ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Selective Knowledge Distillation in Music Cultural Emotion Modeling: A Mechanistic Study of Usability and reliability

Yang Liu[1] *and Hao Wang*[2,*] *and Yukai Liu*[3]

[1] School of Cruise and Art Design, Jiangsu Maritime Institute, Nanjing, 211170, China
[2] School of Music and Dance, Nanjing Normal University of Special Education, Nanjing, 211170, China
[3] School of Computer Science and Engineering of University of New South Wales, Sydney 2052, New South Wales, Australia

SUMMARY: Music cultural emotion modeling faces the problems of strong subjectivity of labeling, obvious differences in cultural contexts, and high cost of complex model deployment. In order to improve the usability and reliability of lightweight models in music emotion recognition, this paper proposes a selective knowledge distillation method. The method takes audio spectrogram, lyrics semantics and cultural labels as multi-source inputs, generates emotional soft labels and hidden layer representations by the teacher model, and dynamically adjusts the distillation weight by combining prediction confidence, modal consistency and category reliability, so that the student model preferentially absorbs high-confidence emotional knowledge. Experimental results show that the proposed method achieves an Accuracy of 0.837, a Macro-F1 of 0.819 and an AUROC of 0.914 in the comprehensive test, and the ECE is reduced to 0.052. The performance is close to the full teacher model with 19.2M parameters and 9.1ms inference delay. In the cross-dataset validation, the AUROC of the model remains between 0.837 and 0.876, and the CCC reaches 0.657 under the low-label condition. The results show that selective knowledge distillation can effectively coordinate model compression, emotion discrimination and confidence calibration, and provide reliable technical support for intelligent music recommendation, digital aesthetic education and human-computer emotional interaction.

KEYWORDS: music cultural emotion modeling; Selective knowledge distillation; Model reliability; Lightweight emotion recognition

1 Introduction

Music cultural emotion modeling reveals the differences in emotional expression and perception of music works in different audience groups by analyzing the emotional cues in music acoustic structure, lyrics semantics, aesthetic experience and cultural context. It is an important research content of music information retrieval, intelligent recommendation, digital aesthetic education and human-computer emotional interaction. Compared with general music emotion recognition, music cultural emotion modeling not only focuses on the computable features such as pitch, rhythm, harmony and timbre, but also emphasizes the influence of lyric imagery, musical style tradition, regional aesthetic and cultural symbols on emotion judgment. In recent years, methods such as deep neural networks, transformers, and multi-modal

*hw090022@163.com
https://doi.org/10.65102/is2026910
Published: 30 April, 2026

representation learning have promoted the performance improvement of music emotion recognition. However, large-scale models often face problems such as large parameters, high inference costs, and limited deployment conditions in practical applications, which are difficult to directly adapt to mobile terminals, online teaching platforms, and lightweight music analysis systems.

There are still two types of prominent contradictions in the existing research. On the one hand, complex teacher model can learn rich emotional knowledge from multi-source music data, but its output is not always stable and reliable. When the training samples have cultural label bias, emotional annotation divergence, or semantic drift of cross-lingual lyrics, the teacher model may pass local noise, weakly related cultural symbols, or overconfident classification results to the student model. On the other hand, traditional knowledge distillation usually regards the teacher output as a unified soft label, and rarely distinguishes the knowledge quality of different samples, different modalities and different emotion categories, which leads to the fact that although the student model obtains high computational efficiency in the compression process, it may sacrifice the generalization ability and confidence credibility across datasets. For music cultural emotion modeling, the usability of the model is not only reflected in the recognition accuracy and inference speed, but also reflected in the stable response in resource-constrained scenarios. Reliability involves handling uncertain samples, adapting to cultural differences and calibrating prediction confidence.

In response to the above problems, this paper constructs a selective knowledge distillation framework for music cultural emotion modeling, which integrates usability and reliability into the knowledge transfer process at the same time. Methodically, the teacher model extracts emotion representations from audio spectrograms, lyrics semantics and cultural tags respectively, and evaluates the credibility level of distilled knowledge through confidence, sample consistency and cross-modal matching. The student model receives the filtered and weighted emotion knowledge in the lightweight structure to avoid the interference of low-quality soft labels on the decision boundary of the model. The main innovation of this paper is to construct a credibility evaluation mechanism for music cultural emotion knowledge, design a dynamic distillation weight to adjust the knowledge transfer strength of different samples, and verify the mechanism of usability and reliability from the dimensions of recognition performance, model size, inference efficiency, cross-dataset generalization and confidence calibration. This research can provide method reference for lightweight and credible music emotion understanding system.

2 Related work

Early research on music emotion recognition mainly relied on acoustic features and traditional classifiers, and used indicators such as rhythm, pitch, timbral, spectral centroid, Mel-frequency cepstral coefficient to depict music emotional tendency. Such methods have certain effects in small-scale data with clear category boundaries, but their feature expression ability is obviously limited in the face of complex emotions, cultural style differences and subjective annotation differences. As deep learning methods enter the field of music information retrieval, convolutional neural networks, recurrent neural networks and transformers are gradually used for music emotion modeling. Louro et al. [1] compared the performance of various deep learning methods in music emotion recognition and pointed out that deep models can improve the acoustic pattern extraction ability, but the more complex the model structure, the more reasoning overhead and deployment difficulty will increase. Lima Louro et al. [2] constructed the MERGE bimodal audio-lyrics dataset, which provides the

basis for the collaborative modeling of audio and text for static music emotion recognition. Gomez-Canon et al. [3] proposed the TROMPA-MER open dataset to further emphasize the importance of listener differences and cultural perception in personalized music emotion recognition.

In recent years, researchers have begun to expand music emotion modeling from two directions: multimodal and cultural context. The Music4All-Onion dataset released by Moscati et al. [4] integrates music content, metadata and multi-aspect annotations, which provides data conditions for large-scale music recommendation and emotion analysis. Strauss et al. [5] constructed a database of EMMA emotional music clips so that music emotion research could be carried out in a more systematic framework of emotion categories. Turchet and Pauwels[6] compared the different understandings of music emotion among composers, performers, listeners and machines, and showed that music emotion was not solely determined by acoustic signals, but was influenced by expression intention, listening experience and cultural background. Turchet et al. [7] further combined EEG, ECG and acoustic signals to identify players' emotions, and showed that there was a modelable emotional association between physiological signals and acoustic features.

Table 1: Related research context and concerns of this paper

Research Direction	Representative Literature	Main Contribution	Remaining Problem
Deep music emotion recognition	Louro et al. [1]	Compared the recognition performance of different deep learning models	High deployment cost of complex models
Audio–lyric bimodal modeling	Lima Louro et al. [2]	Constructed a bimodal music emotion dataset	Insufficient use of cultural context
Personalization and cultural awareness	Gómez-Cañón et al. [3], Turchet and Pauwels [6]	Emphasized listener differences and emotional subjectivity	Cross-dataset generalization remains unstable
Knowledge distillation and model compression	Aslam et al. [8], Brown et al. [9], Malihi et al. [10]	Improved the learning efficiency of lightweight models	Lack of screening for knowledge credibility

As shown in Table 1, musical emotion recognition has expanded from single acoustic modeling to multi-source fusion of audio, lyrics, cultural tags, and user perception, but the practical usability of complex models with output reliability has not been addressed synchronously. In terms of model compression, knowledge distillation provides an important path for lightweight emotion models. Aslam et al. [8] introduced privileged knowledge distillation into continuous emotion recognition in the wild, and proved that the teacher model could transfer richer emotion representation to the student model. Brown et al. [9] evaluated the efficiency of Transformer knowledge distillation and pointed out that distillation can reduce the model size while preserving performance. Malihi and Heidemann[10] combined sequential knowledge distillation with pruning to improve the controllability of the compression process. Dantas et al. [11] systematically reviewed machine learning model compression techniques and considered distillation, pruning, and quantization as key methods in resource-constrained scenarios. However, traditional distillation usually assumes that the teacher output has consistent reliability, and rarely judges the credibility of soft labels on different samples, different modalities and different emotion categories.

Trustworthy artificial intelligence provides theoretical support for solving the above

problems. Kaur et al. [12] discussed the framework of trusted AI from the perspective of transparency, reliability and fairness, and Radclyffe et al. [13] reviewed the list of trusted AI evaluations and put forward suggestions for improvement. Haque et al. [14] summarized the research of explainable artificial intelligence from the perspective of users, emphasizing that the model output needs to be understood and verified by users. Vashistha and Farahi[15] proposed that reliability, capability and confidence jointly affect the credibility of model decision-making. Fakour et al. [16] pointed out that the uncertainty of machine learning would directly affect the judgment stability of the model in complex environments. It can be seen that the distillation process in music cultural emotion modeling should not only pursue compression rate and accuracy, but also consider whether the teacher's knowledge is reliable, whether the student model is suitable for deployment, and whether the prediction confidence is calibrated. Based on this research gap, we introduce selective knowledge distillation into music cultural emotion modeling, and establish an interpretable collaborative optimization path between availability and reliability through credibility evaluation and dynamic weight generation mechanism.

3 Methods

This paper constructs a selective knowledge distillation framework for music cultural emotion modeling. The framework takes music audio, lyric text, cultural labels and artificial emotion annotations as input, and forms a unified multi-source emotion feature representation in the preprocessing stage. In the distillation stage, emotion soft labels, hidden layer representations and credibility scores are generated by the complex teacher model, and then the transfer strength of different samples and different modal knowledge to the student model is controlled by selective distillation weights. The student model uses a lightweight structure to complete the emotion category output and confidence calibration, so as to balance model usability and prediction reliability. The overall process is shown in Figure 1.

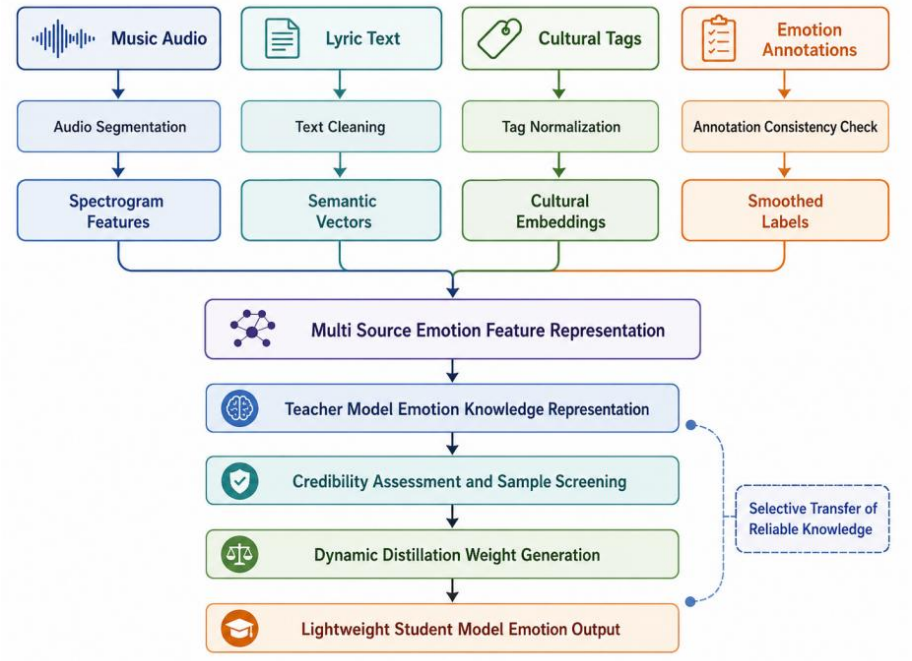

Figure 1: Modeling process of music cultural emotion driven by selective knowledge distillation

3.1 Music cultural emotion feature extraction and preprocessing

Music cultural emotion feature extraction is the pre-link of selective knowledge distillation, and its quality directly affects the stability of the knowledge representation of the teacher model. In this paper, each music piece is represented as a sample unit consisting of audio clips, lyrics text, cultural description labels and emotion annotations. The audio part mainly reflects the acoustic cues such as melody, rhythm, timbre and dynamic intensity. The lyrics part carries emotional imagery, narrative object and semantic tendency. Cultural tags are used to describe external information such as music style, regional aesthetic, singing context, traditional elements and communication scenes. There are obvious differences between the three types of information in data form and time granularity, so it is necessary to complete unified cleaning, alignment and numerical processing before entering the teacher model.

In the audio processing stage, the original music signal is uniformly resampled to 22050 Hz and segmented into segments of 30 s length. For music samples less than 30 s, the mute filling method was used to keep the input length consistent. For samples longer than 30 s, verses, choruses or segments with obvious emotional changes are selected according to the energy distribution, so as to reduce the interference of invalid prelude and long repetitive segments on emotional modeling. The audio signal is then pre-emphasized, framed, windowed and short-time Fourier transformed to generate time-spectrum representation, which is then mapped to Mel spectrogram. Compared with the original waveform, the Mel spectrogram is closer to the human ear's perception of frequency, and can retain the key low-frequency energy, rhythm fluctuation and high-frequency brightness change in music emotion recognition. In this paper, the audio features of the ith sample are expressed as follows.

$$a_i = \mathrm{CNN}(\log(\mathrm{Mel}(\mathrm{STFT}(x_i)) + \epsilon)) \tag{1}$$

where, x_i represents the audio signal of the ith music, STFT(·) represents the short-time Fourier transform, Mel(·) represents the Mel filter mapping, ϵ is the smoothing constant to prevent logarithm operation anomaly, and CNN(·) is used to extract hierarchical acoustic features in the local time-frequency structure. This formulation Bridges audio preprocessing and deep acoustic representations to provide stable inputs for subsequent distillation source models.

The lyrics processing stage adopts the steps of sentence segmentation, removing abnormal symbols, unifying case, and filtering repetitive words without emotional meaning. Considering the existence of metaphor, irony and cultural imagery in lyrics expression, this paper does not use simple word frequency statistics, but uses a pre-trained language model to generate contextual semantic vectors. For music without lyrics, empty text tags are set and style and culture tags are introduced to compensate for the lack of semantics, so as to avoid the samples without lyrics being directly eliminated in the multi-source feature space. In the cultural label processing stage, the music style, language, regional element, singing form, era style and aesthetic scene are converted into discrete coding, and then mapped into low-dimensional cultural vector through the embedding layer, so that the model can obtain cultural context cues in addition to acoustic and text.

To ensure that features from different sources are comparable on a numerical scale, we standardize the audio vector, lyrics vector and culture vector respectively, and construct a unified feature representation at the sample level:

$$z_i = [\mathrm{Norm}(a_i); \mathrm{Norm}(l_i); \mathrm{Norm}(c_i)] \tag{2}$$

where, l_i represents the semantic vector of lyrics, c_i represents the embedding vector of cultural labels, [;] Denotes the concatenation operation, and $\mathrm{Norm}(\cdot)$ denotes the normalization function. This representation is not directly used as the final classification result, but as the basic input for the teacher model to learn emotion knowledge, evaluate sample credibility, and generate distillation weights. Table 2 lists the main parameters adopted in the stage of music cultural emotion feature extraction in this paper.

Table 2: Preprocessing parameter Settings of musical cultural emotional features

Feature Source	Processing Method	Parameter or Configuration	Function Description
Audio signal	Resampling	22050 Hz	Unify the sampling frequency across different datasets
Audio clip	Fixed-length clipping	30 s	Keep the model input length consistent
Time-frequency transformation	Short-time Fourier transform	FFT size = 1024	Extract local spectral variations
Perceptual spectrogram	Mel filtering	64 Mel filters	Enhance frequency representation consistent with auditory perception
Lyric text	Semantic encoding	Pre-trained language model	Capture contextual emotional meanings in lyrics
Cultural labels	Embedding mapping	32-dimensional embedding vector	Represent genre, context, and cultural symbols
Emotion annotations	Consistency screening	Down-weight samples with annotation disagreement	Reduce interference from subjective emotion noise

3.2 Emotion knowledge representation and credibility evaluation of distillation source model

3.2.1 Multi-modal emotion knowledge representation of distilled source model

After the unified preprocessing of music audio, lyrics semantics and cultural labels, we further construct a distilled source model for generating transferable emotional knowledge. Different from the directly trained lightweight model, the distilled source model acts as a "high-capacity emotion interpreter". Its input is not a single acoustic spectrogram, but a multi-source representation composed of audio time-frequency features, lyric context semantic vectors, and cultural context embeddings. Because music cultural emotions are often affected by melody direction, lyric imagery and aesthetic background at the same time, the teacher model needs to learn complex emotional boundaries in a wider representation space, and then transfer the stable and effective knowledge to the student model.

In this paper, a multi-branch teacher network is used for emotion knowledge extraction. In the audio branch, convolutional network and temporal attention module were used to extract rhythm intensity, timbral change and melody fluctuation features. The lyrics branch uses the pre-trained language model to obtain the contextual semantic representation, which is used to recognize metaphors, emotional words and narrative mood. The cultural branch maps labels

such as music style, language, regional element, and singing form into cultural context vectors. After linear projection, the three types of features enter the shared emotion space, and the hidden layer emotion representation of the teacher model is generated by attention fusion. The calculation process can be expressed as follows.

$$h_i^T = \mathrm{Attn}(W_a a_i, W_l l_i, W_c c_i) \tag{3}$$

where, h_i^T represents the fusion emotion representation of the ith sample in the teacher model, a_i, l_i, c_i represent audio features, lyrics semantic features and cultural label features respectively, W_a, W_l, W_c are learnable projection matrices, and $\mathrm{Attn}(\cdot)$ represents the multi-source attention fusion function. This formulation emphasizes that the teacher model does not simply concatenate features, but dynamically adjusts the contribution of each modality according to the content differences of different music samples. As shown in Figure 2, the distillation source model output includes two types of knowledge: one is the probability distribution of emotion categories, which is the teacher soft label. The other category is the emotion representation of the hidden layer after fusion, which is used to describe the relative position of the samples in the continuous emotion space.

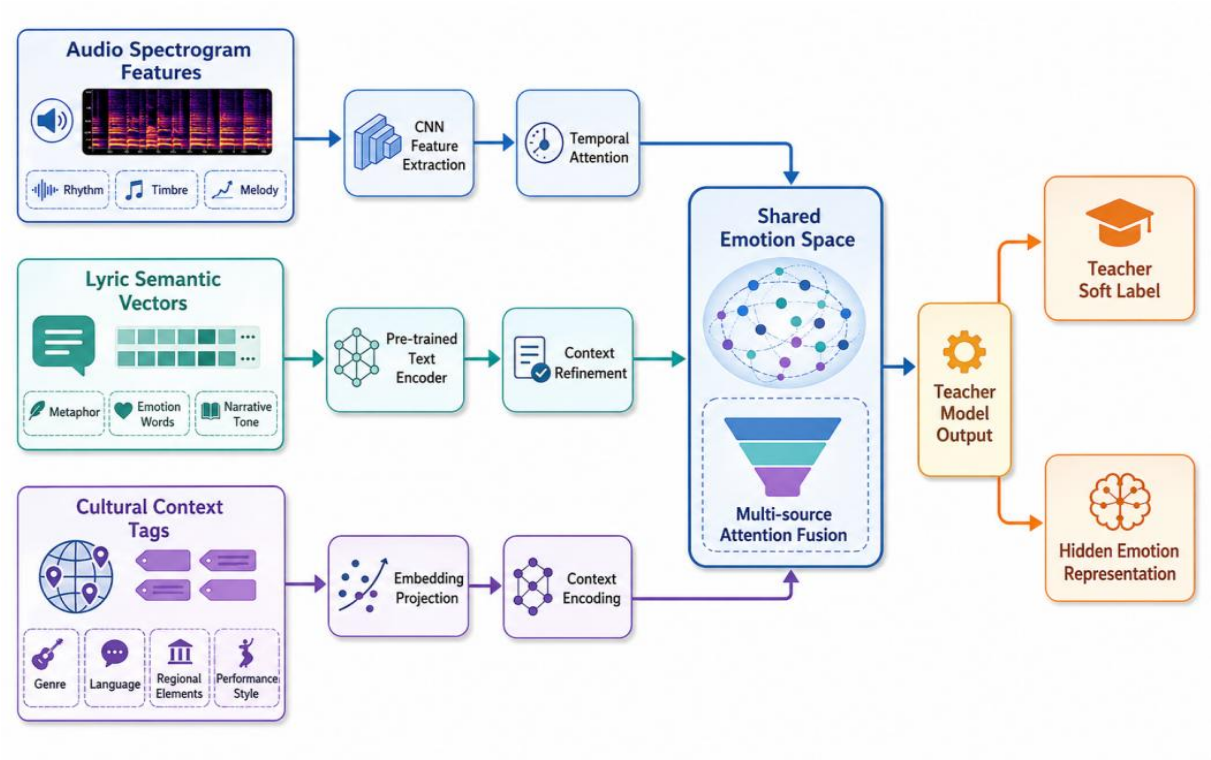

Figure 2: Multi-modal emotion knowledge representation structure of distilled source model

3.2.2 Credibility evaluation and uncertainty screening of emotional knowledge

Although the teacher model has strong representation ability, its output is not necessarily suitable for all transfers. The cultural emotion of music is obviously subjective, and the same work may obtain different emotional judgments due to different audience experience, lyrics understanding and cultural background. If all the soft labels of the teacher model are passed to the student model without distinction, low-confidence samples, cross-modal conflict samples, and labeled disagreement samples will amplify the instability of the student model. Therefore, this paper sets up a credibility evaluation link before distillation to screen and weight the

teacher knowledge of each sample.

The credibility assessment consists of three components: prediction certainty, cross-modal consistency, and annotation reliability. The prediction certainty is used to measure whether the teacher's output is concentrated. If the class probability is too scattered, it means that the model has high uncertainty for the sample. Cross-modal consistency is used to determine whether audio branches, lyrics branches and cultural branches point to similar emotional regions. The labeling reliability is based on the manual labeling consistency rate or labeling variance to judge whether the sample label is stable. The three together form the sample-level credibility score:

$$r_i = \alpha(1 - H(p_i^T)) + \beta S(a_i, l_i, c_i) + \gamma q_i \tag{4}$$

where, r_i represents the distillation credibility of the ith sample, p_i^T represents the emotion probability distribution output by the teacher model, $H(\cdot)$ represents the normalized information entropy, $S(\cdot)$ represents the multi-modal consistency function, q_i represents the reliability of manual labeling, and α, β, γ are the weight coefficients. This design enables the samples with high certainty, high modal consistency and high labeling stability to obtain stronger distillation weights, while the samples with high ambiguity, conflict or noise are suppressed.

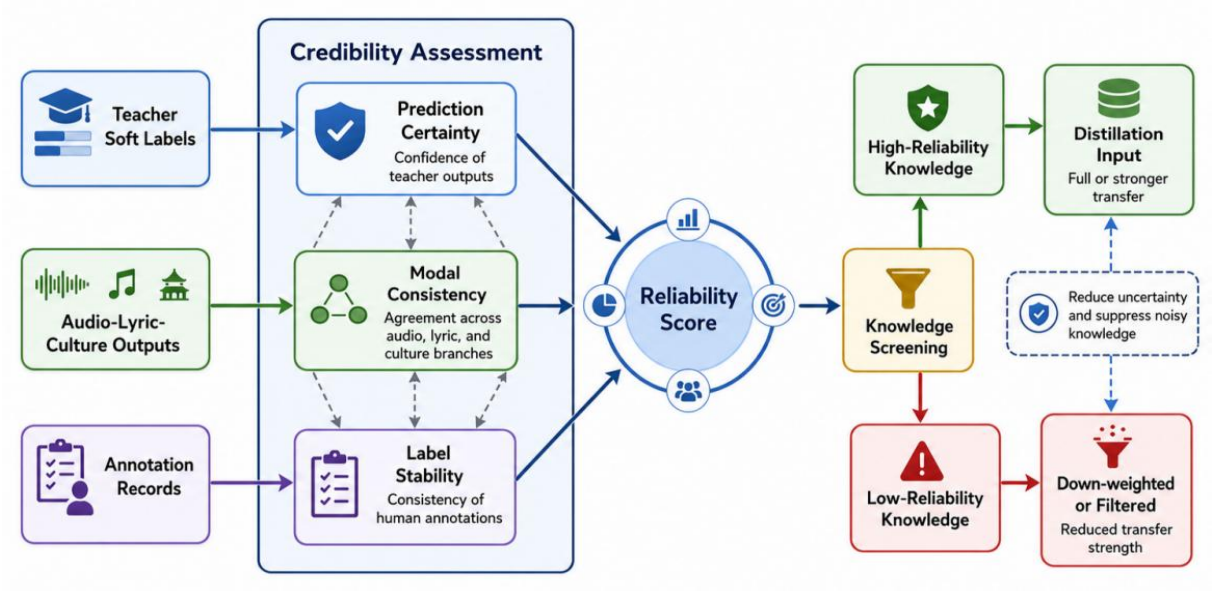

Figure 3: Evaluation and screening process of emotional knowledge credibility

Figure 3 illustrates the screening path of emotion knowledge from teacher output to distilled input. For the samples with high confidence and cross-modal consistency, the student model receives more complete representations of soft labels and hidden layers. For samples with obvious conflicts between lyrics emotion and acoustic emotion, unclear cultural label hints, or large differences in manual labeling, the model reduces its knowledge transfer intensity. Through this mechanism, the distillation process no longer simply pursues the full reproduction of the output of the teacher model, but turns to the selective absorption of effective emotional knowledge. This processing can reduce the interference of uncertain samples on the lightweight student model, and also provides a computable basis for the subsequent collaborative optimization of availability and reliability.

3.3 Dynamic generation of selective knowledge distillation weights

The core of selective knowledge distillation is not to simply compress the teacher model, but to judge which emotional knowledge is worth transferring and to what extent. In music cultural emotion samples, there are often situations in which the semantic meaning of lyrics is not completely consistent with the acoustic atmosphere, the cultural label pointing is ambiguous, and the manual labeling has large differences. If the fixed distillation coefficient is used, the student model will be forced to learn the unstable judgment results of the teacher model. Therefore, this paper introduces a dynamic weight generation mechanism after the output of teacher emotion knowledge, so that the distillation strength is jointly determined by sample credibility, modal consistency and category uncertainty.

In the specific calculation, the teacher output of the ith sample does not directly enter the loss function, but first passes through the weight mapping layer. This layer receives the credibility score r_i obtained in the previous section, the teacher prediction entropy e_i, and the consistency score s_i between the three categories of audio-lyric-culture modalities, and generates the sample-level distillation weights through a nonlinear function. Its expression is:

$$\omega_i = \sigma(\theta_1 r_i + \theta_2 s_i - \theta_3 e_i + b_\omega) \tag{5}$$

where, ω_i represents the selective distillation weight of the ith sample, which ranges from 0 to 1. $\sigma(\cdot)$ is the Sigmoid function. The $\theta_1, \theta_2, \theta_3$ are learnable parameters, and b_ω is the bias term. This formula reflects the corresponding relationship between distillation strength and knowledge quality: when the teacher model judgment is centralized, the modal expression is consistent and the labeling is reliable, ω_i will increase, and the student model will absorb more inter-category relationships in the teacher's soft label. When the samples are clearly ambiguous or the teacher output is scattered, ω_i decreases, and the model instead relies more on the true label and its own discriminative learning.

In order to avoid the sample-level weight being too coarse, this paper further introduces a class-level adjustment term. The emotional categories of music culture are not completely independent. For example, "nostalgia" and "sadness", "calm" and "warmth" often have a close distance in the acoustic and semantic space, and the boundary between "excitement" and "sadness" is relatively clear. The soft distribution of the teacher model on similar categories has certain transfer value, but the excessive diffusion on low confidence categories will weaken the discrimination boundary of the student model. Therefore, this paper generates class distillation coefficients based on teacher probability distribution and class reliable vector:

$$\lambda_{ik} = \frac{\exp\left(\omega_i p_{ik}^T u_k\right)}{\sum_{j=1}^{K} \exp\left(\omega_i p_{ij}^T u_j\right)} \tag{6}$$

where, λ_{ik} represents the distillation regulation coefficient of the ith sample on the KTH emotion, p_{ik}^T is the prediction probability of the teacher model for this category, u_k is the reliability prior of the KTH emotion, and k is the number of emotion categories. Through this processing, the model can control whether to distil at the sample level and control the distillation focus at the category level, so that the student model no longer learns all emotion probabilities on average, but preferentially absorbs structured knowledge in credible categories. Based on the above weights, the selective distillation loss is composed of soft label transfer, true label supervision, and hidden layer representation constraints.

$$\mathcal{L}_{\mathrm{skd}} = \sum_{i=1}^{N} \omega_i \, \tau^2 \mathrm{KL}\left(p_i^{T,\tau} \| p_i^{S,\tau}\right) + \sum_{i=1}^{N} (1-\omega_i)\mathrm{CE}\left(y_i, p_i^{S}\right) + \mu \sum_{i=1}^{N} \omega_i \left\| g_i^{T} - g_i^{S} \right\|_2^2 \quad (7)$$

where $\mathcal{L}_{\mathrm{skd}}$ is the total loss of selective knowledge distillation, τ is the temperature coefficient, $\mathrm{KL}(\cdot)$ represents the divergence loss, $\mathrm{CE}(\cdot)$ represents the cross-entropy loss, p_i^S is the output of the student model, g_i^T and g_i^S represent the intermediate emotion representation of the teacher model and the student model, respectively, and μ is the representation constraint coefficient. This objective function makes high-confidence samples more participate in teacher knowledge transfer, and low-confidence samples more retain hard label supervision, so as to reduce the cumulative influence of noisy soft labels in lightweight training.

With dynamic weight generation, the distillation process moves from "full imitation" to "selective absorption". For music samples with clear cultural semantics and stable acoustic emotions, the student model can fully learn the fine-grained emotional relationships of the teacher model. For samples with ironic lyrics, cross-cultural imagery deviation or large annotation differences, the model automatically reduces the distillation intensity to avoid solidification of unreliable judgments into the lightweight model. This mechanism lays the foundation for subsequent student models to strike a balance between usability and reliability.

3.4 Emotion output of student model for Usability and reliability

After completing the selective knowledge distillation weight generation, the student model needs to transform the absorbed high-confidence emotion knowledge into deployable, interpretable emotion output with stable confidence. Compared with the teacher model, the student model does not pursue a complex multi-branch deep structure, but uses a lightweight temporal encoder, a low-dimensional fusion layer, and a calibrated output layer to form a compact inference path. The goal of this design is not to simply reduce the number of parameters, but to retain the main discriminant boundaries required for music culture emotion recognition at low computational cost, so that the model can be applied to resource-constrained scenarios such as online music analysis, digital aesthetic education platform and mobile terminal emotion recommendation.

The student model receives the multi-source feature sequence after selective distillation constraints and generates a time-step level emotion representation. Since musical emotions have continuous changing characteristics, high activation in local segments does not necessarily represent the dominant emotion of the whole piece. For example, strong emotional peaks may occur in the chorus, but the prelude, interplay, and cofinale still affect the overall cultural emotional judgment. Therefore, in this paper, reliable-guided weighted pooling is used to compress the time-series sentiment representation into a global vector. The calculation process is as follows.

$$v_i^S = \sum_{t=1}^{T_i} \rho_{it} \, u_{it}^S, \quad \rho_{it} = \frac{\exp\left(w_p^T \tanh\left(W_u u_{it}^S + W_r r_i\right)\right)}{\sum_{m=1}^{T_i} \exp\left(w_p^T \tanh\left(W_u u_{im}^S + W_r r_i\right)\right)} \quad (8)$$

where, v_i^S represents the global emotion vector of the student model at the ith sample, u_{it}^S represents the hidden layer representation of the student model at the TTH time step, ρ_{it} is the time weight guided by reliability, r_i is the sample credibility score, W_u, W_r and w_p are the learnable parameters. Compared with ordinary average pooling, the proposed method is able to reduce the impact of noisy segments or segments with unclear emotional expression,

while preserving the response strength to key emotional passages.

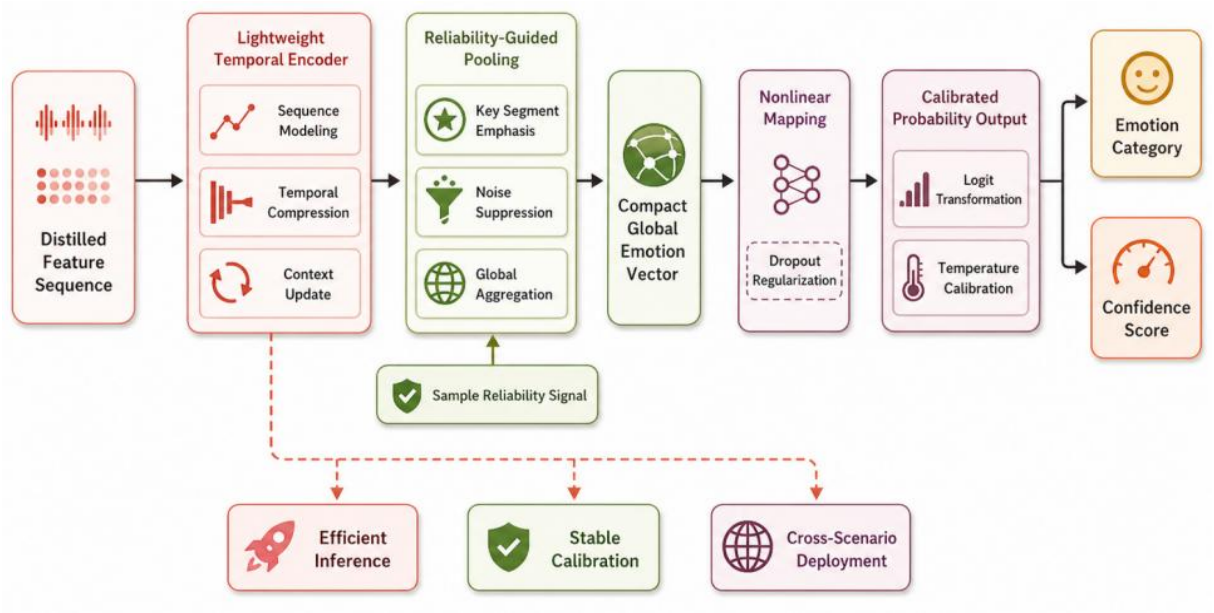

Figure 4: Emotion output process of student model for usability and reliability

As shown in Figure 4, the output path of the student model consists of lightweight coding, reliable-guided compression, nonlinear mapping, and probabilistic calibration. Reliability guided pooling is responsible for converting variable length music clips into fixed dimension vectors, nonlinear mapping layer further characterizes the boundary relationship between different emotion categories, and Dropout is used to reduce the tendency of overfitting in small sample training. This structure not only reduces the redundant computation in the inference link, but also enables the subsequent confidence calibration to obtain a more stable input basis.

In the emotion category output stage, the global emotion vector is first mapped to the category score space through the fully connected layer, and then the probability distribution is generated by temperature calibration. Music cultural sentiment is ambiguous and overlapping, and the uncalibrated Softmax output is prone to overconfidence, especially in cross-dataset testing or cultural label incomplete samples. To this end, this paper introduces calibration temperature to smooth the output distribution:

$$\hat{p}_{ik}^{S} = \frac{\exp\left(o_{ik}^{S}/\tau_c\right)}{\sum_{j=1}^{K} \exp\left(o_{ij}^{S}/\tau_c\right)} \tag{9}$$

where, $\hat{p}_{ik}^{S}$ represents the calibrated probability of the student model that the ith sample belongs to the KTH emotion class, o_{ik}^{S} is the unnormalized score of the corresponding category, τ_c is the calibrated temperature, and k is the number of emotion categories. When τ_c is appropriately increased, the model output distribution tends to be smooth, which can reduce the excessive concentration of a single category. When the emotional boundary of the sample is clear, the calibrated probability can still maintain a high discrimination.

After the above output mechanism, the student model not only gives the final emotion category, but also synchronously provides confidence information that can be used for reliability analysis. For high confidence samples, the output probability usually presents a more concentrated distribution, indicating that the student model inherits the stable emotional

knowledge from the teacher model. For samples with cross-cultural semantic ambiguity or strong acoustic-lyric conflict, the probability distribution after calibration is relatively flat, which can prompt the system to reduce the intensity of automatic decision-making. Therefore, the output of the student model is no longer limited to the classification results, but serves the recognition accuracy, reasoning efficiency, confidence calibration and cross-scenario deployment stability evaluation at the same time, and provides a unified calculation interface for the effectiveness verification of subsequent methods.

4 Evaluation of the effectiveness of the method

4.1 Experimental Data

The experimental data consists of a public music emotion dataset and supplementary cultural context annotations. In order to verify the usability and reliability of selective knowledge distillation in different music scenes, this paper selects MERGE, TROMPA-MER, Music4All-Onion and EMMA as the basic data sources, and filters out the samples with duplicate tracks, low-quality audio, samples with serious missing lyrics and samples with excessive conflict of emotion annotation. After cleaning, a total of 4,860 valid music samples were obtained, covering pop, folk, classical, electronic, rock, jazz and other styles. A 30 s representative segment of each music was uniformly captured, the sampling rate was adjusted to 22050 Hz, and it was converted into a 64-channel Mel spectrogram. After cleaning, the lyrics were input into the pre-trained language model to generate semantic vectors. Cultural labels are coded according to music style, language, singing form and aesthetic scene. The dataset is divided into training set, validation set and test set according to 7:1:2, and the same song and its different versions do not appear in the training and test stages at the same time to avoid sample leakage. Table 3 lists the basic composition of the experimental data.

Table 3: Experimental data composition and preprocessing Settings

Data Source	Valid Samples	Main Information Type	Processing Method	Experimental Use
MERGE	1,240	Audio, lyrics, emotion labels	Audio clipping, lyric cleaning, label unification	Audio–text distillation training
TROMPA-MER	960	Audio, personalized emotion annotations	Annotation consistency screening, emotion category mapping	Reliability and subjective difference validation
Music4All-Onion	1,720	Audio, metadata, cultural labels	Genre and contextual label normalization	Cultural emotion feature modeling
EMMA	940	Emotion-evoking clips, category labels	Emotion category alignment, low-quality sample removal	Cross-dataset generalization testing
Total	4,860	Multi-source music emotion information	Unified sampling, standardization, encoding mapping	Comprehensive experimental evaluation

4.2 Analysis of modeling effect of music cultural emotional characteristics

In order to test the effectiveness of music cultural emotion feature extraction and preprocessing, this paper analyzes the time-frequency response of audio and the spatial separation of multi-source features. At the audio feature level, four representative emotion samples of calm, pleasure, sadness and tension were selected, and their normalized response intensities in eight Mel bands were calculated. At the multi-source feature level, the intra-class distance, inter-class distance and contour coefficient of single audio, audio-lyrics, audio-culture label and joint audio-lyric-culture representation are compared to determine whether cultural context and lyrics semantics can improve the clarity of emotional boundaries.

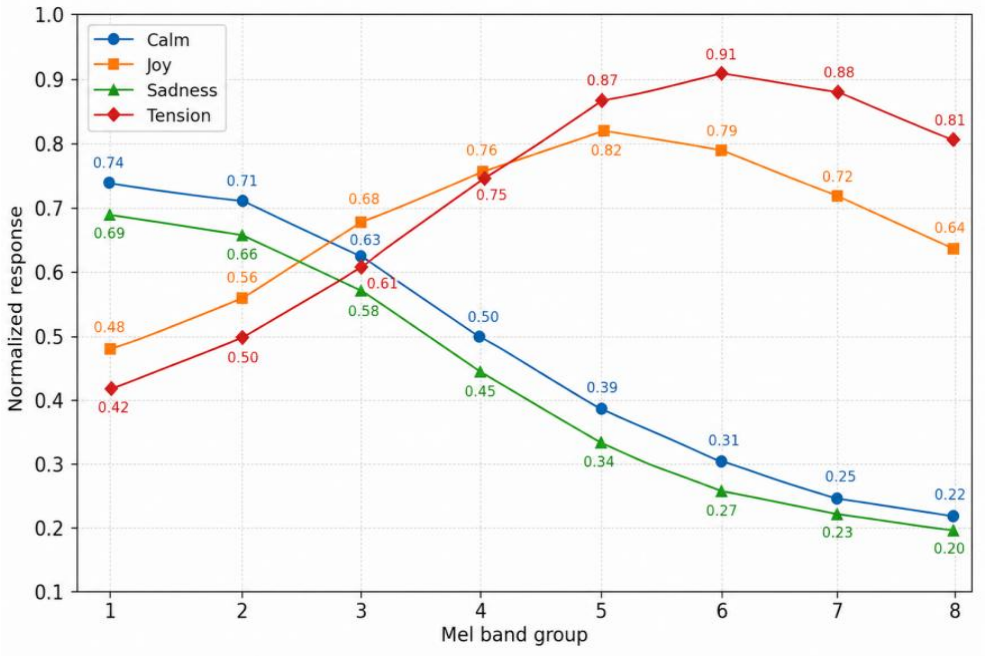

Figure 5: Mel band response characteristics of musical mood samples

As shown in Figure 5, calm samples have a high response in the low frequency band. The response value of Mel band in the first group reaches 0.74, and then gradually decreases with the increase of frequency band, indicating that this type of music mainly relies on stable low frequency, slow rhythm and weak high frequency fluctuation to form emotional atmosphere. Sad samples and calm samples have similar advantages in low frequency, but the overall response is lower, and the lowest frequency is 0.20, indicating that the acoustic expression of sad samples is more convergent, and the dynamic brightness is insufficient. The pleasant samples are significantly enhanced in the middle and high frequency bands, and the Mel band of the fifth group reaches 0.82, indicating that the rhythm activity, timbre brightness and melody upward features have strong contributions to the recognition of pleasant emotions. The Mel frequency band of the stress samples in the sixth group reaches 0.91, and the high frequency response keeps high, which reflects that the fast onset, strong dynamic change and dense spectrum structure can strengthen the separability of stress. The results show that the Mel spectrogram can better retain the frequency band differences related to auditory perception in music emotion, and provide an acoustic basis for the subsequent teacher model to extract stable emotional knowledge.

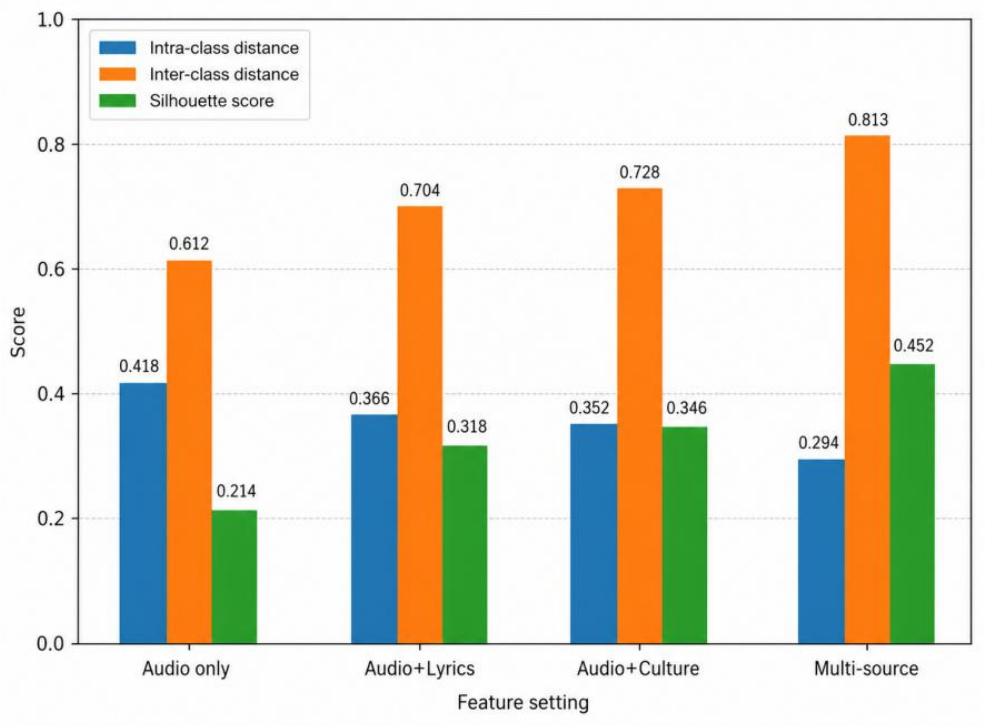

Figure 6: Comparison of emotion category separation under multi-source feature modeling

Figure 6 further demonstrates the effect of different feature combinations on the spatial structure of emotion categories. When only using audio features, the intra-class distance is 0.418, the inter-class distance is 0.612, and the silhouette coefficient is only 0.214, indicating that a single acoustic feature can capture the basic emotional differences, but it is insufficient to distinguish samples with similar cultural contexts and acoustic structures. After adding lyric semantics, the intra-class distance decreases to 0.366, and the inter-class distance increases to 0.704, indicating that lyric imagery and semantic tendency can supplement the emotional orientation that is difficult to express with acoustic features. After adding the cultural label, the inter-class distance further increases to 0.728, indicating that music style, language, singing form and aesthetic scene help to correct the cultural bias in emotional judgment. When the joint representation of audio, lyrics and culture is used, the intra-class distance is reduced to 0.294, the inter-class distance is increased to 0.813, and the contour coefficient is 0.452, which is about 111.21% higher than the single audio feature. This result shows that multi-source cultural emotion features do not simply increase the input dimension, but form complementary constraints among acoustic, semantic and cultural cues, which makes the same emotion samples more aggregated and the boundaries between different emotion samples clearer.

4.3 Emotion recognition performance comparison under different distillation strategies

In order to evaluate the influence of selective knowledge distillation on the performance of music cultural emotion recognition, this paper sets up five sets of comparison strategies: Student-only model without distillation, traditional knowledge distillation (KD), decoupled knowledge distillation (DKD), confidence weighted distillation (CKD), and the proposed selective knowledge distillation method (SKD). All models use the same training set, validation set and test set, and keep the audio, lyrics and cultural label inputs consistent to

ensure that the performance difference mainly comes from the distillation mechanism itself. The evaluation metrics include Accuracy, Macro-F1, AUROC, and Expected Calibration Error (ECE), where lower values of ECE indicate smaller deviations between model confidence and true accuracy.

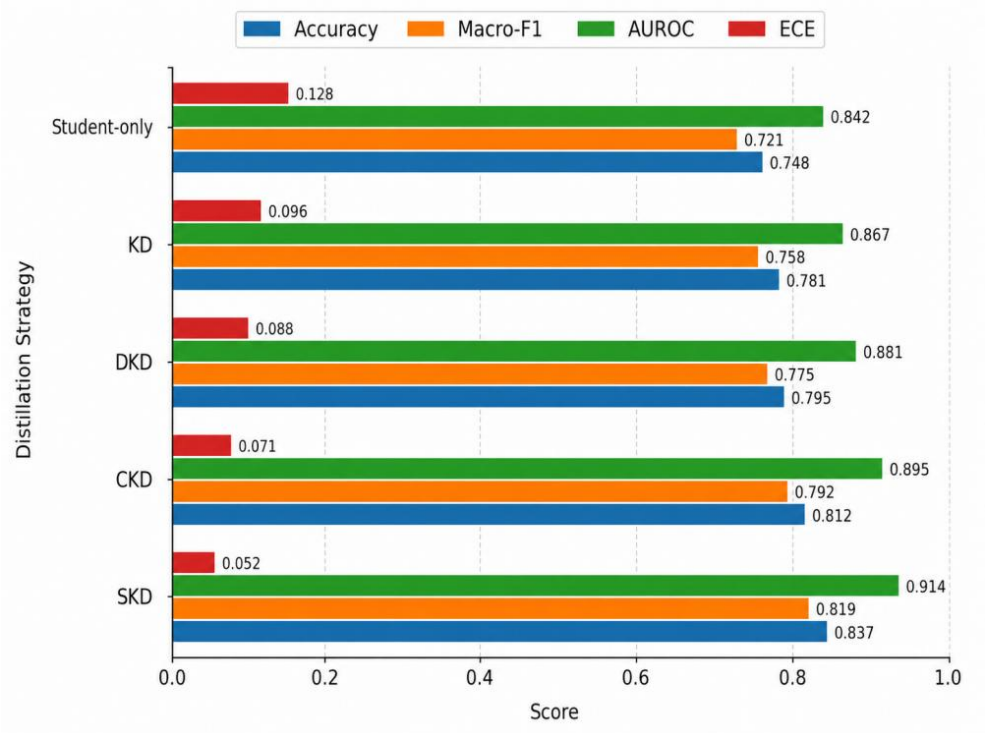

Figure 7: Emotion recognition performance comparison under different distillation strategies

Figure 7 shows that the Accuracy and Macro-F1 of the Student-only model are 0.748 and 0.721, respectively, indicating that the lightweight model has limited ability to depict the emotional boundaries of complex music culture when the teacher's knowledge guidance is lacking. The traditional KD improves the Accuracy to 0.781, but the ECE is still 0.096, indicating that the unified soft label can improve the recognition performance, but it is difficult to fully suppress the uncertain knowledge in the teacher model. DKD further improves the ability of inter-category relationship learning, and the AUROC reaches 0.881, but the adaptation to samples with semantic conflict and cultural label deviation is still not sufficient. CKD reduces the calibration error by the confidence constraint, and the ECE is reduced to 0.071, and the reliability of the model output is improved.

In this paper, the SKD method achieves the best comprehensive performance, with Accuracy, Macro-F1 and AUROC reaching 0.837, 0.819 and 0.914, respectively, and ECE decreasing to 0.052. This result shows that selective distillation does not simply increase the intensity of teacher supervision, but jointly filter transferable knowledge through sample credibility, modal consistency and category reliability, so that the student model can absorb more stable emotional boundaries and reduce the misdirection caused by low-quality soft labels. Compared with the traditional KD, the Macro-F1 of SKD is increased by 0.061, indicating that it has a more balanced recognition ability for different emotion categories. The ECE decrease of 0.044 indicates that the model can maintain a more reasonable confidence distribution while outputting the emotion categories. It can be seen that selective knowledge distillation is able to improve both recognition accuracy and prediction reliability in the lightweight student model.

4.4 Evaluation of the impact of availability constraints on the performance of lightweight models

In order to test the practical usability of selective knowledge distillation in resource-constrained scenarios, this paper constructs a lightweight deployment constraint experiment to compare the complete teacher model, the undistilled lightweight student model, the traditional knowledge distillation student model, and the selective knowledge distillation student model of this paper. The experiments are carried out on the same test set, and mainly evaluate the number of parameters, inference delay, video memory occupation, Accuracy and Macro-F1. The inference delay is calculated as the average processing time of a single 30 s music clip, and the video memory occupancy is taken as the peak record during batch inference. This setting is used to simulate the application environment which is sensitive to response speed and computing resources, such as online music emotion analysis, digital aesthetic education platform and mobile terminal music recommendation.

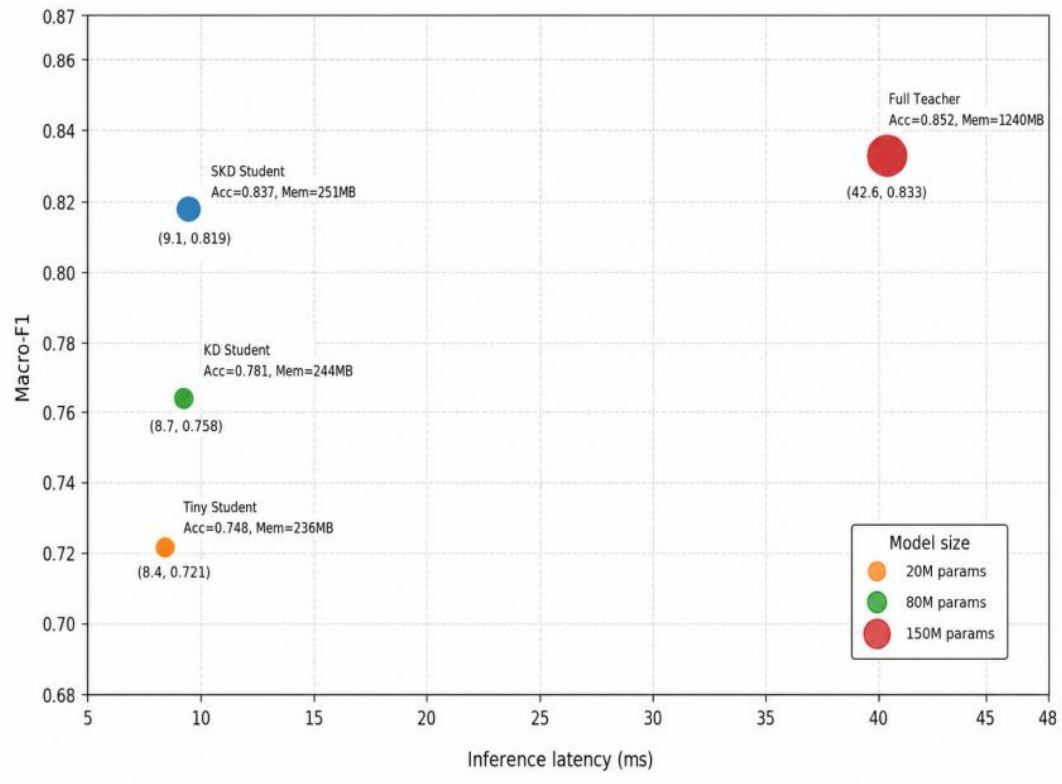

Figure 8: Comparison of model performance and resource consumption under availability constraints

Figure 8 shows that the complete teacher model has high recognition performance, Accuracy and Macro-F1 reach 0.852 and 0.833 respectively, but its parameter number is 148.0M, inference delay is 42.6 ms, and video memory consumption is 1240 MB, which is difficult to meet the requirements of lightweight deployment. The parameter quantity of the undistilled student model is reduced to 18.5M, the inference delay is shortened to 8.4 ms, and the video memory footprint is only 236 MB, but the Accuracy and Macro-F1 are reduced to 0.748 and 0.721, respectively, indicating that the simple compression model structure will weaken its ability to express the emotional boundary of music culture.

The traditional knowledge distillation student model improves the Accuracy to 0.781 and Macro-F1 to 0.758 while the resource consumption is basically unchanged, indicating that

teacher soft labels can help the lightweight model recover part of the emotion recognition ability. However, this method still does not distinguish the teacher knowledge quality, and the emotional noise in the low-confidence samples may enter the student model with the distillation process. In contrast, the parameter quantity of the SKD student model in this paper is 19.2M, the inference delay is 9.1 ms, and the video memory occupation is 251 MB, which is close to that of the ordinary student model, but the Accuracy and Macro-F1 reach 0.837 and 0.819, respectively, which are close to the level of the full teacher model.

4.5 Validation of generalization ability across datasets under reliability conditions

In order to verify the reliability of the model under changes in data distribution, this paper uses zero-shot transfer experiments across data sets. In the experiment, the MERGE dataset was used as the training domain, and the trained model was directly deployed to the TROMPA-MER, EMMA and Music4All-Onion test subsets without fine-tuning and parameter retraining. The three external test domains are different in music style, emotion annotation method, cultural context label completeness and lyrics availability, which can better simulate the generalization problems caused by music style changes, audience subjective differences and cultural background deviation in practical applications. AUROC, Cohen's Kappa and ECE were selected as evaluation indexes, where AUROC reflected the overall discrimination ability of the model under multi-threshold conditions, Kappa was used to measure the consistency between the model prediction and manual labeling, and ECE was used to evaluate the deviation between the prediction confidence and the actual accuracy.

Table 4: Validation of generalization ability across datasets under reliability conditions

Model	Test Domain	AUROC	Kappa	ECE
Student-only	TROMPA-MER	0.781	0.342	0.139
	EMMA	0.754	0.318	0.151
	Music4All-Onion	0.736	0.291	0.164
KD Student	TROMPA-MER	0.816	0.386	0.108
	EMMA	0.792	0.354	0.119
	Music4All-Onion	0.771	0.327	0.132
CKD Student	TROMPA-MER	0.842	0.421	0.083
	EMMA	0.821	0.392	0.091
	Music4All-Onion	0.804	0.368	0.104
SKD Student	TROMPA-MER	0.876	0.468	0.061
	EMMA	0.858	0.443	0.067
	Music4All-Onion	0.837	0.416	0.074

As can be seen from Table 4, in the three external test domains, the performance of the undistilled student model decreases significantly, especially in the Music4All-Onion test subset, the AUROC is only 0.736 and Kappa is 0.291, indicating that the lightweight model is difficult to adapt to more complex cultural labels and music style changes when the lack of teacher knowledge constraints. The traditional KD Student model is improved compared with Student-Only, but its ECE remains between 0.108 and 0.132, indicating that although unified soft label can improve the class discrimination ability, it cannot fully correct the confidence bias under the condition of cross-dataset. CKD reduces the output overconfidence problem by the confidence constraint, and the ECE is reduced to 0.091 on the EMMA test domain, and the prediction reliability is enhanced.

The SKD student model in this paper achieves the best results on three sets of external data. Its AUROC on TROMPA-MER reaches 0.876 and Kappa reaches 0.468, indicating that the model has strong adaptability to personalized emotion labeling. On EMMA, the AUROC is 0.858 and Kappa is 0.443, indicating that the model can stably transfer to the emot-induced scene. On Music4All-Onion, AUROC remains 0.837 and ECE is controlled at 0.074, although affected by the complexity of music style and cultural label. This result shows that selective knowledge distillation can make the student model maintain a more stable emotional boundary and a more reasonable confidence distribution in the external data domain by screening out low-credible teacher output and strengthening the transfer weight of cross-modal consistent samples. Cross-dataset experiments further show that the reliability of the proposed method does not depend on the fitting effect within a single dataset, but reflects the transferable judgment ability in the face of changes in music cultural context.

4.6 Subjective consistency and confidence calibration assessment

In order to further test the reliability of the model under the condition of insufficient subjective emotion annotation, this paper designs a low-labeled resource experiment, using 30%, 50% and 70% manually labeled samples to participate in the training, respectively, and keeping the complete test set unchanged. Music cultural emotion has strong subjectivity, and the emotional judgment of different annotators on the same music segment may be affected by aesthetic experience, lyrics understanding and cultural background. Therefore, it is difficult to fully reflect the output quality of the model by relying on the classification accuracy alone. In this paper, Consistency correlation coefficient (CCC) is introduced to measure the statistical consistency between model emotion prediction and manual annotation, Inter-annotator Consistency ratio (ICR) is introduced as the reference benchmark for human subjective consistency, and Expected Calibration Error (ECE) is used to evaluate the deviation between model confidence and true accuracy.

Table 5: Assessment of subjective consistency and confidence calibration

Model	Annotation Ratio	CCC	ICR	ECE
Student-only	30%	0.518	0.602	0.157
	50%	0.591	0.602	0.139
	70%	0.646	0.602	0.124
KD Student	30%	0.563	0.602	0.121
	50%	0.634	0.602	0.104
	70%	0.691	0.602	0.093
CKD Student	30%	0.602	0.602	0.089
	50%	0.681	0.602	0.076
	70%	0.728	0.602	0.068
SKD Student	30%	0.657	0.602	0.063
	50%	0.731	0.602	0.054
	70%	0.782	0.602	0.047

As can be seen from Table 5, the CCC of each model shows an upward trend as the labeling ratio increases from 30% to 70%, indicating that more manual labeling can enhance the fitting ability of the model to the distribution of subjective emotions. The CCC of the undistilled student model under 30% labeling condition is only 0.518, which is lower than the consensus reference value of 0.602 for manual annotators, indicating that the lightweight model is difficult to stably learn the music cultural emotional boundary under low resource

conditions. The CCC of the traditional KD student model is increased to 0.691 under the 70% labeling condition, indicating that the teacher's soft label can supplement part of the emotional structure information, but its ECE is still 0.093, which has a certain overconfidence problem.

The CKD student model improves the calibration effect by the confidence constraint, and the ECE is reduced to 0.076 under the 50% labeling ratio, which indicates that the confidence information can suppress the interference of uncertain samples on the output distribution. In this paper, the SKD student model achieves better results under three labeling ratios. When the labeling ratio is 30%, the CCC reaches 0.657, which is higher than the ICR reference value. When the labeling ratio increases to 70%, CCC reaches 0.782 and ECE drops to 0.047. The results show that selective knowledge distillation can preferentially transfer high-credible teacher knowledge when manual labeling is insufficient, so that the student model can obtain the emotion output closer to human subjective judgment, and maintain a reasonable confidence distribution. Overall, SKD not only improves the consistency between music cultural emotion prediction and manual annotation, but also enhances the calibration reliability of the model in low-annotated resource scenarios.

5 Discussion

The advantage of the proposed method does not simply come from the compression transfer of the teacher model to the student model, but from the screening and adjustment of the quality of emotional knowledge. Traditional distillation methods usually regard teacher outputs as stable soft labels, which are easy to transfer cultural semantic deviations, labeling disagreements, and cross-modal conflicts to the student model. In this paper, the dynamic weights are constructed by credibility score, modal consistency and category reliability, so that the high-confidence samples play a stronger role in distillation, and the uncertain samples rely more on the true label constraints. This mechanism explains why the Accuracy of SKD student model reaches 0.837, Macro-F1 reaches 0.819 and AUROC reaches 0.914 on the test set, and its performance is significantly better than 0.781, 0.758 and 0.867 of the traditional KD model. This indicates that selective transfer can help the lightweight model retain more stable emotion discrimination boundaries.

From the perspective of usability, although the complete teacher model achieves an Accuracy of 0.852 and a Macro-F1 of 0.833, the parameter number reaches 148.0M and the inference delay is 42.6 ms, which is difficult to adapt to real-time music analysis and mobile terminal deployment scenarios. The parameter quantity of the SKD student model is only 19.2M, the inference delay is controlled at 9.1 ms, and the video memory occupation is 251 MB, but the Macro-F1 is maintained at 0.819, which is close to the level of the teacher model. The results show that selective distillation does not rely on expanding the model scale to improve the performance, but improves the expression efficiency of the student model by compressing high-value emotional knowledge, so that the model achieves a reasonable balance between recognition accuracy and deployment cost.

From the perspective of reliability, music cultural sentiment is affected by lyric imagery, style tradition, aesthetic experience and annotation subjectivity, which is prone to confidence distortion in cross-dataset transfer. The AUROC of the proposed method on the three external test domains of TROMPA-MER, EMMA and Music4All-Onion reached 0.876, 0.858 and 0.837, respectively, and the ECE decreased to 0.061, 0.067 and 0.074, respectively. It shows that it can still maintain a relatively stable output under the change of cultural context and the difference of annotation system. In the low-labeling experiment, the CCC of SKD reached 0.657 under 30% labeling ratio, which was higher than the manual consistency reference value of 0.602, which further indicated that the high-credible teacher's knowledge could

supplement the emotional structure information when the annotation was insufficient. However, the proposed method is still affected by the quality of cultural labels, the availability of lyrics, and the consistency of the emotion category system. For lyricless music, cross-lingual lyrics, or works with highly implicit cultural semantics, existing tag embeddings may still be insufficient to fully express the emotional context. Subsequent research can further introduce multilingual semantic models, user personalized feedback, and online calibration mechanisms to make music cultural emotion modeling remain usable, stable, and interpretable in more complex application environments.

6 Conclusions

In this paper, a selective knowledge distillation method is proposed to solve the problems of insufficient model availability, unbalanced distillation knowledge reliability and unstable cross-context emotion judgment in music cultural emotion modeling. The method takes audio spectrogram, lyrics semantics and cultural labels as multi-source inputs, generates emotional soft labels and hidden layer representations through the teacher model, and introduces credibility score, modal consistency and category reliability to dynamically adjust the distillation strength, so that the student model can preferentially absorb high-value emotional knowledge and reduce the interference of low-credible soft labels on the decision boundary. The experimental results show that the SKD student model in this paper achieves 0.837 Accuracy, 0.819 Macro-F1 and 0.914 AUROC in the comprehensive test, which are better than the traditional KD and confidence weighted distillation methods. Its ECE decreases to 0.052, indicating a higher consistency between the confidence of the model output and the true recognition results. In the usability evaluation, the parameter quantity of the SKD student model is 19.2M, the inference delay is 9.1 ms, and the video memory occupation is 251 MB, which maintains the recognition performance close to that of the teacher model under the condition of close to the resource consumption of the lightweight model. Cross-dataset experiments further show that the AUROC of the proposed method on TROMPA-MER, EMMA and Music4All-Onion reaches 0.876, 0.858 and 0.837, respectively, showing good transfer stability. Under the condition of low annotation, the CCC of SKD reaches 0.657 at 30% annotation ratio, which is higher than the manual consistency reference value of 0.602, which verifies its adaptability to the scene of insufficient subjective emotion annotation. The results show that selective knowledge distillation can coordinate model compression, emotion discrimination and confidence calibration in music culture emotion recognition, and provide a deployable technical path for intelligent music recommendation, digital aesthetic education, music psychological auxiliary analysis and human-computer emotion interaction.

Author's Profile

Yang Liu was born in Hexian, Anhui, P.R. China in 1995. He received his DMA degree from the St. Petersburg State Conservatory of Music in Russia. He works as a teacher at Jiangsu Maritime Institute, and his research focuses on music aesthetic education and vocal performance.

Wang Hao graduated from Wuhan Conservatory of Music with a master's degree and now works at Nanjing Normal University of Special Education. His research interests are aesthetic education and Chinese music history.

Yukai Liu was born in Anhui, China, in 2001. He received his bachelor’s degree from South-Central Minzu University, China. He is currently pursuing a master’s degree at the

School of Computer Science and Engineering, The University of New South Wales (UNSW Sydney). My main research direction is machine learning and natural language processing.

Funding

This work was supported by “Music Aesthetic Education in Higher Education Driven by New Quality Productive Forces” (Project No. 2024BSKY07).

References

[1] Louro P L, Redinho H, Malheiro R, et al. A comparison study of deep learning methodologies for music emotion recognition[J]. Sensors, 2024, 24(7): 2201.

[2] Lima Louro P, Redinho H, Santos R, et al. MERGE--A Bimodal Audio-Lyrics Dataset for Static Music Emotion Recognition[J]. arXiv e-prints, 2024: arXiv: 2407.06060.

[3] Gómez-Cañón J S, Gutiérrez-Páez N, Porcaro L, et al. TROMPA-MER: an open dataset for personalized music emotion recognition[J]. Journal of Intelligent Information Systems, 2023, 60(2): 549-570.

[4] Moscati M, Parada-Cabaleiro E, Deldjoo Y, et al. Music4All-Onion--A Large-Scale Multi-faceted Content-Centric Music Recommendation Dataset[C]//Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022: 4339-4343.

[5] Strauss H, Vigl J, Jacobsen P O, et al. The Emotion-to-Music Mapping Atlas (EMMA): A systematically organized online database of emotionally evocative music excerpts[J]. Behavior Research Methods, 2024, 56(4): 3560-3577.

[6] Turchet L, Pauwels J. Music emotion recognition: intention of composers-performers versus perception of musicians, non-musicians, and listening machines[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 30: 305-316.

[7] Turchet L, O'Sullivan B, Ortner R, et al. Emotion recognition of playing musicians from EEG, ECG, and acoustic signals[J]. IEEE Transactions on Human-Machine Systems, 2024, 54(5): 619-629.

[8] Aslam M H, Zeeshan M O, Pedersoli M, et al. Privileged knowledge distillation for dimensional emotion recognition in the wild[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 3338-3347.

[9] Brown N, Williamson A, Anderson T, et al. Efficient transformer knowledge distillation: A performance review[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track. 2023: 54-65.

[10] Malihi L, Heidemann G. Efficient and controllable model compression through sequential knowledge distillation and pruning[J]. Big Data and Cognitive Computing, 2023, 7(3): 154.

[11] Dantas P V, Sabino da Silva Jr W, Cordeiro L C, et al. A comprehensive review of model

compression techniques in machine learning: PV Dantas et al[J]. Applied Intelligence, 2024, 54(22): 11804-11844.

[12] Kaur D, Uslu S, Rittichier K J, et al. Trustworthy artificial intelligence: a review[J]. ACM computing surveys (CSUR), 2022, 55(2): 1-38.

[13] Radclyffe C, Ribeiro M, Wortham R H. The assessment list for trustworthy artificial intelligence: A review and recommendations[J]. Frontiers in artificial intelligence, 2023, 6: 1020592.

[14] Haque A K M B, Islam A K M N, Mikalef P. Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research[J]. Technological Forecasting and Social Change, 2023, 186: 122120.

[15] Vashistha R, Farahi A. U-trustworthy models. reliability, competence, and confidence in decision-making[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(18): 19956-19964.

[16] Fakour F, Mosleh A, Ramezani R. A structured review of literature on uncertainty in machine learning & deep learning[J]. arXiv preprint arXiv:2406.00332, 2024.

[17] Bogdanov D, Lizarraga Seijas X, Alonso-Jiménez P, et al. MusAV: A dataset of relative arousal-valence annotations for validation of audio models[J]. 2022.

[18] Hashem A, Arif M, Alghamdi M. Speech emotion recognition approaches: A systematic review[J]. Speech Communication, 2023, 154: 102974.

[19] Parchami-Araghi A, Böhle M, Rao S, et al. Good teachers explain: Explanation-enhanced knowledge distillation[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 293-310.

[20] Hebbalaguppe R, Baranwal M, Anand K, et al. Calibration transfer via knowledge distillation[C]//Proceedings of the Asian Conference on Computer Vision. 2024: 513-530.