2.3 Lyrics-based MER

Text-based MER began with bag-of-words models and affective lexicons, and then moved to word embeddings combined with recurrent networks. Hu and Downie (2010) showed that lyrics can outperform audio for certain mood categories, and Malheiro et al. (2018) built emotionally relevant lyric features for both classification and regression. Large pretrained language models, notably BERT (Devlin et al. 2019) and its successor RoBERTa (Liu et al. 2019), now handle long-range dependencies, negation and figurative language much better than earlier encoders. A lyrics-only model still has little to work with for instrumental passages, for ambient genres, and for energetic songs built on repeated words, which is why a second modality remains necessary.

2.4 Multimodal fusion

Multimodal learning in general offers several fusion strategies, ranging from simple concatenation through bilinear and tensor-based fusion (Zadeh et al. 2017) to cross-modal transformers in which one modality attends to another (Tsai et al. 2019); see Baltrušaitis et al. (2019) for a survey. In MER, Louro et al. (2026) set the MERGE baselines with feedforward networks on concatenated audio and lyric features and found that bimodal models consistently beat unimodal ones. Later studies introduced audio–MIDI cross-attention (Xiao et al. 2025), hierarchical cross-modal attention networks (Zhao et al. 2022), and tri-modal alignment of lyrics, vocals and accompaniment (Li et al. 2026b). Among the work we reviewed, none combines physiologically motivated dual-view acoustics with a contextual language model, and none adds a self-training regulariser against feature drift.

2.5 Pseudo-labelling and contrastive regularisation

Pseudo-labelling (Lee 2013) trains on a model's own confident predictions, and confidence-thresholded variants such as FixMatch (Sohn et al. 2020) are widely used. The risk is confirmation bias: a wrong but confident prediction is fed back as a target. Our PCL module reduces this risk by requiring agreement between two independent acoustic views before a pseudo-label is accepted, using the Jensen–Shannon divergence (Lin 1991) as the disagreement measure. On the representation side, contrastive objectives based on noise-contrastive estimation in its InfoNCE form (van den Oord et al. 2018), momentum-updated memories (He et al. 2020) and label-aware positives (Khosla et al. 2020) keep embeddings well organised. SAML borrows from these ideas and applies them to emotion prototypes.

3 Problem Formulation and Dataset

3.1 Affective space

Following the circumplex model, each song i has a quadrant label $y_i \in \{Q1, Q2, Q3, Q4\}$ and a coordinate vector $\mathbf{r}_i = [v_i, a_i] \in [-1, 1]^2$, where v_i is valence and a_i is arousal. The task is multi-task. For every song the model outputs a predicted quadrant $\hat{y}_i$ and predicted coordinates $\hat{\mathbf{r}}_i$.

3.2 The MERGE benchmark and our split

All experiments use MERGE (Music Emotion Recognition next GEneration) (Louro et al. 2026), a large quality-controlled public bimodal MER

corpus. It provides 3,554 audio tracks and 2,568 lyrics, of which 2,216 songs have complete bimodal data confirmed by certified music specialists. To avoid quadrant skew we draw a balanced bimodal subset of exactly 2,000 songs, 500 per quadrant. The subset is divided into 1,400 training, 300 validation and 300 test songs, and every partition contains the four quadrants in equal numbers (350, 75 and 75 songs per quadrant respectively).

Because the split is balanced, a classifier that always predicts one quadrant would reach 25% accuracy.

4 Model Architecture

4.1 Overview

The full architecture is drawn in Fig. 1. Information passes through five stages: (1) dual-view acoustic preprocessing, (2) lyric embedding, (3) DSAF, (4) audio–lyrics cross-attention, and (5) multi-task prediction supported by the PCL and SAML modules.

4.2 Audio preprocessing and dual-view extraction

Every recording is resampled to $f_s = 22{,}050$ Hz and RMS-normalised so that loudness differences between tracks do not influence the result. We want to sample several parts of a song (verse, chorus, bridge) without heavy computation. We therefore take $N_{\text{seg}} = 5$ non-overlapping 3-second segments per track on a linear schedule, and pre-emphasise each segment to offset spectral tilt. A 2,048-sample window with a 512-sample hop gives $T = 87$ frames per segment in two representations:

- a log-mel spectrogram $\mathbf{M} \in \mathbb{R}^{F_m \times T}$ with $F_m = 128$ triangular mel filters, which follows perceived pitch;
- a log-cochleagram $\mathbf{C} \in \mathbb{R}^{F_c \times T}$ with $F_c = 84$ gammatone filters spaced on the ERB scale, which follows cochlear frequency selectivity.

The impulse response of a gammatone filter is $g(t) = t^{\,n-1} e^{-2\pi b t} \cos(2\pi f_c t + \varphi)$ where the bandwidth parameter b increases with the filter's centre frequency f_c following the ERB scale. Both views are converted to a logarithmic power scale with an 80 dB dynamic range relative to the peak power $P_{\max}$:

$$X_{\text{dB}} = 10 \log_{10}\big(\max(P, \varepsilon)\big), \quad \varepsilon = P_{\max} \cdot 10^{-8}, \tag{1}$$

where P is the spectral power.

4.3 Contextual lyric encoding

Each lyric text is cleaned before encoding. Section tags such as [Verse] and [Chorus], artist metadata and duplicated choruses are removed, which keeps repetition from inflating the token count. We pass the cleaned text l through RoBERTa-base (Liu et al. 2019), a 12-layer encoder with hidden size 768 whose weights stay frozen. Some songs are longer than its 512-token limit, so the text is split into sliding windows of $W = 126$ tokens and at most $N_w = 12$ windows are kept. Each window is mean-pooled into a 768-dimensional vector and linearly projected by f_{proj} to the shared dimension $d = 128$. A mask $\mathbf{m} \in \{0,1\}^{N_w}$ marks the windows that contain real text:

$$\mathbf{L} = f_{\text{proj}}\big(\text{pool}(\text{RoBERTa}(l))\big) \in \mathbb{R}^{N_w \times d}. \tag{2}$$

Keeping the language model frozen limits the number of trainable parameters, which matters with only 1,400 training songs.

4.4 Module 1: Dual-Stream Attention Fusion

Both spectrograms are first turned into token sequences over time:

$$\bar{\mathbf{M}} = \mathbf{M}^\top \mathbf{W}_m + \mathbf{b}_m, \qquad \bar{\mathbf{C}} = \mathbf{C}^\top \mathbf{W}_c + \mathbf{b}_c, \tag{3}$$

with $\mathbf{W}_m \in \mathbb{R}^{128 \times d}$, $\mathbf{W}_c \in \mathbb{R}^{84 \times d}$ and $d = 128$. We write $\text{CA}(\mathbf{X}, \mathbf{Y}, \mathbf{Y})$ for multi-head cross-attention in which $\mathbf{X}$ supplies the queries and $\mathbf{Y}$ supplies the keys and values. With h heads of size $d_k = d/h$, each head computes

$$\text{softmax}\left(\frac{(\mathbf{X}\mathbf{W}_j^Q)(\mathbf{Y}\mathbf{W}_j^K)^\top}{\sqrt{d_k}}\right)\mathbf{Y}\mathbf{W}_j^V, \tag{4}$$

and the heads are concatenated and projected (Vaswani et al. 2017). DSAF runs two rounds of bidirectional cross-attention between the two views (Fig. 1):

$$\mathbf{M}^{(1)} = \text{CA}(\bar{\mathbf{M}}, \bar{\mathbf{C}}, \bar{\mathbf{C}}), \tag{5}$$

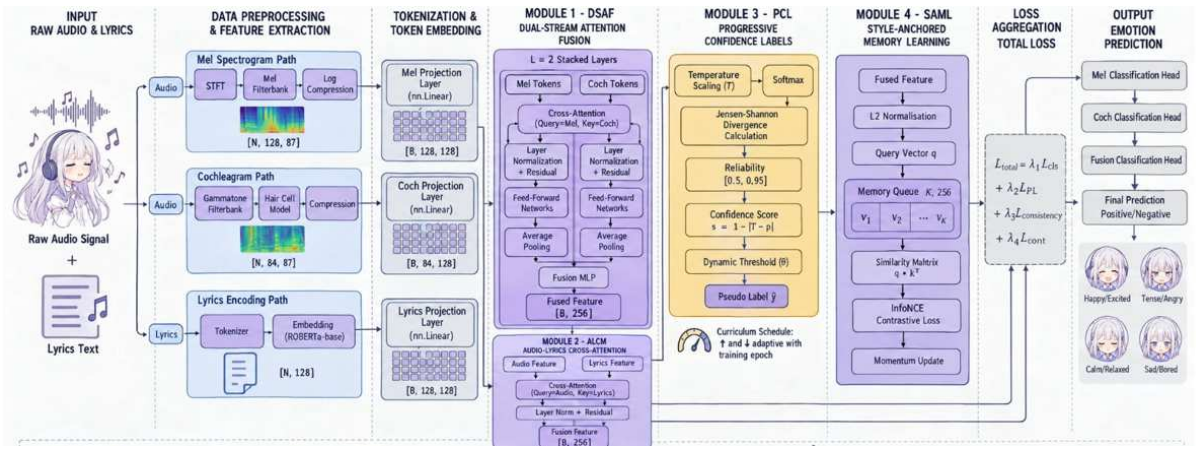


Fig. 1 Architecture of the proposed system. Audio and lyrics first undergo feature extraction and token embedding. The audio is processed along a log-mel path and a cochleagram path, while the lyrics go through a frozen RoBERTa-base encoder. Module 1 (DSAF) merges the two acoustic streams using stacked cross-attention layers, and Module 2 applies cross-attention between audio and lyrics. Module 3 (PCL) chooses pseudo-labels using the Jensen–Shannon divergence and a changing confidence threshold, and Module 4 (SAML) adds a momentum-queue contrastive loss. The combined loss trains the emotion prediction heads.

$$\mathbf{C}^{(1)} = \mathrm{CA}(\bar{\mathbf{C}}, \bar{\mathbf{M}}, \bar{\mathbf{M}}), \tag{6}$$

$$\mathbf{M}^{(2)} = \mathrm{CA}(\mathbf{M}^{(1)}, \mathbf{C}^{(1)}, \mathbf{C}^{(1)}), \tag{7}$$

$$\mathbf{C}^{(2)} = \mathrm{CA}(\mathbf{C}^{(1)}, \mathbf{M}^{(1)}, \mathbf{M}^{(1)}). \tag{8}$$

As Fig. 1 indicates, each attention layer is followed by a residual connection with layer normalisation and by a feed-forward network. The token sequences of the five segments are then concatenated along the time axis into $T_{\mathrm{tot}} = 435$ frames. Temporal average pooling gives the summaries $\mathbf{z}_{\mathrm{mel}}$ and $\mathbf{z}_{\mathrm{coch}}$, which are concatenated and compressed by a two-layer MLP f_{fuse}:

$$\mathbf{z}_{\mathrm{fuse}} = f_{\mathrm{fuse}}\big([\mathbf{z}_{\mathrm{mel}}; \mathbf{z}_{\mathrm{coch}}]\big) \in \mathbb{R}^d. \tag{9}$$

The cross-modal stage receives the averaged sequence $\mathbf{Z}_a = 0.5\,(\mathbf{M}^{(2)} + \mathbf{C}^{(2)}) \in \mathbb{R}^{T_{\mathrm{tot}} \times d}$.

4.5 Module 2: Audio–lyrics cross-attention

Cross-attention is applied in both directions between the acoustic tokens $\mathbf{Z}_a$ and the lyric tokens $\mathbf{L}$:

$$\mathbf{H}_{a \leftarrow l} = \mathrm{CA}(\mathbf{Z}_a, \mathbf{L}, \mathbf{L}), \tag{10}$$

$$\mathbf{H}_{l \leftarrow a} = \mathrm{CA}(\mathbf{L}, \mathbf{Z}_a, \mathbf{Z}_a). \tag{11}$$

In Eq. (10) the audio asks which parts of the lyrics account for the energy it carries. In Eq. (11) the lyrics ask which sounds accompany which words. After masked pooling, which ignores padded lyric windows, both outputs are concatenated with $\mathbf{z}_{\mathrm{fuse}}$ and projected by f_{bi} into the multimodal embedding:

$$\mathbf{z}_{\mathrm{bi}} = f_{\mathrm{bi}}\big([\mathrm{pool}(\mathbf{H}_{a \leftarrow l}); \mathrm{pool}(\mathbf{H}_{l \leftarrow a}); \mathbf{z}_{\mathrm{fuse}}]\big) \in \mathbb{R}^d. \tag{12}$$

Including $\mathbf{z}_{\mathrm{fuse}}$ gives the head a direct acoustic path, so the model still works reasonably when a song has little usable text.

4.6 Multi-task heads

The embedding $\mathbf{z}_{\mathrm{bi}}$ feeds two heads: a quadrant classifier $\mathbf{o} = \mathbf{W}_q \mathbf{z}_{\mathrm{bi}} + \mathbf{b}_q \in \mathbb{R}^4$ and a valence–arousal regressor $\hat{\mathbf{r}} = \mathbf{W}_v \mathbf{z}_{\mathrm{bi}} + \mathbf{b}_v \in \mathbb{R}^2$. Two auxiliary classifiers on $\mathbf{z}_{\mathrm{mel}}$ and $\mathbf{z}_{\mathrm{coch}}$ supply unimodal supervision and also provide the two views that PCL compares.

4.7 Module 3: Progressive Confidence Labelling

A pseudo-label is trustworthy when independent views of the same song agree. The temperature-scaled softmax outputs $\mathbf{p}_{\mathrm{mel}}$ and $\mathbf{p}_{\mathrm{coch}}$ of the auxiliary heads should therefore be close for a

reliable sample, and we measure their distance with the Jensen–Shannon divergence (Lin 1991), which is symmetric and bounded. The divergence becomes a reliability weight

$$\rho_i = \exp\big(-\mathrm{JS}(\mathbf{p}_{\mathrm{mel}} \,\|\, \mathbf{p}_{\mathrm{coch}})\big), \tag{13}$$

and the confidence is $c_i = \rho_i \max(\mathbf{p}_{\mathrm{avg}})$, where $\mathbf{p}_{\mathrm{avg}}$ is the averaged prediction. A threshold $\theta(t)$ rises linearly from $\theta_{\min} = 0.80$ to $\theta_{\max} = 0.90$ over the training epochs. A sample with $c_i \geq \theta(t)$ receives the pseudo-label $\tilde{y}_i = \arg\max(\mathbf{p}_{\mathrm{avg}})$ and joins the accepted set $\mathcal{A}_t$. The rising threshold holds the model to a stricter standard as training proceeds.

4.8 Module 4: Style-Anchored Memory Learning

SAML maintains a queue $\mathcal{Q} = \{\mathbf{k}_j\}_{j=1}^{K}$ of $K = 512$ L2-normalised representations. It is updated by momentum, $\mathbf{k} \leftarrow \eta\mathbf{k} + (1-\eta)\mathbf{q}$ with $\eta = 0.95$, in the spirit of momentum-contrast memories (He et al. 2020). For the query $\mathbf{q} = \mathbf{z}_{\mathrm{bi}}/\|\mathbf{z}_{\mathrm{bi}}\|_2$, a supervised InfoNCE loss with temperature τ_m, in the style of Khosla et al. (2020) and van den Oord et al. (2018), raises the similarity between $\mathbf{q}$ and the entries $\mathbf{k}_+$ that carry the same emotion label:

$$\mathcal{L}_{\mathrm{cont}} = -\log \frac{\exp(\mathbf{q}\cdot\mathbf{k}_+/\tau_m)}{\sum_{j=1}^{K}\exp(\mathbf{q}\cdot\mathbf{k}_j/\tau_m)}. \tag{14}$$

Because the entries change slowly, they act as stable emotion prototypes.

4.9 Training objective

The training loss combines four weighted components. The first, $\mathcal{L}_{\mathrm{cls}}$, bundles supervised quadrant classification and valence–arousal squared error. The others are the pseudo-label cross-entropy $\mathcal{L}_{\mathrm{pl}}$ on the accepted set, a cross-view consistency term $\mathcal{L}_{\mathrm{cons}}$ that keeps the two acoustic views aligned, and the contrastive memory term $\mathcal{L}_{\mathrm{cont}}$ of Eq. (14):

$$\mathcal{L}_{\mathrm{total}} = 1.0\,\mathcal{L}_{\mathrm{cls}} + 0.8\,\mathcal{L}_{\mathrm{pl}} + 0.2\,\mathcal{L}_{\mathrm{cons}} + 0.1\,\mathcal{L}_{\mathrm{cont}}, \tag{15}$$

$$\mathcal{L}_{\mathrm{cls}} = \mathrm{CE}(\hat{y}, y) + 1.0\,\|\hat{\mathbf{r}} - \mathbf{r}\|_2^2. \tag{16}$$

Algorithm 1 Training procedure of the audio–lyrics model

Require: training set $\mathcal{D}_{\mathrm{train}}$, validation set $\mathcal{D}_{\mathrm{val}}$, epochs E, thresholds $\theta_{\min}, \theta_{\max}$, queue size K, momentum η

Initialise DSAF, lyric projection, fusion layers, heads and an empty queue $\mathcal{Q}$
for $t = 1$ to E **do**
  $\theta(t) \leftarrow$ linear schedule from $\theta_{\min}$ to $\theta_{\max}$
  for each mini-batch (audio, lyrics, y, $\mathbf{r}$) **do**
    $\mathbf{M}, \mathbf{C} \leftarrow$ log-mel and cochleagram of 5 segments
    $\mathbf{Z}_a, \mathbf{z}_{\mathrm{mel}}, \mathbf{z}_{\mathrm{coch}}, \mathbf{z}_{\mathrm{fuse}} \leftarrow \mathrm{DSAF}(\mathbf{M}, \mathbf{C})$
    $\mathbf{L} \leftarrow f_{\mathrm{proj}}(\mathrm{pool}(\mathrm{RoBERTa}(\mathrm{lyrics})))$ ▷ frozen
    $\mathbf{z}_{\mathrm{bi}} \leftarrow$ fusion of $\mathbf{Z}_a, \mathbf{L}, \mathbf{z}_{\mathrm{fuse}}$ ▷ Eqs. (10)–(12)
    predict logits $\mathbf{o}$ and coordinates $\hat{\mathbf{r}}$ from $\mathbf{z}_{\mathrm{bi}}$
    $\rho_i \leftarrow \exp(-\mathrm{JS}(\mathbf{p}_{\mathrm{mel}}\|\mathbf{p}_{\mathrm{coch}}))$; $c_i \leftarrow \rho_i \max(\mathbf{p}_{\mathrm{avg}})$
    pseudo-label samples with $c_i \geq \theta(t)$
    $\mathcal{L}_{\mathrm{total}} \leftarrow$ Eq. (15)
    backpropagate; update parameters; update $\mathcal{Q}$ by momentum
  end for
  compute validation Macro-F1; keep the checkpoint if it improves
end for
return checkpoint with best validation Macro-F1

4.10 Training algorithm

Algorithm 1 shows how the components interact during training.

5 Experimental Setup

5.1 Implementation details

Our code is written in PyTorch (Paszke et al. 2019), with torchaudio for signal processing and the Transformers library from HuggingFace (Wolf et al. 2020) for the language model. Table 1 lists the main settings. Tokens have dimension $d = 128$, attention uses 4 heads and dropout is 0.2. Optimisation relies on Adam (Kingma and Ba 2015), starting from a learning rate of 10^{-4} with an L2 penalty of 10^{-5}, mini-batches of 32 songs and gradient-norm clipping at 5.0. Training lasts 25

Table 1 Hyperparameter configuration.

Parameter	Value
Acoustic front end	
Sampling rate	22,050 Hz
Segments per track	5×3 s
STFT window / hop	2,048 / 512
Mel / cochleagram bands	128 / 84
Dynamic range	80 dB
Lyric encoder	
Backbone (frozen)	RoBERTa-base
Window length / max windows	126 / 12
Fusion and heads	
Token dimension d	128
Attention heads	4
Dropout	0.2
PCL and SAML	
Threshold $\theta(t)$	$0.80 \rightarrow 0.90$
Queue size K	512
Momentum η	0.95
Loss weights (cls, pl, cons, cont)	1.0, 0.8, 0.2, 0.1
Training	
Optimiser	Adam
Initial learning rate	10^{-4}
L2 penalty	10^{-5}
Songs per batch	32
Training epochs	25
Gradient clipping	5.0

epochs on a workstation with an NVIDIA GPU. With 1,400 training songs and a batch size of 32, one epoch is 44 parameter updates, so the full schedule is about 1,100 updates. Early stopping monitors validation Macro-F1, which peaked at epoch 11. Every experiment was repeated with three random seeds (42, 43 and 44), and we also evaluated a soft-voting ensemble of the three models.

5.2 Evaluation metrics

Following the MERGE protocol, quadrant classification is judged by accuracy and by Macro-F1, which is the plain average over quadrants of $F_1 = 2PR/(P + R)$, with P and R the precision and recall of that quadrant. To judge how well the predicted scores rank the classes we also report one-vs-rest AUC-ROC (area under the ROC curve) and average precision (AP). The continuous valence–arousal outputs are evaluated with RMSE, the determination coefficient R^2 and the concordance correlation coefficient (CCC) of Lin (1989),

$$\mathrm{CCC} = \frac{2\rho\,\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2}, \tag{17}$$

which penalises both weak correlation and differences in scale and offset between predictions and targets. The symbol ρ denotes the Pearson correlation, while μ and σ stand for the mean and standard deviation of the targets and of the predictions.

6 Results and Discussion

6.1 Training behaviour

Figure 2 traces the best run. The training loss falls from 2.92 to 1.70, while the validation loss levels off near 2.05 and does not rise. Training accuracy reaches 71.2%. Validation Macro-F1 peaks at 0.673 at epoch 11, the checkpoint retained by early stopping. Over the 25 epochs the loss terms behave as follows: the supervised and pseudo-label losses decrease steadily, the consistency loss keeps the two views aligned, and the SAML InfoNCE loss stays close to 6.25. The last value deserves a comment. With $K = 512$ entries, a loss that cannot tell the positive from the other entries would equal $\ln 512 \approx 6.24$, so the flat curve shows that SAML acts mainly as a regulariser that anchors the embeddings rather than as a term that is driven to small values.

6.2 Classification performance

Table 2 gives test results on the 300 held-out songs. The best seed reaches 73.33% accuracy and 73.50% Macro-F1. Averaged over the three seeds, Macro-F1 is 72.97%, and the soft-voting ensemble scores 72.90%.

The published MERGE baselines (Louro et al. 2026) are listed for orientation only. They were obtained on the complete MERGE split rather than on our balanced subset, so the numbers cannot be compared directly, and in particular they should not be read as showing that our model is better or worse than the published bimodal baseline of 81.74% Macro-F1. A comparison on the same split is available only through our own ablation (Section 6.6): the audio-only configuration reaches 70.84% Macro-F1 and the full model is

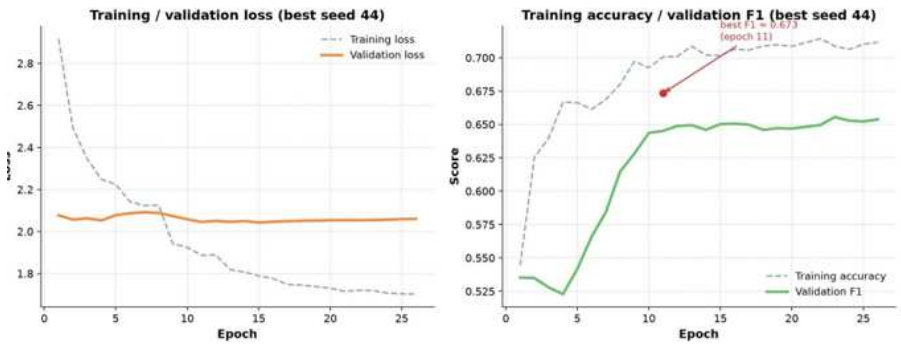


Fig. 2 Training dynamics of the best run (seed 44). Left: training and validation loss. Right: training accuracy and validation Macro-F1, which peaks at 0.673 at epoch 11.

Table 2 Quadrant classification on the MERGE test split.

Model	Acc. (%)	Macro-F1 (%)
MERGE complete split, published (Louro et al. 2026)		
Audio-only (mel)	71.30	70.84
Lyrics-only (BERT)	75.40	75.05
Bimodal concatenation	82.10	81.74
MERGE balanced split, this work		
Proposed (best seed)	73.33	73.50
Proposed (3-seed mean)	–	72.97
Proposed (3-seed ensemble)	–	72.90

Table 3 Per-quadrant metrics on the test split (best seed).

Quadrant	Prec.	Rec.	F1	AUC	AP
Q1 Exuberance	0.71	0.75	0.73	0.914	0.750
Q2 Distress/Anger	0.86	0.88	0.87	0.985	0.960
Q3 Depression	0.73	0.71	0.72	0.917	0.758
Q4 Serenity	0.63	0.60	0.62	0.845	0.620
Average	0.733	0.735	0.735	0.915	0.772

2.66 points higher, which supports adding lyrics through cross-attention.

Statistical uncertainty.

A test set of 300 songs limits how precisely any score can be stated. One misclassified song changes accuracy by 0.33 percentage points. For the best seed, 220 of 300 songs are correct, and a 95% Wilson confidence interval for the accuracy is roughly 68.1% to 78.0%. Differences of one or two points between configurations, including most of those in Table 6, are therefore within the range that sampling alone can produce, and they should be read as indicative rather than conclusive.

6.3 Per-quadrant analysis

Table 3 breaks the results down by quadrant. Q2 (distress and anger) is recognised best, with an F1 of 0.87 (precision 0.86, recall 0.88). In this quadrant, sharp percussive attacks and high RMS energy tend to co-occur with hostile lyrics, so both modalities point the same way. Q1 reaches an F1 of 0.73 and Q3 an F1 of 0.72. Q4 (serenity) is the weakest class, with an F1 of 0.62 and a recall of 0.60.

The confusion matrix (Fig. 3) shows that 19 of the 75 calm songs (25%) are labelled sad, and 15 of the 75 sad songs (20%) are labelled calm. Q2 has the highest true-positive rate, at 88%. Calm and sad music share slow tempo, restricted dynamics and low pitch variance, so the acoustic evidence is nearly the same. When a calm song uses introspective or sombre wording, the lyric stream adds to the confusion instead of resolving it, and the model sometimes settles on negative valence. Averaged over the four quadrants the precision is 0.733, the recall 0.735, the F1 0.735, the AUC-ROC 0.915 and the average precision 0.772. The one-vs-rest curves in Fig. 3 give AUCs of 0.985 for Q2, 0.917 for Q3, 0.914 for Q1 and 0.845 for Q4, and the AP for Q2 is 0.960.

6.4 Error decomposition along the affective axes

Because the quadrants are defined by the signs of valence and arousal, every misclassification can be assigned to one of three types: the valence sign is wrong, the arousal sign is wrong, or both are

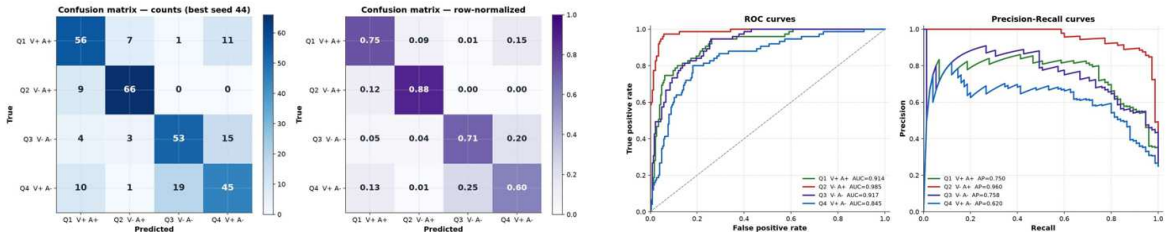


Fig. 3 Left: confusion matrix on the 300-song test split (75 songs per quadrant), as counts (220 correct, 73.33% accuracy) and row-normalised, showing the highest true-positive rate for Q2 (88%) and the Q3/Q4 confusion. Right: one-vs-rest ROC curves (mean AUC 0.915) and precision–recall curves (mean AP 0.772).

Table 4 Decomposition of the 80 test errors by the affective axis that is wrong (derived from Fig. 3).

Error type	Songs	Share of errors
Valence sign only	50	62.5%
Arousal sign only	24	30.0%
Both signs	6	7.5%
Total errors	80	100%

Table 5 Valence–arousal regression (best seed).

Target	RMSE	R^2	CCC
Valence	0.298	0.544	0.762
Arousal	0.189	0.382	0.676

wrong. Reading the counts of Fig. 3 in this way gives Table 4. Of the 80 errors, 50 (62.5%) change only the valence sign, 24 (30.0%) change only the arousal sign, and 6 (7.5%) change both. Treated as two binary decisions, valence has a sign accuracy of 81.3% (244 of 300 songs) and arousal one of 90.0% (270 of 300).

The pattern agrees with the motivation of the work. Arousal is the easier axis for a model that hears the music, and valence is where mistakes concentrate: the two largest single confusions, Q4 to Q3 (19 songs) and Q3 to Q4 (15 songs), both flip valence while keeping arousal low. Lyrics are the natural source of evidence for resolving these cases, and the cases in which they fail to do so are those in which the wording is sombre despite a calm, positive overall impression.

6.5 Valence–arousal regression

Table 5 reports the continuous results. The model reaches a valence CCC of 0.762 (R^2 = 0.544, RMSE = 0.298), which is consistent with lexical sentiment reaching the regression head through cross-attention. Arousal reaches a CCC of 0.676 (R^2 = 0.382, RMSE = 0.189), in line with the close link between acoustic energy and arousal. Figure 4 plots the predictions against the true coordinates. Note that CCC depends on the spread of the targets as well as on the size of the errors, so a higher CCC for valence than for arousal does not by itself mean that valence is estimated more accurately in absolute terms: its RMSE is in fact larger.

6.6 Ablation study

We removed each component in turn while keeping all other hyperparameters unchanged (Table 6).

- **Without lyrics.** Dropping the lyric stream leaves only the dual-view acoustic DSAF branch. Macro-F1 falls by 2.66 points to 70.84%, so the lyrics are needed to resolve valence.
- **Without cross-attention.** Replacing cross-attention with simple concatenation costs 2.30 points. The gain therefore comes from attention-based querying and not merely from the presence of lyrics.
- **Without PCL or SAML.** Removing PCL costs 1.85 points and removing SAML 1.40 points, so both modules help to stabilise training.

Given the sampling uncertainty described above, the lyric and cross-attention effects are the more convincing; the PCL and SAML contributions are smaller.

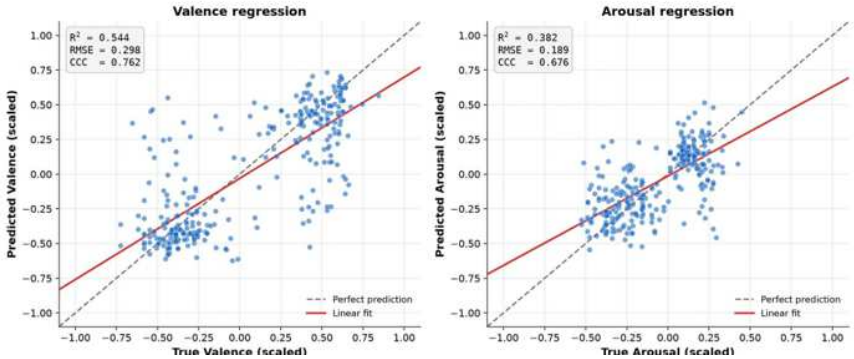


Fig. 4 Predicted versus true valence (left) and arousal (right) on the test split. The dashed line marks ideal prediction and the solid line the fitted linear trend.

Table 6 Ablation study (Macro-F1 on the test split).

Configuration	Macro-F1 (%)	Δ
Full model	73.50	0.00
Without SAML queue	72.10	−1.40
Without PCL curriculum	71.65	−1.85
Without lyrics (audio only)	70.84	−2.66
Concatenation, no cross-attention	71.20	−2.30

6.7 Error and confidence analysis

Eighty of the 300 songs land in the wrong quadrant, which agrees with the 73.33% accuracy. The model is more confident when it is right. The mean maximum softmax probability is 0.826 for correct predictions and 0.682 for misclassified ones, so the confidence score is informative about reliability. Judged by true quadrant, the mean confidence is 0.84 for both Q1 and Q2, 0.74 for Q3 and 0.73 for Q4, which mirrors the per-quadrant F1 ordering.

At the song level, the ten most accurate predictions have a Euclidean valence–arousal error below 0.04, and the ten worst exceed 0.80. These songs are marked by artistic irony and by a clash between acoustic and lexical cues.

7 Deployment

To check that the model works outside the benchmark, we integrated it into a mood-aware music exploration web application. A FastAPI and PyTorch backend runs an asynchronous pipeline that accepts uploaded audio (.mp3, .wav) together with lyric text. Worker processes compute the 128-band mel spectrogram, the 84-band cochleagram and the RoBERTa token embeddings in under 180 ms. The interface shows spectrogram heatmaps, reports the predicted quadrant with calibrated confidence (Guo et al. 2017), and places the song on an interactive circumplex. Recommendation retrieves tracks whose predicted valence–arousal coordinates lie nearest, in Euclidean distance, to the mood the user selects.

8 Limitations

The system has several known limitations. The text encoder is an English-only RoBERTa model, so global catalogues would need a multilingual alternative such as XLM-RoBERTa. Predictions are static, so emotional changes within a track are not followed, and for purely instrumental tracks the model falls back on the acoustic DSAF branch and loses cross-modal modulation. Finally, our results come from a balanced 2,000-song subset and are not directly comparable with results on the full published MERGE split; training the published baselines on our split is the natural way to remove this ambiguity. The test set is small (300 songs), so the ablation differences are comparable to the sampling uncertainty.

9 Conclusion and Future Work

We presented a dual-view cross-attention architecture for music emotion recognition. It joins log-mel spectrograms and gammatone cochleagrams with contextual RoBERTa lyric embeddings through bidirectional cross-attention, and it uses PCL and

SAML to stabilise multi-task training. On the balanced MERGE split the best seed reaches 73.33% accuracy and 73.50% Macro-F1, with a valence CCC of 0.762 and an arousal CCC of 0.676. The analysis shows that most errors concern the valence axis, and that the confusion between calm and sad music is the main remaining weakness. Our next steps are to track how emotion evolves within a song, to test multilingual cross-modal transfer on the DEAM dataset (Aljanaki et al. 2017), to repeat the ablations over several seeds with baselines trained on the same split, and to apply low-bit quantisation for mobile deployment.

Acknowledgements. The authors are grateful to the Department of Artificial Intelligence and Data Science at Mepco Schlenk Engineering College, Sivakasi, which made its GPU resources available for this study, and to the MERGE team for releasing their dataset publicly.

Declarations

- **Funding:** No external funding supported the work reported here.
- **Competing interests:** None of the authors holds a financial or personal interest that could be seen as influencing this article.
- **Data availability:** MERGE is openly distributed by its creators (Louro et al. 2026); the identifiers of the songs in our balanced split can be obtained from the corresponding author on request.
- **Code availability:** The implementation can be requested from the corresponding author.
- **Author contributions:** Model implementation, experiments and the first draft were carried out by M.B.C.S. and Y.D. D.M. supervised the project and revised the text. All authors approved the submitted version.

References

Aljanaki A, Yang YH, Soleymani M (2017) Developing a benchmark for emotional analysis of music. PLOS ONE 12(3):e0173392

Baltrušaitis T, Ahuja C, Morency LP (2019) Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(2):423–443

Devlin J, Chang MW, Lee K, et al (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp 4171–4186

Eerola T, Vuoskoski JK (2011) A comparison of the discrete and dimensional models of emotion in music. Psychology of Music 39(1):18–49

Glasberg BR, Moore BCJ (1990) Derivation of auditory filter shapes from notched-noise data. Hearing Research 47(1–2):103–138

Guo C, Pleiss G, Sun Y, et al (2017) On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, pp 1321–1330

He K, Fan H, Wu Y, et al (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9729–9738

Hu X, Downie JS (2010) When lyrics outperform audio for music mood classification: A feature analysis. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pp 619–624

Juslin PN, Laukka P (2004) Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. Journal of New Music Research 33(3):217–238

Khosla P, Teterwak P, Wang C, et al (2020) Supervised contrastive learning. In: Advances in Neural Information Processing Systems, pp 18661–18673

Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (ICLR)

Lee DH (2013) Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICML Workshop on Challenges in Representation Learning

Li Q, Chen CLP, Zhang T (2026a) Memo2496: Expert-annotated dataset and dual-view adaptive framework for music emotion recognition. IEEE Transactions on Affective Computing https://doi.org/10.1109/TAFFC.2026.3715195

Li W, Deng Y, Liao S, et al (2026b) Music emotion recognition using tri-modal fusion of lyrics, vocals, and accompaniment with cross-attention. Multimedia Tools and Applications 85:559. https://doi.org/10.1007/s11042-026-21719-3

Lin J (1991) Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory 37(1):145–151

Lin LIK (1989) A concordance correlation coefficient to evaluate reproducibility. Biometrics 45(1):255–268

Liu Y, Ott M, Goyal N, et al (2019) RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:190711692

Louro PL, Redinho H, Ribeiro TFR, et al (2026) MERGE: A bimodal audio-lyrics dataset for static music emotion recognition. IEEE Transactions on Affective Computing 17(2):2058–2075. https://doi.org/10.1109/TAFFC.2026.3668037

Malheiro R, Panda R, Gomes P, et al (2018) Emotionally-relevant features for classification and regression of music lyrics. IEEE Transactions on Affective Computing 9(2):240–254

van den Oord A, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv preprint arXiv:180703748

Panda R, Malheiro R, Paiva RP (2023) Audio features for music emotion recognition: A survey. IEEE Transactions on Affective Computing 14(1):68–88. https://doi.org/10.1109/TAFFC.2020.3032373

Paszke A, et al (2019) PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems

Russell JA (1980) A circumplex model of affect. Journal of Personality and Social Psychology 39(6):1161–1178

Sohn K, Berthelot D, Carlini N, et al (2020) FixMatch: Simplifying semi-supervised learning with consistency and confidence. In: Advances in Neural Information Processing Systems, pp 596–608

Tsai YHH, Bai S, Liang PP, et al (2019) Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 6558–6569

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Advances in Neural Information Processing Systems, pp 5998–6008

Wolf T, et al (2020) Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp 38–45

Xiao Y, Ruan H, Zhao X, et al (2025) An efficient bi-modal fusion framework for music emotion recognition. IEEE Transactions on Affective Computing 16(2):999–1015. https://doi.org/10.1109/TAFFC.2024.3486340

Yang YH, Chen HH (2012) Machine recognition of music emotion: A review. ACM Transactions on Intelligent Systems and Technology 3(3):40

Zadeh A, Chen M, Poria S, et al (2017) Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp 1103–1114

Zhao J, Ru G, Yu Y, et al (2022) Multimodal music emotion recognition with hierarchical cross-modal attention network. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), pp 1–6, https://doi.org/10.1109/ICME52920.2022.9859812