Bridging Categorical and Dimensional Affect: The MVEmo Multi-Task Benchmark for Music-Related Emotion Recognition

Jiaxing Yu, Ziyi Huang, Shuyu Li, Songruoyao Wu, Shulei Ji, Kejun Zhang*, *Member, IEEE*

***Abstract*—Music-related emotion recognition (MRER) aims to automatically predict emotional states based on different musical forms (e.g., lyrics, music, and music videos). Despite notable advancements in the field, MRER still faces the following challenges: 1) heterogeneity in emotion representations across datasets, including categorical and dimensional labels; 2) scarcity of comprehensive modalities and emotion annotations (e.g., static and dynamic) in existing music video emotion datasets; and 3) absence of a benchmark that contains various tasks and evaluation metrics for MRER. In this paper, we first propose a unified emotion representation consisting of emotion category and intensity, along with correlated conversion strategies to integrate disparate labels. Building upon the unified representation, we introduce an innovative emotion annotation framework MVAnno, which employs a hierarchical continual fine-tuning process on multiple modalities to obtain accurate emotion annotations. We also construct a large-scale music video emotion dataset MVEmo, which comprises 11K samples, with 11K static emotion labels and 5M dynamic emotion labels. Finally, we present MVEmo-Bench, a multi-task benchmark with evaluation metrics specifically designed for MRER tasks and the natural language outputs generated by large language models (LLMs). Our work makes a significant contribution to MRER by addressing key challenges and providing robust foundations for future research.**

***Index Terms*—Affective computing, emotion recognition, emotion representation, music video dataset, multi-task benchmark**

I. Introduction

Music plays a crucial role in modulating human emotions [1] and social interactions [2]. In recent years, Music-Related Emotion Recognition (MRER) has advanced rapidly, utilizing different musical forms, such as lyrics [3], [4], [5], music [6], [7], [8] and music videos [9], [10], [11], [12], to analyze affective content. The practical impact of MRER is evident through its widespread applications across various domains, including emotion-aware music recommendation [13], [14], music generation [15], [16], and music therapy [17], [18]. Despite these advancements, the field faces three key challenges: heterogeneity of emotion representations, scarcity of comprehensive music video emotion datasets, and lack of unified multi-task benchmarks.

1) The first challenge arises from the growing development of MRER datasets. These datasets can generally be divided into two types based on their representation models: categorical and dimensional. Categorical models focus on classifying emotions into discrete categories, such as happy and sad, providing static and interpretable descriptions. And dimensional models, such as Russell's circumplex model of affect [45], represent emotions along continuous axes of valence and arousal, allowing for more dynamic and nuanced expressions. As shown in Table I, although these datasets provide a large number of valuable resources, their heterogeneous representations pose a significant challenge to combine them effectively. Previous studies [46], [47], [48], [49], [50], [51], [52] primarily focused on mapping methods between categorical and dimensional labels to integrate MRER datasets, while [8] proposed a shared model architecture compatible with both types. However, these approaches rely on the original representations, which are constrained by their inherent limitations. Consequently, it is essential to design a unified emotion representation that harmonizes different emotion labels across datasets.

2) The second challenge is the incomprehensiveness of emotion annotations in existing music video datasets, which are typically categorized into static and dynamic annotations. Static annotations represent global emotional states, enabling an overall understanding of the dataset, while dynamic annotations capture time-varying emotional changes, providing more subtle and accurate temporal information. Since most existing datasets [43], [41], [40], [42], [53] only contain static emotion annotations, the MuVi dataset [44] incorporates dynamic emotion annotations on music videos every 0.5-seconds. However, the quantity of samples with fine-grained, time-varying emotion labels remains limited. Therefore, constructing a comprehensive music video dataset that provides both high-quality static and dynamic emotion annotations is essential.

3) The third challenge is the absence of a unified benchmark capable of systematically evaluating different MRER tasks. Currently, the field encompasses a variety of tasks that focus on specific modalities, ranging from unimodal emotion recognition [4], [5], [7], [8], [11], [12] to various multimodal combinations [54], [55], [56], [57], [58]. However, these tasks are typically explored independently, with each employing

This work was supported by the National Natural Science Foundation of China (No.62272409).

Jiaxing Yu, Ziyi Huang, Shuyu Li, and Songruoyao Wu are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China (e-mail: yujx@zju.edu.cn; ziyihuang1016@zju.edu.cn; lsyxary@zju.edu.cn; wsry@zju.edu.cn).

Shulei Ji and Kejun Zhang are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China and Innovation Center of Yangtze River Delta, Zhejiang University, Jiaxing 314100, China (e-mail: shuleiji@zju.edu.cn; zhangkejun@zju.edu.cn).

TABLE I
SUMMARY OF EXISTING UNIMODAL AND MULTIMODAL EMOTION DATASETS

Dataset	Size	Modality				Annotation		Representation		Phase
		Lyrics	Music	Image	Video	Static	Dynamic (T)	Categorical	Dimensional	
MoodyLyrics [19]	2,595	✓				✓		✓ (4Q)		U3
MER Lyrics Dataset [20]	771	✓				✓		✓ (4Q)	✓ (VA)	U3
LED Dataset[1]	1,160	✓				✓		✓ (9 labels)		U3
MTG-Jamendo [21]	17,982		✓			✓		✓ (56 labels)		U2
Emotify [22]	400		✓			✓		✓ (9 labels)		U2
YM2413-MDB [23]	669		✓			✓		✓ (19 labels)		U2
4Q Audio Emotion Dataset [24]	900		✓			✓		✓ (4Q)		U2
EMOPIA [25]	1,087		✓			✓		✓ (4Q)		U2
Emo-Soundscapes [26]	1,213		✓			✓			✓ (VA)	U2
CCMED - WCMED [27]	800		✓			✓			✓ (VA)	U2
Moodo [28]	200		✓			✓			✓ (VA)	U2
MusAV [29]	2,092		✓			✓			✓ (VA)	U2
EmoMusic [30]	1,000		✓				✓ (0.23±0.09s)		✓ (VA)	U2
DEAM [31]	1,802		✓			✓	✓ (0.5s)		✓ (VA)	U2
OASIS [32]	900			✓		✓			✓ (VA)	U1
EmoMadrid [33]	1,200			✓		✓			✓ (VA)	U1
GAPED [34]	730			✓		✓			✓ (VA)	U1
SMID [35]	2,941			✓		✓			✓ (VA)	U1
Cowen's Dataset [36]	2,185				✓	✓		✓ (27 labels)		U4
MoodsMIREX [37]	903	✓	✓			✓		✓ (28 labels)		M2
MERGE [38]	3,554	✓	✓			✓		✓ (4Q)		U2, U3, M2
PMEmo [39]	794	✓	✓			✓	✓ (0.5s)		✓ (VA)	U2, U3, M2
MVED [40]	3,438		✓		✓	✓		✓ (6 labels)		U4
Extended MVED [41]	5,743		✓		✓	✓		✓ (6 labels)		M1
EmoMV [42]	5,986		✓		✓	✓		✓ (5 labels)		U2, U4, M1
DEAP [43]	120		✓		✓	✓			✓ (VAD)	M1
MuVi [44]	81		✓		✓	✓	✓ (0.5s)	✓ (27 labels)	✓ (VA)	U2, U4, M1

T refers to the temporal interval between consecutive dynamic labels. Phase refers to the specific stage at which the dataset is utilized during MVAnno fine-tuning, where U denotes the unimodal fine-tuning phase, M denotes the multimodal fine-tuning phase, and the number denotes the step within each phase.

distinct datasets and evaluation metrics, making it difficult to compare the performance of models. Additionally, a unified benchmark helps to assess the impact of different modalities on multimodal emotion recognition across various models. As a result, it is crucial to establish a multi-task benchmark that integrates diverse modalities and MRER tasks with specialized evaluation metrics for evaluation.

Specifically, to address the challenge of heterogeneous emotion representations, we introduce a unified representation that combines both categorical and dimensional perspectives, ensuring compatibility with various annotation formats. The representation consists of two dimensions: emotion category and emotion intensity. Drawing inspiration from Russell's circumplex model of affect [45], emotion category is defined as a fixed set of 28 discrete words to describe qualitative emotional states, while emotion intensity is represented as a continuous scale ranging from 0 to 1 with a step size of 0.1 to measure quantitative emotional certainty. To facilitate the alignment of our unified emotion representation with multiple annotation formats, we map both emotion category and emotion intensity onto a polar coordinate system, with each emotion category corresponding to a unique angular value and emotion intensity denoted by the radial distance. Furthermore, we develop three conversion strategies, each tailored for one of the most common emotion labels. Based on the unified representation, we then propose MVAnno, an emotion annotation framework that leverages the powerful multimodal foundation model, Phi-4-Multimodal [59]. It undergoes a hierarchical continual fine-tuning process on both static and dynamic emotions, first on single modalities and then on their multimodal combinations, to acquire accurate annotations.

To handle the challenge of dataset scarcity, we construct MVEmo, a large-scale multimodal dataset that consists of 11,764 music video samples with both static and dynamic emotion annotations for MRER. We propose data collection and processing strategies for various modalities to acquire time-synchronized lyrics, audio music, symbolic music, and video. Then, we employ MVAnno to annotate static and dynamic emotions across the entire MVEmo dataset. Static emotions are labeled for each sample, while dynamic emotions are annotated every 0.5 seconds starting from 2.5 seconds, which results in 11,764 and 5,673,670 pairs of emotion categories and intensities, respectively. We further conduct detailed statistics and analysis, including basic, modality-specific, and emotion-related analysis, to provide a comprehensive understanding of the dataset for future multimodal emotion research.

Building upon the unified emotion representation and the MVEmo dataset, we introduce MVEmo-Bench, an evaluation benchmark designed to cover a range of MRER tasks. We systematically structure the benchmark to contain both unimodal (lyrics, music, and video) and multimodal recognition of static and dynamic emotions. In terms of baseline selection, we evaluate task-specific models and generic LLMs for unimodal recognition, while employing MLLMs capable of

processing text, audio, and video inputs for multimodal text-based emotion prediction. Furthermore, we implement a set of standard and specialized evaluation metrics, considering the characteristics of LLM-generated text outputs to ensure accurate evaluation of the performance across all baseline models.

Our contributions are summarized as follows:

- We design a unified emotion representation and corresponding conversion strategies that effectively tackle the challenges of heterogeneous emotion labels.
- We propose an emotion annotation framework MVAnno, which adopts the unified representation and employs a hierarchical continual fine-tuning strategy to acquire high-quality annotations.
- We construct a large-scale music video emotion dataset MVEmo, which contains various temporally aligned music-related modalities with static and dynamic emotion annotations, as well as in-depth dataset statistics and analysis. The dataset is publicly available.[2]
- We build a multi-task benchmark MVEmo-Bench, which encompasses typical MRER tasks with tailored metrics to provide comprehensive evaluation results for emotion recognition.

II. Related Work

A. Emotion Representation Models: Categorical and Dimensional

In recent years, music-related emotion datasets have garnered increasing attention. These datasets predominantly adopt categorical or dimensional emotion representation models, exhibiting inherent differences that hinder their effective combination.

To tackle the challenge, previous works primarily focus on representation mapping, which refers to the transformation of categorical and dimensional models into a unified model to integrate different datasets consistently. [46], [47], [49], and [52] created dimensional emotion-to-word mapping lexicons by collecting ratings or comparative annotations from multiple annotators. [50] utilized the lexicons to map categorical emotion labels into dimensional labels, and [60] further addressed the many-to-one mapping problem through averaging or GMMs. These direct mapping methods rely on the quality of predefined sets, which may result in inaccurate or incomplete emotion labels when the lexicons are biased or insufficient. To address this issue, [48] employed the KNN regression to achieve bidirectional mapping between the two representation types. [51] adopted acoustic features as a bridge to facilitate the conversion from category labels to AV values through a regression model. [61] introduced an emotion framework that used a GAN-based model to map various emotion representations into a new 3D representation. Although these works are more flexible in mapping, their performance is limited by the model's robustness, making it difficult to sufficiently learn the correlation between different representations. [62] handled this problem by leveraging LLM to align categorical labels from multiple datasets into a common semantic space for zero-shot inference. [63] proposed a multi-source learning framework that integrated dimensional datasets such as PMEmo [39], EmoMusic [30], and the Bi-Modal Emotion Dataset [64]. These two approaches only enable the combination of datasets within a single representation type. The recent study [8] presented a unified multi-task learning framework that supported both representations by employing a shared architecture with task-specific output branches, further leveraging knowledge distillation [65] to enhance cross-dataset generalization. However, it still cannot fully address the inherent limitations of the original representations [66], such as the low resolution of categorical models and the reduced interpretability of dimensional models.

In this paper, we propose MVAnno, a novel emotion annotation framework that harmonizes heterogeneous emotion labels across datasets through a unified emotion representation, including emotion category and intensity, and leverages a two-stage continual fine-tuning strategy to acquire high-quality annotation.

B. Music Video Emotion Datasets

Music video (MV) emotion datasets are essential for music-related emotion recognition, as they enable both the independent analysis of emotional states within single modalities and the simultaneous, cross-modal analysis of emotions across these modalities. Table I demonstrates existing available MV emotion datasets that contain at least music and video modalities. These datasets can be broadly divided into two categories: static emotion annotation and dynamic emotion annotation. Most datasets employ static emotion annotation, where a single emotion label is assigned to the entire sample. Although these annotations partially reflect the predominant emotion of the samples, the absence of dynamic emotion annotations limits the ability to capture the evolving nature of emotions throughout the sample. To address this gap, some datasets, such as MuVi [44], implement dynamic emotion annotation by tracking emotion changes over time every 0.5 seconds. However, the number of samples with time-varying labels remains limited, and these datasets are also constrained by the absence of lyric modality. In this paper, we present MVEmo, a comprehensive, large-scale dataset that includes aligned lyrics, video, audio music, and symbolic music, along with static and dynamic emotion labels based on our proposed unified representation. Additionally, we provide detailed statistics and analysis of the dataset for future research.

C. Music-Related Emotion Recognition

Music-related emotion recognition (MRER) has significantly advanced the field of emotion-aware music research. Currently, MRER can be categorized into two types based on the number of modalities involved: unimodal emotion recognition and multimodal emotion recognition.

1) Unimodal Emotion Recognition: For unimodal emotion recognition, we review existing works from three distinct modalities: lyrics, music, and video.

[2]https://nextlab-zju.github.io/mvemo

Lyrics: Lyrics provide rich semantic information for emotion recognition [67]. Early research utilized lexicon-based methods [68], [69], matching lyric words to affective dictionaries, but they were limited by vocabulary size and a lack of semantic awareness. Later statistical methods, such as Bag-of-Words [70] and TF-IDF [71], enhanced robustness yet still ignored semantics. To address this challenge, semantic embeddings (e.g., Word2Vec [72], GloVe [73], and BERT [74]) were proposed and widely used in deep learning models [75], [76]. Notably, large language models (LLMs) [77] achieved state-of-the-art performance in natural language understanding. However, their capabilities in lyrics emotion recognition (LER) remain unexplored.

Music: Music signals contain rich emotional cues [78] which are important for music emotion recognition (MER). Early methods focused on handcrafted acoustic features [79], [80], [81] which were summarized (e.g., mean, variance) and then passed into classifiers [82], [83] for emotion prediction. These methods often require much human effort in feature engineering and are limited by their inability to capture complex musical patterns. To tackle this issue, convolutional neural network (CNN)-based methods [84] were used to process acoustic features for automatic extraction of timbral and temporal patterns, while recurrent neural network (RNN)-based architectures, such as long-short term memory networks (LSTMs) [85] and gated recurrent units (GRUs) [86] were effective in capturing temporal information. Recently, large-scale pretraining-based methods have seen a rise in popularity. [7] proposed a music understanding model that leveraged a combination of acoustic and musical teacher models for effective pretraining, and achieved promising results in MER.

Video: Video emotion recognition (VER) leverages visual and auditory information to gain deep insights into the emotional state of music videos. Previous works [40], [41] focused on feature extraction from video frames and audio spectrograms, with fusion strategies to integrate both modalities. Later advancements in spatio-temporal modeling, such as long-range correlation networks [87] and multi-level attention mechanisms [9], further enhanced the accuracy of emotion prediction. In addition to task-specific architectures, large-scale vision–language models, such as LLaVA-Video [88], InternVL3 [11] and Qwen2.5-VL [12], have demonstrated powerful capabilities in general-purpose video understanding, which could also be adapted for emotion detection in music videos.

2) Multimodal Emotion Recognition: Multimodal emotion recognition aims to integrate multiple music-related modalities for a more comprehensive understanding of the emotional expression. Since research on Multi-ER is relatively limited, we focus on multimodal large language models (MLLMs), including AnyGPT [54], EMOVA [55], VITA-1.5 [56], MiniCPM-o-2.6 [57], and Qwen2.5-Omni [58], which have outperformed in various multimodal tasks and can be applied to multimodal emotion recognition. Qwen2.5-Omni adopted an end-to-end framework that can concurrently process diverse inputs and generate responses in a streaming manner. AnyGPT employed a modality-agnostic framework that enabled seamless fusion of music, image, and text, and was pretrained on music-specific datasets. EMOVA and VITA-1.5, with their advanced emotional speech processing strategies, effectively leveraged multimodal information from music videos for enhanced emotion recognition.

Despite significant progress in both unimodal and multimodal emotion recognition, there is a lack of benchmarks that encompasses the diverse range of tasks involved in MRER. In this paper, we propose MVEmo-Bench, a benchmark designed to integrate multiple tasks from both unimodal and multimodal recognition, covering static and dynamic emotions across lyrics, music, and video modalities.

III. MVAnno: Unified Emotion Annotation Framework

In this section, we first propose a unified emotion representation that transforms disparate labels into a common format. Building upon the unified representation, we then introduce MVAnno, an emotion annotation framework that employs a two-stage continual fine-tuning strategy across multiple modalities to acquire accurate emotion annotations for the MVEmo dataset.

A. Unified Emotion Representation

Inspired by Russell's circumplex model of affect [45], we design a unified emotion representation that integrates categorical and dimensional perspectives and harmonizes diverse annotation formats across datasets. It consists of two dimensions: emotion category and emotion intensity.

Emotion category refers to semantic labels that describe the qualitative type of perceived emotional experience [45]. In our unified representation, it is defined as a fixed set of 28 discrete emotion words from Russell's model: *happy*, *delighted*, *excited*, *astonished*, *aroused*, *tense*, *alarmed*, *angry*, *afraid*, *annoyed*, *distressed*, *frustrated*, *miserable*, *sad*, *gloomy*, *depressed*, *bored*, *droopy*, *tired*, *sleepy*, *calm*, *relaxed*, *satisfied*, *at ease*, *content*, *serene*, *glad*, and *pleased*. These words provide a balanced and comprehensive coverage of emotional states, facilitating both fine-grained sentiment distinctions and cross-dataset compatibility. Moreover, their psychological foundation [45], [89] enhances interpretability and ensures consistency in emotion labeling, thereby supporting reliable emotion recognition across various multimodal contexts.

Emotion intensity denotes the quantitative degree of perceived certainty in the emotional experience [90], with higher values indicating greater clarity and recognizability of the underlying emotional state. It is measured on a normalized continuous scale ranging from 0 to 1, with increments of 0.1. This fine-grained scale enables the modeling of subtle variations in emotional expression within the same category.

To facilitate the application of our unified emotion representation across diverse datasets, we map both the emotion category and intensity onto a polar coordinate system. In this system, each emotion category is represented by a unique angular value (in degrees) θ, while the emotion intensity is denoted by the radial distance r. To calculate the angular values, we refer to [45], which provides the frequency of placement of 28 words into eight reference categories, including pleasure,

excitement, arousal, distress, misery, depression, sleepiness, and contentment. The reference categories are assigned the following angular values: $\theta_{pleasure} = 0°$, $\theta_{excitement} = 45°$, $\theta_{arousal} = 90°$, $\theta_{distress} = 135°$, $\theta_{misery} = 180°$, $\theta_{depression} = 225°$, $\theta_{sleepiness} = 270°$, and $\theta_{contentment} = 315°$. For our proposed emotion categories, their angular values are derived based on their relative position within the reference categories using the weighted circular mean method. Given an emotion category c, let $A = \{\theta_1, \theta_2, \ldots, \theta_8\}$ denote the set of angular values corresponding to the reference categories, and let $F = \{f_1, f_2, \ldots, f_8\}$ denote the set of frequencies, where each f_i represents the frequency of w placed into the reference category i. The angular value (in degrees) θ_c for emotion category c is computed as follows:

$$X_i = f_i \cos(\theta_i) \tag{1}$$

$$Y_i = f_i \sin(\theta_i) \tag{2}$$

$$\tilde{\theta} = \operatorname{atan2}\left(\sum_{i=1}^{8} Y_i, \sum_{i=1}^{8} X_i\right) \tag{3}$$

$$\theta_c = \begin{cases} \tilde{\theta}, & \tilde{\theta} \geq 0 \\ \tilde{\theta} + 360°, & \tilde{\theta} < 0 \end{cases} \tag{4}$$

where X_i and Y_i represent the horizontal and vertical components of the weighted vector corresponding to each reference category. The function $\mathrm{atan2}$ computes the angle (in degrees) of the vector formed by summing all X_i and Y_i components, taking into account the correct quadrant of the angle. The resulting angle $\tilde{\theta}$ is then normalized to the range $[0°, 360°)$ to obtain the final angular value θ_c of the emotion category.

Based on this polar coordinate mapping, we further develop three conversion strategies, each tailored to a specific type of label in Table I.

For categorical annotations, existing labels can be divided into two primary types: 4-quadrant (4Q) and discrete emotion-word. The 4Q type is a coarse-grained representation of affect, where emotions are categorized based on their positions in the valence-arousal space. Following [91], we adopt a mapping approach where each quadrant is associated with a representative emotion word located at the center of its respective region. Specifically, Q1 (high valence, high arousal) is mapped to excited, Q2 (low valence, high arousal) to distressed, Q3 (low valence, low arousal) to depressed, and Q4 (high valence, low arousal) to relaxed. For each of these mappings, the emotion intensity is uniformly assigned a default value of 0.5 to reflect a moderate level of emotional certainty in the absence of finer-grained annotations. For the emotion-word type, we first collect all emotion words that appear across the existing datasets, resulting in a total of 70 unique terms. Given the heterogeneity of these labels and the potential for semantic overlap or ambiguity, an accurate mapping is required to align them with our predefined set of 28 emotion categories. To this end, we employ Qwen-Max [92] to select the semantically closest word for each source label from the target categories as the emotion label for mapping. To ensure the reliability of the process, all mappings generated by the model are manually reviewed and verified. Similar to the 4Q type, a default intensity value of 0.5 is assigned to represent moderate emotional certainty.

For dimensional annotations, we focus on the valence-arousal (V–A) labels, as it is widely adopted in affective computing research. However, the valence and arousal values are often measured using different scales across existing datasets, as shown in Table I, making it difficult to achieve consistent interpretation and integration. To address this challenge, we apply min-max normalization to rescale the raw valence and arousal values into a unified range of $[-1, 1]$. Formally, given a raw valence (or arousal) value v, its normalized counterpart $\tilde{v}$ is computed as:

$$\tilde{v} = 2 \cdot \frac{v - v_{\text{min}}}{v_{\text{max}} - v_{\text{min}}} - 1 \tag{5}$$

where v_{min} and v_{max} denote the minimum and maximum values observed in the dataset. After normalization, the resulting V–A pair $(\tilde{v}, \tilde{a})$ is converted into the polar coordinate system defined by our unified emotion representation. Specifically, we compute the angular value (in degrees) θ and radial distance r as follows:

$$\tilde{\theta} = \operatorname{atan2}(\tilde{a}, \tilde{v}) \tag{6}$$

$$\theta = \begin{cases} \tilde{\theta}, & \tilde{\theta} \geq 0 \\ \tilde{\theta} + 360°, & \tilde{\theta} < 0 \end{cases} \tag{7}$$

$$\tilde{r} = \sqrt{\tilde{v}^2 + \tilde{a}^2} \tag{8}$$

$$r_{\text{max}} = \min\left(\frac{1}{|\cos\theta|}, \frac{1}{|\sin\theta|}\right) \tag{9}$$

$$r = \frac{\tilde{r}}{r_{\text{max}}} \tag{10}$$

where $\tilde{r}$ represents the Euclidean norm of the V–A pair, r_{max} represents the maximum possible projection of a vector pointing in direction θ within the normalized space, and r represents the final normalized emotion intensity, ensuring that the vector lies within the unit circle and preserving the relative magnitude of emotional certainty along that direction. This procedure allows dimensional annotations to be consistently represented within the same polar coordinate system as categorical labels, enabling unified modeling across annotation types.

B. Emotion Annotation Framework

On top of the proposed unified emotion representation, we build MVAnno upon Phi-4-Multimodal [59], a powerful multimodal foundation model, to perform emotion annotation on the MVEmo dataset. We adopt a two-stage continual fine-tuning strategy consisting of unimodal fine-tuning followed by multimodal fine-tuning on both static and dynamic branches, conducted across a set of carefully selected datasets covering diverse modalities and emotion sources. We also evaluate the annotation capabilities of MVAnno through quantitative metrics, demonstrating its effectiveness in delivering accurate and consistent emotion labels across complex multimodal inputs.

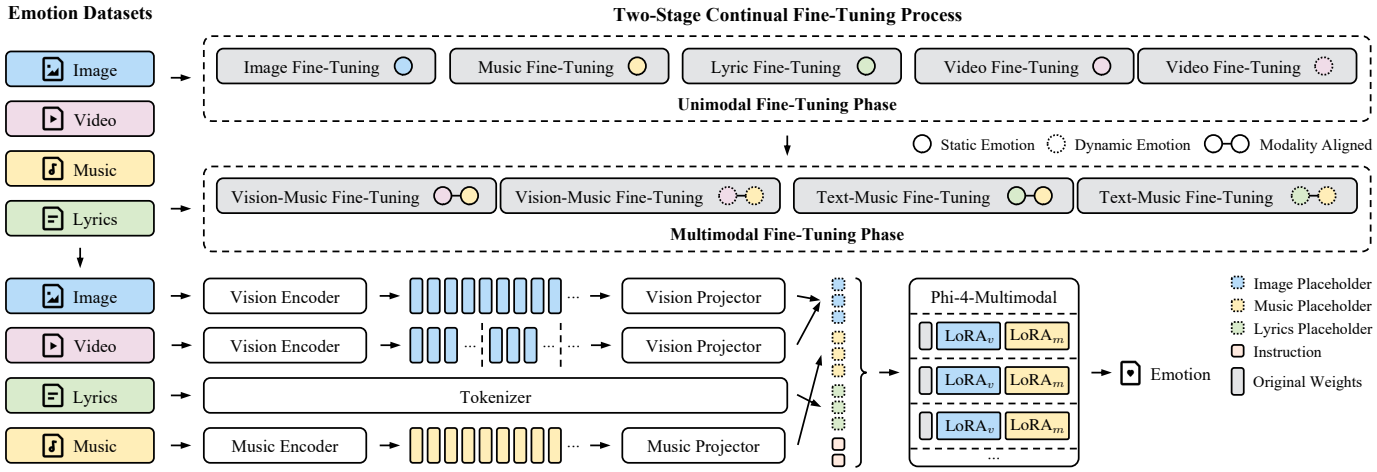


Fig. 1. Illustration of the MVAnno. We employ a two-stage continual fine-tuning process (top) for the MVAnno to ensure accurate static and dynamic emotion annotations. During the unimodal and multimodal fine-tuning phases, we adopt the mixture-of-LoRAs design (bottom), specifically LoRA$_v$ and LoRA$_m$, to integrate different modalities.

1) Model Architecture: MVAnno leverages the core architecture of Phi-4-Multimodal, which is built upon Phi-4-Mini [59], a decoder-only Transformer [93] language model with 32 layers and 3072 hidden dimensions. To support long-context generation and multilingual representation, Phi-4-Mini employs Group Query Attention (GQA) [94] with 24 query heads and 8 key/value heads, enabling efficient KV caching, as well as fractional RoPE [95] to enhance position encoding robustness over long sequences.

For multimodal integration, MVAnno inherits the mixture-of-LoRAs design from Phi-4-Multimodal, retaining separate LoRA modules that are subsequently fine-tuned to handle interactions between different modalities. This design enables MVAnno to flexibly support a variety of input configurations, including unimodal and multimodal scenarios.

2) Modality Details: We apply modality-specific modeling strategies to enable effective integration across the vision, music, and text modalities.

Vision Modality: The vision modality in MVAnno encompasses both image and video inputs. Following [59], it is implemented using a SigLIP-400M [96] image encoder fine-tuned with LLM2CLIP [97], a two-layer MLP projector for aligning vision and language embeddings, and a LoRA$_v$ module integrated into the language decoder for continual fine-tuning.

To maintain consistency with the LLM2CLIP training, all image inputs are resized to a fixed resolution of 448 × 448. For video inputs, MVAnno adopts distinct sampling strategies based on the granularity required by different emotion annotation tasks to convert videos into sequences of images. For static emotion fine-tuning, we employ a sparse sampling strategy, extracting 1 frame every 5 seconds, with up to 64 frames per video to cover the length of common music videos. For dynamic emotion fine-tuning, a sequence of 11 frames is extracted for each target timestamp. The frames are sampled at 0.5-second intervals within a 5-second window, extending from 2.5 seconds before to 2.5 seconds after the target timestamp.

Music Modality: The music modality is implemented using a pre-trained encoder with 3 convolutional layers and 24 Conformer blocks [98], a two-layer MLP projector for mapping 1024-dimensional audio features to the language embedding space, and a LoRA$_m$ module applied to the language decoder. For static emotion fine-tuning, the full-length music is used. For dynamic emotion fine-tuning, a 5-second music segment is extracted, centered on the target timestamp (2.5 seconds before and after). These music inputs are represented as 80-dimensional log-Mel spectrogram features with a 10 ms frame rate and are then fed into the encoder.

Text Modality: The text modality in MVAnno primarily handles lyrics inputs. We utilize the o200k-base tiktoken tokenizer[3] with a vocabulary size of 200,064 to convert the text into discrete token sequences. To facilitate effective multimodal alignment, we apply both vision-specific (LoRA$_v$) and music-specific (LoRA$_m$) LoRA modules to the language decoder instead of introducing a separate text-specific module. This design leverages the semantic information already captured by modality-specific adapters, with text serving as a bridge across visual and musical contexts [99], [100], [101], thereby enabling a more efficient integration of cross-modal information for music-related emotion recognition.

3) Two-Stage Continual Fine-Tuning Pipeline: As shown in Fig. 1, the MVAnno framework employs a two-stage continual fine-tuning strategy, including unimodal fine-tuning phase and multimodal fine-tuning phase. Based on this strategy, we fine-tune two models: MVAnno$_{\text{static}}$, which is designed for static emotion annotation, and MVAnno$_{\text{dynamic}}$, which extends MVAnno$_{\text{static}}$ through additional fine-tuning to capture temporal features required for dynamic emotion annotation.

Unimodal Fine-Tuning Phase: The unimodal fine-tuning phase builds upon the original LoRA weights and consists of four-stage processes, each targeting a specific modality: image, music, lyrics, and video. 1) Image Fine-Tuning stage: The image encoder, projector, and LoRA$_v$ are fine-tuned using

[3]https://github.com/openai/tiktoken

image emotion datasets. This stage enhances the model's ability to align features derived from visual inputs, such as facial expressions, color composition, and scene semantics, with corresponding emotional expression. 2) Music Fine-Tuning stage: In this stage, the music encoder, projector, and LoRA$_m$ are fine-tuned using audio music emotion datasets. Unlike symbolic music, audio captures richer and more nuanced musical elements [78] which are critical for emotion modeling. 3) Lyrics Fine-Tuning stage: Lyrics emotion datasets are used to simultaneously fine-tune both LoRA$_v$ and LoRA$_m$, leveraging the semantic knowledge acquired in previous stages. The textual modality serves as a bridge between visual and musical domains, enabling the model to achieve coherent and efficient cross-modal integration for emotion annotation. 4) Video Fine-Tuning stage: Finally, video fine-tuning is conducted on the LoRA$_v$ in two steps. First, the model is fine-tuned on sparsely sampled frames from the entire video to obtain MVAnno$_{\text{static}}$, which is capable of unimodal static emotion annotation. Based on MVAnno$_{\text{static}}$, we then further fine-tune the model using densely sampled frames centered around target timestamps to acquire MVAnno$_{\text{dynamic}}$, allowing the model to perform unimodal dynamic emotion annotation by capturing temporal changes. All four stages are fine-tuned for 50k steps using the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-7}$) with a learning rate of 4e-5, a batch size of 4, and a weight decay of 0.01.

Multimodal Fine-Tuning Phase: The multimodal fine-tuning phase is conducted after the unimodal fine-tuning process and comprises two stages based on the availability of multimodal emotion datasets. 1) Vision-Music Joint Fine-Tuning stage: In this stage, the image encoder, projector, and LoRA$_v$ are further fine-tuned using paired video and music data. Specifically, sparsely sampled video frames and full-length music are used for the static emotion annotation model MVAnno$_{\text{static}}$, while densely sampled frames centered around the target timestamp, along with corresponding music segments, are used to fine-tune the dynamic emotion annotation model MVAnno$_{\text{dynamic}}$. This joint training allows the model to capture correlations between visual and musical emotional cues, facilitating more robust multimodal emotion annotation. 2) Text-Music Joint Fine-Tuning stage: In the second stage, LoRA$_v$ and LoRA$_m$ are jointly fine-tuned using paired lyrics and music data. Similar to the first stage, lyrics are used together with full-length music for MVAnno$_{\text{static}}$, and with corresponding music segments for MVAnno$_{\text{static}}$. The multimodal fine-tuning phase is performed using the same hyperparameters as in the unimodal phase.

4) Fine-Tuning Datasets: Based on the above fine-tuning pipeline, we construct a large-scale emotion dataset that contains both unimodal and multimodal emotion data, as presented in Table I. The original emotion data combines categorical and dimensional labels, constituting a more comprehensive and diverse dataset, thereby enhancing the model's robustness and adaptability. To obtain high-quality data, we perform data processing for separate modalities, including lyrics, video, and music. The processing strategies are consistent with those used in constructing the MVEmo dataset, as detailed in Section IV-A. Additionally, all emotion labels from different sources are unified using the method outlined in Section III-A to ensure consistency. After data processing, the final dataset comprises 123,882 emotion samples, of which 73,009 are unimodal and 50,873 are multimodal. We randomly divide the dataset into training, validation, and test sets at a ratio of 8:1:1.

TABLE II
RESULTS OF MVANNO ACROSS UNIMODAL AND MULTIMODAL FINE-TUNING PHASES FOR STATIC AND DYNAMIC BRANCHES

Model	Unimodal Phase				Multimodal Phase	
	I	M	L	V	V-M	L-M
MVAnno$_{\text{static}}$	0.2669	0.2404	0.2361	0.2378	0.2218	0.2170
MVAnno$_{\text{dynamic}}$	-	-	-	0.2517	0.2453	0.2396

I refers to image. M refers to music. L refers to lyrics. V refers to video.

5) Annotation Evaluation: To ensure the reliability of MVAnno, we conduct an evaluation of the model's annotation quality. Since our goal is to enable the model to capture the direction and certainty of emotion in the affective space, we propose a new metric, Emotion Distance (D_e), which measures the Euclidean distance between the original and predicted emotion labels in the polar coordinate system using our representation, with values ranging from 0 to 2. The emotion category is first mapped to the central angle of its corresponding sector, while intensity is represented as the radial coordinate. Given the original label as (r_o, θ_o) and the predicted label as (r_p, θ_p), D_e is defined as:

$$D_e = \sqrt{r_o^2 + r_p^2 - 2r_o r_p \cos(\theta_o - \theta_p)} \tag{11}$$

Table II shows the results for both the unimodal fine-tuning phase and the multimodal fine-tuning phase. Across all stages, the D_e values of MVAnno$_{\text{static}}$ and MVAnno$_{\text{dynamic}}$ indicate that the predicted emotions are closely aligned with the ground-truth labels in the affective space, demonstrating the effectiveness of our models for high-quality annotation.

IV. MVEMO: DATASET CONSTRUCTION

In this section, we first introduce the data collection and processing strategies for different modalities, including lyrics, videos, audio music, and symbolic music. Then, we describe the annotation of static and dynamic emotion in detail. Finally, statistics and analysis of the dataset will be given.

A. Data Collection and Processing

Online music video sites provide the opportunity to obtain a large amount of data from various artists. The proposed MVEmo dataset is constructed by music videos from two primary platforms: TheoryTab[4] and YouTube[5]. TheoryTab offers structured music metadata, such as title, artist, gerne, tonality, and links to corresponding music videos on YouTube. As TheoryTab indexes music pieces rather than complete tracks, we meticulously filter its database for all valid YouTube links and download the full-length music videos at the highest

[4]https://www.hooktheory.com/theorytab
[5]https://www.youtube.com

available resolution. This approach allows us to utilize TheoryTab for its comprehensive music metadata and YouTube for its extensive collection of video content. To obtain high-quality multimodal data, we perform distinct collection and processing strategies for each modality, as detailed below:

1) Lyrics: For the acquisition of lyrics, our primary approach is to retrieve official lyrics through a metadata-based search. We query online lyric databases, specifically QQ Music[6], NetEase Cloud Music[7], and Genius[8], using the song title and artist. If the lyrics are not retrievable, we employ an automated transcription method on the audio from the music videos, beginning with voice separation by Demucs[9], followed by multilingual speech recognition using Whisper[10]. This dual strategy ensures that we can obtain sentence-level lyrics for each music video in our dataset. Then following [102], we clean the lyrics by retaining only words and a specific set of punctuation marks, including quotes, commas, colons, semicolons, periods, question marks, and exclamation marks.

2) Video: For each downloaded music video, we sample frames at a fixed interval of every five seconds to generate an image sequence denoted as $X = \{x_1, x_2, \ldots, x_N\}$, where x_t represents the t^{th} sampled frame and N is the total number of frames in the sequence. To filter out visually static videos, such as those that exclusively display album art or a still image, we implement a content similarity check based on the Mean Squared Error (MSE). This involves calculating the average inter-frame MSE, denoted as M, across the entire video sequence:

$$\mathcal{M} = \frac{1}{N-1} \sum_{t=1}^{N-1} \text{MSE}(x_t, x_{t+1}) \tag{12}$$

Videos with an average MSE below a predefined threshold ($\mathcal{M} \leq 2$) are consequently discarded. All remaining videos are further transcoded into the MPEG-4 format and uniformly resampled to 30 fps to ensure consistency across the entire MVEmo dataset.

3) Music: The domain of music is commonly categorized into two modalities: audio music and symbolic music. Audio music conveys perceived affect through performance nuances, whereas symbolic music outlines the structural correlates of emotion [78]. Given that each modality offers unique advantages for affective analysis, our dataset is specifically designed to incorporate both.

Symbolic Music: For symbolic music, we employ a lead sheet-style representation that contains melodies, chords, and core music attributes including key, tempo, position, pitch, duration, and velocity. This format is widely adopted in Music Information Retrieval (MIR) for understanding and generation tasks [15], [103], [104], [105], [106] due to its capacity to provide structured and comprehensive music information. We utilize Sheetsage[11], a state-of-the-art transcription tool, to convert audio music into lead sheets. To enhance transcription accuracy, the tempo of each song in beats per minute (BPM), is first automatically estimated using the Librosa library[12] and then provided to Sheetsage. For the velocity of each note, we compute the short-term root mean square (RMS) energy and map it to the corresponding velocity value. If the energy is zero, the note is considered a rest note. Specifically, given an audio signal $y(n)$, the energy L_m within each frame is calculated as:

$$L_m = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} y^2[mH + n]} \tag{13}$$

Here,L_m represents the RMS energy of the m^{th} frame, N is the frame length, and H is the hop length. The velocity value of the note at m^{th} frame, obtained by normalizing the energy value L_m to the standard range $[40, 127]$, is calculated as follows:

$$V = 87 \cdot \frac{L_m - min(L)}{max(L) - min(L)} + 40 \tag{14}$$

The lower bound of 40 ensures that velocity values remain within a musical range, thereby enhancing overall auditory perception and improving the reliability of the transcribed lead sheets.

B. Emotion Annotation

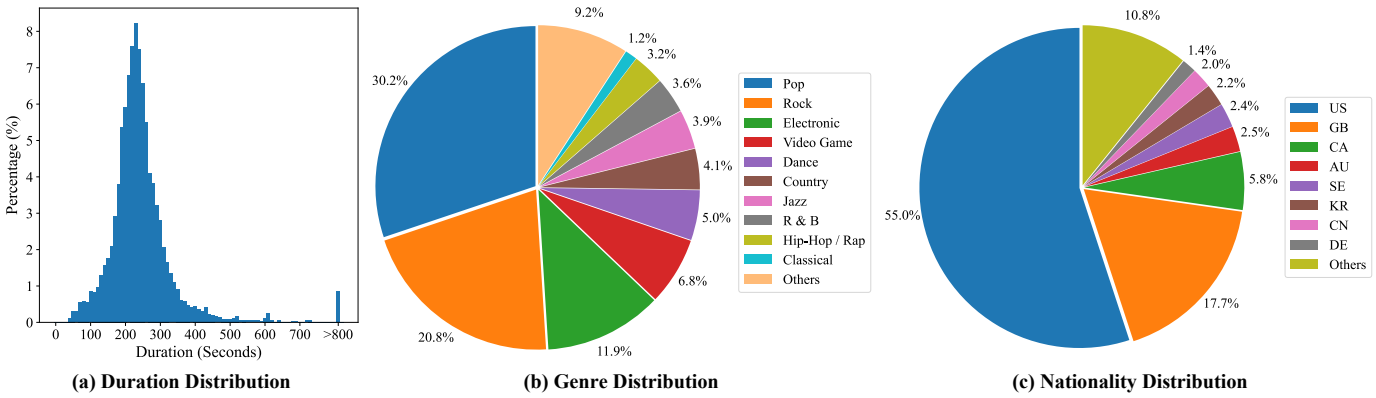
We employ $\text{MVAnno}_{\text{static}}$ and $\text{MVAnno}_{\text{dynamic}}$ to annotate the static and dynamic emotions of MVEmo, respectively. For different input modalities, we adopt the same processing strategies used during model fine-tuning, as detailed in Section III-B2. Although lyrics are not available for a small portion of MVEmo samples, our continuous cross-modal fine-tuning strategy effectively accommodates this condition and maintains reliable annotation performance. Both static and dynamic emotion annotations are carried out on the entire dataset, providing a comprehensive basis for music-related emotion recognition.

C. Dataset Statistics and Analysis

The MVEmo dataset comprises 11,764 music video samples (7,923 with lyrics and 3,841 without lyrics), establishing it as a large-scale resource for multimodal emotion research. We conduct comprehensive analysis, including basic, modality-specific, and emotion-related analysis, based on the statistics of the MVEmo dataset.

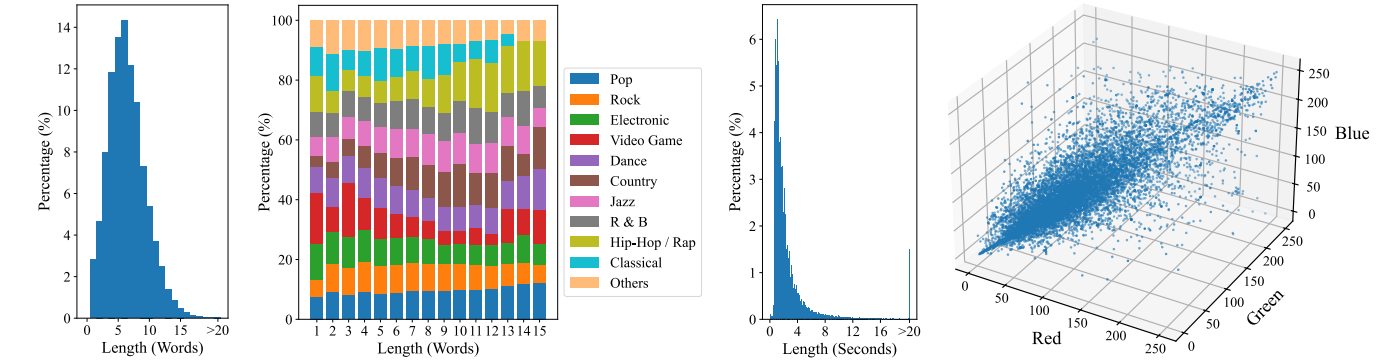
The first part of the analysis focuses on basic statistics, specifically examining the duration, genre and nationality distribution of the samples. The duration of each sample ranges from 35.94 to 4687.11 seconds, with an average of 246.19 seconds (approximately 4.10 minutes) and a total of 804.49 hours. Fig. 2(a) displays the duration distribution of the MVEmo dataset. It can be observed that 80% of the durations fall between 151.24 and 321.29 seconds, indicating that the distribution is consistent with the duration of typical music videos, while the broad span of minimum and maximum

[6]https://y.qq.com
[7]https://music.163.com
[8]https://genius.com
[9]https://github.com/facebookresearch/demucs
[10]https://github.com/openai/whisper
[11]https://github.com/chrisdonahue/sheetsage
[12]https://github.com/librosa/librosa

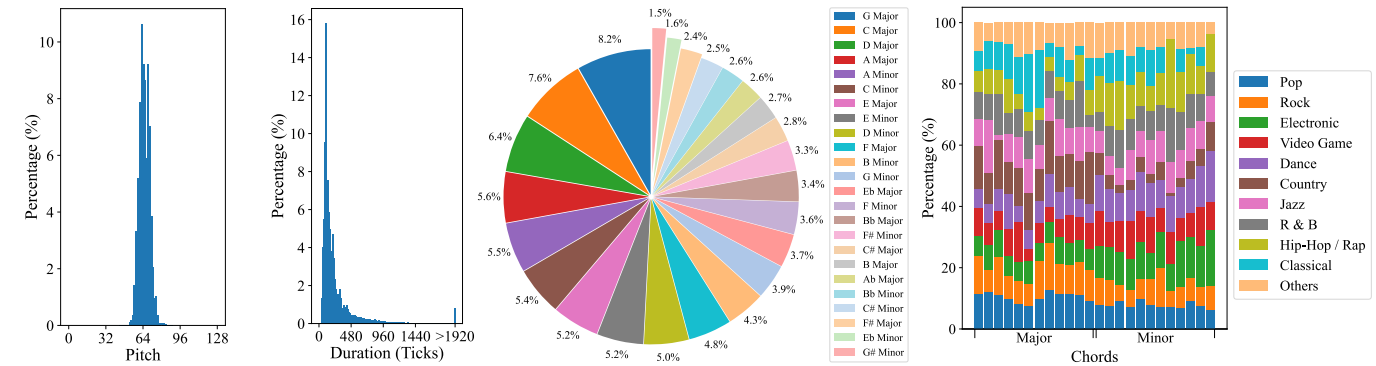


(a) Duration Distribution **(b) Genre Distribution** **(c) Nationality Distribution**

Fig. 2. Basic analysis of the MVEmo dataset: (a) duration distribution, (b) genre distribution, and (c) nationality distribution.



(a) Lyrics Length Distribution **(b) Lyrics Length Distribution By Genre** **(c) Video Segment Length Distribution** **(d) Video Color Distribution**



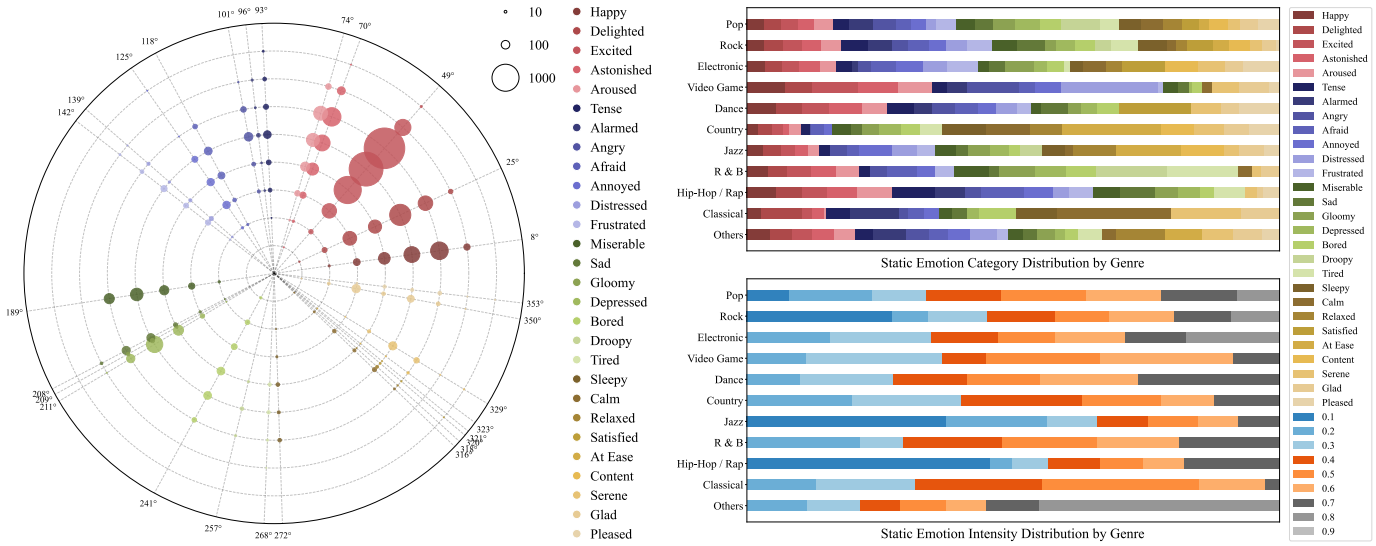
(e) Music Pitch Distribution **(f) Music IOI Distribution** **(g) Music Tonality Distribution** **(h) Music Tonality Distribution**

Fig. 3. Modality-specific analysis of the MVEmo dataset: (a-b) lyrics-related distributions, (c-d) video-related distributions, and (e-h) music-related distributions.

values further highlights the dataset's diversity. Fig. 2(b) outlines the genre distribution of all samples. We carefully select 10 commonly used genre categories and classify the original genre labels accordingly, with unclassified genres placed in the 'Others' category, as detailed in Table III. It is evident that our dataset encompasses multiple musical styles, with no single genre predominating. Fig. 2(c) illustrates the dataset's nationality distribution, presenting the top 10 countries by sample count. The distribution reflects the dataset's cultural richness, while indicating that the majority of samples are in English.

To gain further insights into different modalities and emotions of the MVEmo dataset, we conduct analysis from the perspectives of lyrics, video, and music, and then extend our investigation to both static and dynamic emotions.

1) Lyrics: Fig. 3(a) shows the sentence length (in words) distribution of the lyrics. 80% of the sentences fall between 3 and 10 words, indicating a tendency toward concise lyrical structures, while longer sentences exceeding 15 words occur infrequently. This distribution suggests that most lyrics favor brevity and repetition, which is a common feature in contemporary music [107]. Furthermore, we explore the variations in sentence length across different music genres. We select sentence lengths ranging from 1 to 15 words, ensuring that each length has more than 1000 occurrences. Then, we calculate the proportion of each sentence length across different genres, and analyze the distribution of the proportion within each length, as shown in Fig. 3(b). It can be seen that some



(a) Static Emotion Category Distribution **(b) Static Emotion Distribution By Genre**

Fig. 4. Static emotion analysis of the MVEmo dataset: (a) static emotion category distribution, and (b) static emotion distribution by genre.

genres exhibit distinct patterns, such as Hip-Hop / Rap and Country, which are relatively more prominent in longer lyrics distribution, while Video Game and Classical genres tend to feature shorter lyrics. However, genres such as Pop, Rock, Electronic, R & B, Dance, and Jazz, do not show significant variation in specific sentence lengths compared to other genres.

2) Video: For the video modality, we analyze the distribution of segment lengths and dominant colors, which provide a comprehensive understanding of the visual patterns in the MVEmo dataset. We utilize PySceneDetect[13] for transition point detection and segment the videos based on these points. The segment length distribution is shown in Fig. 3(c). It can be observed that the majority of video clips are short, with a significant portion of them under 2 seconds. This reveals that the video content is often composed of rapid transitions, likely reflecting fast-paced or visually dynamic scenes typical in music videos. The dominant color of each video is extracted by applying k-means clustering to the HSV values of the video frames, as it better aligns with human perception of color attributes [108] such as hue, saturation, and brightness. For visualization, we convert the clustered data into RGB format and compute the color distribution for the entire dataset, as shown in Fig. 3(d). The distribution indicates that the data points cluster along the main diagonal of the RGB color space, suggesting a balanced magnitude across all three color channels.

3) Music: To analyze the musical attributes of the dataset, we examine the distributions of pitch, Inter Onset Interval (IOI) [109], and tonality. Fig. 3(e) illustrates that the majority of note pitches fall within the central range of the MIDI scale, between 56 and 72. The IOI distribution, shown in Fig. 3(f), follows a long-tailed pattern, with a notable peak at the lower end, around 120 ticks. This corresponds to common note durations, such as eighth notes or quarter notes, in various time signatures. The rapid drop-off indicates that shorter rhythmic notes are more prevalent than longer ones. For tonality distribution, Fig. 3(g) outlines the diversity of key signatures, with major keys being more commonly used than minor ones. We also analyze the relative distribution of tonality across different genres and divide them according to major and minor keys, as shown in Fig. 3(h). Although major keys predominated overall, genre-specific variations are evident, demonstrating distinct tonal preferences within each genre. Specifically, genres such as Pop, Rock, Country, and Classical show a higher relative proportion of major keys, while Electronic, Dance, R & B, and Hip-Hop / Rap tend to favor minor keys.

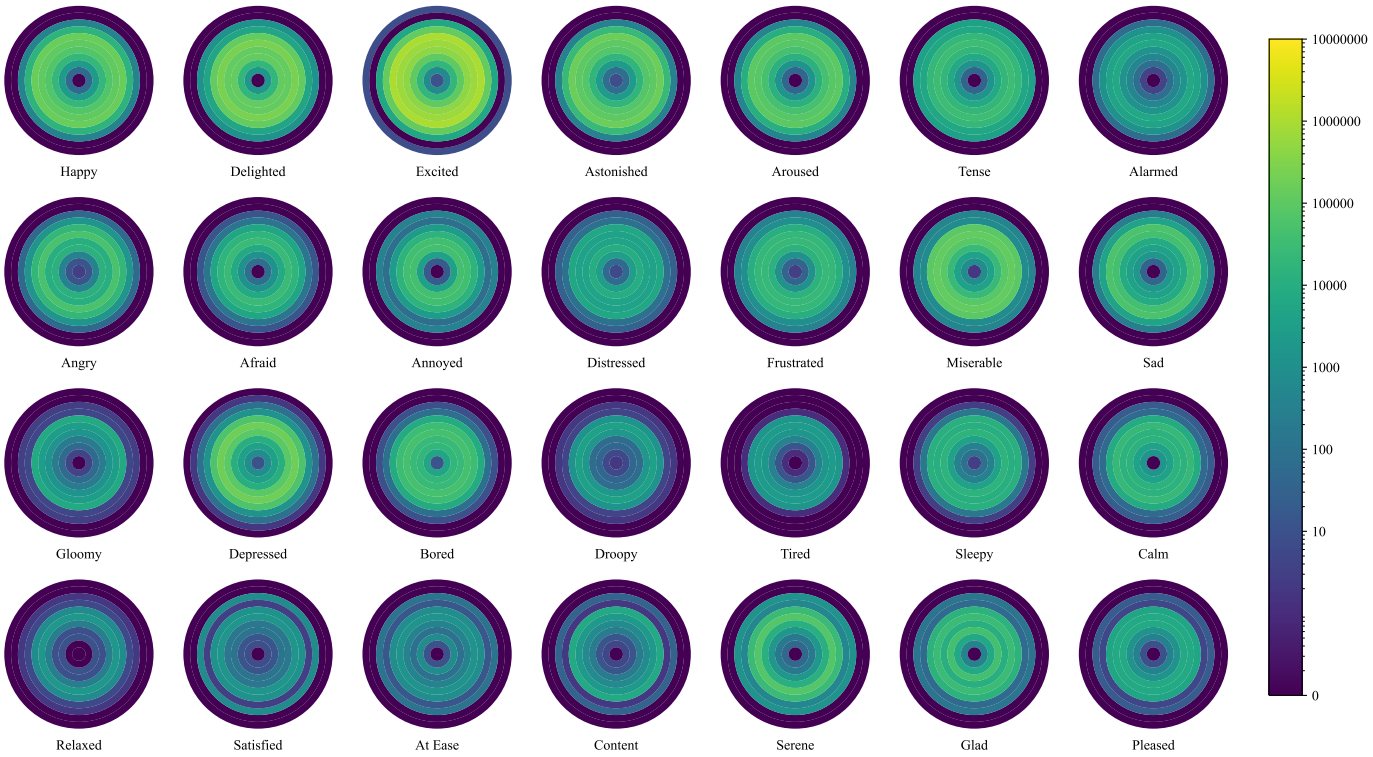
[13]https://www.scenedetect.com

TABLE III
CLASSIFICATION OF GENRE LABELS INTO GENRE CATEGORIES

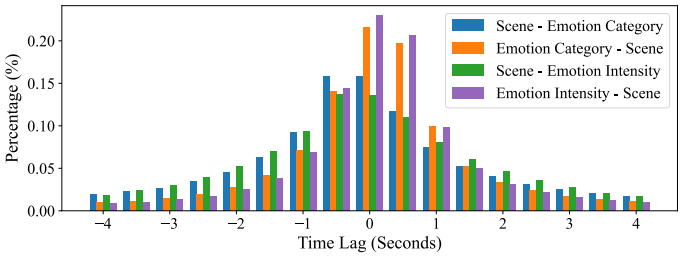
Genre Category	Genre Label
Pop	Pop, K-Pop, J-Pop, Singer Songwriter
Rock	Rock, Punk, Indie, Metal, Alternative
Electronic	Electronic, House, Techno
Video Game	Video Game
Dance	Dance
Country	Country, Alt-Country, Folk
Jazz	Jazz, Blues, Soul
R & B	R & B
Hip-Hop / Rap	Hip-Hop / Rap
Classical	Classical
Others	Children's, Disney, Experimental, Holiday, Latin, Reggae, Soundtrack, Vocal, World, Worship

4) Emotion: In emotion-related analysis, we aim to examine the underlying patterns of emotion itself and the connections between different attributes and emotion from both static and dynamic perspectives.

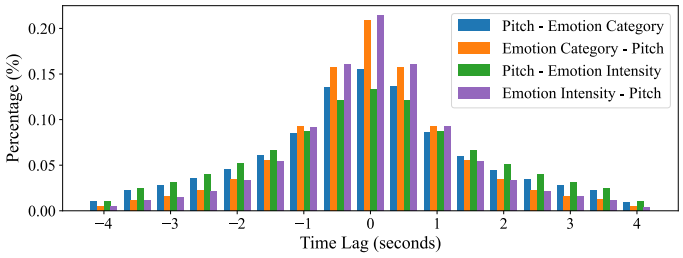
For static emotion, the MVEmo dataset contains 11,764 pairs of emotion categories and intensities, corresponding to the total number of samples in the dataset. Fig. 4(a) illustrates the distribution of static emotions in the polar coordinate system, where the number of each pair's occurrence is pro-



(a) Dynamic Emotion Intensity Distribution By Category



(b) Time Lag Distribution Between Video Scene and Dynamic Emotion



(c) Time Lag Distribution Between Music Pitch and Dynamic Emotion

Fig. 5. Dynamic emotion analysis of the MVEmo dataset: (a) dynamic emotion intensity distribution by category, (b) time lag distribution between video scene and dynamic emotion, and (c) time lag distribution between music pitch and dynamic emotion.

portional to the area of its corresponding circular marker, with larger areas indicating higher frequencies. The distribution shows a pronounced concentration of emotion categories in the first quadrant, which represents emotions with high valence and high arousal, such as *happy*, *delighted*, and *excited*. Following this dominant region, the third quadrant, which corresponds to low valence and low arousal states, is notable as the second most frequent region. Regarding intensity, the majority of the pairs cluster within the intermediate intensity range of 0.4 to 0.6. This suggests that the most annotated static emotions are perceived at moderate strength. The occurrence of both very low and very high intensity samples is comparably rare, as evidenced by the consistently smaller marker sizes in those regions.

To further analyze the relationship between static emotions and music genres, we compute the distribution in a two-step process. First, for each emotion category or intensity, we calculate its relative proportion across different genres. Then, within each genre, we normalize these values to obtain the distribution of proportions among all categories and intensities. This method allows us to highlight not only the dominance of particular emotions within a genre but also the extent to which certain genres emphasize specific emotions relative to others. Fig. 4(b) presents the results of this analysis. The top part indicates that Hip-Hop / Rap and Video Game are primarily associated with high-arousal emotions, which may be attributed to their energetic musical characteristics. In contrast, R & B is dominated by low-arousal and low-valence emotions, such as *dropy*, *tired*, and *miserable*, while Classical, Country, and Jazz display a higher concentration of low-arousal but relatively high-valence emotions, including *calm*, *relaxed*, and *satisfied*. Additionally, Rock, Pop, and Dance exhibit a more balanced distribution across different emotion categories. The bottom part illustrates the distribution of emotion intensities across genres with notable differences. Pop, Rock, and Electronic are characterized by higher intensity levels (0.7 to 0.9), suggesting that emotions evoked by these genres tend to be perceived as stronger. Jazz and Hip-Hop / Rap present more prominent

distributions in the lower intensity range (0.1 to 0.3), indicating that a greater proportion of their samples are associated with relatively mild emotions. Meanwhile, Classical, R & B, and Country concentrate within the intermediate range (0.4 to 0.6), reflecting moderate emotion intensity.

For dynamic emotion, the MVEmo dataset contains 5,673,670 emotion category and intensity pairs, with an average of approximately 482 pairs per sample. Fig. 5(a) illustrates the distribution of intensity levels across all 28 emotion categories. Each subplot corresponds to a specific emotion, visualized as a series of concentric rings, where the radius represents the intensity score—ranging from 0 at the center to 1.0 at the outermost edge. It is evident that distinct patterns emerge with respect to different emotional states. Similar to static emotion distribution, categories in the first quadrant exhibit the highest density of annotations, as evidenced by the bright yellow and light green bands. And emotions in the third quadrant also demonstrate a robust number of pairs. Furthermore, regarding the intra-category intensity distribution, while the aggregate trend reveals a concentration in the moderate range, deviations exist depending on the nature of the emotion. High-arousal categories exhibit a richer distribution that is slightly skewed toward the outer rings, with peak densities shifting to higher intensity intervals (0.6 to 0.8). This suggests that these emotions are typically identified only when they manifest strongly. Conversely, low-arousal states such as Tired and Relaxed present a more diffuse distribution that extends noticeably into lower intensity bands (0.2 to 0.4). Unlike the peaked distribution of high-arousal emotions, these categories maintain high sample counts even at lower intensities, indicating that the annotation model is able to perceive and label these states even when their expression is mild or implicit. This diversity in intensity profiles confirms that the MVEmo dataset captures the nuanced dynamics of emotion, covering both acute, high-intensity and lingering, low-intensity emotional states.

We further explore the temporal relationship between visual and emotional contents by analyzing the alignment between video scene transition points, emotion category points, and emotion intensity points. Fig. 5(b) presents the distribution of time lags between these events. Specifically, it illustrates four directional relationships: the time lag from a video scene point to the nearest emotion category or intensity point, and conversely, the time lag from an emotion category or intensity point to the nearest video scene point. The highest probabilities are concentrated at 0-second time lag, followed closely by -0.5 and +0.5 seconds. Given that the minimum annotation granularity for the dataset is 0.5 seconds, the distribution indicates a strong alignment between visual and emotional content. However, a notable asymmetry exists between the visual-to-emotion and emotion-to-visual alignments, with the probabilities for "Emotion Category - Scene" and "Emotion Intensity - Scene" at the zero-lag mark being higher ($>$0.20) than their inverse counterparts ($<$0.16). It suggests that when an emotion changes (either in category or intensity), a video scene transition is likely to occur simultaneously. In contrast, video scene transitions can predict emotion changes, but the correlation is weaker. This reflects that emotion changes typically require a new shot, whereas many scene transitions (e.g., shifts in camera angle within the same dialogue) occur without altering the emotional state.

A similar pattern is presented in the auditory domain regarding the relationship between music pitch peaks and emotion transitions. As shown in Fig. 5(c), the distributions reflect the visual domain's patterns, exhibiting significant synchronization centered at zero lag and a distinct asymmetry where emotion transitions are typically accompanied by pitch peaks.

V. MVEmo-Bench: Evaluation Benchmark

In this section, we introduce MVEmo-Bench, a comprehensive evaluation benchmark that covers a variety of music-related emotion recognition tasks. We further develop specialized evaluation metrics to rigorously assess model performance across these tasks.

A. Unimodal Emotion Recognition

Unimodal emotion recognition focuses on predicting emotions from a single type of music-related input, such as lyrics, videos, and music. In MVEmo-Bench, this task is further extended to include both static emotion recognition and dynamic emotion recognition.

We define unimodal emotion recognition as the task of mapping a single modality input x^m to its corresponding emotion representation, where $m \in \{\text{lyrics}, \text{video}, \text{music}\}$. The emotion representation is unified across tasks and consists of two discrete outputs: the emotion category label y_c and the emotion intensity label y_i. This task can be formulated as a conditional probability modeling problem as follows.

For static emotion recognition, the input is the entire modality sequence x^m, and the objective is to predict its overall emotion. It is modeled as a joint conditional probability distribution for the emotion category and intensity, defined as:

$$p(y_c, y_i \mid x^m; \theta) = p(y_c \mid x^m; \theta) \cdot p(y_i \mid x^m; \theta) \tag{15}$$

where $y_c \in \mathcal{Y}_c$ denotes the emotion category, $y_i \in \mathcal{Y}_i$ denotes the emotion intensity, and θ denotes the model parameters.

For dynamic emotion recognition, the task is extended to predict emotions at each time step, based on a sliding window of modality information. The input at each time step t is a local context window $x^m_{[t-\Delta,\, t+\Delta]}$, where $\Delta = 2.5$ seconds, and predictions are made at fixed intervals of 0.5 seconds. The model outputs the predicted emotion category and intensity at each time step, formulated as:

$$\begin{aligned} p(y_c^t, y_i^t \mid x^m_{[t-\Delta,t+\Delta]}; \theta) = p(y_c^t \mid x^m_{[t-\Delta,t+\Delta]}; \theta) \\ \cdot\, p(y_i^t \mid x^m_{[t-\Delta,t+\Delta]}; \theta) \end{aligned} \tag{16}$$

where $y_c^t \in \mathcal{Y}_c$ and $y_i^t \in \mathcal{Y}_i$ represent the emotion category and intensity at time step t, respectively.

B. Multimodal Emotion Recognition

Multimodal emotion recognition aims to predict emotions by simultaneously analyzing multiple music-related modalities. This task is more complex than unimodal recognition due

to the challenge of effectively combining different emotional cues from each modality. In MVEmo-Bench, we focus on evaluating the performance of MLLMs in both static and dynamic emotion recognition.

We define multimodal emotion recognition as the task of mapping multimodal inputs $\{x^{\text{lyrics}}, x^{\text{video}}, x^{\text{music}}\}$ to a corresponding natural language output. The model generates the emotion category y_c and emotion intensity y_i in a fixed format, specifically "Emotion Category: y_c, Emotion Intensity: y_i". The task involves not only generating this natural language response but also extracting the emotion category and intensity from the output, formulated as:

$$r = \mathcal{M}\left(x^{\text{lyrics}}, x^{\text{video}}, x^{\text{music}}; \theta\right) \tag{17}$$

$$y_c, y_i = \mathcal{P}\left(r\right) \tag{18}$$

where $\mathcal{M}$ represents the multimodal emotion recognition model, r denotes the natural language response of the model and $\mathcal{P}$ represents the parsing function that extracts the emotion category and intensity from the generated text.

For dynamic emotion recognition, similar to the unimodal task, the model predicts emotion at each time step t using a sliding window of multimodal inputs $x^{\text{modality}}_{[t-\Delta,t+\Delta]}$, as follows:

$$r^t = \mathcal{M}\left(x^{\text{lyrics}}_{[t-\Delta,t+\Delta]}, x^{\text{video}}_{[t-\Delta,t+\Delta]}, x^{\text{music}}_{[t-\Delta,t+\Delta]}; \theta\right) \tag{19}$$

$$y_c^t, y_i^t = \mathcal{P}\left(r^t\right) \tag{20}$$

C. Evaluation Metrics

Inspired by previous works [110], we adopt a set of metrics to comprehensively assess the performance of different models. Both static and dynamic recognition tasks are evaluated using $Precision$ (P), $Recall$ (R), F_1 Score, and Emotion Distance (D_e). For emotion category and intensity, $Precision$, $Recall$ and F_1 Score are computed separately, while D_e serves as a complementary measure of annotation consistency.

1) $Precision$ (P), $Recall$ (R) and F_1 Score: P represents the proportion of correctly predicted labels, and R measures the model's ability to capture all ground truth. The F_1 Score, defined as the harmonic mean of P and R, offers a balanced evaluation of both the accuracy and completeness of the model's predictions.

2) Emotion Distance (D_e): In addition to standard classification metrics, we introduce Emotion Distance (D_e), to quantify the difference between the model's predictions and the original annotations in terms of emotion alignment, as described in Section III-B5.

Additionally, given the characteristics of natural language outputs, we introduce a new metric, Error Rate (ER) to quantify the proportion of outputs that are either ill-formatted or non-compliant with the predefined sets. The metric is formulated as:

$$Fmt(r) = \begin{cases} 1, & \text{if } r \text{ conforms to the specified format} \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

$$ER = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\neg Fmt(r) \vee y_c \notin \mathcal{Y}_c \vee y_i \notin \mathcal{Y}_i\right) \tag{22}$$

where $Fmt(\cdot)$ denotes the format-check function and $\mathbb{I}(\cdot)$ is the indicator function that outputs 1 if the condition is true and 0 otherwise.

For static emotion recognition, these metrics are calculated over the entire dataset. For dynamic emotion recognition, the metrics are computed for each sample, with the final result being the average value across the dataset.

VI. Experiments

In this section, we utilize MVEmo-Bench to evaluate existing methods in unimodal and multimodal emotion recognition, present experimental results and conduct in-depth discussions.

A. Unimodal Emotion Recognition

1) Baselines and Settings: For unimodal emotion recognition, we systematically compare the performance of baseline models across three modalities: lyrics, music, and video. We first evaluate models designed for lyrics emotion recognition (LER), including Bi-LSTM [4], Bi-GRU [3], and XLM-EMO [5]. Due to the scarcity of available models in this domain, we extend the comparison to generic LLMs, namely Baichuan-2-8B [111], Llama-3.1-8B [112], and Qwen3-8B [92], to explore their potential for the task. Then, we select a range of representative models, including Music2Emo [8], MERT [7], M3BERT [6], UIBK-DBIS [114], and Mirable [113] to evaluate their capabilities on music emotion recognition (MER). For video emotion recognition (VER), we incorporate both specialized models and VLMs, including VAANet [9], CTEN [115], LLaVA-Video-7B [88], InternVL3-8B [11], and Qwen2.5-VL-7B [12].

All tasks, except for the LER, are evaluated for both static and dynamic emotions. The LER task specifically focuses on the overall emotion of the lyrics. For static emotion recognition, the evaluation is conducted on the entire dataset of the MVEmo-Bench. Due to the large number of dynamic emotion labels associated with each sample, the dynamic emotion recognition task is performed by randomly selecting 1000 samples from the MVEmo-Bench, with each sample containing complete dynamic emotion labels. To ensure a robust evaluation, the dataset is split into training, validation, and test sets at a ratio of 8:1:1.

2) Results and Discussion: Table IV presents the unimodal emotion recognition results of baseline models under static and dynamic settings. In the LER task, which is evaluated only under static setting, the results reveal a significant performance difference between traditional task-specific methods and LLMs. As shown in Table IV, the most recent models, Qwen3-8B and Llama-3.1-8B, demonstrate comparable results and decisively outperform all other baseline models. Qwen3-8B achieves the highest F_1 Scores for both category (F_c = 0.4806) and intensity (F_i = 0.8481). Meanwhile, Llama-3.1-8B exhibits the lowest (most optimal) emotion distance (D_e = 0.2926), followed by Qwen3-8B (0.3032) and Baichuan-2-8B (0.3426), which also maintain strong performance in this metric. The results highlights a notable progress compared to earlier specialized models. For example, the XLM-EMO model shows the best results (F_c = 0.2871, F_i = 0.2276, D_e =

TABLE IV
RESULTS OF BASELINE MODELS ON THE MVEMO-BENCH

Model	Static Emotion								Dynamic Emotion							
	Category			Intensity			$D_e \downarrow$	$ER \downarrow$	Category			Intensity			$D_e \downarrow$	$ER \downarrow$
	$P \uparrow$	$R \uparrow$	$F_1 \uparrow$	$P \uparrow$	$R \uparrow$	$F_1 \uparrow$			$P \uparrow$	$R \uparrow$	$F_1 \uparrow$	$P \uparrow$	$R \uparrow$	$F_1 \uparrow$		
Bi-LSTM [4]	0.2169	0.2027	0.2059	0.2506	0.2214	0.1767	0.5415	-	-	-	-	-	-	-	-	-
Bi-GRU [3]	0.2235	0.2309	0.2069	0.2451	0.2593	0.1836	0.4119	-	-	-	-	-	-	-	-	-
XLM-EMO [5]	0.2726	0.3061	0.2871	0.2378	0.3505	0.2276	0.3719	-	-	-	-	-	-	-	-	-
Baichuan-2-8B [111]	0.3432	0.4743	0.3740	0.7777	0.8724	0.8223	0.3426	0.0000	-	-	-	-	-	-	-	-
Llama-3.1-8B [112]	**0.4699**	0.5314	0.4683	0.7914	**0.8800**	0.8299	**0.2926**	0.0000	-	-	-	-	-	-	-	-
Qwen3-8B [92]	0.4501	**0.5333**	**0.4806**	**0.8428**	0.8552	**0.8481**	0.3032	0.0000	-	-	-	-	-	-	-	-
Mirable [113]	0.2678	0.2716	0.2417	**0.3818**	**0.3636**	0.3224	0.4227	-	0.2726	0.2955	0.2567	0.3361	0.3258	0.2826	0.4118	-
UIBK-DBIS [114]	0.2366	0.2576	0.2289	0.2458	0.3052	0.2557	0.4415	-	0.2499	0.2987	0.2562	0.2368	0.2944	0.2452	0.4207	-
M3BERT [6]	0.2461	0.2652	0.2393	0.2381	0.3041	0.2561	0.4348	-	0.2190	0.2576	0.2221	0.2272	0.2890	0.2450	0.4434	-
MERT [7]	0.2609	0.2911	0.2494	0.2966	0.3496	0.2817	0.4086	-	0.3115	**0.3019**	0.2609	0.3093	**0.3625**	0.2916	0.4032	-
Music2Emo [8]	**0.3036**	**0.3254**	**0.3111**	0.3802	0.3333	**0.3418**	**0.3657**	-	**0.3358**	0.2980	**0.3080**	**0.3387**	0.3276	**0.3159**	**0.3670**	-
VAANet [9]	0.1921	0.1723	0.1790	0.2984	0.3148	0.2919	0.4102	-	0.2023	0.2180	0.2011	0.2982	0.3591	0.3157	0.4083	-
CTEN [115]	0.2174	0.2676	0.2117	0.2767	0.2841	0.2744	0.3575	-	0.2121	0.2334	0.2135	0.3260	0.3262	0.3199	0.4060	-
LLaVA-Video-7B [10]	0.1986	0.2798	0.2268	0.2587	0.2575	0.2514	0.4012	0.0000	0.2621	0.3173	0.2809	**0.4099**	**0.4356**	**0.4085**	**0.3553**	0.0003
InternVL3-8B [11]	**0.2402**	0.3064	0.2510	0.2811	**0.3384**	**0.2991**	0.3578	0.0000	0.2729	0.3301	0.2926	0.3455	0.3909	0.3506	0.3720	0.0001
Qwen2.5-VL-7B [12]	0.2359	**0.3579**	**0.2647**	**0.3068**	0.3162	0.2949	**0.3421**	0.0017	**0.2846**	**0.3320**	**0.2933**	0.3504	0.3621	0.3445	0.3716	0.0000
AnyGPT-7B [54]	0.3232	**0.4009**	**0.3191**	**0.5049**	**0.5156**	**0.5073**	**0.2952**	0.0000	0.2620	0.3683	0.2902	**0.4809**	**0.4720**	**0.4372**	0.3555	0.0005
EMOVA-7B [55]	0.2114	0.3606	0.2487	0.2731	0.3179	0.2930	0.3617	0.0000	0.1670	0.2443	0.1960	0.2762	0.3124	0.2922	0.4281	0.0004
VITA-1.5-7B [56]	0.2694	0.2267	0.2059	0.2990	0.2489	0.2393	0.4659	0.0026	0.3034	**0.3911**	**0.3274**	0.3931	0.4089	0.3962	0.3490	0.0000
MiniCPM-o-2.6-8B [57]	0.2780	0.3178	0.2848	0.2974	0.2754	0.2791	0.3729	0.0034	**0.3050**	0.3852	0.3238	0.4280	0.4485	0.4233	**0.3401**	0.0001
Qwen2.5-Omni-7B [58]	**0.3425**	0.3399	0.3149	0.3620	0.2982	0.2878	0.3557	0.0000	0.2465	0.3689	0.2641	0.2986	0.3228	0.2960	0.3881	0.0000

The four parts respectively present the results of lyric, music, video, and multimodal emotion recognition, where bold values refer to the best performance and underlined values refer to the worst performance.

0.3719) among traditional approaches, such as Bi-LSTM and Emotion-Detection-RNN, but still falls short in comparison to LLMs. The performance differential can be attributed to general-purpose pre-training and superior semantic reasoning capabilities of LLMs, which allow them to capture overall emotional states in text, including emotion category and intensity, more effectively than previous methods.

In the domain of MER, the Music2Emo model demonstrates promising performance in both static and dynamic emotions. For static prediction, Music2Emo achieves the highest category and intensity F_1 Scores (F_c = 0.3111, F_i = 0.3418), along with the lowest emotion distance (D_e = 0.3657). This superiority extends to the dynamic setting, where Music2Emo also leads the baseline models with the strongest F_c (0.3080), F_i (0.3159), and D_e (0.3670), showing its robust capability in continuous emotion recognition. Other models, such as MERT, M3BERT, UIBK-DBIS, and Mirable, vary across the evaluation metrics, with each model exhibiting strengths in particular aspects. The performance gap suggests that Music2Emo's architecture, which incorporates knowledge distillation within a multitask framework, is more adept at capturing the complex music features related to emotion category and intensity. Nonetheless, it is evident that the emotion distances for MER are considerably higher than those for LER, emphasizing the inherent challenges in abstracting emotional states from musical signals [116].

For the VER task, the results in Table IV illustrate a clear advantage of VLMs over specialized models. In the static evaluation, the Qwen2.5-VL-7B model demonstrates the best overall performance, attaining the highest F_1 Scores (F_c = 0.2647, F_i = 0.2949) and the lowest emotion distance (D_e = 0.3421). Notably, while other VLMs, such as InternVL3-8B and LLaVA-Video-7B, outperform the specialized CTEN model, the older VAANet model remains competitive, particularly in predicting emotion intensity (F_i = 0.2919), nearly matching Qwen2.5-VL-7B's performance. However, VAANet's relatively poor category prediction (F_c = 0.1790) has a negative impact on the emotion distance (D_e = 0.4102), despite its reasonable capability in intensity. In the dynamic evaluation, the results are more nuanced: LLaVA-Video-7B achieves the optimal emotion distance (D_e = 0.3553) and intensity score (F_i = 0.4085), while Qwen2.5-VL-7B obtains the highest category score (F_c = 0.2933). These results highlight the effectiveness of VLMs in capturing emotion expressions from visual information, facilitating more accurate emotion predictions.

B. Multimodal Emotion Recognition

1) Baselines and Settings: For multimodal emotion recognition, we select a range of MLLMs that are capable of processing text, audio, and video inputs, and generating text-based emotion predictions. These models include AnyGPT-7B [54], EMOVA-7B [55], VITA-1.5-7B [56], MiniCPM-o-2.6-8B [57], and Qwen2.5-Omni-7B [58]. To guarantee a fair comparison, we use the official pretrained weights of each model and provide the corresponding multimodal inputs for fine-tuning and inference. The models are required to output the predicted emotion category and intensity in a fixed format. As with the unimodal tasks, the static emotion recognition task is evaluated on the entire dataset, while dynamic emotion recognition is performed on a randomly selected subset of

TABLE V
RESULTS OF BASELINE MODELS UNDER THE DIMENSIONAL (V–A) EMOTION REPRESENTATION AND THE UNIFIED EMOTION REPRESENTATION

Model	Dimensional (V–A)		Unified	
	D_e (S) ↓	D_e (D) ↓	D_e (S) ↓	D_e (D) ↓
Llama-3.1-8B [112]	0.4091	-	**0.3358**	-
Qwen3-8B [92]	0.4203	-	**0.3472**	-
MERT [7]	0.3801	**0.3899**	**0.3766**	0.3908
Music2Emo [8]	**0.3207**	**0.3421**	0.3379	0.3502
InternVL3-8B [11]	0.4073	0.4162	**0.3714**	**0.3805**
Qwen2.5-VL-7B [12]	0.3889	0.3867	**0.3601**	**0.3795**
AnyGPT-7B [54]	0.3491	**0.3640**	**0.3372**	0.3714
MiniCPM-o-2.6-8B [57]	0.3804	0.3667	**0.3683**	**0.3428**

S refers to static emotion. D refers to dynamic emotion.

1000 samples. The dataset split also follows the same procedure as described in Section VI-A1, with training, validation, and test sets divided at a ratio of 8:1:1.

2) Results and Discussion: The experimental results for static and dynamic multimodal emotion recognition are presented in Table IV. In the static emotion recognition task, AnyGPT-7B demonstrates superior performance, achieving the highest F_1 Scores (F_c = 0.3191, F_i = 0.5073) and the lowest emotion distance (D_e = 0.2952). The performance can be attributed to AnyGPT-7B's extensive pre-training on diverse audio music datasets, allowing it to effectively capture the rich acoustic cues inherent in the music clips. In the dynamic emotion recognition task, the performance varies among different baseline models. VITA-1.5-7B, which underperforms in the static setting (F_c = 0.2059, D_e = 0.4659), emerges as the leading model for dynamic prediction. It obtains the highest category score (F_c = 0.3274) and a remarkable emotion distance (D_e = 0.3490), highlighting its architectural optimization for real-time visual and speech interaction. Meanwhile, MiniCPM-o-2.6-8B maintains high stability (D_e of 0.3729 in static and 0.3401 in dynamic) across both tasks, showing a balanced capability in processing global and temporal information.

C. Method Analysis

We analyze the effectiveness of the designed unified emotion representation through systematic comparisons across lyrics, music, video, and multimodal emotion recognition tasks. For each modality, we select two well-performing baseline models: Llama-3.1-8B and Qwen3-8B for lyrics, MERT and Music2Emo for music, InternVL3-8B and Qwen2.5-VL-7B for video, and AnyGPT-7B and MiniCPM-o-2.6-8B for multimodal tasks. The experiments are conducted on datasets that originally employ the dimensional (V-A) emotion representation: MER Lyrics Dataset for lyrics, PMEmo for music, and MuVi for video and multimodal tasks. For each dataset, models are evaluated using both the dimensional (V-A) representation and our proposed unified emotion representation, with performance measured by the D_e metric for static and dynamic emotion recognition.

The results presented in Table V reveal distinct performance differences of the proposed representation method across different model architectures. Our unified emotion representation demonstrates superior effectiveness when applied to language models, including LLMs for lyrics, VLMs for video, and MLMs for multimodal tasks. We attribute this improvement to the emotional interpretability of our representation. Unlike the abstract continuous values of the V-A model, the explicit definitions of emotion category and intensity provide concrete semantic foundations, allowing language models to leverage their pre-trained reasoning capabilities to understand and predict emotional states more accurately. Additionally, regarding traditional task-specific models in music emotion recognition, our unified representation achieves performance comparable to the V–A model. The results show that our representation generally outperforms in static emotion recognition, while the V–A model maintains a slight advantage in dynamic emotion recognition, due to its continuous characteristics, which is inherently better suited for capturing fine-grained temporal variations in emotion.

VII. CONCLUSION

In this paper, we propose a unified emotion representation that consists of emotion category and intensity to bridge the gap between categorical and dimensional labels. Based on the representation, we build an emotion annotation framework MVAnno, which adopts a hierarchical continual fine-tuning process on multiple modalities to guarantee accurate emotion annotations. We also construct a large-scale music video emotion dataset MVEmo with both static and dynamic emotion labels. Building on these efforts, we introduce MVEmo-Bench, a comprehensive evaluation benchmark that contains typical MRER tasks and specialized metrics to assess the performance of baseline models. In the future, we plan to extend our MVEmo-Bench to include more diverse multimodal emotion understanding and generation tasks, and explore the application of MVEmo to emotion-aware lyrics, music and video generation.

REFERENCES

[1] L. B. Meyer, *Emotion and meaning in music*. University of chicago Press, 2024.

[2] A. D'Ausilio, G. Novembre, L. Fadiga, and P. E. Keller, "What can music tell us about social interaction?" *Trends in cognitive sciences*, vol. 19, no. 3, pp. 111–114, 2015.

[3] A. Seyeditabari, N. Tabari, S. Gholizadeh, and W. Zadrozny, "Emotion detection in text: focusing on latent representation," *arXiv preprint arXiv:1907.09369*, 2019.

[4] A. JIDDY, A. IBNU, and F. A. W. YANUAR, "Emotion classification of song lyrics using bidirectional lstm method with glove word representation weighting," *JURNAL RESTI (REKAYASA SISTEM DAN TEKNOLOGI INFORMASI)*, vol. 4, no. 4, pp. 723–729, 2020.

[5] F. Bianchi, D. Nozza, and D. Hovy, "Xlm-emo: Multilingual emotion prediction in social media text," in *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 2022, pp. 195–203.

[6] T. Greer, X. Shi, B. Ma, and S. Narayanan, "Creating musical features using multi-faceted, multi-task encoders based on transformers," *Scientific Reports*, vol. 13, no. 1, p. 10713, 2023.

[7] L. Yizhi, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos *et al.*, "Mert: Acoustic music understanding model with large-scale self-supervised training," in *The Twelfth International Conference on Learning Representations*, 2023.

[8] J. Kang and D. Herremans, "Towards unified music emotion recognition across dimensional and categorical models," *arXiv preprint arXiv:2502.03979*, 2025.

[9] S. Zhao, Y. Ma, Y. Gu, J. Yang, T. Xing, P. Xu, R. Hu, H. Chai, and K. Keutzer, "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 01, 2020, pp. 303–311.

[10] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li, "Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models," *arXiv preprint arXiv:2407.07895*, 2024.

[11] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao *et al.*, "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025.

[12] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[13] Z. Fu, Z. Zhang, J. Zheng, K. Lin, and D. Li, "Eamr: An emotion-aware music recommender method via mel spectrogram and arousal-valence model," in *2022 International Conference on Frontiers of Artificial Intelligence and Machine Learning (FAIML)*. IEEE, 2022, pp. 57–64.

[14] H. Tran, T. Le, A. Do, T. Vu, S. Bogaerts, and B. Howard, "Emotion-aware music recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 13, 2023, pp. 16 087–16 095.

[15] D. Makris, K. R. Agres, and D. Herremans, "Generating lead sheets with affect: A novel conditional seq2seq framework," in *2021 international joint conference on neural networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[16] J. Kang, S. Poria, and D. Herremans, "Video2music: Suitable music generation from videos using an affective multimodal transformer model," *Expert Systems with Applications*, vol. 249, p. 123640, 2024.

[17] T. H. Zhou, W. Liang, H. Liu, L. Wang, K. H. Ryu, and K. W. Nam, "Eeg emotion recognition applied to the effect analysis of music on emotion changes in psychological healthcare," *International Journal of Environmental Research and Public Health*, vol. 20, no. 1, p. 378, 2022.

[18] Z. Qiu, R. Yuan, W. Xue, and Y. Jin, "Generated therapeutic music based on the iso principle," in *Summit on Music Intelligence*. Springer, 2023, pp. 32–45.

[19] E. Çano and M. Morisio, "Moodylyrics: A sentiment annotated lyrics dataset," in *Proceedings of the 2017 international conference on intelligent systems, metaheuristics & swarm intelligence*, 2017, pp. 118–124.

[20] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva, "Emotionally-relevant features for classification and regression of music lyrics," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 240–254, 2016.

[21] P. Tovstogan, D. Bogdanov, and A. Porter, "Mediaeval 2021: Emotion and theme recognition in music using jamendo." in *MediaEval*, 2021.

[22] A. Aljanaki, F. Wiering, and R. Veltkamp, "Dataset on induced musical emotion from game with a purpose emotify," *Technical report, Institute for Logic, Language and Computation (ILLC)*, 2015.

[23] E. Choi, Y. Chung, S. Lee, J. Jeon, T. Kwon, and J. Nam, "Ym2413-mdb: A multi-instrumental fm video game music dataset with emotion annotations," *arXiv preprint arXiv:2211.07131*, 2022.

[24] R. Panda, R. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 614–626, 2018.

[25] H.-T. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y.-H. Yang, "Emopia: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation," *arXiv preprint arXiv:2108.01374*, 2021.

[26] J. Fan, M. Thorogood, and P. Pasquier, "Emo-soundscapes: A dataset for soundscape emotion recognition," in *2017 Seventh international conference on affective computing and intelligent interaction (ACII)*. IEEE, 2017, pp. 196–201.

[27] J. Fan, Y.-H. Yang, K. Dong, and P. Pasquier, "A comparative study of western and chinese classical music based on soundscape models," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 521–525.

[28] M. Pesek, P. Godec, M. Poredos, G. Strle, J. Guna, E. Stojmenova, M. Pogacnik, and M. Marolt, "Introducing a dataset of emotional and color responses to music." in *ISMIR*, 2014, pp. 355–360.

[29] D. Bogdanov, X. Lizarraga Seijas, P. Alonso-Jiménez, and X. Serra, "Musav: A dataset of relative arousal-valence annotations for validation of audio models," 2022.

[30] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, 2013, pp. 1–6.

[31] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PloS one*, vol. 12, no. 3, p. e0173392, 2017.

[32] B. Kurdi, S. Lozano, and M. R. Banaji, "Introducing the open affective standardized image set (oasis)," *Behavior research methods*, vol. 49, no. 2, pp. 457–470, 2017.

[33] L. Carretié, M. Tapia, S. López-Martín, and J. Albert, "Emomadrid: An emotional pictures database for affect research," *Motivation and Emotion*, vol. 43, no. 6, pp. 929–939, 2019.

[34] E. S. Dan-Glauser and K. R. Scherer, "The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance," *Behavior research methods*, vol. 43, no. 2, pp. 468–477, 2011.

[35] D. L. Crone, S. Bode, C. Murawski, and S. M. Laham, "The socio-moral image database (smid): A novel stimulus set for the study of social, moral and affective processes," *PloS one*, vol. 13, no. 1, p. e0190954, 2018.

[36] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proceedings of the national academy of sciences*, vol. 114, no. 38, pp. E7900–E7909, 2017.

[37] A. F. Ehmann[1], "The 2007 mirex audio mood classification task: Lessons learned," in *ISMIR 2008: Proceedings of the 9th International Conference of Music Information Retrieval*. Lulu. com, 2008, p. 462.

[38] P. L. Louro, H. Redinho, R. Santos, R. Malheiro, R. Panda, and R. P. Paiva, "Merge–a bimodal dataset for static music emotion recognition," *arXiv preprint arXiv:2407.06060*, 2024.

[39] K. Zhang, H. Zhang, S. Li, C. Yang, and L. Sun, "The pmemo dataset for music emotion recognition," in *Proceedings of the 2018 acm on international conference on multimedia retrieval*, 2018, pp. 135–142.

[40] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 2887–2905, 2021.

[41] Y. R. Pandeya, B. Bhattarai, and J. Lee, "Deep-learning-based multimodal emotion classification for music videos," *Sensors*, vol. 21, no. 14, p. 4927, 2021.

[42] H. T. P. Thao, G. Roig, and D. Herremans, "Emomv: Affective music-video correspondence learning datasets for classification and retrieval," *Information Fusion*, vol. 91, pp. 64–79, 2023.

[43] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[44] P. Chua, D. Makris, D. Herremans, G. Roig, and K. Agres, "Predicting emotion from music videos: exploring the relative contribution of visual and auditory information to affective responses," *arXiv preprint arXiv:2202.10453*, 2022.

[45] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[46] M. M. Bradley and P. J. Lang, "Affective norms for english words (anew): Instruction manual and affective ratings," Technical report C-1, the center for research in psychophysiology ..., Tech. Rep., 1999.

[47] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[48] S. Buechel and U. Hahn, "A flexible mapping scheme for discrete and dimensionalemotion representations: Evidence from textual stimuli," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 39, 2017.

[49] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2018, pp. 174–184.

[50] S. Park, J. Kim, S. Ye, J. Jeon, H. Y. Park, and A. Oh, "Dimensional emotion detection from categorical emotion," in *Proceedings of the 2021 conference on empirical methods in natural language processing*, 2021, pp. 4367–4380.

[51] M. Trnka, S. Darjaa, M. Ritomskỳ, R. Sabo, M. Rusko, M. Schaper, and T. H. Stelkens-Kobsch, "Mapping discrete emotions in the dimensional space: An acoustic approach," *Electronics*, vol. 10, no. 23, p. 2950, 2021.

[52] S. M. Mohammad, "Nrc vad lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms," *arXiv preprint arXiv:2503.23547*, 2025.

[53] X. Wu, J. Wang, J. Yu, T. Zhang, and K. Zhang, "Popular hooks: A multimodal dataset of musical hooks for music understanding and

generation," in *2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 2024, pp. 1–6.

[54] J. Zhan, J. Dai, J. Ye, Y. Zhou, D. Zhang, Z. Liu, X. Zhang, R. Yuan, G. Zhang, L. Li *et al.*, "Anygpt: Unified multimodal llm with discrete sequence modeling," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 9637–9662.

[55] K. Chen, Y. Gou, R. Huang, Z. Liu, D. Tan, J. Xu, C. Wang, Y. Zhu, Y. Zeng, K. Yang *et al.*, "Emova: Empowering language models to see, hear and speak with vivid emotions," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5455–5466.

[56] C. Fu, H. Lin, X. Wang, Y.-F. Zhang, Y. Shen, X. Liu, H. Cao, Z. Long, H. Gao, K. Li *et al.*, "Vita-1.5: Towards gpt-4o level real-time vision and speech interaction," *arXiv preprint arXiv:2501.01957*, 2025.

[57] O. M.-o. Team, "Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone," 2025.

[58] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang *et al.*, "Qwen2. 5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[59] A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, N. Bach, J. Bao, A. Benhaim, M. Cai, V. Chaudhary, C. Chen *et al.*, "Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras," *arXiv preprint arXiv:2503.01743*, 2025.

[60] T. Takahashi and M. Barthet, "Emotion-driven harmonisation and tempo arrangement of melodies using transfer learning." in *ISMIR*, 2022, pp. 741–748.

[61] R. Paskaleva, M. Holubakha, A. Ilic, S. Motamed, L. Van Gool, and D. Paudel, "A unified and interpretable emotion representation and expression generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2447–2456.

[62] R. Liu, A. Roy, and D. Herremans, "Leveraging llm embeddings for cross dataset label alignment and zero shot music emotion prediction," *arXiv preprint arXiv:2410.11522*, 2024.

[63] S. Cazzaniga, F. Gasparini, A. Saibene *et al.*, "A multi-source deep learning model for music emotion recognition," in *CEUR WORKSHOP PROCEEDINGS*, vol. 3903. CEUR-WS, 2024, pp. 33–43.

[64] R. Malheiro, R. Panda, P. Gomes, and R. Paiva, "Bi-modal music emotion recognition: Novel lyrical features and dataset." 9th International Workshop on Music and Machine Learning–MML'2016–in …, 2016.

[65] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[66] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of music*, vol. 39, no. 1, pp. 18–49, 2011.

[67] P. N. Juslin and P. Laukka, "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening," *Journal of new music research*, vol. 33, no. 3, pp. 217–238, 2004.

[68] Y. Hu, X. Chen, and D. Yang, "Lyric-based song emotion detection with affective lexicon and fuzzy clustering method." in *ISMIR*, 2009, pp. 123–128.

[69] X. Hu and J. S. Downie, "When lyrics outperform audio for music mood classification: A feature analysis." in *Ismir*, 2010, pp. 619–624.

[70] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International journal of machine learning and cybernetics*, vol. 1, no. 1, pp. 43–52, 2010.

[71] S. Qaiser and R. Ali, "Text mining: use of tf-idf to examine the relevance of words to documents," *International journal of computer applications*, vol. 181, no. 1, pp. 25–29, 2018.

[72] Y. Goldberg and O. Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[73] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[74] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[75] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, "Music mood detection based on audio and lyrics with deep neural net," *arXiv preprint arXiv:1809.07276*, 2018.

[76] Y. Agrawal, R. G. R. Shanker, and V. Alluri, "Transformer-based approach towards music emotion recognition from lyrics," in *European conference on information retrieval*. Springer, 2021, pp. 167–175.

[77] S. H. Muhammad, N. Ousidhoum, I. Abdulmumin, S. M. Yimam, J. P. Wahle, T. L. Ruas, M. Beloucif, C. De Kock, T. D. Belay, I. S. Ahmad *et al.*, "Semeval-2025 task 11: Bridging the gap in text-based emotion detection," in *Proceedings of the 19th international workshop on semantic evaluation (SemEval-2025)*, 2025, pp. 2558–2569.

[78] J. Yu, S. Wu, G. Lu, Z. Li, L. Zhou, and K. Zhang, "Suno: potential, prospects, and trends," *Frontiers of Information Technology & Electronic Engineering*, vol. 25, no. 7, pp. 1025–1030, 2024.

[79] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling." in *Ismir*, vol. 270. Plymouth, MA, 2000, pp. 1–11.

[80] B. Logan and A. Salomon, "A music similarity function based on signal analysis." in *ICME*, vol. 1, 2001, pp. 745–748.

[81] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[82] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[83] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[84] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 2392–2396.

[85] W. Chen *et al.*, "A novel long short-term memory network model for multimodal music emotion analysis in affective computing," *Journal of Applied Science and Engineering*, vol. 26, no. 3, pp. 367–376, 2022.

[86] N. Niu, "Music emotion recognition model using gated recurrent unit networks and multi-feature extraction," *Mobile Information Systems*, vol. 2022, no. 1, p. 5732687, 2022.

[87] Y. Yi, J. Zhou, H. Wang, P. Tang, and M. Wang, "Emotion recognition in user-generated videos with long-range correlation-aware network," *IET Image Processing*, vol. 18, no. 12, pp. 3288–3301, 2024.

[88] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li, "Video instruction tuning with synthetic data," *arXiv preprint arXiv:2410.02713*, 2024.

[89] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant." *Journal of personality and social psychology*, vol. 76, no. 5, p. 805, 1999.

[90] R. T. Ross, "A statistic for circular series." *Journal of Educational Psychology*, vol. 29, no. 5, p. 384, 1938.

[91] Z. Wang, L. Ma, C. Zhang, B. Han, Y. Xu, Y. Wang, X. Chen, H. Hong, W. Liu, X. Wu *et al.*, "Remast: Real-time emotion-based music arrangement with soft transition," *IEEE Transactions on Affective Computing*, 2024.

[92] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[93] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[94] J. Ainslie, J. Lee-Thorp, M. De Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai, "Gqa: Training generalized multi-query transformer models from multi-head checkpoints," *arXiv preprint arXiv:2305.13245*, 2023.

[95] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

[96] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.

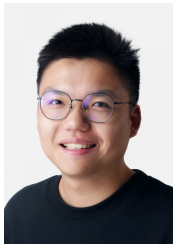
[97] A. Wu, Y. Yang, X. Luo, Y. Yang, C. Wang, L. Hu, X. Dai, D. Chen, C. Luo, L. Qiu *et al.*, "Llm2clip: Powerful language model unlock richer visual representation," in *NeurIPS 2024 Workshop: Self-Supervised Learning-Theory and Practice*.

[98] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[99] Y.-D. Tsai, T.-Y. Yen, P.-F. Guo, Z.-Y. Li, and S.-D. Lin, "Text-centric alignment for multi-modality learning," *arXiv preprint arXiv:2402.08086*, 2024.

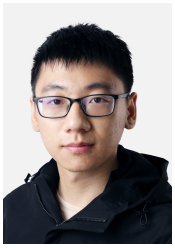
[100] Y.-D. Tsai, T.-Y. Yen, K.-T. Liao, and S.-D. Lin, "Enhance modality robustness in text-centric multimodal alignment with adversarial prompting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 26, 2025, pp. 27 740–27 747.

[101] Z. Luo, N. Liu, X. Yang, S. Khan, R. M. Anwer, H. Cholakkal, F. S. Khan, and J. Han, "Tavis: Text-bridged audio-visual segmentation with foundation models," *arXiv preprint arXiv:2506.11436*, 2025.
[102] J. Yu, X. Wu, Y. Xu, T. Zhang, S. Wu, L. Ma, and K. Zhang, "Songglm: Lyric-to-melody generation with 2d alignment encoding and multi-task pre-training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 24, 2025, pp. 25 742–25 750.
[103] Y. Zou, P. Zou, Y. Zhao, K. Zhang, R. Zhang, and X. Wang, "Melons: generating melody with long-term structure using transformers and structure graph," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 191–195.
[104] S. Wu, D. Yu, X. Tan, and M. Sun, "Clamp: Contrastive language-music pre-training for cross-modal symbolic music information retrieval," *arXiv preprint arXiv:2304.11029*, 2023.
[105] X. Wu, Z. Huang, K. Zhang, J. Yu, X. Tan, T. Zhang, Z. Wang, and L. Sun, "Melodyglm: Multi-task pre-training for symbolic melody generation," *arXiv preprint arXiv:2309.10738*, 2023.
[106] J. Wu, N. Zhang, C. Zhong, B. Chen, H. Liu, and J. Yan, "Melody structure transfer network: Generating music with separable self-attention," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
[107] E. Parada-Cabaleiro, M. Mayerl, S. Brandl, M. Skowron, M. Schedl, E. Lex, and E. Zangerle, "Song lyrics have become simpler and more repetitive over the last five decades," *Scientific Reports*, vol. 14, no. 1, p. 5531, 2024.
[108] S. Sural, G. Qian, and S. Pramanik, "Segmentation and histogram generation using the hsv color space for image retrieval," in *Proceedings. international conference on image processing*, vol. 2. IEEE, 2002, pp. II–II.
[109] L.-C. Yang and A. Lerch, "On the evaluation of generative models in music," *Neural Computing and Applications*, vol. 32, no. 9, pp. 4773–4784, 2020.
[110] Z. Lian, H. Chen, L. Chen, H. Sun, L. Sun, Y. Ren, Z. Cheng, B. Liu, R. Liu, X. Peng *et al.*, "Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models," *arXiv preprint arXiv:2501.16566*, 2025.
[111] A. Yang, B. Xiao, B. Wang, B. Zhang, C. Bian, C. Yin, C. Lv, D. Pan, D. Wang, D. Yan *et al.*, "Baichuan 2: Open large-scale language models," *arXiv preprint arXiv:2309.10305*, 2023.
[112] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.
[113] H. H. Tan, "Semi-supervised music emotion recognition using noisy student training and harmonic pitch class profiles," *arXiv preprint arXiv:2112.00702*, 2021.
[114] M. Mayerl, M. Vötter, A. Peintner, G. Specht, and E. Zangerle, "Recognizing song mood and theme: Clustering-based ensembles." in *MediaEval*, 2021.
[115] Z. Zhang, L. Wang, and J. Yang, "Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18 888–18 897.
[116] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proc. ismir*, vol. 86, 2010, pp. 937–952.



Jiaxing Yu received the BS degree in Computer Science and Technology from Zhejiang University, Hangzhou, China, in 2022. He is currently working toward the PhD degree at the College of Computer Science and Technology, Zhejiang University, Hangzhou, China. His research interests include affective computing, speech and audio processing, and multimodal music understanding and generation.

Ziyi Huang received the BS degree in Computer Science and Technology from Nanjing University of Science and Technology, Nanjing, China, in 2024. She is currently working toward the PhD degree at the College of Computer Science and Technology, Zhejiang University, Hangzhou, China. Her research interests include computational music generation, music–media alignment, and interactive music systems.



Shuyu Li received the BS degree in Artificial Intelligence from Xi'an Jiaotong University, Xi'an, China, in 2024. He is currently working toward the PhD degree at the College of Computer Science and Technology, Zhejiang University, Hangzhou, China. His research interests include applications of deep learning models, intelligent music systems, and multimodal algorithms.

Songruoyao Wu received the BS degree in Design from Zhejiang University, Hangzhou, China, in 2020. She is currently working toward the PhD degree at the College of Computer Science and Technology, Zhejiang University, Hangzhou, China. Her research interests include AI-aided art creation and human-computer interaction.

Shulei Ji received the PhD degree in Computer Science and Technology from Xi'an Jiaotong University, Xi'an, China, in 2024. She is currently a postdoctoral researcher at the College of Computer Science and Technology, Zhejiang University, Hangzhou, China and Innovation Center of Yangtze River Delta, Zhejiang University, Jiaxing 314100, China. Her research interests include AI music generation, music emotion recognition, and multimodal learning.

Kejun Zhang (Member, IEEE) received the PhD degree in Computer Science and Technology from Zhejiang University, Hangzhou, China. He is currently a professor at the College of Computer Science and Technology, Zhejiang University, Hangzhou, China. His research interests include affective computing, music information retrieval, artificial intelligence and machine learning.