

EthnoPulse: A Multi-Task Deep Learning Framework for Traditional Music Emotion Recognition and Online Engagement Prediction

Cheng Zhang

2024123930@jou.edu.cn

Nizhny Novgorod State Conservatory named after M.I. Glinka

Research Article

Keywords: Traditional music, Music emotion recognition, Multi-task learning, Deep learning, Online engagement, Cross-cultural transfer

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10908620/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

EthnoPulse: A Multi-Task Deep Learning Framework for Traditional Music Emotion Recognition and Online Engagement Prediction

Cheng Zhang^{1*}

^{1*}M. I. Glinka Novosibirsk State Conservatory, Sovetskaya Ulitsa 31,
Novosibirsk, 630000, Novosibirsk Oblast, Russia.

Corresponding author(s). E-mail(s): 2024123930@jou.edu.cn;

Abstract

Traditional-music emotion recognition and online engagement are usually studied separately, although both depend partly on acoustic structure. EthnoPulse formulates the two problems as a partial-label multi-task setting in which disjoint affect and engagement corpora can share an encoder without imputing missing labels. The model combines dual-resolution log-Mel inputs, a multi-scale residual encoder, squeeze-and-excitation recalibration, task-specific gates, uncertainty weighting, an emotion–valence/arousal consistency term, and CORAL alignment. CTIM and FMA define the raw-audio tasks, with composition- and artist-disjoint splits used to reduce identity leakage. CCMED and WCMED provide an independent feature-level test of the evaluation protocol. Under composition-grouped five-fold validation, the lowest arousal MAEs are 0.286 on CCMED and 0.282 on WCMED, and the strongest rank correlations are 0.773 and 0.806, respectively. Random excerpt splitting raises CCMED valence correlation from 0.318 to 0.527. Cross-domain arousal transfer reaches $\rho = \mathbf{0.808}$, whereas valence transfer is weaker and the effect of CORAL depends on both direction and target. The results underline the importance of group-disjoint evaluation and dimension-specific reporting in cross-cultural music analysis.

Keywords: Traditional music, Music emotion recognition, Multi-task learning, Deep learning, Online engagement, Cross-cultural transfer

1 Introduction

Music emotion recognition (MER) is a central task in music information retrieval because affective descriptors support retrieval, recommendation, catalog organization[1], and human-centered music interfaces[2]. Recent surveys describe a rapid shift toward deep, attention-based, and pretrained music-audio models.[3] Recent IJCIS work likewise illustrates the continued use of deep architectures for music-emotion classification.[4] However, the empirical basis of MER remains culturally uneven. Widely used benchmarks are dominated by Western popular or art music, whereas traditional instrumental repertoires express affect through distinctive combinations of timbre, ornamentation, articulation, register, tempo, density, and culturally learned performance conventions. A model that performs well on a generic music corpus therefore cannot be assumed to preserve the same error structure when applied to Chinese traditional music.

The availability of culturally focused datasets now makes this problem more tractable. Wu *et al.*[5] introduced the Chinese Traditional Instrumental Music (CTIM) dataset, which contains 273 ten-second naturalistic excerpts annotated for valence, arousal, familiarity, and seven discrete emotions. Xu *et al.*[6] subsequently showed that timbral descriptors explain substantial variation in CTIM affect ratings and that several feature–emotion relationships differ from patterns commonly reported for Western music. These results indicate that traditional-music MER is not simply a smaller version of generic MER: both the acoustic realization of affect and the evaluation protocol require explicit cultural consideration.

A second gap is that MER is typically studied as an isolated content-analysis task. On digital platforms, music is also listened to, revisited, saved, discussed, and circulated, producing observable engagement signals. Engagement is not equivalent to emotion, musical quality, or causal dissemination. It is influenced by exposure, artist history, recommendation systems, release timing, network effects, and other non-acoustic factors.[7] Nevertheless, acoustic content remains relevant to cold-start modeling because a new or weakly observed track may have little behavioral history. FMA is useful in this setting because it provides open audio together with track meta-data and play-related counters at substantially larger scale than culturally focused emotion corpora.[8]

A practical complication is that the two forms of supervision do not coexist in the same recordings. CTIM provides culturally specific affect annotations but no directly comparable platform-engagement target, whereas FMA provides engagement-related counters but no CTIM-style seven-class emotion labels. Treating the problem as ordinary multi-task learning would therefore require either discarding useful data or fabricating missing targets. We instead formulate the setting as *partial-label multi-task learning*: source-specific batches update a shared acoustic encoder, while only the losses associated with genuinely observed labels are activated. The objective is not to assert that emotion causes engagement, but to test whether affect recognition and content-based engagement prediction can benefit from partially shared representations of timbre, rhythm, dynamics, spectral balance, and temporal structure.

EthnoPulse is a compact multi-task architecture built around dual-resolution log-Mel inputs, a parallel multi-scale convolutional stem, residual acoustic encoding,

squeeze-and-excitation (SE) recalibration, and task-specific feature gates.[9] Three output branches predict seven-class emotion, continuous valence–arousal, and standardized engagement. The heterogeneous task losses are balanced with homoscedastic uncertainty weighting.[10] Recent deep multi-task learning reviews identify adaptive loss balancing as an important mechanism when task scales differ.[11] An emotion–VA consistency term connects discrete and dimensional affect. CORAL is used as a lightweight second-order alignment term between the affect and engagement sources.[12] Recent domain-generalization work also cautions that alignment should be evaluated under distribution shift rather than assumed to help every target.[13] The architecture is kept compact so that the contribution of individual components can be examined directly.

The evaluation protocol is defined at the same time as the model because excerpts from the same composition, performance, or artist are not statistically independent. Randomly distributing them across training and test sets can reward memorization of recording, performer, instrumentation, or production signatures. Accordingly, CTIM is evaluated with composition/performance grouping and FMA with artist grouping. The importance of this decision is tested independently with the matched Chinese Classical Music Emotion Dataset (CCMED) and Western Classical Music Emotion Dataset (WCMED), which provide 400 excerpts per domain in a common 467-dimensional acoustic feature space.[14] These resources also enable controlled Chinese–Western transfer experiments and a direct assessment of whether CORAL alignment is consistently helpful.

The study makes four contributions:

- We formulate traditional-music emotion recognition and content-based online-engagement prediction as a partial-label multi-task problem, allowing disjoint corpora to train a shared encoder without imputing unavailable labels.
- We propose EthnoPulse, which integrates dual-resolution log-Mel inputs, multi-scale residual convolution, SE recalibration, task-specific gates, uncertainty weighting, and emotion–VA consistency within a compact end-to-end architecture.
- We introduce a leakage-aware evaluation protocol that separates CTIM data by composition or performance and FMA data by artist. Using matched external datasets, we further quantify the performance inflation caused by random excerpt splitting.
- We evaluate cross-cultural generalization with CCMED and WCMED, showing a consistent asymmetry between arousal and valence and demonstrating that covariance alignment is target- and direction-dependent rather than uniformly beneficial.

2 Related Work

2.1 Music Emotion Recognition and Acoustic Representation Learning

MER has evolved from pipelines based on handcrafted descriptors toward end-to-end and pretrained representation learning. A recent review summarizes the rapid expansion of deep MER architectures while noting persistent problems in label subjectivity,

dataset scale, and evaluation design.[1] Comparative benchmarking reaches a similar conclusion across modern deep-learning pipelines.[2] Broader surveys of AI in music also emphasize the growing role of transferable audio representations.[3] Recent MER studies continue to explore multimodal and deep architectures under different label spaces and evaluation settings.[15] A recent IJCIS study provides another example of this trend.[4] CNNs remain useful because spectrogram-like inputs preserve local time–frequency structure. Residual learning provides a standard mechanism for training deeper convolutional feature extractors.[16] Convolutional-recurrent models extend this representation with longer temporal dependencies.[17] More recent pretrained music models show that learned audio representations can transfer across multiple music-information-retrieval tasks.[18] MuQ further develops this direction through self-supervised music representation learning.[19]

Recent systems increasingly combine local spectral analysis with longer-range temporal modeling. Time- and frequency-masking augmentation, popularized by SpecAugment, is also widely used to regularize time–frequency models.[?]] A 2025 instrumental-music study, for example, integrates convolutional, recurrent, and attention components for emotion recognition.[?]] Attention-based MER models also continue to combine local spectral structure with broader contextual information.[20] Transformer-based mood-recognition work provides a complementary recent example.[21] Li[22] explored deep convolutional MER, while Su *et al.*[23] used multi-resolution spectrograms to recover complementary affective cues that a single analysis window may miss. At a broader representation level, MERT demonstrates large-scale self-supervised music representation learning.[18] MuQ extends this line with Mel residual vector quantization.[19] Multi-level self-supervised audio representations have also been evaluated on music segmentation and related MIR tasks.[24] These developments motivate the dual-resolution front end of EthnoPulse while providing a contemporary context for transferable acoustic representations. In contrast to architectures aimed primarily at maximizing single-dataset MER accuracy, our focus is on a compact representation that can be shared across affect and engagement tasks under incomplete supervision.

Attention also provides a mechanism for selective feature emphasis without requiring a full self-attention layer over every time–frequency location. SE blocks summarize channel responses globally and learn channel-wise recalibration weights.[9] Recent explainable and theory-guided work further highlights the value of architectures whose acoustic contributions can be inspected rather than hidden inside an unconstrained predictor.[25] This consideration is relevant to traditional instrumental music, where learned channels may respond differently to spectral envelopes, transient density, harmonic structure, and instrument-dependent timbral cues. EthnoPulse therefore uses SE recalibration at higher residual stages rather than introducing a much larger attention backbone.

2.2 Traditional and Cross-Cultural Music Emotion

Cultural specificity remains an important limitation of current MER benchmarks. A shared valence–arousal coordinate system is useful for comparison,[26] but it does not imply that identical acoustic cues map to affect in the same way across repertoires. A

recent experiment comparing Chinese pentatonic modes with Western major/minor modes found culture- and training-related differences in emotion induction.[?] A 2025 standardized emotional-music database that combines Chinese and Western instrumental excerpts further underscores the need for culture-aware affective validation.[?] Recent listening evidence also suggests measurable in-group advantages in emotion recognition.[27] Cross-cultural MER should therefore distinguish between transfer of an affective ordering and complete invariance of the underlying feature distribution.

The matched CCMED and WCMED resources are particularly useful for this distinction because they provide common acoustic descriptors and valence/arousal ratings for Chinese and Western classical-music excerpts.[14] This reduces confounding introduced when unrelated datasets are compared only through published benchmark scores. CTIM complements these resources by focusing specifically on Chinese traditional instrumental music. Wu *et al.*[5] provide both discrete and dimensional annotations, while Xu *et al.*[6] report strong timbral associations with affect, especially arousal. These datasets motivate an evaluation strategy that treats traditional music as a culturally structured domain and reports valence and arousal separately rather than collapsing them into a single aggregate score.

2.3 Multi-Task Learning, Partial Labels, and Domain Alignment

Multi-task learning (MTL) was originally introduced as a way to improve generalization by sharing information across related objectives.[28] Recent reviews show that modern deep MTL still depends strongly on task relatedness and loss balancing.[11] In MER, recent personalized models combine shared representations with task- or listener-adaptive mechanisms rather than assuming that one representation is equally suitable for every target.[29] Negative transfer can arise when tasks differ in label semantics, data scale, or gradient magnitude. This issue is pronounced in the present setting because CTIM and FMA differ simultaneously in repertoire, recording conditions, dataset size, and supervision type.

EthnoPulse addresses this heterogeneity in three ways. First, partial-label optimization activates only the losses supported by each source, avoiding fabricated supervision. Second, task-specific gates allow the shared embedding to be filtered differently for emotion, dimensional affect, and engagement. Third, homoscedastic uncertainty weighting adapts the relative scale of classification and regression objectives rather than fixing arbitrary coefficients.[10] These mechanisms provide a middle ground between hard sharing, which can promote negative transfer, and completely independent models, which cannot exploit common acoustic structure.

The source mismatch also motivates explicit but conservative domain regularization. Deep CORAL aligns second-order feature statistics without adversarial training [12], while recent domain-generalization surveys emphasize that distribution shift remains a central obstacle to reliable out-of-domain performance. [13] However, domain alignment is not guaranteed to improve every task: if domain-specific structure is itself predictive, aggressive invariance can remove useful information. We therefore use CORAL as an ablation-sensitive regularizer on the shared embedding rather than as a requirement that CTIM and FMA become distributionally indistinguishable. The

matched CCMED/WCMED transfer experiments provide an independent test of this assumption.

2.4 Online Music Engagement Prediction and Research Gap

Music engagement and popularity prediction differ from MER because the target is generated by both content and a social platform. Recommendation exposure, artist history, peer influence, marketing, and network position interact with acoustic properties, so an audio-only model cannot be interpreted as a complete causal account of dissemination. Contemporary recommendation research therefore combines interaction structure with content features.[30] Recent song-popularity work also shows that homophily and social influence materially affect prediction.[7] These findings support treating engagement as a contextual platform signal rather than as a direct acoustic property of a track.

For the present study, FMA provides a reproducible compromise: it offers open audio, standard MIR metadata, and track-level play-related counters within a public research framework.[8] We use engagement as a content-based prediction target rather than as a proxy for intrinsic musical value. This distinction is important for cold-start analysis, where acoustic evidence can be available before rich interaction histories accumulate.

Prior work leaves three connected gaps. First, culturally focused traditional-music MER remains underrepresented relative to Western/popular benchmarks. Second, affect recognition and online engagement are usually modeled independently even though they can share acoustic structure, while available labels often reside in different corpora. Third, random excerpt splitting and unconditional domain alignment can lead to overly optimistic or overly simplified claims about generalization. EthnoPulse addresses these gaps through partial-label multi-task learning, leakage-aware group splitting, and target-sensitive cross-cultural validation. Its objective is deliberately limited to learning transferable acoustic structure for affect and content-based engagement; it does not assume that emotion alone determines online dissemination.

3 Methodology

3.1 Problem Formulation and Data Preparation

EthnoPulse addresses a partial-label multi-task setting in which traditional-music affect labels and online-engagement labels are observed in different corpora. This separation is fundamental to the proposed formulation. Rather than imputing missing targets or constructing artificial one-to-one correspondences between datasets, the model shares an acoustic encoder across sources and activates only the task losses supported by the labels available in the current mini-batch. Consequently, the framework can learn a common acoustic representation while preserving the distinct supervision structure of emotion recognition, dimensional affect estimation, and engagement prediction.

Let

$$\mathcal{D}_E = \{(x_i, y_i^E, \mathbf{y}_i^{VA})\}_{i=1}^{N_E} \quad (1)$$

denote the affect corpus and

$$\mathcal{D}_G = \{(x_j, y_j^G)\}_{j=1}^{N_G} \quad (2)$$

denote the engagement corpus. Here, x is an audio excerpt, $y^E \in \{1, \dots, C\}$ is the discrete emotion label with $C = 7$ for CTIM, $\mathbf{y}^{VA} = [y^V, y^A]^\top \in [-1, 1]^2$ contains the continuous valence and arousal targets when available, and $y^G \in \mathbb{R}$ is a standardized engagement target. The two corpora are disjoint; therefore, no assumption is made that the three target types coexist for the same recording.

The primary raw-audio experiments use CTIM as the culturally focused affect source and FMA as the engagement source. CTIM contains 273 naturalistic ten-second excerpts of Chinese traditional instrumental music rated by 168 participants for valence, arousal, familiarity, and seven discrete emotion dimensions. The released discrete labels are retained directly rather than reconstructed from valence–arousal quadrants so that culturally meaningful categories such as solemnness and transcendence are preserved. When dimensional ratings are expressed on a 1–7 scale, they are mapped to $[-1, 1]$ by

$$\tilde{y} = \frac{y - 4}{3}. \quad (3)$$

FMA contains 106,574 tracks from 16,341 artists and provides audio, genre meta-data, tags, play counters, and precomputed MIR features. The engagement experiment retains Folk, International, and closely related tracks for which both audio and the required interaction counters are available. Because platform counters are strongly right-skewed, each counter is transformed by $\log(1 + x)$ and standardized using statistics estimated only from the training partition. When both listens L and interest I are available, the primary engagement target is

$$y^G = \frac{1}{2} [Z_{\text{tr}}\{\log(1 + L)\} + Z_{\text{tr}}\{\log(1 + I)\}], \quad (4)$$

where $Z_{\text{tr}}(\cdot)$ denotes training-set standardization. This variable is interpreted as observed platform engagement rather than a causal or universal measure of musical popularity.

Two safeguards are imposed before model training. First, all transformations used to construct standardized targets are fitted on the training partition only. Second, data splitting is group-aware. CTIM excerpts originating from the same performance or composition are kept in the same partition, while FMA uses artist identity as the grouping variable. These restrictions reduce leakage from persistent performer, instrumentation, production, mastering, or local musical signatures. CCMED and WCMED are not used to update the raw-audio model; they are reserved for independent feature-level validation of grouped evaluation and cross-cultural transfer.

For an input waveform x , the complete model first constructs a two-channel time–frequency representation $\mathbf{S}(x)$ and then maps it through a shared encoder f_θ :

$$\mathbf{z} = f_\theta(\mathbf{S}(x)) \in \mathbb{R}^d, \quad d = 256. \quad (5)$$

Table 1 Datasets and their methodological roles in the study.

Dataset	Domain and analysis unit	Supervision used	Methodological role
CTIM	Chinese traditional instrumental music; 273 ten-second excerpts	Seven discrete emotions, valence, and arousal	Primary emotion and dimensional-affect source
FMA	Folk/International and related open music; track-level samples after filtering	Listens, interest, and available interaction metadata	Primary online-engagement source
CCMED	Chinese classical/traditional music; 400 excerpts from 62 compositions	Valence, arousal, and 467 acoustic features	External Chinese-domain validation
WCMED	Western classical music; 400 excerpts from 61 compositions	Valence, arousal, and 467 acoustic features	Cross-cultural external control

To reduce negative transfer among heterogeneous objectives, each branch applies a task-specific sigmoid gate,

$$\mathbf{z}_t = g_t(\mathbf{z}) = \mathbf{z} \odot \sigma(\mathbf{W}_t^g \mathbf{z} + \mathbf{b}_t^g), \quad t \in \{E, VA, G\}, \quad (6)$$

where $\odot$ denotes the Hadamard product. The corresponding outputs are

$$\hat{\mathbf{p}}^E = \text{softmax}(h_E(\mathbf{z}_E)), \quad (7)$$

$$\hat{\mathbf{y}}^{VA} = \tanh(h_{VA}(\mathbf{z}_{VA})), \quad (8)$$

$$\hat{y}^G = h_G(\mathbf{z}_G). \quad (9)$$

Thus, the network simultaneously supports seven-class emotion recognition, continuous valence–arousal estimation, and content-based engagement prediction, while allowing the three tasks to share only the representation components that are useful to them.

3.2 Multi-Resolution Acoustic Representation and Shared Encoder

All primary audio signals are converted to mono and resampled to $f_s = 22.05$ kHz. CTIM excerpts are already 10 s long. For FMA, a deterministic 10-s evaluation crop is retained for every selected track so that all compared models are evaluated on identical acoustic content. Training-time variability is introduced by augmentation rather than by changing held-out excerpts.

A single time–frequency resolution imposes a fixed compromise between temporal localization and frequency precision. EthnoPulse therefore computes two log-Mel representations using FFT sizes $N_1 = 1024$ and $N_2 = 2048$ with a common hop length

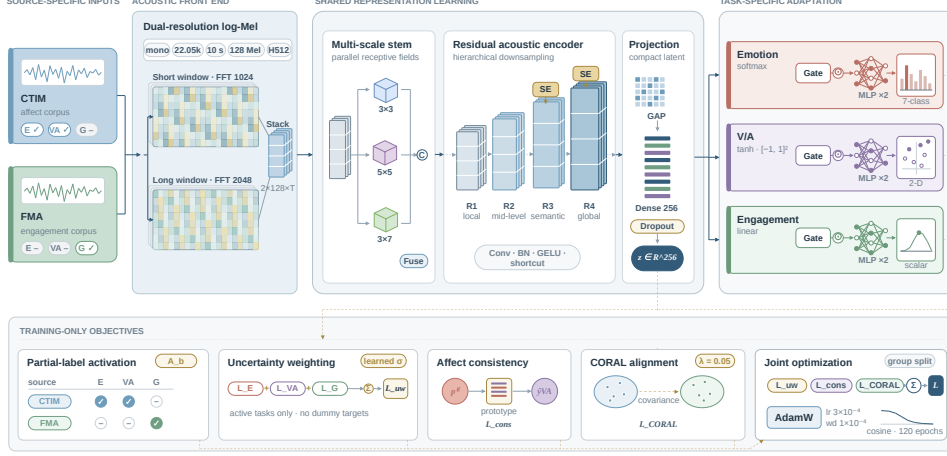


Fig. 1 Overview of the EthnoPulse framework. CTIM and FMA are processed by a common dual-resolution log-Mel front end and a shared multi-scale residual acoustic encoder. SE recalibration and global pooling produce a compact shared representation, which is adapted by task-specific gates for emotion classification, valence–arousal regression, and online-engagement prediction. Partial-label masking activates only the losses supported by the current data source. Emotion–VA consistency links the two affective outputs, while CORAL regularizes second-order discrepancies between CTIM and FMA embeddings. Dashed paths denote training-time regularization rather than inference-time operations.

$H = 512$. For resolution $r \in \{1, 2\}$, the short-time Fourier transform is

$$X^{(r)}(m, k) = \sum_{n=0}^{N_r-1} x[n + mH] w_r[n] \exp\left(-j \frac{2\pi kn}{N_r}\right), \quad (10)$$

where $w_r[n]$ denotes the analysis window. The power spectrum is projected onto $M = 128$ Mel filters using $\mathbf{B}^{(r)}$, giving

$$S_{m,q}^{(r)} = \log\left(\varepsilon + \sum_k B_{qk}^{(r)} \left|X^{(r)}(m, k)\right|^2\right), \quad q = 1, \dots, M. \quad (11)$$

After excerpt-level normalization, the aligned maps are stacked as

$$\mathbf{S}(x) = \text{Stack}\left[\text{Norm}(\mathbf{S}^{(1)}), \text{Norm}(\mathbf{S}^{(2)})\right] \in \mathbb{R}^{2 \times M \times T}. \quad (12)$$

The shorter analysis window provides finer localization of onsets and rapid articulation, while the longer window gives a more stable description of harmonic and timbral structure.

The shared encoder begins with a parallel multi-scale stem using

$$\mathcal{K} = \{(3, 3), (5, 5), (3, 7)\}. \quad (13)$$

For each kernel $k_q \in \mathcal{K}$,

$$\mathbf{U}^{(q)} = \phi[\text{BN}(\text{Conv}_{k_q}(\mathbf{S}))], \quad (14)$$

and the branch outputs are concatenated:

$$\mathbf{U}_0 = \text{Concat}[\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}]. \quad (15)$$

The 3×3 branch models compact local patterns, the 5×5 branch captures broader joint time–frequency structure, and the 3×7 branch extends temporal context while retaining a relatively narrow frequency extent. The branches therefore impose complementary inductive biases rather than merely increasing channel count.

The fused representation is then processed by four residual convolutional stages. With residual transform $\mathcal{F}_\ell(\cdot)$, stage ℓ is written as

$$\mathbf{U}_\ell = \phi[\mathcal{F}_\ell(\mathbf{U}_{\ell-1}) + \mathcal{P}_\ell(\mathbf{U}_{\ell-1})], \quad \ell = 1, \dots, 4, \quad (16)$$

where $\mathcal{P}_\ell$ is either an identity shortcut or a projection when channel dimensions or spatial resolution change. Residual learning facilitates optimization while preserving lower-level information across deeper stages.

The final two residual stages are followed by squeeze-and-excitation (SE) recalibration. For $\mathbf{U} \in \mathbb{R}^{C \times F \times T}$, channel statistics are obtained by

$$s_c = \frac{1}{FT} \sum_{f=1}^F \sum_{t=1}^T U_{cft}, \quad c = 1, \dots, C, \quad (17)$$

and the excitation network computes

$$\mathbf{a} = \sigma[\mathbf{W}_2^{se} \delta(\mathbf{W}_1^{se} \mathbf{s} + \mathbf{b}_1^{se}) + \mathbf{b}_2^{se}]. \quad (18)$$

The recalibrated feature tensor is

$$\tilde{U}_{cft} = a_c U_{cft}. \quad (19)$$

Compared with unconstrained attention over every time–frequency bin, SE introduces only modest additional complexity and allows the network to emphasize channels associated with informative spectral envelopes, transient density, harmonic structure, and other learned acoustic patterns.

Global average pooling, a dense projection, and dropout with rate 0.30 produce the 256-dimensional shared embedding in Eq. (5). The task gates then provide lightweight task adaptation without duplicating the encoder. The emotion head uses a two-layer MLP followed by softmax, the VA head uses a two-layer MLP with a two-dimensional tanh output, and the engagement head uses a two-layer MLP with a linear scalar output.

During training, the time–frequency inputs are augmented using random time shift, gain perturbation, and frequency/time masking inspired by SpecAugment. All augmentation is disabled during validation and testing.

Table 2 EthnoPulse architecture for the primary raw-audio multi-task experiment.

Component	Operation and role
Audio front end	22.05-kHz mono audio; 10-s segment; two 128-band log-Mel maps with FFT 1024/2048 and hop 512
Multi-scale stem	Parallel 3×3 , 5×5 , and 3×7 convolutions followed by channel-wise fusion
Residual encoder	Four convolutional residual stages with normalization and hierarchical downsampling
SE recalibration	Channel-wise adaptive feature weighting after the higher-level residual stages
Shared projection	Global average pooling, dense projection, dropout 0.30; $d = 256$ embedding
Task adaptation	Three sigmoid feature gates for emotion, VA, and engagement branches
Prediction heads	Seven-class softmax, two-dimensional tanh, and scalar linear output
Regularization	Partial-label masking, uncertainty weighting, emotion-VA consistency, and CORAL

3.3 Partial-Label Multi-Task Learning and Regularization

The principal optimization challenge is that CTIM and FMA provide different forms of supervision. EthnoPulse therefore defines an active-task set $\mathcal{A}_b$ for each batch b . For CTIM, the active tasks are discrete emotion recognition and VA regression; for FMA, the active task is engagement regression. Missing targets are never replaced with dummy labels.

For emotion classification, the cross-entropy loss is

$$\mathcal{L}_E = -\frac{1}{B_E} \sum_{i=1}^{B_E} \sum_{c=1}^C \mathbb{I}(y_i^E = c) \log \hat{p}_{ic}^E. \quad (20)$$

For continuous affect and engagement, SmoothL1 is used because it remains quadratic for small residuals but is less sensitive than squared error to large deviations. For error e ,

$$\rho_\delta(e) = \begin{cases} \frac{e^2}{2\delta}, & |e| < \delta, \\ |e| - \frac{\delta}{2}, & \text{otherwise.} \end{cases} \quad (21)$$

The VA and engagement losses are

$$\mathcal{L}_{VA} = \frac{1}{2B_E} \sum_{i=1}^{B_E} \sum_{r \in \{V, A\}} \rho_\delta(\hat{y}_{ir} - y_{ir}), \quad (22)$$

$$\mathcal{L}_G = \frac{1}{B_G} \sum_{j=1}^{B_G} \rho_\delta(\hat{y}_j^G - y_j^G). \quad (23)$$

Because the three objectives have different numerical scales, manually fixed weights can cause one task to dominate optimization. We therefore use learned homoscedastic uncertainty weighting. Let $s_t = \log \sigma_t^2$ denote a trainable log-variance parameter for task t . The active supervised objective is

$$\mathcal{L}_{\text{uw}} = \frac{1}{|\mathcal{A}_b|} \sum_{t \in \mathcal{A}_b} \left[\eta_t \exp(-s_t) \mathcal{L}_t + \frac{1}{2} s_t \right], \quad (24)$$

where $\eta_E = 1$ for cross-entropy and $\eta_{VA} = \eta_G = 1/2$ for the regression objectives. A task absent from the current batch is omitted from both the data-loss and uncertainty terms, ensuring that unavailable supervision does not influence the gradient.

CTIM provides both categorical emotion labels and continuous VA ratings. EthnoPulse uses this joint information through a soft consistency regularizer. For every emotion class c , an affect prototype is estimated exclusively from the training fold:

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{i: y_i^E = c} \mathbf{y}_i^{VA}. \quad (25)$$

The emotion posterior then induces a probability-weighted VA expectation,

$$\tilde{\mathbf{y}}_i^{VA} = \sum_{c=1}^C \hat{p}_{ic}^E \boldsymbol{\mu}_c, \quad (26)$$

and the consistency penalty is

$$\mathcal{L}_{\text{cons}} = \frac{1}{B_E} \sum_{i=1}^{B_E} \text{SmoothL1}(\hat{\mathbf{y}}_i^{VA}, \tilde{\mathbf{y}}_i^{VA}). \quad (27)$$

This formulation does not force discrete categories into rigid VA quadrants. Instead, it encourages the two affective views to remain compatible while preserving culturally specific emotion classes.

The two primary sources also differ in repertoire, recording conditions, scale, and label-generation mechanism. To reduce excessive representation mismatch without enforcing complete domain invariance, CORAL is applied to the shared embeddings as a lightweight second-order regularizer. For source-specific embedding matrices $\mathbf{Z}_E \in \mathbb{R}^{B_E \times d}$ and $\mathbf{Z}_G \in \mathbb{R}^{B_G \times d}$, let

$$\mathbf{C}_q = \frac{1}{B_q - 1} \bar{\mathbf{Z}}_q^\top \bar{\mathbf{Z}}_q, \quad q \in \{E, G\}, \quad (28)$$

where $\bar{\mathbf{Z}}_q$ is row-centered. The alignment penalty is

$$\mathcal{L}_{\text{CORAL}} = \frac{1}{4d^2} \|\mathbf{C}_E - \mathbf{C}_G\|_F^2. \quad (29)$$

CORAL is evaluated only when both source mini-batches contain enough samples to estimate covariance. In the principal benchmark, $\lambda_{\text{coral}} = 0.05$ is fixed before test-set evaluation and its contribution is assessed through ablation rather than test-set tuning.

The complete EthnoPulse objective is therefore

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{uw}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{coral}}\mathcal{L}_{\text{CORAL}} \quad (30)$$

with unavailable regularizers set to zero for batches that do not satisfy their supervision requirements. The three terms have distinct roles: $\mathcal{L}_{\text{uw}}$ fits the observed tasks, $\mathcal{L}_{\text{cons}}$ couples categorical and dimensional affect, and $\mathcal{L}_{\text{CORAL}}$ limits excessive second-order discrepancy between the emotion and engagement domains.

3.4 Optimization and Overall Training Algorithm

The network is optimized using AdamW with an initial learning rate of 3×10^{-4} and weight decay of 1×10^{-4} . The nominal mini-batch size is 32 per source when memory permits. Training is limited to 120 epochs, and the learning rate follows a cosine-annealing schedule. Early stopping is determined exclusively from a group-disjoint validation partition. All statistics used for target standardization, affect prototypes, and other preprocessing transformations are estimated from the corresponding training fold and then frozen for validation and test samples.

At each update, an emotion mini-batch and an engagement mini-batch are sampled when both sources are available. Each batch passes through the same acoustic front end and shared encoder. The CTIM batch activates the emotion, VA, and consistency terms; the FMA batch activates only the engagement loss. If both batches are present, their shared embeddings are additionally used to compute CORAL. The uncertainty-weighted supervised loss and the active regularizers are then combined according to Eq. (30), and all trainable parameters are updated jointly.

Algorithm 1 summarizes the full training procedure. The formulation makes the partial-label mechanism explicit and ensures that the shared encoder receives information from both corpora without introducing fabricated supervision. Model selection is performed only on the group-disjoint validation data, and the selected checkpoint is subsequently evaluated unchanged on the held-out fold.

4 Evaluation

4.1 Experimental Setup

We evaluate EthnoPulse along four axes: *task performance*, *component contribution*, *leakage-aware generalization*, and *cross-cultural robustness*. The primary benchmark uses the CTIM emotion source and the FMA engagement source under the partial-label setting defined in Sec. 3. CTIM is split by performance/composition and FMA by artist so that no group appears in both training and test partitions. This group-disjoint protocol is essential for short music excerpts because neighboring segments can share recording conditions, performer identity, instrumentation, and local musical

Algorithm 1 EthnoPulse partial-label multi-task training

Require: Affect corpus $\mathcal{D}_E$, engagement corpus $\mathcal{D}_G$, group-disjoint train/validation partitions, λ_{cons} , λ_{coral} , maximum epochs T_{max}

Ensure: Best model parameters Θ^*

```
1: Fit engagement normalization and affect prototypes using training data only
2: Initialize encoder, SE modules, task gates, task heads, uncertainty parameters, AdamW,
   and learning-rate scheduler
3: for  $e = 1$  to  $T_{\text{max}}$  do
4:   for paired source mini-batches  $(\mathcal{B}_E, \mathcal{B}_G)$  do
5:     Construct augmented dual-resolution log-Mel inputs
6:     Compute available shared embeddings  $\mathbf{Z}_E$  and  $\mathbf{Z}_G$ 
7:     if  $\mathcal{B}_E$  is available then
8:       Predict  $\hat{\mathbf{p}}^E$  and  $\hat{\mathbf{y}}^{VA}$ 
9:       Compute  $\mathcal{L}_E$ ,  $\mathcal{L}_{VA}$ , and  $\mathcal{L}_{\text{cons}}$ 
10:    else
11:       $\mathcal{L}_E, \mathcal{L}_{VA}, \mathcal{L}_{\text{cons}} \leftarrow 0$ 
12:    end if
13:    if  $\mathcal{B}_G$  is available then
14:      Predict  $\hat{\mathbf{y}}^G$  and compute  $\mathcal{L}_G$ 
15:    else
16:       $\mathcal{L}_G \leftarrow 0$ 
17:    end if
18:    Form active task set  $\mathcal{A}_b$  and compute  $\mathcal{L}_{\text{uw}}$ 
19:    if  $\mathbf{Z}_E$  and  $\mathbf{Z}_G$  both contain at least two samples then
20:      Compute  $\mathcal{L}_{\text{CORAL}}$ 
21:    else
22:       $\mathcal{L}_{\text{CORAL}} \leftarrow 0$ 
23:    end if
24:    Compute  $\mathcal{L}_{\text{total}}$  using Eq. (30)
25:    Update  $\Theta$  by back-propagation and AdamW
26:  end for
27:  Evaluate without augmentation on the group-disjoint validation partition
28:  if validation performance improves then
29:    Save  $\Theta^* \leftarrow \Theta$ 
30:  end if
31:  if early stopping is triggered then
32:    break
33:  end if
34: end for
35: return  $\Theta^*$ 
```

material. All neural comparisons use the same outer folds and validation protocol; test folds are never used for hyperparameter selection.

Metrics. Seven-class emotion recognition is evaluated with accuracy, macro- F_1 , and unweighted average recall (UAR). Valence and arousal are reported separately using mean absolute error (MAE), root mean squared error (RMSE), R^2 , and Spearman’s rank correlation coefficient (ρ). Engagement prediction is evaluated with MAE, RMSE,

Table 3 Primary raw-audio training configuration. All settings are fixed before test-set evaluation.

Item	Setting
Sampling rate	22.05 kHz
Clip duration	10 s
Log-Mel input	128 bands per resolution
FFT / hop	1024 and 2048 / 512 samples
Optimizer	AdamW
Initial learning rate	3×10^{-4}
Weight decay	1×10^{-4}
Batch size	32 per source when memory permits
Maximum epochs	120 with early stopping
Scheduler	CosineAnnealingLR
Shared embedding	256 dimensions
Dropout	0.30
CORAL coefficient	0.05 (pre-specified)
Outer validation	Five group-disjoint folds
Split unit	CTIM: composition; FMA: artist

R^2 , and ρ . We emphasize macro- F_1 /UAR for emotion classification and rank correlation for continuous targets because these metrics are less sensitive to class imbalance and scale calibration, respectively.

Baselines. The raw-audio benchmark compares EthnoPulse with matched single-task CNNs, a plain multi-task CNN, a convolutional recurrent neural network (CRNN), [17] and a ResNet-18-style multi-task baseline. [16] The external CCMED/WCMED benchmark uses a mean baseline, Ridge regression, PLS-10, Random Forest, Extra Trees, and histogram gradient boosting (HistGB). The external models are intentionally heterogeneous: linear, latent-factor, bagging, and boosting methods test whether the observed trends depend on one specific estimator family.

4.2 Primary Benchmark and Ablation

The primary experiment is designed to separate the gain from *task sharing* from the gain due to the proposed architecture. The single-task CNNs establish the cost or benefit of sharing representations across emotion and engagement. The plain CNN-MTL baseline tests multi-task learning without the proposed multi-resolution and task-adaptive components. CRNN and ResNet-18 provide stronger temporal and residual baselines under exactly the same folds. EthnoPulse is therefore considered successful only if its improvement is consistent across the principal emotion and engagement metrics rather than being driven by one favorable score.

The ablation study follows the same principle. Each variant removes a single component while keeping the data partition, optimizer, and training budget unchanged. This controls for confounding changes in optimization and directly tests the design assumptions behind the framework.

Table 4 Pre-specified ablation matrix for the primary CTIM/FMA benchmark.

Variant	Question tested
Full EthnoPulse	Complete partial-label multi-task system
– second Mel resolution	Does complementary spectral resolution help?
– multi-scale stem	Are parallel receptive fields necessary?
– SE attention	Does channel recalibration improve prediction?
– task-specific gates	Do gates reduce negative transfer?
– emotion–VA consistency	Does cross-view affect consistency help?
Fixed equal task weights	Is uncertainty weighting beneficial?
– CORAL	Does cross-source alignment help?
– audio augmentation	Is the gain robust to regularization removal?
Single-task models	Does shared representation learning help?

Table 5 Within-domain prediction under composition-grouped five-fold cross-validation. Values are fold mean $\pm$ standard deviation where available.

Dataset	Model	Valence MAE	Valence ρ	Arousal MAE	Arousal ρ
CCMED	Mean baseline	0.506 $\pm$ 0.076	–	0.503 $\pm$ 0.071	–
CCMED	Ridge	0.567 $\pm$ 0.174	0.223	0.331 $\pm$ 0.139	0.753
CCMED	PLS-10	0.607 $\pm$ 0.215	0.179	0.336 $\pm$ 0.127	0.730
CCMED	Random Forest	0.459 $\pm$ 0.055	0.340	0.316 $\pm$ 0.068	0.756
CCMED	Extra Trees	0.460 $\pm$ 0.051	0.318	0.297 $\pm$ 0.062	0.773
CCMED	HistGB	0.474 $\pm$ 0.052	0.292	0.286$\pm$0.054	0.771
WCMED	Mean baseline	0.504 $\pm$ 0.067	–	0.503 $\pm$ 0.050	–
WCMED	Ridge	0.549 $\pm$ 0.106	0.201	0.303 $\pm$ 0.056	0.798
WCMED	PLS-10	0.586 $\pm$ 0.101	0.155	0.290 $\pm$ 0.020	0.806
WCMED	Random Forest	0.488$\pm$0.081	0.201	0.291 $\pm$ 0.014	0.781
WCMED	Extra Trees	0.489 $\pm$ 0.081	0.205	0.283 $\pm$ 0.008	0.786
WCMED	HistGB	0.508 $\pm$ 0.097	0.169	0.282$\pm$0.014	0.771

4.3 Within-Domain External Validation

CCMED and WCMED provide a controlled external test because they share the same 467-dimensional acoustic representation and valence–arousal annotation scheme. We use composition-grouped five-fold cross-validation, so excerpts originating from the same source composition remain in a single fold. This benchmark is not a substitute for the raw-audio CTIM/FMA experiment; it is used to test whether the methodological conclusions hold in a matched feature space.

Figure 2 compares MAE for all six models across the four dataset–target conditions. Arousal is consistently more predictable than valence. On CCMED, HistGB achieves the lowest arousal MAE (0.286 ± 0.054), while Extra Trees reaches $\rho = 0.773$. On WCMED, HistGB reaches an arousal MAE of 0.282 ± 0.014 , and PLS-10 yields the strongest arousal correlation ($\rho = 0.806$). Valence correlations remain modest across all estimators; Random Forest performs best on CCMED valence ($\rho = 0.340$), while

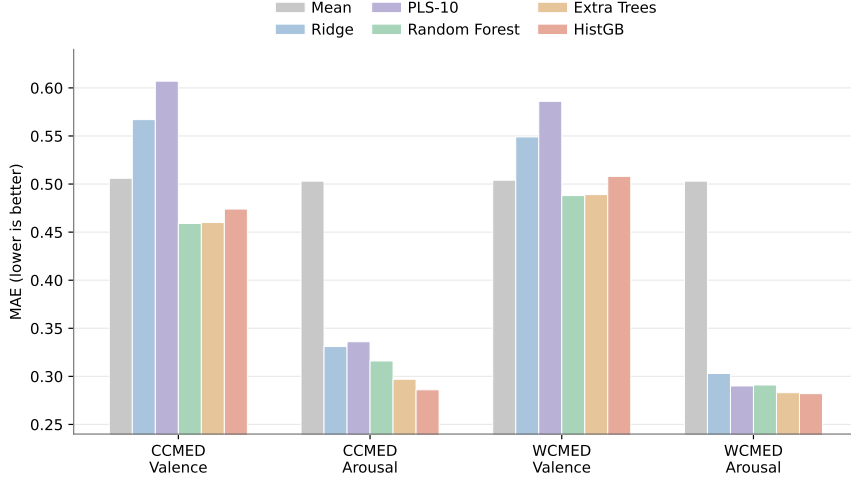


Fig. 2 Within-domain MAE under composition-grouped five-fold cross-validation. The grouped bar chart summarizes six models across four dataset–target conditions. Lower values indicate better prediction. The consistent reduction in arousal error, compared with the smaller gains for valence, reveals a strong target-dependent asymmetry.

Table 6 Effect of random excerpt splitting versus composition-grouped splitting for Extra Trees. Positive $\Delta\rho$ indicates optimistic inflation under random splitting.

Dataset	Target	Group MAE	Random MAE	Group ρ	Random ρ	$\Delta\rho$
CCMED	Valence	0.460	0.404	0.318	0.527	0.209
CCMED	Arousal	0.297	0.254	0.773	0.824	0.051
WCMED	Valence	0.489	0.461	0.205	0.323	0.118
WCMED	Arousal	0.283	0.263	0.786	0.829	0.042

the best WCMED valence correlation is only 0.205. These results justify reporting valence and arousal separately rather than collapsing them into a single affect score.

4.4 Generalization and Statistical Robustness

Leakage analysis. We first test whether ordinary random excerpt splitting produces optimistic estimates. Extra Trees is evaluated under both composition-grouped and shuffled five-fold splits. Random splitting improves every condition. The largest effect occurs for CCMED valence, where ρ rises from 0.318 to 0.527 ($\Delta\rho = 0.209$). WCMED valence increases from 0.205 to 0.323, while arousal also improves but by a smaller margin. These gains are not evidence for a better model; they quantify information leakage from composition-specific acoustic signatures.

Cross-cultural transfer. Models trained on one cultural domain are evaluated on the other without target labels. Direct arousal transfer remains strong in both directions. For CCMED→WCMED, Random Forest reaches $\rho = 0.808$; for WCMED→CCMED,

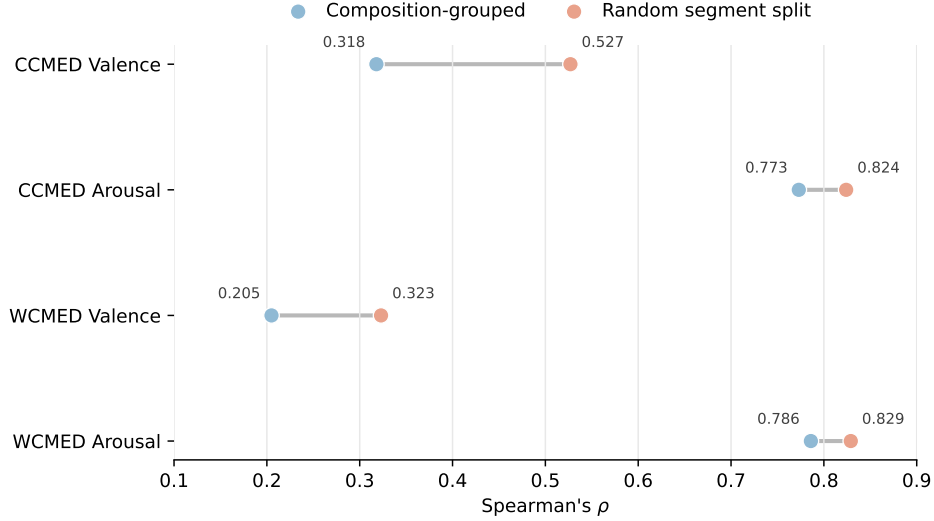


Fig. 3 Leakage diagnostic for composition-grouped and random excerpt splitting. Each pair represents the same dataset–target condition; the displacement directly visualizes the optimism introduced by random segment mixing.

Table 7 Cross-domain transfer between CCMED and WCMED. CORAL uses target features but no target labels.

Direction	Adaptation	Model	Valence MAE	Valence ρ	Arousal MAE	Arousal ρ
CCMED→WCMED	Direct	Ridge	0.496	0.229	0.313	0.805
CCMED→WCMED	Direct	Random Forest	0.476	0.304	0.337	0.808
CCMED→WCMED	Direct	Extra Trees	0.481	0.252	0.327	0.805
CCMED→WCMED	Direct	HistGB	0.501	0.200	0.325	0.802
CCMED→WCMED	CORAL	Ridge	0.505	0.185	0.346	0.701
CCMED→WCMED	CORAL	Random Forest	0.483	0.214	0.396	0.683
CCMED→WCMED	CORAL	Extra Trees	0.482	0.185	0.375	0.701
CCMED→WCMED	CORAL	HistGB	0.502	0.140	0.355	0.644
WCMED→CCMED	Direct	Ridge	0.667	0.167	0.464	0.652
WCMED→CCMED	Direct	Random Forest	0.492	0.130	0.337	0.756
WCMED→CCMED	Direct	Extra Trees	0.519	−0.113	0.326	0.778
WCMED→CCMED	Direct	HistGB	0.522	−0.070	0.316	0.779
WCMED→CCMED	CORAL	Ridge	0.511	0.145	0.344	0.679
WCMED→CCMED	CORAL	Random Forest	0.486	0.267	0.422	0.628
WCMED→CCMED	CORAL	Extra Trees	0.481	0.283	0.394	0.672
WCMED→CCMED	CORAL	HistGB	0.488	0.223	0.363	0.640

HistGB reaches $\rho = 0.779$. Valence is markedly less stable. In the reverse direction, direct Extra Trees and HistGB even yield negative rank correlations. CORAL partially repairs some valence transfer—for example, WCMED→CCMED Extra Trees improves from -0.113 to 0.283 —but simultaneously degrades arousal from 0.778 to 0.672 . Alignment is therefore target-dependent rather than universally beneficial.

Statistical reliability. Composition-level absolute errors are compared with the mean baseline using paired one-sided Wilcoxon signed-rank tests. Extra Trees and

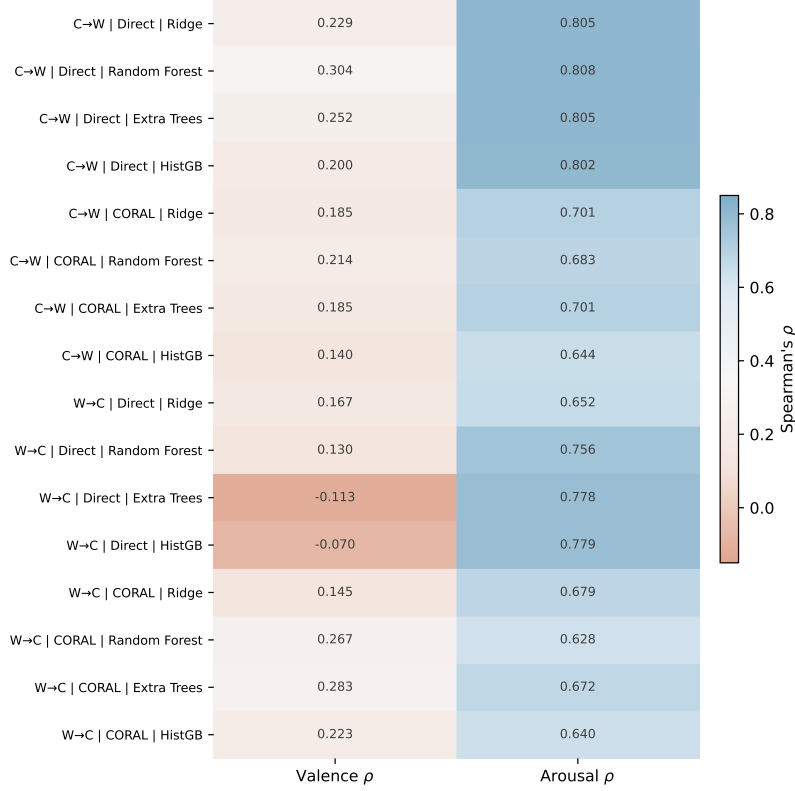


Fig. 4 Cross-domain Spearman correlations under direct transfer and CORAL alignment. Arousal remains highly transferable across domains, whereas valence is weaker, asymmetric, and more sensitive to adaptation.

HistGB significantly reduce arousal error on both datasets ($p < 0.001$). CCMED valence also improves significantly, whereas WCMED valence does not. Bootstrap confidence intervals support the same conclusion: arousal effects are both larger and more stable, while valence estimates remain uncertain in WCMED.

4.5 Discussion and Limitations

The external analyses show a stable difference between the two affect dimensions. Arousal is predicted more accurately than valence across estimator families, datasets, and transfer directions. The split strategy also matters: random excerpt splitting raises correlation, especially for valence, indicating that composition-specific acoustic information can leak across folds. Cross-cultural transfer is similarly uneven. Arousal transfers well in both directions, whereas valence is weaker and more sensitive to adaptation. CORAL improves some valence settings but reduces arousal performance in others, so alignment is best treated as a task-dependent regularizer rather than a universal correction.

Table 8 Paired one-sided Wilcoxon tests on composition-level absolute error versus the mean baseline.

Dataset	Target	Model	Songs	Median error	Baseline	p
CCMED	Valence	Extra Trees	62	0.440	0.490	0.001
CCMED	Valence	HistGB	62	0.462	0.490	0.011
CCMED	Arousal	Extra Trees	62	0.254	0.515	< 0.001
CCMED	Arousal	HistGB	62	0.235	0.515	< 0.001
WCMED	Valence	Extra Trees	61	0.469	0.482	0.178
WCMED	Valence	HistGB	61	0.474	0.482	0.299
WCMED	Arousal	Extra Trees	61	0.282	0.477	< 0.001
WCMED	Arousal	HistGB	61	0.284	0.477	< 0.001

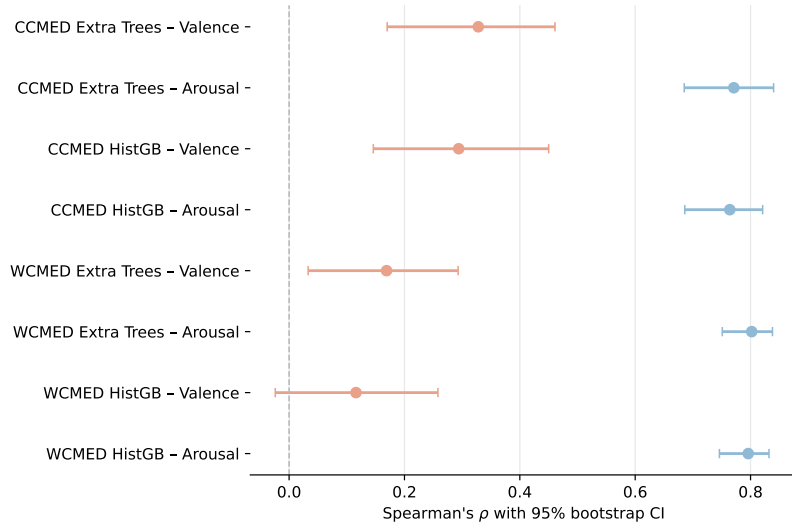


Fig. 5 Bootstrap 95% confidence intervals for selected grouped-CV Spearman correlations. The forest plot shows that arousal effects are consistently large and precise, whereas valence estimates are smaller and less stable.

These findings should be read with two limitations. CTIM is small for seven-class deep learning, and FMA engagement reflects exposure, recommendation, artist history, and network effects in addition to acoustic content. The engagement head therefore represents a content-based cold-start estimate rather than a causal model of dissemination. In addition, the CCMED/WCMED experiments use released engineered features. They provide an external check on grouped evaluation and cross-cultural transfer, but they are not an end-to-end evaluation of the raw-audio CTIM/FMA model. A complete assessment of EthnoPulse therefore requires the planned CTIM/FMA benchmark with fold- and seed-level uncertainty reported under the same group-disjoint protocol.

5 Conclusion

EthnoPulse was developed for a setting in which affect and engagement labels come from separate music corpora. The model shares a multi-resolution acoustic encoder while retaining task-specific gates and losses, and it uses uncertainty weighting, emotion–valence/arousal consistency, and CORAL to control interactions among the tasks and sources.

The completed CCMED/WCMED analyses provide two clear methodological results. First, arousal is more stable than valence both within domain and under Chinese–Western transfer. Second, random excerpt splitting produces more optimistic estimates than composition-grouped evaluation. CORAL is not uniformly beneficial: it can improve valence in selected transfer settings while reducing arousal performance. These results favor group-disjoint validation, separate reporting of valence and arousal, and cautious use of domain alignment. The remaining step is to evaluate the full raw-audio CTIM/FMA system under the same protocol; until that benchmark is completed, claims about the end-to-end multi-task model should remain limited.

Declarations

Data availability

CTIM, FMA, CCMED, and WCMED are third-party research datasets and remain subject to the access conditions and licenses specified by their original providers. No restricted or license-controlled audio is redistributed with this manuscript. Derived CCMED/WCMED analysis files can be provided by the corresponding author on reasonable request.

Author contributions

C.Z. was responsible for the conceptualization, methodology, data analysis, visualization, and preparation of the manuscript. The author reviewed and approved the final manuscript.

Funding

This work was supported by the project “Development and Research of an AI-Based Audio–Visual Synesthesia Psychological Regulation System” (Project No. BY20251490).

Competing interests

The author declares no competing interests.

References

- [1] Jiang, X., Zhang, Y., Lin, G., Yu, L.: Music emotion recognition based on deep learning: A review. *IEEE Access* **12**, 157716–157745 (2024) <https://doi.org/10.1109/ACCESS.2024.3484470>

- [2] Louro, P.L., Redinho, H., Malheiro, R., Paiva, R.P., Panda, R.: A comparison study of deep learning methodologies for music emotion recognition. *Sensors* **24**(7), 2201 (2024) <https://doi.org/10.3390/s24072201>
- [3] Mycka, J., Mańdziuk, J.: Artificial intelligence in music: Recent trends and challenges. *Neural Computing and Applications* **37**, 801–839 (2025) <https://doi.org/10.1007/s00521-024-10555-x>
- [4] Wu, J.: Emotion classification in music: Leveraging machine learning for music therapy and emotional response analysis. *International Journal of Computational Intelligence Systems* **19**, 287 (2026) <https://doi.org/10.1007/s44196-026-01304-0>
- [5] Wu, D., Jia, X., Rao, W., Dou, W., Li, Y., Li, B.: Construction of a chinese traditional instrumental music dataset: A validated set of naturalistic affective music excerpts. *Behavior Research Methods* **56**, 3757–3778 (2024) <https://doi.org/10.3758/s13428-024-02411-6>
- [6] Xu, L., Liu, Y., Jiang, Z., Tang, Y., Wu, X.: How does chinese traditional instrumental music convey emotions? the power of affective timbres. *Humanities and Social Sciences Communications* **12**, 1931 (2025) <https://doi.org/10.1057/s41599-025-06211-x>
- [7] Reisz, N., Servedio, V.D.P., Thurner, S.: Quantifying the impact of homophily and influencer networks on song popularity prediction. *Scientific Reports* **14**, 8929 (2024) <https://doi.org/10.1038/s41598-024-58969-w>
- [8] Defferrard, M., Benzi, K., Vandergheynst, P., Bresson, X.: FMA: A dataset for music analysis. In: *Proceedings of the 18th International Society for Music Information Retrieval Conference*, pp. 316–323 (2017)
- [9] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141 (2018)
- [10] Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491 (2018)
- [11] Abdelsamie, M.M., Azab, S.S., Hefny, H.A.: Deep multi-task learning: A review of concepts, methods, and cross-domain applications. *International Journal of Data Science and Analytics* **21**, 77 (2026) <https://doi.org/10.1007/s41060-025-00892-y>
- [12] Sun, B., Saenko, K.: Deep CORAL: Correlation alignment for deep domain adaptation. In: *Computer Vision – ECCV 2016 Workshops*, pp. 443–450. Springer, ??? (2016)
- [13] Khoei, A.G., Yu, Y., Feldt, R.: Domain generalization through meta-learning:

A survey. *Artificial Intelligence Review* **57**, 285 (2024) <https://doi.org/10.1007/s10462-024-10922-z>

- [14] Fan, J., Yang, Y.-H., Dong, K., Pasquier, P.: A comparative study of western and chinese classical music based on soundscape models. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (2020). <https://doi.org/10.1109/ICASSP40776.2020.9052994>
- [15] Wang, J., Sharifi, A., Gadekallu, T.R., Shankar, A.: MMD-MII model: A multi-layered analysis and multimodal integration interaction approach revolutionizing music emotion classification. *International Journal of Computational Intelligence Systems* **17**, 99 (2024) <https://doi.org/10.1007/s44196-024-00489-6>
- [16] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
- [17] Choi, K., Fazekas, G., Sandler, M., Cho, K.: Convolutional recurrent neural networks for music classification. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2392–2396 (2017)
- [18] Li, Y., Yuan, R., Zhang, G., *et al.*: MERT: Acoustic music understanding model with large-scale self-supervised training. In: *Proceedings of the 12th International Conference on Learning Representations (ICLR)* (2024)
- [19] Zhu, H., Zhou, Y., Chen, H., Yu, J., Ma, Z., Gu, R., Luo, Y., Tan, W., Chen, X.: MuQ: Self-supervised music representation learning with mel residual vector quantization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **33**, 3653–3664 (2025) <https://doi.org/10.1109/TASLPRO.2025.3602320>
- [20] Du, X.: BLNN: A muscular and tall architecture for emotion prediction in music. *Soft Computing* **28**, 11855–11871 (2024) <https://doi.org/10.1007/s00500-024-09922-6>
- [21] Sun, Y.: Feature centric based deep learning approach for music mood recognition with HuBERT transformer model. *Scientific Reports* **15**, 45460 (2025) <https://doi.org/10.1038/s41598-025-29200-1>
- [22] Li, T.: Music emotion recognition using deep convolutional neural networks. *Journal of Computational Methods in Science and Engineering* **24**(4–5), 3063–3078 (2024) <https://doi.org/10.3233/JCM-247551>
- [23] Su, Y., Chai, R., Yang, H., Wu, X., Sun, W.: Multi-resolution spectrogram based multi-branch hybrid attention network for music emotion recognition. *Journal of Audio, Speech, and Music Processing* **2026**, 25 (2026) <https://doi.org/10.1186/s13636-026-00460-7>

- [24] Buisson, M., McFee, B., Essid, S., Crayencour, H.C.: Self-supervised learning of multi-level audio representations for music segmentation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **32**, 2141–2152 (2024) <https://doi.org/10.1109/TASLP.2024.3379894>
- [25] Fong, H., Kumar, V., Sudhir, K.: A theory-based explainable deep learning architecture for music emotion. *Marketing Science* **44**(1), 196–219 (2025) <https://doi.org/10.1287/mksc.2022.0323>
- [26] Russell, J.A.: A circumplex model of affect. *Journal of Personality and Social Psychology* **39**(6), 1161–1178 (1980)
- [27] Lyu, M., Egermann, H.: A cross-cultural study in chinese and western music: Cultural advantages in recognition of emotions. *Psychology of Music* **54**(1), 63–82 (2026) <https://doi.org/10.1177/03057356241303857>
- [28] Caruana, R.: Multitask learning. *Machine Learning* **28**, 41–75 (1997)
- [29] Zhang, D., You, W., Liu, Z., Sun, L., Chen, P.: Personalized dynamic music emotion recognition with dual-scale attention-based meta-learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 1629–1637 (2025). <https://doi.org/10.1609/aaai.v39i2.32155>
- [30] Bevec, M., Tkalcic, M., Pesek, M.: Hybrid music recommendation with graph neural networks. *User Modeling and User-Adapted Interaction* **34**, 1891–1928 (2024) <https://doi.org/10.1007/s11257-024-09410-4>