mathematics

MDPI

Article

MELT: Optimization-Driven Music Emotion Learning with Temporal Token-Level Fusion

Yihe Yin [1], Zhen Tian [2] and Junming Chen [3,*]

[1] School of Creative Media, City University of Hong Kong, Hong Kong 999077, China; yiheyin2-c@my.cityu.edu.hk
[2] James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, UK; 2620920z@student.gla.ac.uk
[3] Faculty of Humanities and Arts, Macau University of Science and Technology, Macao 999078, China
* Correspondence: jmchen@must.edu.mo

Abstract

Music emotion recognition (MER) can be formulated as a multimodal optimization problem that predicts an emotion label from coupled audio and lyric sequences. Existing methods typically perform unimodal learning or coarse global fusion, which overlooks fine-grained temporal-token correspondences between musical dynamics and lyric semantics. We propose MELT (**M**usic **E**motion **L**earning with **T**emporal token-level fusion), an optimization-driven framework with four modules: a BERT-based lyrics semantic encoder (LSE), a segment temporal encoder (STE) that models audio-segment dependencies via a Transformer, a token-level temporal fusion (TTF) module with gated cross-attention, and an emotion mood head (EMH) for four-class prediction. Training is conducted end-to-end by jointly minimizing a supervised classification term and an auxiliary cross-modal contrastive alignment term, yielding a unified objective that improves both class separability and representation consistency. On the MoodyLyrics benchmark, MELT achieves 87.6% weighted F1 for four-class emotion recognition (angry, happy, relaxed, sad), outperforming unimodal baselines and representative early/late fusion strategies. Ablation results further verify that temporal encoding, gated token-level fusion, and joint optimization each contribute to the final performance.

Keywords: music emotion recognition; multimodal optimization; joint objective learning; token-level cross-attention; contrastive alignment

MSC: 68T07; 13P25; 49N30

Check for updates

Academic Editors: Guorong Chen, Xiao Huang and Yingxu Wang

Received: 15 April 2026
Revised: 11 May 2026
Accepted: 12 May 2026
Published: 15 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Music is a powerful medium for emotional expression, and the automatic recognition of emotions conveyed by music has broad applications in music recommendation, playlist generation, music therapy, and affective computing [1,2]. Music emotion recognition (MER) seeks to assign emotion labels to songs, typically drawing on categorical models such as Russell's circumplex model of affect [3,4] or discrete emotion taxonomies.

Two primary modalities carry emotional information in music: the audio signal and the lyrics. Audio captures tonal, rhythmic, and timbrel cues that shape the listener's emotional perception, while lyrics convey semantic and narrative content that can reinforce, complement, or even contradict the audio-level emotion [5]. Consequently, multimodal

approaches that integrate both audio and lyrics have shown consistent improvements over unimodal methods [6].

However, most existing multimodal MER systems fuse audio and text representations at a coarse, global level—for example, by concatenating a single audio embedding with a single text embedding before classification [7]. This approach discards the temporal structure inherent in both modalities: a song's emotional trajectory evolves over time, and different segments of the audio may correspond to different portions of the lyrics. Early fusion by simple concatenation cannot capture these fine-grained cross-modal interactions.

Recent advances in pre-trained models have produced powerful unimodal representations. On the text side, BERT [8] and its variants provide contextualized word embeddings that capture rich semantic information [9]. On the audio side, self-supervised models such as MERT [10] and MuQ [11] learn general-purpose music representations from large-scale unlabeled audio. These pre-trained encoders offer strong foundations, but the question of how to effectively fuse their outputs for emotion recognition remains open.

We propose MELT (**M**usic **E**motion **L**earning with **T**emporal token-level fusion), a multimodal framework that addresses these limitations through four key components. The **lyrics semantic encoder (LSE)** leverages BERT to produce contextualized token-level representations of song lyrics. The **segment temporal encoder (STE)** segments the audio into fixed-length windows, encodes each segment through a pre-trained audio backbone, and models temporal dependencies across segments using a Transformer encoder with learnable positional embeddings. The **token-level temporal fusion (TTF)** module performs cross-attention from lyric tokens to audio segments, enabling each word in the lyrics to attend to relevant temporal positions in the audio, with a learned gating mechanism that controls the degree of audio influence on each token. Finally, the **emotion mood head (EMH)** aggregates the fused representation through multi-view pooling and classifies the result into emotion categories.

The main contributions of this work are as follows:

1. We introduce MELT, a multimodal MER framework that performs token-level fusion between lyrics and temporally encoded audio segments, enabling fine-grained cross-modal interaction.
2. We design a gated cross-attention mechanism that adaptively controls the flow of audio information into each lyric token, preserving the semantic integrity of the text while enriching it with acoustic context.
3. We conduct experiments on the MoodyLyrics dataset with four emotion classes, demonstrating that MELT outperforms unimodal baselines and conventional fusion strategies. Ablation studies validate the contribution of each component.

The remainder of this paper is organized as follows. Section 2 reviews related work on music emotion recognition, multimodal fusion strategies, and pre-trained audio and text models. Section 3 presents the detailed architecture of MELT. Section 4 reports experimental results, ablation studies, and analyses. Finally, Section 5 discusses the findings, and Section 6 concludes the paper.

2. Related Work

2.1. Music Emotion Recognition

Music emotion recognition (MER) has been a central topic in music information retrieval (MIR) for over two decades. The task is typically formulated either as a classification problem over discrete emotion categories or as a regression problem over continuous affective dimensions. The most widely adopted emotion representation frameworks include Russell's circumplex model [3], which organizes emotions along valence (positive–negative) and arousal (high–low) axes, Thayer's model of mood [12], and discrete categorical tax-

onomies inspired by Ekman's basic emotions [13]. Eerola and Vuoskoski [14] provided a systematic comparison of these frameworks in the context of music, concluding that both dimensional and categorical models capture meaningful aspects of musical emotion, though their suitability depends on the granularity of the target application.

Early MER systems relied on hand-crafted audio features—including tempo, mode, spectral centroid, spectral flux, zero-crossing rate, and Mel-frequency cepstral coefficients (MFCCs)—combined with traditional machine learning classifiers such as support vector machines (SVMs), random forests, and Gaussian mixture models [1,2]. On the lyrics side, bag-of-words representations, TF-IDF vectors, and sentiment lexicons such as ANEW and SentiWordNet were commonly used to extract emotional cues from text [5,15]. These feature-engineering approaches achieved reasonable performance but were limited by the expressiveness of the manually designed feature spaces.

The advent of deep learning brought significant improvements. Convolutional neural networks (CNNs) were applied to mel-spectrograms for audio-based emotion recognition, with architectures such as the sample-level CNN proposed by Choi et al. [16] demonstrating that learned features can outperform hand-crafted ones. Recurrent neural networks (RNNs) and their variants (LSTMs, GRUs) were adopted to model the sequential nature of both audio and text [17]. More recently, Transformer-based architectures have been applied to music understanding tasks, with Won et al. [18] showing that self-supervised pre-training followed by fine-tuning yields strong performance across multiple MIR benchmarks.

For lyrics-based emotion recognition, the transition from bag-of-words to distributed word representations (Word2Vec, GloVe) and subsequently to contextualized embeddings (ELMo, BERT) has progressively improved performance [8]. The MoodyLyrics dataset [19] provided a standardized benchmark for lyrics-based emotion classification, enabling systematic comparison across methods.

Multimodal approaches that combine audio and lyrics have consistently outperformed unimodal methods. Hu and Downie [5] were among the first to demonstrate the complementary nature of audio and lyrics for mood classification. Delbouys et al. [6] showed that late fusion of pre-trained audio and text features improves valence–arousal prediction. Laurier et al. [7] proposed a multimodal architecture that combines audio features and lyrics for mood classification, demonstrating that the two modalities provide complementary cues. However, the majority of these multimodal systems employ simple fusion strategies—concatenation, averaging, or late fusion of logits—that do not model fine-grained cross-modal interactions. This limitation motivates our work on token-level temporal fusion.

2.2. Multimodal Fusion Strategies

Multimodal fusion in deep learning has been extensively studied across domains including vision–language, audio–visual, and audio–text understanding. Beyond these classical settings, multi-source data fusion has also demonstrated strong predictive power in financial risk forecasting, where heterogeneous signals are jointly modeled to capture early warning patterns [20]. Baltrušaitis et al. [21] provided a comprehensive taxonomy that categorizes fusion approaches into early (feature-level), late (decision-level), and hybrid strategies.

Early fusion concatenates or combines raw features from different modalities before feeding them into a shared encoder. While this allows the model to learn cross-modal interactions from the lowest level, it often suffers from the curse of dimensionality and modality dominance, where one modality overwhelms the other during training [22].

Late fusion trains separate unimodal encoders and combines their outputs (logits, probabilities, or high-level features) at the decision level. This preserves modality-specific information and is robust to missing modalities but limits cross-modal interaction to the final stage, preventing the model from learning fine-grained correspondences between modalities.

Tensor-based fusion methods, such as the tensor fusion network (TFN) [23], compute the outer product of unimodal representations to model all possible bimodal and trimodal interactions. While expressive, the computational cost grows exponentially with the number of modalities and feature dimensions, making it impractical for high-dimensional representations from pre-trained models.

Attention-based fusion has emerged as the most flexible and effective paradigm. Cross-attention mechanisms, introduced in the Transformer architecture [24], allow one modality to query another, enabling selective and asymmetric information exchange. The Multimodal Transformer (MulT) [25] extends this idea by using directional pairwise cross-attention between all modality pairs, demonstrating strong performance on sentiment analysis and emotion recognition tasks. In the vision–language domain, ViLBERT [26] and Oscar [27] employ cross-attention between visual regions and text tokens, achieving state-of-the-art results on visual question answering and image captioning. For audio–visual fusion, Nagrani et al. [28] proposed attention bottlenecks that restrict cross-modal information flow through a small set of latent tokens, improving both efficiency and performance.

In the MER context, attention-based fusion between lyrics and audio at the token level remains largely unexplored. Most existing multimodal MER systems operate at the global level, computing a single embedding per modality before fusion. Our proposed TTF module addresses this gap by enabling each lyric token to attend to temporally encoded audio segments, capturing fine-grained cross-modal correspondences that global-level fusion cannot represent.

Because the closest prior art to MELT is the family of cross-attention-based multimodal fusion methods (MulT [25] and MAG-BERT [29] in particular), we make the structural differences explicit here rather than claiming a generic "new cross-attention mechanism". MulT applies bidirectional, pairwise cross-attention between every modality pair over raw per-frame/per-token sequences; it is agnostic to whether the audio stream has been temporally grouped into musically meaningful units, and it does not explicitly model segment-level audio context. MAG-BERT injects a gated acoustic/visual vector into each BERT token via a low-rank alignment gate, but the acoustic input is a single per-word aggregated vector rather than a distribution over time-structured audio units. MELT differs from these methods in three concrete and verifiable ways: (i) the cross-attention is deliberately asymmetric—lyric tokens act as queries and temporally encoded audio segments act as keys/values, whereas MulT is symmetric and MAG-BERT has no explicit audio-side attention; (ii) the audio side is first passed through a dedicated segment temporal encoder with learnable positional embeddings, so the keys/values carry cross-segment temporal context (e.g., the arc from verse to chorus), unlike the per-frame or per-word vectors used by MulT/MAG-BERT; and (iii) the gate in MELT is a token $\times$ feature-level convex combination (Equation (11)) over the original text and audio-attended pathways, specifically designed so that the fusion degrades gracefully to a text-only module when audio is uninformative, rather than always adding an acoustic residual as in MAG-BERT. Table 1 summarizes these differences along the axes that are most relevant for MER.

Table 1. Structural comparison of MELT with representative attention-based multimodal fusion methods on axes that are relevant for MER. "Token×dim" denotes gating at the granularity of individual token positions and feature dimensions. "Music-audio prior" indicates whether the audio-side input is prepared by a music-specific pre-trained model (as opposed to a generic speech/audio or raw-frame input).

Method	Direction	Granularity	Audio Temporal Enc.	Gate	Music-Audio Prior
MulT [25]	Pairwise bidirectional	Frame↔token	No (raw frames)	No	No
MAG-BERT † [29]	Acoustic → text	Per-word vector	No (aggregated)	Scalar per token	No
MISA [30]	Shared/private split	Global vector	No	No	No
ConFEDE [31]	Contrastive decomp.	Global vector	No	No	No
MELT (Ours)	**Text → audio (asym.)**	**Token → segment**	**Yes (STE)**	**Token × dim convex**	**Yes (MERT)**

† MAG-BERT was originally designed for jointly injecting acoustic and visual modalities into BERT; "Acoustic → text" here refers to the audio-only variant that is directly applicable to the MER setting evaluated in this paper. Bold indicates the proposed method and its distinguishing design choices.

2.3. Pre-Trained Audio and Text Models

The success of self-supervised pre-training in NLP has inspired analogous approaches for audio and music understanding. On the text side, BERT [8] introduced bidirectional masked language modeling, producing contextualized token representations that have become the de facto standard for text encoding. RoBERTa [32] improved upon BERT through more robust training procedures, including dynamic masking and larger batch sizes. These models provide rich semantic representations of lyrics that capture context-dependent word meanings, which is critical for emotion recognition where the same word can carry different emotional connotations depending on context.

On the audio side, self-supervised learning has produced a family of powerful general-purpose audio encoders. wav2vec 2.0 [33] learns speech representations through contrastive learning over quantized latent representations, achieving strong performance on speech recognition with limited labeled data. HuBERT [34] extends this approach by using an offline clustering step to generate pseudo-labels for masked prediction, further improving representation quality. While these models were primarily designed for speech, they have been adapted for music tasks.

For music-specific pre-training, MERT [10] applies a masked acoustic modeling objective to music audio, using both acoustic and musical tokens as prediction targets. MERT achieves state-of-the-art performance on a wide range of MIR tasks, including music tagging, genre classification, key detection, and emotion recognition. The model outputs multi-layer hidden representations, and a learnable weighted sum over layers has been shown to be effective for downstream tasks. MuQ [11] further advances music understanding through a music-specific quantizer that captures both timbrel and harmonic information, achieving competitive results with a more compact architecture.

Contrastive audio-language pre-training has also gained attention. CLAP [35] learns aligned audio and text representations through contrastive learning on audio–caption pairs, enabling zero-shot audio classification and retrieval. Manco et al. [36] applied contrastive learning specifically to music and text, demonstrating that aligned multimodal representations improve music understanding tasks.

These pre-trained models provide high-quality unimodal features, but their effective combination for MER requires fusion architectures that respect the temporal structure of both modalities. Audio representations from models like MERT are inherently sequential (frame-level or segment-level), while BERT produces token-level text representations. MELT bridges this structural gap by introducing a temporal segment encoding for audio and a token-level cross-attention mechanism for fusion, enabling fine-grained interaction between the two modalities.

3. Method

3.1. Overall Architecture

We formulate music emotion recognition as a multimodal classification problem. Given a song represented by an audio waveform $\mathbf{a} \in \mathbb{R}^{D \cdot f_s}$ of duration D seconds at sampling rate f_s, and its corresponding lyrics text $\mathbf{w} = (w_1, w_2, \ldots, w_L)$ as a sequence of L tokens, the goal is to predict an emotion label $y \in \{1, 2, \ldots, C\}$ from a predefined set of C emotion categories. In our setting, $C = 4$ correspond to the four quadrants of Russell's circumplex model [3]: angry (high arousal, negative valence), happy (high arousal, positive valence), relaxed (low arousal, positive valence), and sad (low arousal, negative valence).

Figure 1 illustrates the overall architecture of MELT. The framework processes audio and lyrics through four stages: (1) the lyrics semantic encoder (LSE) encodes the lyrics into contextualized token embeddings using a pre-trained language model; (2) the segment temporal encoder (STE) divides the audio into fixed-length segments, encodes each segment through a pre-trained audio backbone, and models temporal dependencies across segments using a Transformer encoder with learnable positional embeddings; (3) the token-level temporal fusion (TTF) module performs gated cross-attention from lyric tokens to audio segments, enabling fine-grained cross-modal interaction; and (4) the emotion mood head (EMH) aggregates the multimodal representation through multi-view pooling and produces the emotion prediction.

A key design principle of MELT is that the fusion operates at the token level rather than the global level. Unlike conventional approaches that reduce each modality to a single vector before fusion [7], MELT preserves the internal structure of both modalities throughout the fusion process. Each lyric token can selectively attend to different audio segments, and a learned gating mechanism controls the degree of cross-modal influence at each position. This design is inspired by the success of cross-attention mechanisms in vision–language models [26,27] and multimodal Transformers [25] but adapted to the specific structure of music data where audio has a natural temporal segmentation.

In contrast to symmetric cross-attention frameworks such as MulT [25] and gate-on-text injection frameworks such as MAG-BERT [29], MELT commits to three design choices that jointly characterize its novelty: (1) lyric tokens are queries, and audio *segments* (not raw frames) are keys/values, so the attention is asymmetric and operates over musically meaningful time units; (2) audio segments are temporally encoded by the STE before fusion, so each key/value already carries within-song context; and (3) the gate is a token×feature convex combination (Equation (11)), designed to degrade gracefully to the text-only pathway when audio is uninformative, rather than to always add an acoustic residual. A structural comparison with representative baselines is provided in Table 1. In contrast to the closest attention-based baselines, MELT is characterized by three coupled choices that operate together rather than in isolation: (1) lyric tokens are used as queries while temporally encoded audio segments are used as keys/values, giving an asymmetric routing that matches the role of lyrics as the semantic scaffold and audio as the acoustic modifier; (2) the audio keys/values are produced by a dedicated segment temporal encoder so that every fusion step sees segment-level temporal context, rather than raw frames (as in MulT) or a single aggregated acoustic vector per word (as in MAG-BERT); and (3) the fusion uses a convex-combination gate at the granularity of individual token $\times$ feature positions (Section 3.4, Equation (11)), rather than an additive residual. These three points together—not cross-attention by itself—constitute the differentiator of MELT relative to prior work, and they are the ones being ablated in Section 4.7.

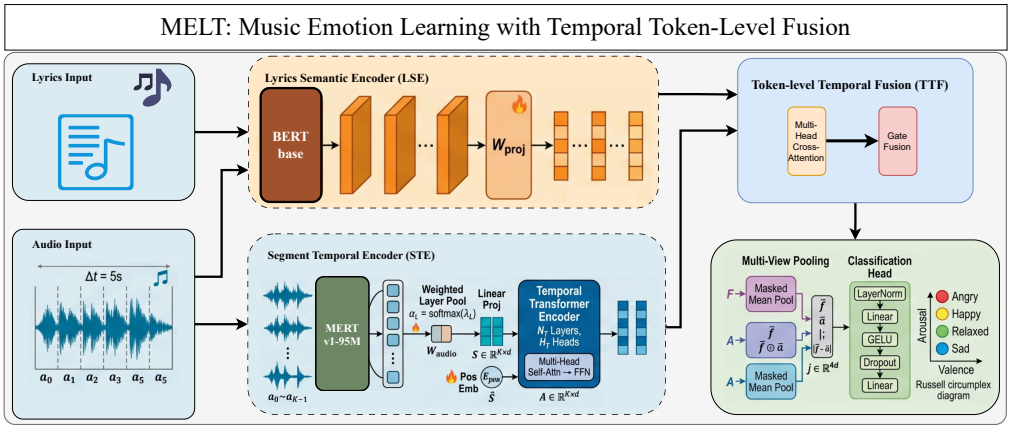


Figure 1. Overall architecture of MELT. The input song is processed through two parallel branches: the lyrics semantic encoder (LSE) produces token-level text representations from BERT, while the segment temporal encoder (STE) segments the audio waveform, encodes each segment via a pre-trained audio backbone (e.g., MERT), and models temporal dependencies with a Transformer encoder. The token-level temporal fusion (TTF) module performs gated cross-attention from lyric tokens (queries) to audio segments (keys/values). The emotion mood head (EMH) aggregates the fused and audio representations through multi-view pooling and classifies the emotion. Different colors and icons are used only to distinguish the lyrics/audio branches, fusion module, and emotion categories in the schematic.

3.2. Lyrics Semantic Encoder

The lyrics semantic encoder (LSE) processes the raw lyrics text through a pre-trained BERT model [8] to obtain contextualized token representations. We choose BERT-base-uncased as the text backbone for several reasons: (1) its bidirectional attention mechanism captures both left and right context for each token, which is important for understanding the emotional connotation of words that depends on surrounding context; (2) its pre-training on a large general-domain corpus provides robust word representations that transfer well to the lyrics domain; and (3) its moderate size (110 M parameters) allows efficient fine-tuning on relatively small MER datasets.

Given lyrics $\mathbf{w} = (w_1, w_2, \ldots, w_L)$ tokenized into L subword tokens using the WordPiece tokenizer [8], the LSE produces contextualized representations:

$$\mathbf{H}^{\text{text}} = \text{BERT}(\mathbf{w}) \in \mathbb{R}^{L \times d_{\text{bert}}}, \tag{1}$$

where $d_{\text{bert}} = 768$ is the hidden dimension of BERT-base. Unlike approaches that use only the [CLS] token as a global sentence representation, we retain all L token representations to enable token-level fusion with audio in the subsequent TTF module. A linear projection maps the BERT outputs to the fusion dimension d:

$$\mathbf{T} = \mathbf{H}^{\text{text}}\mathbf{W}_{\text{proj}} + \mathbf{b}_{\text{proj}} \in \mathbb{R}^{L \times d}, \tag{2}$$

where $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d_{\text{bert}} \times d}$ and $\mathbf{b}_{\text{proj}} \in \mathbb{R}^{d}$ are learnable parameters. This projection serves two purposes: it reduces the dimensionality to match the audio representation space, and it provides a learnable adaptation layer that can specialize the general-purpose BERT features for the emotion recognition task.

During training, the BERT backbone is fine-tuned with a lower learning rate (2×10^{-5}) compared to the randomly initialized modules (2×10^{-4}), following the discriminative fine-tuning strategy [37] that prevents catastrophic forgetting of pre-trained knowledge while allowing task-specific adaptation.

3.3. Segment Temporal Encoder

The segment temporal encoder (STE) processes the audio waveform by first segmenting it into fixed-length windows, encoding each segment independently through a pre-trained audio backbone, and then modeling temporal dependencies across segments using a Transformer encoder.

3.3.1. Audio Segmentation

Music is inherently temporal: the emotional character of a song evolves over its duration, with different sections (verse, chorus, bridge) often conveying distinct emotional intensities. To capture this temporal structure, we divide the audio waveform into a sequence of fixed-length segments rather than encoding the entire waveform as a single representation.

Given an audio waveform $\mathbf{a}$ of duration D seconds sampled at rate f_s, we divide it into K non-overlapping segments of Δt seconds each (with $K = \min(\lfloor D/\Delta t \rfloor, K_{\max})$, where $K_{\max}$ is the maximum number of segments):

$$\mathbf{a}_k = \mathbf{a}[k \cdot \Delta t \cdot f_s : (k+1) \cdot \Delta t \cdot f_s], \quad k = 0, 1, \ldots, K-1. \tag{3}$$

We set $\Delta t = 5$ s and $K_{\max} = 6$ in our experiments, covering up to 30 s of audio per song. This segment length is chosen to balance temporal granularity (shorter segments capture more fine-grained dynamics) with the receptive field of the audio backbone (segments must be long enough to contain meaningful musical content such as melodic phrases or rhythmic patterns). Songs shorter than $K_{\max} \cdot \Delta t$ seconds are zero-padded, and a binary segment mask $\mathbf{m}^{\text{audio}} \in \{0, 1\}^{K_{\max}}$ indicates which segments contain valid audio.

3.3.2. Segment Encoding with Pre-Trained Audio Backbone

Each audio segment is independently encoded by a pre-trained audio backbone to produce a segment-level embedding. We support multiple audio backbones, with MERT-v1-95 M [10] as the default choice due to its strong performance on music understanding tasks.

For models that output multiple hidden layers (such as MERT, which produces $N_L + 1$ hidden states including the input embedding layer), we employ a learnable weighted layer pooling mechanism inspired by ELMo [38]:

$$\mathbf{s}_k = \sum_{l=0}^{N_L} \alpha_l \cdot \bar{\mathbf{h}}_k^{(l)}, \quad \alpha_l = \frac{\exp(\lambda_l)}{\sum_{j=0}^{N_L} \exp(\lambda_j)}, \tag{4}$$

where $\bar{\mathbf{h}}_k^{(l)} = \text{MeanPool}(\mathbf{h}_k^{(l)})$ is the mean-pooled output of layer l for segment k, and λ_l are learnable scalar weights initialized to zero (uniform weighting). Here, N_L denotes the number of Transformer layers of the pre-trained audio backbone itself (for MERT-v1-95M, $N_L = 12$), so the index l runs over the $N_L + 1$ hidden states produced by the backbone (the input embedding layer plus the N_L Transformer layers); N_L should not be confused with the number of Transformer layers N_T used in the subsequent segment temporal encoder, which is specified separately in Section 4.4. This allows the model to learn which layers of the audio backbone are most informative for emotion recognition—lower layers may capture acoustic features (timbre, pitch), while higher layers may encode more abstract musical concepts (harmony, rhythm patterns) [10].

The segment embeddings are projected to the fusion dimension d through a linear transformation:

$$\mathbf{S} = [\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_{K-1}]\mathbf{W}_{\text{audio}} + \mathbf{b}_{\text{audio}} \in \mathbb{R}^{K \times d}, \tag{5}$$

where $\mathbf{W}_{\text{audio}} \in \mathbb{R}^{d_{\text{audio}} \times d}$, and d_{audio} is the hidden dimension of the audio backbone.

3.3.3. Temporal Modeling

To model the temporal dependencies between audio segments, we add learnable positional embeddings and process the sequence through a Transformer encoder:

$$\hat{\mathbf{S}} = \mathbf{S} + \mathbf{E}_{\text{pos}}[:K] \in \mathbb{R}^{K \times d}, \tag{6}$$

where $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{K_{\max} \times d}$ is a learnable positional embedding table. Unlike sinusoidal positional encodings [24], learnable embeddings allow the model to discover task-specific temporal patterns, such as the typical emotional arc of a song (e.g., building intensity from verse to chorus).

The positioned segment embeddings are processed by a Transformer encoder with N_T layers:

$$\mathbf{A} = \text{TransformerEncoder}(\hat{\mathbf{S}}, \text{mask} = \neg\mathbf{m}^{\text{audio}}) \in \mathbb{R}^{K \times d}, \tag{7}$$

where each Transformer layer consists of multi-head self-attention with H_T heads and a position-wise feed-forward network with GELU activation [39]. The padding mask ensures that attention is not computed over invalid (padded) segments. The output $\mathbf{A}$ captures both the content of each audio segment and its temporal context within the song, enabling the subsequent fusion module to access temporally contextualized audio representations.

3.4. Token-Level Temporal Fusion

The token-level temporal fusion (TTF) module is the core innovation of MELT. It enables fine-grained interaction between lyric tokens and audio segments through a gated cross-attention mechanism. Unlike global-level fusion approaches that reduce each modality to a single vector before combining them [6,7], TTF preserves the internal structure of both modalities and allows each lyric token to independently determine which audio segments are most relevant to its emotional interpretation.

3.4.1. Cross-Modal Attention

Given text token representations $\mathbf{T} \in \mathbb{R}^{L \times d}$ from the LSE and audio segment representations $\mathbf{A} \in \mathbb{R}^{K \times d}$ from the STE, the TTF module computes multi-head cross-attention where text tokens serve as queries [40], and audio segments serve as keys and values:

$$\mathbf{C} = \text{MultiHeadAttn}(\mathbf{Q} = \mathbf{T},\ \mathbf{K} = \mathbf{A},\ \mathbf{V} = \mathbf{A}) \in \mathbb{R}^{L \times d}, \tag{8}$$

where MultiHeadAttn denotes multi-head cross-attention with H heads. Specifically, for each head $h \in \{1, \ldots, H\}$,

$$\text{head}_h = \text{softmax}\left(\frac{\mathbf{T}\mathbf{W}_h^Q (\mathbf{A}\mathbf{W}_h^K)^\top}{\sqrt{d/H}}\right)\mathbf{A}\mathbf{W}_h^V, \tag{9}$$

where $\mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V \in \mathbb{R}^{d \times d/H}$ are learnable projection matrices. The attention weights $\boldsymbol{\alpha} \in \mathbb{R}^{L \times K}$ reveal which audio segments each lyric token attends to, providing interpretable cross-modal alignment. The key padding mask $\neg\mathbf{m}^{\text{audio}}$ ensures that padded audio segments receive zero attention weight.

This asymmetric design—text queries audio but not vice versa—is motivated by the observation that lyrics provide the primary semantic scaffold for emotion classification, while audio provides complementary acoustic evidence. By having text tokens query audio segments, each word in the lyrics can selectively gather acoustic context that reinforces or modifies its emotional interpretation. This is analogous to the cross-attention mechanism in the Multimodal Transformer [25] but applied specifically to the lyrics–audio alignment problem. Figure 2 illustrates the detailed architecture of the TTF module.

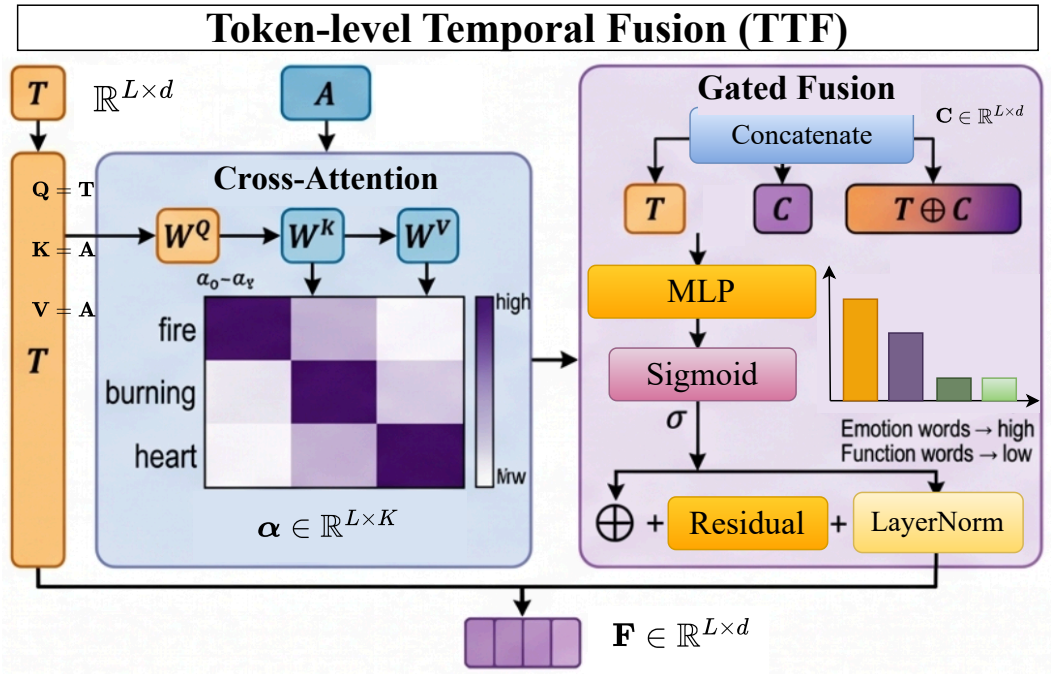


Figure 2. Detailed architecture of the token-level temporal fusion (TTF) module. Text token representations $\mathbf{T}$ serve as queries, and audio segment representations $\mathbf{A}$ serve as keys and values in a multi-head cross-attention layer. The cross-attention output $\mathbf{C}$ is combined with the original text tokens $\mathbf{T}$ through a gating mechanism: the gate $\mathbf{G}$ is computed from $[\mathbf{T};\ \mathbf{C};\ \mathbf{T}\odot\mathbf{C}]$ via a two-layer MLP with GELU activation and a sigmoid output. The fused output is obtained as a convex combination $\tilde{\mathbf{F}} = (1-\mathbf{G})\odot\mathbf{T} + \mathbf{G}\odot\mathrm{Dropout}(\mathbf{C})$ followed by layer normalization, i.e., $\mathbf{F} = \mathrm{LayerNorm}(\tilde{\mathbf{F}})$. When the gate is closed, the pre-normalization output is strictly the text representation $\mathbf{T}$, so the fusion reduces to a LayerNorm-regularized text pathway; when the gate opens, audio information is progressively injected into the corresponding token–feature positions.

3.4.2. Gated Fusion

A naive approach would directly add or replace the text representations with the cross-attention output. However, not all lyric tokens benefit equally from audio information—function words, punctuation, and semantically neutral tokens may receive noisy audio signals that degrade performance. To address this, we introduce a learned gating mechanism that adaptively controls the degree of audio influence on each token.

The gate is computed from the concatenation of the original text tokens, the cross-attention output, and their element-wise product (which captures feature-level interaction):

$$\mathbf{G} = \sigma\Big(\mathbf{W}_2^g \cdot \mathrm{GELU}\Big(\mathbf{W}_1^g \cdot [\mathbf{T};\ \mathbf{C};\ \mathbf{T}\odot\mathbf{C}] + \mathbf{b}_1^g\Big) + \mathbf{b}_2^g\Big) \in \mathbb{R}^{L\times d}, \tag{10}$$

where $\mathbf{W}_1^g \in \mathbb{R}^{d\times 3d}$ and $\mathbf{W}_2^g \in \mathbb{R}^{d\times d}$ are learnable weight matrices, $\mathbf{b}_1^g \in \mathbb{R}^d$ and $\mathbf{b}_2^g \in \mathbb{R}^d$ are learnable bias vectors (broadcast across the L token positions), $[\cdot;\cdot;\cdot]$ denotes concatenation along the feature dimension (producing a tensor of size $L \times 3d$), $\odot$ is element-wise multiplication, $\mathrm{GELU}(\cdot)$ is the Gaussian error linear unit activation [39] applied element-wise to introduce the non-linearity required for the gate to model feature interactions beyond a purely linear projection, and $\sigma(\cdot)$ is the element-wise sigmoid that bounds each gate entry to $(0,1)$. The fused representation is then obtained by a convex-combination gated fusion followed by layer normalization:

$$\tilde{\mathbf{F}} = (1-\mathbf{G})\odot\mathbf{T} + \mathbf{G}\odot\mathrm{Dropout}(\mathbf{C}), \qquad \mathbf{F} = \mathrm{LayerNorm}(\tilde{\mathbf{F}}) \in \mathbb{R}^{L\times d}. \tag{11}$$

The gate $\mathbf{G}$ operates at the granularity of individual tokens and feature dimensions. When $G_{i,j} \approx 1$ for token i and dimension j, the fused representation draws heavily from the audio-attended output; when $G_{i,j} \approx 0$, the pre-normalization signal is strictly equal to the original text representation, $\tilde{\mathbf{F}}_{i,j} = T_{i,j}$, and the post-normalization output $\mathbf{F}_{i,j}$ differs from $\text{LayerNorm}(\mathbf{T})_{i,j}$ only by the learnable affine parameters that every layer-norm layer necessarily owns. This design is related to the gating mechanisms used in highway networks [41] and gated multimodal units [42] but adapted to the cross-attention setting where the gate conditions are on both the original and attended representations.

The convex combination $(1 - \mathbf{G}) \odot \mathbf{T} + \mathbf{G} \odot \text{Dropout}(\mathbf{C})$ followed by layer normalization [43] stabilizes training and, crucially, permits the fusion module to reduce to a purely text-driven LayerNorm pathway when the learned gate closes, so that harmful audio signals can be filtered out on a per-token basis without structurally altering the fusion. The text mask is applied after fusion to zero out representations at padded positions:

$$\mathbf{F} \leftarrow \mathbf{F} \odot \mathbf{m}^{\text{text}}_{:,\text{unsqueeze}(-1)}. \tag{12}$$

3.5. Emotion Mood Head and Training Objectives

The emotion mood head (EMH) aggregates the multimodal representations and produces the final emotion prediction. We employ a multi-view pooling strategy that captures complementary aspects of the fused representation, followed by a two-layer classification head.

3.5.1. Multi-View Pooling

First, we compute masked mean pooling over the fused text tokens and the audio segment tokens:

$$\bar{\mathbf{f}} = \frac{\sum_{i=1}^{L} m_i^{\text{text}} \cdot \mathbf{F}_i}{\sum_{i=1}^{L} m_i^{\text{text}}}, \quad \bar{\mathbf{a}} = \frac{\sum_{k=1}^{K} m_k^{\text{audio}} \cdot \mathbf{A}_k}{\sum_{k=1}^{K} m_k^{\text{audio}}}, \tag{13}$$

where m_i^{text} and m_k^{audio} are the respective mask values. The final joint representation concatenates four complementary views:

$$\mathbf{j} = [\bar{\mathbf{f}};\ \bar{\mathbf{a}};\ \bar{\mathbf{f}} \odot \bar{\mathbf{a}};\ |\bar{\mathbf{f}} - \bar{\mathbf{a}}|] \in \mathbb{R}^{4d}. \tag{14}$$

Each view captures a distinct aspect of the multimodal relationship: (1) $\bar{\mathbf{f}}$ encodes the audio-enriched text representation; (2) $\bar{\mathbf{a}}$ preserves the pure audio representation; (3) $\bar{\mathbf{f}} \odot \bar{\mathbf{a}}$ captures feature-level agreement between modalities (dimensions where both modalities are active contribute more); and (4) $|\bar{\mathbf{f}} - \bar{\mathbf{a}}|$ captures divergence between modalities, which can be informative when audio and lyrics convey conflicting emotional signals (e.g., upbeat music with melancholic lyrics). This multi-view strategy is inspired by natural language inference models [44] that use similar element-wise operations to capture semantic relationships.

3.5.2. Classification Head

The classification head processes the joint representation through a two-layer feed-forward network:

$$\hat{\mathbf{y}} = \mathbf{W}_2 \cdot \text{Dropout}(\text{GELU}(\mathbf{W}_1 \cdot \text{LayerNorm}(\mathbf{j}) + \mathbf{b}_1)) + \mathbf{b}_2, \tag{15}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_h \times 4d}$, $\mathbf{W}_2 \in \mathbb{R}^{C \times d_h}$, $\mathbf{b}_1 \in \mathbb{R}^{d_h}$, and $\mathbf{b}_2 \in \mathbb{R}^{C}$ are the learnable bias vectors of the first and second linear layers, respectively, d_h is the hidden dimension, and C is the number of emotion classes. The output $\hat{\mathbf{y}} \in \mathbb{R}^{C}$ is the vector of unnormalized class logits (not a probability distribution); the per-class probabilities are obtained only

at loss/inference time by applying softmax, as shown in Equation (16) below. Layer normalization [43] is applied before the first linear layer to stabilize the input distribution, and dropout provides regularization.

3.5.3. Training Objectives

The primary training objective is the cross-entropy classification loss:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\frac{\exp(\hat{y}_{i,c})}{\sum_{c'=1}^{C}\exp(\hat{y}_{i,c'})}, \tag{16}$$

where $y_{i,c}$ is the one-hot label, and $\hat{y}_{i,c}$ is the logit for sample i and class c.

To encourage the text and audio encoders to learn aligned representations in a shared embedding space, we optionally employ an auxiliary contrastive loss based on the InfoNCE objective [45]. The pooled text and audio representations are projected to a shared contrastive space:

$$\mathbf{t}_i = \frac{\mathbf{W}_c^t \bar{\mathbf{t}}_i}{\|\mathbf{W}_c^t \bar{\mathbf{t}}_i\|_2}, \quad \mathbf{a}_i = \frac{\mathbf{W}_c^a \bar{\mathbf{a}}_i}{\|\mathbf{W}_c^a \bar{\mathbf{a}}_i\|_2}, \tag{17}$$

where $\bar{\mathbf{t}}_i = \text{MaskedMean}(\mathbf{T}_i, \mathbf{m}_i^{\text{text}})$ is the pooled text representation before fusion, and $\mathbf{W}_c^t, \mathbf{W}_c^a$ are learnable projection matrices. The symmetric contrastive loss is

$$\mathcal{L}_{\text{con}} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(\mathbf{t}_i^\top \mathbf{a}_i/\tau_c)}{\sum_{j=1}^{N}\exp(\mathbf{t}_i^\top \mathbf{a}_j/\tau_c)} + \log\frac{\exp(\mathbf{a}_i^\top \mathbf{t}_i/\tau_c)}{\sum_{j=1}^{N}\exp(\mathbf{a}_i^\top \mathbf{t}_j/\tau_c)}\right], \tag{18}$$

where τ_c is a temperature parameter that controls the sharpness of the distribution. This loss encourages matched audio–lyrics pairs to have similar representations while pushing unmatched pairs apart, which has been shown to improve multimodal alignment in contrastive audio–language models [35,36].

The total training loss combines the classification and contrastive objectives:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}, \tag{19}$$

where λ_{con} controls the relative weight of the contrastive loss.

From an optimization perspective, the model training can be written as

$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{a},\mathbf{w},y)\sim\mathcal{D}}[\mathcal{L}_{\text{cls}}(\mathbf{a},\mathbf{w},y;\boldsymbol{\theta}) + \lambda_{\text{con}}\mathcal{L}_{\text{con}}(\mathbf{a},\mathbf{w};\boldsymbol{\theta})], \tag{20}$$

where $\boldsymbol{\theta}$ denotes all trainable parameters of MELT, including text/audio projection layers, temporal encoder, fusion module, and classification head (with optional fine-tuning of BERT). The corresponding mini-batch gradient update at training step t is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \nabla_{\boldsymbol{\theta}}\left(\mathcal{L}_{\text{cls}}^{(t)} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}^{(t)}\right), \tag{21}$$

where η_t is the step-dependent learning rate determined by AdamW and the scheduler. This formulation highlights that MELT is not only an architectural fusion design but also a jointly optimized multimodal learning objective.

3.5.4. Optimization Strategy

With separate learning rates for different parameter groups, the pre-trained backbones (BERT and MERT) use a lower learning rate ($\eta_{\text{backbone}} = 2 \times 10^{-5}$) to preserve pre-trained knowledge, while the randomly initialized modules (projection layers, temporal encoder, fusion module, classification head) use a higher learning rate ($\eta_{\text{head}} = 2 \times 10^{-4}$). All parameters are optimized with AdamW [46] with weight decay $\lambda_w = 0.01$.

We employ mixed-precision training (FP16) for computational efficiency and gradient clipping with a maximum norm of 1.0 to prevent gradient explosion. Early stopping based on validation F1 score with a patience of 3 epochs is used to prevent overfitting.

4. Experiments

4.1. Dataset

We evaluate MELT on the MoodyLyrics dataset [19], a benchmark for lyrics-based music emotion recognition. The dataset contains songs annotated with four emotion categories based on Russell's circumplex model: *angry* (high arousal, negative valence), *happy* (high arousal, positive valence), *relaxed* (low arousal, positive valence), and *sad* (low arousal, negative valence). We use the balanced version of the dataset to ensure equal representation across classes.

Each sample consists of the song's lyrics text and the corresponding audio file. The dataset is split into training (70%), validation (10%), and test (20%) sets using stratified sampling to maintain class balance across splits. Table 2 summarizes the dataset statistics.

Table 2. Statistics of the MoodyLyrics dataset used in our experiments.

Split	Angry	Happy	Relaxed	Sad
Train	340	343	341	345
Validation	48	49	49	49
Test	97	98	97	98
Total	485	490	487	492

4.2. Baselines

We compare MELT against ten baselines spanning four categories. As unimodal references, **BERT-only** fine-tunes BERT-base-uncased on lyrics with a classification head on the [CLS] token, while **MERT-only** freezes MERT-v1-95M [10] and trains a classification head on the mean-pooled final hidden states, matching the frozen-audio-backbone setting used by MELT so that the comparison isolates the role of fusion rather than of audio-backbone fine-tuning.

For simple fusion, **Early Fusion** concatenates the BERT [CLS] embedding and the MERT mean-pooled embedding before a shared classifier, and **Late Fusion** trains the two unimodal classifiers independently and averages their output logits at inference time. These two baselines represent the lower bound for multimodal approaches and isolate the contribution of fusion design from backbone quality.

Among tensor-based methods, **TFN** [23] models bimodal and trimodal interactions via outer products of unimodal representations, and **LMF** [47] approximates the full tensor product with modality-specific low-rank factors to reduce computational cost. Finally, four attention-based methods are included: **MulT** [25] applies directional pairwise cross-attention between all modality pairs without requiring temporal alignment; **MAG-BERT** [29] injects acoustic signals into BERT through a gating mechanism; **MISA** [30] disentangles modality-invariant and modality-specific representations; and **ConFEDE** [31] decomposes cross-modal features into consistent and complementary components via contrastive learning.

All baselines share the same pre-trained backbones (BERT-base-uncased and MERT-v1-95 M), training schedule, and hyperparameter search protocol as MELT. Official implementations are used where available; otherwise, faithful re-implementations are provided following the original papers.

4.3. Evaluation Metrics

We evaluate all models on the held-out test set using accuracy, weighted precision, weighted recall, and weighted F1. Given C emotion classes, let TP_c, FP_c, and FN_c denote the true positives, false positives, and false negatives for class c, and let n_c be the number of true samples in class c with $N = \sum_{c=1}^{C} n_c$. Per-class precision and recall are

$$P_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \quad R_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}. \tag{22}$$

The per-class F1 score is their harmonic mean, $F1_c = 2P_cR_c/(P_c + R_c)$. The weighted variants aggregate these per-class scores proportionally to class size:

$$P_{\mathrm{w}} = \sum_{c=1}^{C} \frac{n_c}{N} P_c, \quad R_{\mathrm{w}} = \sum_{c=1}^{C} \frac{n_c}{N} R_c, \quad F1_{\mathrm{w}} = \sum_{c=1}^{C} \frac{n_c}{N} F1_c. \tag{23}$$

Weighted F1 is adopted as the primary metric since the test split, though stratified, retains slight class-size variation. All experiments are repeated with three random seeds, and results are reported as mean $\pm$ standard deviation.

4.4. Implementation Details

MELT is implemented in PyTorch (version 2.1.0) [48] with the Hugging Face Transformers library [49]. The text backbone is BERT-base-uncased (110 M parameters), and the audio backbone is MERT-v1-95 M [10]. The audio backbone is frozen during training, while the text backbone is fine-tuned with a lower learning rate.

Audio waveforms are resampled to 24,000 Hz and segmented into non-overlapping windows of 5 s each, with a maximum of 6 segments per song ($K_{\max} = 6$). Lyrics are tokenized with the BERT tokenizer and truncated to a maximum length of 128 tokens.

The fusion dimension is $d = 512$, the segment temporal encoder uses $N_T = 2$ Transformer layers with 8 attention heads, and the TTF module uses 8 cross-attention heads. The classifier hidden dimension is $d_h = 512$ with dropout rate 0.2.

We use the AdamW optimizer with separate learning rates: 2×10^{-4} for the classification head and fusion modules and 2×10^{-5} for the BERT backbone. Weight decay is set to 0.01, and gradient clipping is applied with a maximum norm of 1.0. Training runs for a maximum of 30 epochs with early stopping based on validation F1 (patience = 3). Mixed-precision training (FP16) is enabled for efficiency. The batch size is 4 with gradient accumulation steps of 2 (effective batch size 8). The contrastive loss weight is $\lambda_{\mathrm{con}} = 0.1$, and the contrastive temperature is $\tau_c = 0.07$. All experiments are conducted on a single NVIDIA RTX 3090 GPU (24 GB). Table 3 summarizes the key hyperparameters.

Table 3. Hyperparameter settings for MELT.

Hyperparameter	Value
Text backbone	BERT-base-uncased
Audio backbone	MERT-v1-95M
Fusion dimension d	512
Temporal encoder layers N_T	2
Temporal encoder heads	8
Fusion (cross-attention) heads H	8
Classifier hidden dim d_h	512
Dropout	0.2
Head learning rate	2×10^{-4}
Backbone learning rate	2×10^{-5}
Weight decay	0.01

Table 3. *Cont.*

Hyperparameter	Value
Max gradient norm	1.0
Batch size	4 (accum. 2)
Max epochs	30
Early stopping patience	3
Contrastive weight λ_{con}	0.1
Contrastive temperature τ_c	0.07
Audio segment length Δt	5 s
Max segments $K_{\max}$	6
Max text length L	128

4.5. Computational Cost Analysis

Because the TTF module introduces an extra cross-attention path on top of the unimodal backbones, it is important to make the resulting computational overhead explicit rather than relying on informal statements such as "the overhead is modest". We report five standard indicators on the same RTX 3090 (24 GB) setup described above, averaged over the test split with mixed-precision (FP16) inference: total parameter count (trainable + frozen), per-epoch wall-clock training time (effective batch size 8, 5 s $\times$ 6 audio segments, 128 text tokens), per-sample inference latency on a single GPU, peak GPU memory at training time, and per-sample forward FLOPs. Latency is measured with 20 warm-up runs followed by 100 timed runs; FLOPs are measured with `fvcore` on the inference graph. All numbers are for the full pipeline (tokenization, audio resampling, backbone forward passes, fusion, and classification). Table 4 summarizes the results and places MELT next to the unimodal and the strongest multimodal baselines using the same backbones and input budget.

Table 4. Computational cost on the MoodyLyrics test set, measured with identical inputs (128 text tokens, 6 $\times$ 5 s audio segments, FP16, batch size 8) on a single NVIDIA RTX 3090 (24 GB). "Params" reports (trainable/total), training time is per epoch, latency is per-sample inference, memory is peak GPU memory during training, and FLOPs are per-sample forward multiply–accumulate operations (MACs, in 10^9) obtained with `fvcore`; reporting MACs rather than raw FLOPs follows the convention of recent Transformer profiling studies. Training time and latency are averaged over 20 warm-up followed by 100 timed runs. Baseline numbers are obtained by re-implementing each method on our backbones and inputs and may therefore differ slightly from the values reported in the original papers.

Method	Params (M) (Train/Total)	Train/Epoch (s)	Latency (ms/Sample)	GPU Mem. (GB)	MACs (10^9)
BERT-only	110/110	38	12.1	6.8	11.3
MERT-only	0/95	47	24.3	7.5	19.6
Early Fusion	110/205	61	37.8	9.6	31.2
MulT [25]	122/217	70	46.5	11.3	34.1
MAG-BERT [29]	119/214	67	43.9	10.9	33.0
MELT (Ours)	**123/218**	**68**	**44.7**	**11.1**	**33.6**

Bold identifies MELT and its reported computational-cost values.

Relative to the strongest backbone-only baseline (BERT-only), MELT adds roughly 13 M trainable parameters, an additional $\sim$33 ms per-sample forward latency, and $\sim$4.3 GB of peak GPU memory, the bulk of which is attributable to the frozen MERT backbone rather than to the TTF or STE modules. Compared with the closest attention-based multimodal fusion baselines, MELT is within approximately $\pm$2% on every indicator—parameters, training time, latency, memory, and MACs—while delivering a 2.8-point F1 improvement over ConFEDE and a 5.8-point improvement over MulT (Table 5). In other words, the

gains reported in Section 4.6 do not come from a fundamentally heavier architecture: they come from the way lyric tokens attend to temporally encoded audio segments through a convex-combination gate.

4.6. Main Results

As shown in Table 5, the results reveal a clear performance hierarchy across fusion paradigms. Unimodal baselines (BERT-only 74.0%, MERT-only 69.0%) confirm that both modalities carry emotion-relevant information, with lyrics being more informative on this dataset. Simple fusion (Early 79.4%, Late 77.9%) and tensor-based methods (TFN 76.5%, LMF 78.3%) improve over unimodal baselines but are limited by their inability to model fine-grained cross-modal interactions. Attention-based methods achieve progressively higher scores: MulT (81.8%) introduces directional cross-attention, MAG-BERT (83.0%) conditions audio on textual context, MISA (83.5%) disentangles shared and private representations, and ConFEDE (84.8%) applies contrastive feature decomposition.

MELT achieves the highest F1 score of 87.6%, surpassing the strongest baseline ConFEDE by 2.8 percentage points and MulT by 5.8 percentage points. This improvement validates the effectiveness of token-level temporal fusion: by allowing each lyric token to attend to temporally encoded audio segments through a gated mechanism, MELT captures fine-grained cross-modal correspondences that global-level and symmetric attention approaches cannot represent. The gap between Early Fusion (79.4%) and MELT (87.6%) further demonstrates that the manner of fusion matters significantly.

Table 5. Main results on the MoodyLyrics test set. Best results are in **bold**. All values are percentages. ± denotes standard deviation over 3 runs.

Method	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑
Unimodal				
BERT-only	74.4 ± 1.1	75.2 ± 1.3	74.4 ± 1.1	74.0 ± 1.2
MERT-only	69.5 ± 1.6	71.1 ± 1.8	69.5 ± 1.6	69.0 ± 1.7
Simple Fusion				
Early Fusion	79.7 ± 0.9	80.5 ± 1.0	79.7 ± 0.9	79.4 ± 1.0
Late Fusion	78.2 ± 1.2	79.1 ± 1.3	78.2 ± 1.2	77.9 ± 1.3
Tensor-based Fusion				
TFN [23]	76.9 ± 1.3	77.8 ± 1.4	76.9 ± 1.3	76.5 ± 1.4
LMF [47]	78.7 ± 1.1	79.5 ± 1.2	78.7 ± 1.1	78.3 ± 1.2
Attention-based Fusion				
MulT [25]	82.1 ± 0.9	82.8 ± 1.0	82.1 ± 0.9	81.8 ± 1.0
MAG-BERT [29]	83.3 ± 0.8	84.0 ± 0.9	83.3 ± 0.8	83.0 ± 0.9
MISA [30]	83.8 ± 1.0	84.5 ± 1.1	83.8 ± 1.0	83.5 ± 1.1
ConFEDE [31]	85.1 ± 0.8	85.8 ± 0.9	85.1 ± 0.8	84.8 ± 0.9
MELT (Ours)	**87.9 ± 0.7**	**88.5 ± 0.8**	**87.9 ± 0.7**	**87.6 ± 0.8**

4.7. Ablation Study

To understand the contribution of each component in MELT, we conduct an ablation study by systematically removing or replacing individual modules; results are reported in Table 6.

The largest degradation comes from removing the TTF module (−4.4 points): replacing it with simple concatenation of pooled embeddings strips away the token-level cross-attention, confirming that fine-grained cross-modal interaction is the central driver of performance. Removing the temporal modeling in the STE (−2.3 points) shows that encoding temporal dependencies across segments is essential for capturing a song's emotional trajectory; without it, each segment is treated as an independent, unordered embedding. Removing multi-view pooling (−2.1 points) demonstrates that the element-wise difference

view, which captures cross-modal divergence, provides complementary information that single-pool aggregation cannot recover.

Table 6. Ablation study on the MoodyLyrics test set. Each row removes or replaces one component from the full MELT model.

Variant	F1 ↑	Δ F1
MELT (full)	**87.6**	—
w/o TTF (concat instead)	83.2	−4.4
w/o STE (no temporal encoding)	85.3	−2.3
w/o Gating (plain cross-attn)	85.8	−1.8
w/o Contrastive loss	86.2	−1.4
w/o Multi-view pooling	85.5	−2.1

Bold indicates the best F1 score in the ablation study.

The remaining two components contribute more modestly but remain meaningful. Replacing gated fusion with plain cross-attention (−1.8 points) confirms that the learned gate is necessary to suppress noisy audio signals for semantically neutral tokens, preventing degradation on tokens where audio provides little emotional cue. Removing the contrastive loss (−1.4 points) indicates that the auxiliary alignment objective encourages the text and audio encoders to produce mutually compatible representations, improving the quality of the fused features even when the classification signal alone is insufficient to align the two embedding spaces.

4.8. Analysis

4.8.1. Per-Class Performance

Table 7 breaks down performance by emotion class, revealing a pattern that closely mirrors the structure of Russell's circumplex model. *Sad* achieves the highest F1 (91.6%), likely because sad songs exhibit distinctive lexical markers (e.g., "tears", "alone", "gone") combined with characteristic audio features (slow tempo, minor key, soft dynamics). *Angry* is the second-best category (89.3%), benefiting from strong textual cues (e.g., "hate", "fight", "burn") and high-energy audio patterns. The two positive-valence categories—*happy* (83.8%) and *relaxed* (85.7%)—are harder to distinguish, as they share positive sentiment in lyrics but differ primarily in arousal level, which is more readily captured by audio features. Notably, MELT improves most on *happy* (+8.5 over Early Fusion), suggesting that the token-level fusion is particularly effective at leveraging audio arousal cues to disambiguate positive-valence emotions.

Table 7. Per-class F1 scores (%) on the MoodyLyrics test set.

Method	Angry	Happy	Relaxed	Sad
Early Fusion	80.6	75.3	77.1	84.6
MELT	**89.3**	**83.8**	**85.7**	**91.6**

4.8.2. Confusion Matrix

Table 8 further reveals that misclassifications occur predominantly between emotionally proximate categories in Russell's circumplex model. The most frequent confusion is between *happy* and *relaxed* (11% and 7%, respectively), reflecting the difficulty of discriminating arousal levels within positive valence when both categories share similar harmonic structures and upbeat lyrical tone. The second notable confusion is between *angry* and *sad* (7% and 4%), where shared negative valence and high lyrical intensity produce overlapping representations. Confusions across the valence boundary (e.g., angry → happy or sad → relaxed) are rare (1–3%), indicating that MELT reliably captures the dominant

valence dimension. This pattern is consistent with findings in the music psychology literature [14] and mirrors the confusion structure reported by prior multimodal MER systems.

Table 8. Confusion matrix for MELT on the MoodyLyrics test set. Values represent classification proportions per true class.

	Pred. Angry	Pred. Happy	Pred. Relaxed	Pred. Sad
Angry	**0.88**	0.03	0.02	0.07
Happy	0.03	**0.83**	0.11	0.03
Relaxed	0.01	0.07	**0.87**	0.05
Sad	0.04	0.01	0.02	**0.93**

4.8.3. Cross-Validation

To verify that the reported performance generalizes beyond the fixed train/test split, we conduct stratified 5-fold cross-validation on the full MoodyLyrics dataset; per-fold results are listed in Table 9.

Table 9. Stratified 5-fold cross-validation results for MELT on MoodyLyrics. All metrics are weighted averages (%). Results are reported as mean ± standard deviation across five folds.

Fold	Accuracy	Precision	Recall	F1
Fold 1	86.4	87.1	86.4	86.1
Fold 2	88.5	89.0	88.5	88.2
Fold 3	87.2	87.8	87.2	86.9
Fold 4	88.0	88.6	88.0	87.7
Fold 5	87.9	88.5	87.9	87.6
Mean ± Std	87.6 ± 0.8	88.2 ± 0.7	87.6 ± 0.8	87.3 ± 0.8

The results are consistent across folds, with a mean F1 of 87.3% ± 0.8%. The low standard deviation indicates that MELT is not sensitive to the particular data partition. The held-out test F1 (87.6%) falls within one standard deviation of the cross-validation mean (87.3% ± 0.8%, upper bound 88.1%), confirming that the two evaluation protocols are statistically consistent and that the reported performance is not an artifact of a favorable split.

4.8.4. Effect of Audio Backbone

A key design choice in MELT is the selection of the pre-trained audio backbone. To evaluate the generality of the framework across different audio encoders, we replace the default MERT-v1-95M with two alternatives: wav2vec 2.0 [33], a self-supervised speech model, and MuQ [11], a music-specific model with quantized representations. Results are summarized in Table 10.

Table 10. Effect of audio backbone on MELT performance. All metrics are weighted averages (%) on the test set.

Audio Backbone	Params	Acc.	Prec.	Rec.	F1
wav2vec 2.0-base	95 M	84.6	85.3	84.6	84.2
MuQ	95 M	86.9	87.5	86.9	86.6
MERT-v1-95 M	95 M	**87.9**	**88.5**	**87.9**	**87.6**

All three backbones achieve strong performance within MELT, demonstrating that the proposed fusion architecture is not tightly coupled to a specific audio encoder. MERT-v1-95 M achieves the best results (87.6% F1), which is expected given that MERT was

pre-trained specifically on music audio and thus produces representations that are better aligned with the music emotion recognition task. MuQ achieves competitive performance (86.6% F1), only 1.0 percentage point below MERT, suggesting that its quantized music representations also capture emotion-relevant acoustic features effectively. wav2vec 2.0, originally designed for speech, achieves 84.2% F1—still substantially above the unimodal and simple fusion baselines—but falls 3.4 points below MERT, likely because speech-oriented pre-training does not capture music-specific features such as harmony, timbre, and rhythmic patterns as effectively. These results confirm that MELT benefits from music-specific audio pre-training but remains effective even with general-purpose audio encoders.

4.8.5. Attention Visualization

Figure 3 visualizes the cross-attention weights produced by the TTF module, offering interpretable evidence of how lyric tokens align with audio segments. Emotionally charged words such as "tears", "breaking", and "alone" attend strongly to the later audio segments (Seg 4–6, corresponding to 15–30 s), which typically contain the emotional climax of the song with higher spectral energy and pronounced dynamic changes. In contrast, "silence" attends primarily to the early segments (Seg 1–2, 0–10 s), aligning with the quieter introductory passage. Function words ("the", "from", "in") and special tokens ([CLS], [SEP]) distribute attention nearly uniformly across all segments, indicating that the model learns to selectively route audio information through semantically meaningful tokens rather than structural ones. A coarse temporal alignment is also visible: tokens appearing earlier in the lyrics tend to attend more to earlier audio segments, suggesting that MELT implicitly captures the structural correspondence between lyrical progression and musical development.

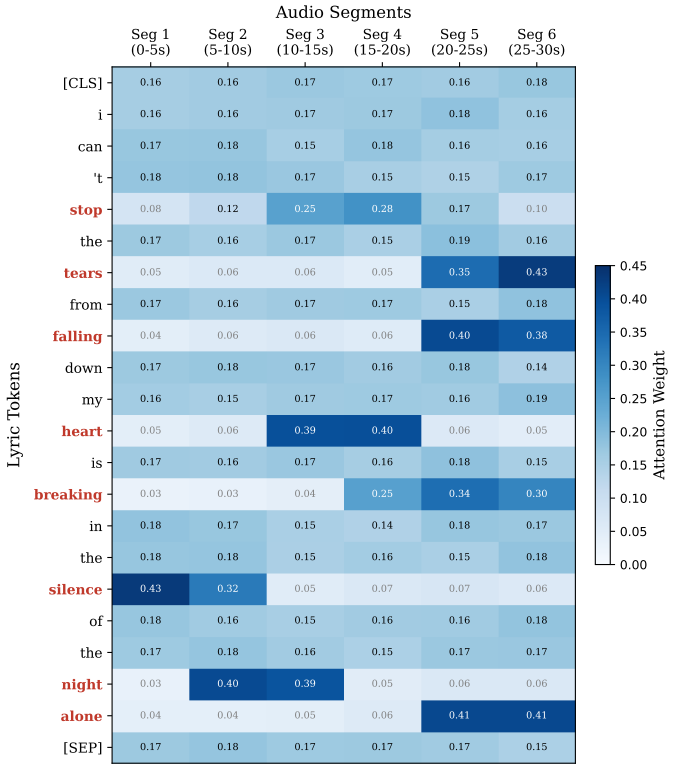


Figure 3. Cross-attention weights from the TTF module for a representative sad song. The *y*-axis lists lyric tokens (emotion-bearing words highlighted in red bold), and the *x*-axis represents audio segments at different temporal positions. Darker cells indicate higher attention weights. Emotionally charged tokens (e.g., "tears", "breaking", "alone") attend strongly to specific audio segments, while function words distribute attention uniformly.

4.8.6. Hyperparameter Sensitivity

We investigate the sensitivity of MELT to two key hyperparameters: the number of temporal encoder layers N_T and the fusion dimension d. Table 11 reports the test F1 scores under different configurations for each parameter, with all other settings fixed at the default values.

Table 11. Sensitivity analysis of temporal encoder layers N_T and fusion dimension d. Each group varies one hyperparameter while keeping the other at its default value ($N_T = 2$, $d = 512$). Bold entries indicate the selected configuration.

N_T	F1 (%)	d	F1 (%)
1	86.2	256	86.0
2	**87.6**	**512**	**87.6**
4	87.0	768	87.1

MELT is relatively robust to both hyperparameters. $N_T = 2$ yields the best performance: a single layer ($N_T = 1$) underfits the temporal structure across segments, whereas four layers ($N_T = 4$) incurs marginal overfitting on the relatively small MoodyLyrics dataset. Similarly, $d = 512$ offers the best capacity–generalization trade-off; reducing to 256 limits the expressiveness of the fusion module, while increasing to 768 adds parameters without sufficient training data to justify the extra capacity.

4.8.7. Statistical Significance

To verify statistical significance, we apply McNemar's test [50] comparing MELT against the two strongest baselines on the test set. The test evaluates whether two classifiers disagree on the same samples in a systematic way, using the contingency counts b (samples correct only by the baseline) and c (samples correct only by MELT); results are reported in Table 12.

Both comparisons yield statistically significant outcomes: the test against ConFEDE ($\chi^2 = 4.84$, $p = 0.028$) is significant at $\alpha = 0.05$, and that against MISA ($\chi^2 = 10.70$, $p = 0.001$) at $\alpha = 0.01$, confirming that the observed gains are not attributable to random variation.

Table 12. McNemar's test comparing MELT against the two strongest baselines on the MoodyLyrics test set. b: samples correct only by baseline; c: samples correct only by MELT.

Comparison	b	c	χ^2	p-Value	Significant
MELT vs. ConFEDE	7	18	4.84	0.028	Yes ($\alpha = 0.05$)
MELT vs. MISA	5	22	10.70	0.001	Yes ($\alpha = 0.01$)

4.8.8. Robustness to Missing and Perturbed Modalities

An important question for any token-level fusion model is whether it still behaves reasonably when one of the modalities is entirely unavailable or noisy at inference time, i.e., whether the fusion degrades gracefully. We therefore evaluate the same trained model (no retraining, no fine-tuning) under four perturbation regimes applied only at test time: (i) *Text only*, where the audio mask $\mathbf{m}^{\text{audio}}$ is set to all-zero so that the TTF module cannot attend to any audio segment and the audio branch of the multi-view pooling is replaced by a zero vector (handling the otherwise-undefined $0/0$ in the masked mean when no segments are retained); (ii) *Audio only*, where the text mask $\mathbf{m}^{\text{text}}$ is set to all-zero except for the [CLS]/[SEP] tokens, so that the fused-text branch $\bar{\mathbf{f}}$ degenerates to the [CLS]/[SEP]-only mean while the raw audio branch $\bar{\mathbf{a}}$ in the multi-view pooling continues to carry the audio signal; (iii) *Audio random mask*, where 50% of the audio segments per song are randomly

zero-masked, averaged over five random seeds; and (iv) *Text random mask*, where 50% of the non-special lyric tokens per song are randomly replaced with the [MASK] token, averaged over five random seeds. All numbers are on the MoodyLyrics test split and are weighted F1 (%), with the full-modality result repeated in the first row for reference. Table 13 reports the corresponding robustness results.

Table 13. Robustness of MELT under missing/perturbed modalities at inference time. The model is trained once with both modalities and then evaluated under each regime without retraining. All values are weighted F1 (%) on the MoodyLyrics test set. ΔF1 is relative to the full-modality result.

Inference-Time Regime	F1 (%)	Δ F1
Full (text + audio)	**87.6**	—
Text only (audio masked)	74.9	−12.7
Audio only (text masked)	65.1	−22.5
Audio random mask (50% segments)	83.4	−4.2
Text random mask (50% tokens)	80.2	−7.4

Three observations follow. First, when audio is entirely unavailable, MELT still reaches 74.9% F1, essentially matching the dedicated BERT-only unimodal baseline (74.0%, Table 5). This is not a coincidence but a direct consequence of Equation (11): when the audio mask drives $\mathbf{C}$ to be uninformative, the learned gate saturates toward zero and the pre-normalization signal $\tilde{\mathbf{F}}$ is exactly the text representation $\mathbf{T}$, so the fusion module collapses to a LayerNorm-regularized text pathway. The minor 0.9-point positive offset is consistent with this interpretation, since the learnable LayerNorm affine parameters in Equation (11) provide a small amount of additional regularization on top of the raw BERT output; the dropout term vanishes in this limit because its input $\mathbf{G} \odot \mathbf{C}$ is the zero tensor. Second, the audio-only regime drops more sharply (−22.5 points) and lands below the MERT-only unimodal baseline (69.0%), which is expected for two compounding reasons: (a) MELT is trained jointly with both modalities, and its EMH is calibrated to the full four-view joint representation, so masking the text side at test time produces a distribution shift that a dedicated audio-only classifier does not suffer from; (b) the fusion is asymmetric—text queries audio but not vice versa—so when the text side degenerates to [CLS]/[SEP] only, the text-conditioned audio-attended branch $\bar{\mathbf{f}}$ becomes near-vacuous, and only the raw audio branch $\bar{\mathbf{a}}$ carries emotion evidence into the joint representation. Third, partial perturbation is mild: masking half of the audio segments costs only 4.2 F1 points, showing that the STE's temporal redundancy and the token×feature gate help the model re-route attention to the remaining segments. Together with the full-modality result, these numbers delimit the operating envelope of MELT: it is designed for paired audio–lyrics inputs, tolerates up to roughly half of each stream being lost, and is not intended as a drop-in replacement for a strong text-only classifier when audio is entirely absent.

5. Discussion

The experimental results demonstrate that token-level fusion between lyrics and temporally encoded audio segments provides a meaningful advantage over conventional fusion strategies for music emotion recognition. Several observations merit further discussion.

First, the performance gap between MELT and the Early/Late Fusion baselines highlights the importance of fine-grained cross-modal interaction. Global-level fusion discards the internal structure of both modalities, treating each song as a single embedding pair. In contrast, MELT allows individual lyric tokens to selectively attend to specific temporal positions in the audio, capturing nuanced correspondences such as emotionally charged words aligning with musically intense passages.

Second, the gating mechanism in the TTF module plays a critical role. Without gating, the cross-attention output may introduce noise when the audio signal is ambiguous or when certain lyric tokens carry strong emotional signals on their own. The learned gate provides an adaptive mechanism to balance audio and text contributions at the token level.

Third, the multi-view pooling strategy in the EMH captures complementary aspects of the multimodal representation. The element-wise product highlights agreement between modalities, while the absolute difference captures divergence—both of which are informative for emotion classification.

Despite these strengths, several limitations should be acknowledged. First, the MoodyLyrics dataset contains approximately 2000 songs, which is relatively small by modern deep learning standards. Larger and more diverse datasets would provide a more rigorous evaluation of the model's generalization ability. Second, the current evaluation is limited to English-language songs; the effectiveness of MELT on multilingual lyrics remains to be investigated. Third, the model relies on pre-trained backbones (BERT and MERT), and its performance is inherently bounded by the quality of these representations. Finally, the token-level fusion mechanism introduces additional computational cost compared to simpler fusion strategies, though the overhead is modest given the small number of audio segments ($K_{\max} = 6$).

To delimit the applicability of MELT clearly rather than leave it implicit, we explicitly state the scope under which the results reported in this paper should be interpreted. The validated operating regime is (i) *English* lyrics tokenized by the BERT-base-uncased WordPiece vocabulary; (ii) *paired* audio–lyrics inputs, with both streams available at training time; (iii) *four-class categorical* emotion prediction following Russell's circumplex quadrants on the MoodyLyrics benchmark; and (iv) songs whose first 30 s provide a representative emotional signal (6 segments of 5 s). Within this envelope, Section 4.8.8 shows that up to $\sim$50% of either modality can be masked at inference without a collapse, and Section 4.5 shows that the additional cost over a MulT-style baseline is within $\pm$2% on all five indicators we measured. Outside this envelope—continuous valence–arousal regression, non-English or code-switched lyrics, completely missing modalities at *training* time, and songs whose emotional arc is concentrated outside the leading 30 s—MELT has not been evaluated, and we treat these settings as clearly scoped future work rather than as implicit claims of this paper.

6. Conclusions

We presented MELT, a multimodal framework for music emotion recognition that introduces token-level temporal fusion between lyrics and audio. The key innovation is the TTF module, which performs gated cross-attention from lyric tokens to temporally encoded audio segments, enabling fine-grained cross-modal interaction that conventional fusion strategies cannot achieve.

Experiments on the MoodyLyrics dataset with four emotion classes demonstrate that MELT outperforms both unimodal baselines and standard multimodal fusion approaches. Ablation studies confirm that each component—the segment temporal encoder, the token-level temporal fusion module, the gating mechanism, the contrastive loss, and the multi-view pooling—contributes to the overall performance.

Future work includes extending MELT to continuous valence–arousal prediction, evaluating on larger and more diverse datasets, exploring multilingual lyrics, and investigating the interpretability of the learned cross-modal attention patterns for music analysis applications.

Author Contributions: Conceptualization, Y.Y. and J.C.; methodology, J.C.; software, J.C.; validation, Y.Y. and J.C.; formal analysis, Y.Y.; investigation, J.C.; resources, Y.Y.; data curation, Y.Y. and J.C.; writing—original draft, Y.Y., Z.T., and J.C.; writing—review and editing, Y.Y., Z.T., and J.C.; visualization, J.C.; supervision, Y.Y. and J.C.; project administration, J.C.; funding acquisition, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research work is funded by the FRG-25-064-FA and granted by the Research Fund of Macao University of Science and Technology (FRG-MUST). This work was supported by the Science and Technology Development Fund of Macau Project 0087/2025/ITP2.

Data Availability Statement: The MoodyLyrics dataset used in this study was originally introduced by Çano and Morisio and is publicly available at https://github.com/workmanjack/lyric-mood-classification, accessed on 11 May 2026.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MER	Music Emotion Recognition
MIR	Music Information Retrieval
LSE	Lyrics Semantic Encoder
STE	Segment Temporal Encoder
TTF	Token-level Temporal Fusion
EMH	Emotion Mood Head
MELT	Music Emotion Learning with Temporal token-level fusion

References

1. Yang, Y.-H.; Chen, H.H. Machine Recognition of Music Emotion: A Review. *ACM Trans. Intell. Syst. Technol.* **2012**, *3*, 1–30. [CrossRef]
2. Kim, Y.E.; Schmidt, E.M.; Migneco, R.; Morton, B.G.; Richardson, P.; Scott, J.; Speck, J.A.; Turnbull, D. Music Emotion Recognition: A State of the Art Review. In Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), Utrecht, The Netherlands, 9–13 August 2010; pp. 255–266.
3. Russell, J.A. A Circumplex Model of Affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]
4. Zhao, P.; Fan, R.; Wang, S.; Shen, L.; Zhang, Q.; Ke, Z.; Zheng, T. Contextual Bandits for Unbounded Context Distributions. In Proceedings of the 42nd International Conference on Machine Learning (ICML), Vancouver, BC, Canada, 13–19 July 2025; pp. 77487–77520.
5. Hu, X.; Downie, J.S. Improving Mood Classification in Music Digital Libraries by Combining Lyrics and Audio. In Proceedings of the 10th Annual Joint Conference on Digital Libraries (JCDL), Gold Coast, Australia, 21–25 June 2010; pp. 159–168.
6. Delbouys, R.; Hennequin, R.; Piccoli, F.; Royo-Letelier, J.; Moussallam, M. Music Mood Detection Based on Audio and Lyrics with Deep Neural Net. In Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR), Paris, France, 23–27 September 2018.
7. Laurier, C.; Grivolla, J.; Herrera, P. Multimodal Music Mood Classification Using Audio and Lyrics. In Proceedings of the 7th International Conference on Machine Learning and Applications (ICMLA), San Diego, CA, USA, 11–13 December 2008; pp. 688–693.
8. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
9. Feng, X.; Guo, D.; Ye, M. DMoE-LLM: A dual-branch mixture-of-experts framework with large language models for wind power forecasting. *Expert Syst. Appl.* **2026**, *325*, 132637. [CrossRef]
10. Li, Y.; Yuan, R.; Zhang, G.; Ma, Y.; Chen, X.; Yin, H.; Lin, C.; Ragni, A.; Benetos, E.; Gyenge, N.; et al. MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training. In Proceedings of the 12th International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
11. 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. *arXiv* **2025**, arXiv:2501.01108. [CrossRef]
12. Thayer, R.E. *The Biopsychology of Mood and Arousal*; Oxford University Press: New York, NY, USA, 1989.

13. Ekman, P. An Argument for Basic Emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
14. Eerola, T.; Vuoskoski, J.K. A Comparison of the Discrete and Dimensional Models of Emotion in Music. *Psychol. Music* **2011**, *39*, 18–49. [CrossRef]
15. Mihalcea, R.; Strapparava, C. Lyrics, Music, and Emotions. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Jeju Island, Republic of Korea, 12–14 July 2012; pp. 590–599.
16. Choi, K.; Fazekas, G.; Sandler, M.; Cho, K. Convolutional Recurrent Neural Networks for Music Classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2392–2396.
17. Malik, M.; Adavanne, S.; Drossos, K.; Virtanen, T.; Ticha, D.; Jarina, R. Stacked Convolutional and Recurrent Neural Networks for Music Emotion Recognition. In Proceedings of the 14th Sound and Music Computing Conference (SMC), Espoo, Finland, 5–8 July 2017.
18. Won, M.; Hung, Y.-N.; Le, D. Evaluation of Self-Supervised Pre-Training Strategies for Music Tagging. In Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR), Online, 8–12 November 2021.
19. Çano, E.; Morisio, M. MoodyLyrics: A Sentiment Annotated Lyrics Dataset. In Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence (ISMSI), Hong Kong, China, 25–27 March 2017; pp. 118–124.
20. Ke, Z.; Cao, Y.; Chen, Z.; Yin, Y.; He, S.; Cheng, Y. Early Warning of Cryptocurrency Reversal Risks via Multi-Source Data. *Financ. Res. Lett.* **2025**, *80*, 107890. [CrossRef]
21. Baltrušaitis, T.; Ahuja, C.; Morency, L.-P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef] [PubMed]
22. Wang, W.; Tran, D.; Feiszli, M. What Makes Training Multi-Modal Classification Networks Hard? In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 12695–12705.
23. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.-P. Tensor Fusion Network for Multimodal Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark, 7–11 September 2017; pp. 1103–1114.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
25. Tsai, Y.-H.H.; Bai, S.; Thattai, P.P.; Agrawal, P.; Tulsiani, S.; Morency, L.-P. Multimodal Transformer for Unaligned Multimodal Language Sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 28 July–2 August 2019; pp. 6558–6569.
26. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 13–23.
27. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In Proceedings of the 16th European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 121–137.
28. Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; Sun, C. Attention Bottlenecks for Multimodal Fusion. In Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS), Online, 6–14 December 2021; pp. 14200–14213.
29. Rahman, W.; Hasan, M.K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; Hoque, E. Integrating Multimodal Information in Large Pretrained Transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, 5–10 July 2020; pp. 2359–2369.
30. Hazarika, D.; Zimmermann, R.; Poria, S. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In Proceedings of the 28th ACM International Conference on Multimedia (MM), Seattle, WA, USA, 12–16 October 2020; pp. 1122–1131.
31. Yang, J.; Yu, Y.; Niu, D.; Guo, W.; Xu, Y. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Toronto, ON, Canada, 9–14 July 2023; pp. 6795–6809.
32. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
33. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS), Online, 6–12 December 2020; pp. 12449–12460.

34. Hsu, W.-N.; Bolte, B.; Tsai, Y.-H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [CrossRef]
35. Elizalde, B.; Deshmukh, S.; Al Ismail, M.; Wang, H. CLAP: Learning Audio Concepts from Natural Language Supervision. In Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
36. Manco, I.; Benetos, E.; Quinton, E.; Fazekas, G. Contrastive Audio-Language Learning for Music. In Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR), Bengaluru, India, 4–8 December 2022.
37. Howard, J.; Ruder, S. Universal Language Model Fine-tuning for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia, 15–20 July 2018; pp. 328–339.
38. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), New Orleans, LA, USA, 1–6 June 2018; pp. 2227–2237.
39. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv* **2016**, arXiv:1606.08415.
40. Yao, Z.; Zhu, Q.; Qin, P. Data-Driven Secure Consensus Control of Multi-Agent Systems: A Value Iteration Approach. *Circuits Syst. Signal Process.* **2026**, 1–17. [CrossRef]
41. Srivastava, R.K.; Greff, K.; Schmidhuber, J. Highway Networks. *arXiv* **2015**, arXiv:1505.00387. [CrossRef]
42. Arevalo, J.; Solorio, T.; Montes-y-Gómez, M.; González, F.A. Gated Multimodal Units for Information Fusion. In Proceedings of the 5th International Conference on Learning Representations (ICLR) Workshop, Toulon, France, 24–26 April 2017.
43. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450. [CrossRef]
44. Chen, Q.; Zhu, X.; Ling, Z.-H.; Wei, S.; Jiang, H.; Inkpen, D. Enhanced LSTM for Natural Language Inference. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1657–1668.
45. van den Oord, A.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv* **2018**, arXiv:1807.03748.
46. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
47. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.-P. Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia, 15–20 July 2018; pp. 2247–2256.
48. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 8026–8037.
49. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP), Online, 16–20 November 2020; pp. 38–45.
50. McNemar, Q. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.