A Tasks with Automatic Annotation

The performance of current algorithms designed for annotating textual description, lyrics, and musical section annotation is not satisfactory due to their reliance on subjective human evaluation. Except these, other kinds of data, such as phonetic alignment, vocal separation, and audio-to-MIDI conversion, do not significantly align with human perception. Annotating these elements manually is particularly challenging, requiring extensive effort and time. However, there are now numerous advanced algorithms that efficiently handle these tasks, as detailed in the **Appendix B**. As a result, we utilize data preprocessing algorithms for the automatic annotation of this content, eliminating the need for manual annotation or intervention, and seamlessly integrate this processed content into our dataset.

Classification	Task
A	Musical Section Annotation
	Lyric Correction
	Lyric Screening
	Rhyme Annotation
B	Professional Music Description
	Amateur Music Description

Table 3: Classification of Annotation Tasks

B Data Preprocessing

- **Music Genre Clustering** To mitigate subjective bias and ensure diverse descriptions across various music genres, it's crucial to distribute a broad spectrum of music genres among annotators, thereby enriching the annotation's diversity. To facilitate this, we utilize MERT [Li *et al.*, 2023], a pre-trained music audio encoder, to process the audio data. Following this, we cluster the encoded data, resulting in 1000 unique audio clusters. From these clusters, we evenly distribute music data, guaranteeing that annotators are presented with a balanced mix of music for labeling. This approach ensures that each music cluster is described by a range of annotators, significantly enhancing the diversity and richness of the annotated data.
- **Vocal & Track Separation** To make the dataset suitable for tasks such as accompaniment generation, melody generation, and vocal synthesis, we apply Demucs [Rouard *et al.*, 2023; Défossez, 2021] to perform vocal separation, separating the vocals from the musical accompaniment in the audio files. Furthermore, considering the requirements of a wider range of music-related tasks, we also separate individual instrument tracks, such as drums and bass.
- **Phonemic Level Alignment in Audio-Lyrics** To prepare audio-lyrics pairs for applications such as vocal synthesis, it's necessary to align them at the phonemic level. We employ the Montreal Forced Aligner (MFA) [McAuliffe *et al.*, 2017] for this task, initially achieving a 67% accuracy rate. While the MFA demonstrates a commendable 95% accuracy for aligning monophonic phonemes to single characters, its performance drops due to inaccuracies in marking the offsets for melismatic phonemes. These phonemes are characterized by multiple pitches sung within a single syllable or note, complicating the alignment process and diminishing the overall accuracy. To address this, we optimized the MFA algorithm with a focus on accurately identifying and aligning melismatic phonemes. Furthermore, we implemented features to recognize and annotate significant pauses and breaths during singing. These enhancements significantly improve our final alignment accuracy to 97%.
- **Automatic Pre-annotation** To improve the efficiency of future manual annotations, we implemented specific software for automatic pre-annotation of certain tasks related to lyric annotation. For annotating rhyme schemes in lyrics, we use a specialized program that pre-annotates the rhyme scheme for each line. For theme annotation in lyrics, we employ a fine-tuned version of Qwen to preliminarily identify the main theme of the lyrics for each piece of music. During the formal annotation phase, these pre-annotations serve as a basis for manual review. Annotators can assess the accuracy of these automatic annotations and adjust them as necessary, or use them as a guideline for their own annotation efforts.
- **Lead Sheet Transcription** To facilitate symbolic music-related tasks using MIDI, we transcribe the audio in the MuChin into lead sheets. These sheets, which are a simplified form of MIDI notation, are created using Sheet Sage [Donahue and Liang, 2021], software that utilizes the encoding model of Jukebox [Dhariwal *et al.*, 2020]. This conversion facilitates the application of MuChin to a wide range of tasks associated with symbolic music.

Classification	Accuracy(%)
I	[90, 100]
II	[70, 90)
III	[60, 70)
IV	[0, 60)

Table 4: Classification of Annotators of Type A Tasks Based on Accuracy

Classification	Score
I	[90, 100]
II	[70, 90)
III	[60, 70)
IV	[0, 60)

Table 5: Classification of Annotators of Type B Tasks Based on Score

C Quality Assurance Mechanisms

In this section, we will provide a detailed introduction to the quality assurance mechanism, including the classification of tasks, scoring guidelines and the classification of individuals.

Dimension	Score	Standard
Expressive Impact (S. & A.)	13	4 for Number of Labels; 4 for Label Relevance; 5 for Innovation
Emotional Impact	13	4 for Number of Labels; 4 for Label Relevance; 5 for Innovation
Textual Description	8	3 for Description Relevance; 5 for Word Counts and Innovation
Musical Genres	8	8 for Level of Detail
Tempo and Rhythm	5	5 for Label Relevance
Instrumentation	12	5 for Number of Labels; 3 for Label Relevance; 2 for Description Relevance; 2 for Description Thoroughness
Song Purpose	6	3 for Label Relevance; 3 for Innovation
Culture and Region	6	3 for Label Relevance; 3 for Innovation
Target Audience	6	3 for Label Relevance; 3 for Innovation
Vocal Components	12	5 for Number of Labels; 3 for Label Relevance; 2 for Description Relevance; 2 for Description Thoroughness
Audio Effects	5	5 for Label Relevance
Lyric Themes	6	3 for Label Relevance; 3 for Innovation
Total	100	-

Table 6: Scoring Guidelines of Professional Music Description

Dimension	Score	Standard
Perception of Uniqueness	8	4 for Label Relevance; 4 for Innovation
Perception of Tempo	5	3 for Label Relevance; 2 for Innovation
Expressive Impact (S.)	13	4 for Number of Labels; 4 for Label Relevance; 5 for Innovation
Emotional Impact (L.)	13	4 for Number of Labels; 4 for Label Relevance; 5 for Innovation
Textual Description	8	3 for Description Relevance; 5 for Word Counts and Innovation
Instrumentation	12	5 for Number of Labels; 3 for Label Relevance; 2 for Description Relevance; 2 for Description Thoroughness
Song Purpose	6	3 for Label Relevance; 3 for Innovation
Culture and Region	6	3 for Label Relevance; 3 for Innovation
Target Audience	6	3 for Label Relevance; 3 for Innovation
Vocal Components	12	5 for Number of Labels; 3 for Label Relevance; 2 for Description Relevance; 2 for Description Thoroughness
Audio Effects	5	5 for Label Relevance
Lyric Themes	6	3 for Label Relevance; 3 for Innovation
Total	100	-

Table 7: Scoring Guidelines of Amateur Music Description

C.1 Classification of Annotation Tasks

We classify annotation tasks into two categories based on their potential for objective evaluation: **Type A**, which can be objectively assessed, and **Type B**, which are subject to subjective assessment. This section exemplifies the classification of each annotation task. To maximize the accuracy and comprehensiveness of each song's annotations, we allocate two annotators to Type A tasks and one annotator to Type B tasks for each song. These tasks are carried out separately, not simultaneously. Additionally, apart from annotators, several quality assurance inspectors are needed to evaluate the annotators' outputs. According to the division into Type A and B, we consolidate Type A tasks into one phase, denoted as the **Structure Annotation Phase**, and Type B tasks into the subsequent phase, denoted as the **Music Description Annotation Phase**. Data must sequentially pass through these two phases before inclusion in the dataset. That is, data must undergo structure annotation and pass quality assurance before proceeding to the music description annotation phase, after which, data that passes quality assurance following music description annotation can be added to the dataset. For Type A tasks, if both annotators provide identical annotations, we consider the annotation accurate. However, when there is a discrepancy, quality assurance inspectors must deliver their judgment to determine which result is correct, or if both are incorrect, provide their own accurate annotation. For Type B tasks, quality assurance inspectors are required to assign a score ranging from 0 to 100 to the annotation results, with the scoring guidelines detailed in Table 6 and 7.

C.2 Classification of Individuals

To ensure diligent performance from annotators, we have implemented a screening mechanism. During the structural annotation phase, the precision of Type A task annotations is assessed through the previously mentioned quality assurance system. In the music description annotation phase, given that Type B tasks involve subjective descriptions challenging to assess objectively, we randomly review 20% of the annotations from each annotator for quality control. Moreover, we evaluate behaviors indicated by backend analytics, such as interaction frequency with the progress bar and task skipping. Annotators showing superficial engagement will be warned. In both phases, annotators are categorized into four groups based on their weekly accuracy rates or average scores, as detailed in Table 4 and 5. Type IV annotators, and those receiving two or more warnings, will be excluded from future tasks, and their data for the current week will be disregarded. Type I annotators will be rewarded, while Type III annotators may incur penalties.

C.3 Other Quality Assurance Measures

Annotators are responsible for screening the data (Type A & B). For songs that contain languages other than Chinese, have poor audio quality, or involve pornography or violence,

therefore unsuitable for inclusion in the dataset, annotators can mark these for exclusion and skip their annotation.

When annotating musical sections of Type A, annotators must repeatedly listen to a music piece. Consequently, the dedication to their annotation tasks is assessed by the amount of time they spend on the annotation page, their frequency of interactions with the progress bar, and the frequency of their play/pause button clicks.

In the textual description annotation (Type B), to ensure that annotators listen to each song attentively and provide thoughtful music descriptions, we stipulate that annotators must listen to the entire song in one sitting before adjusting the progress bar and playback speed. They must compose a textual description of no fewer than 50 words, and are prohibited from writing the description within the first 30 seconds of the song's playback, as well as from copying and pasting any content.

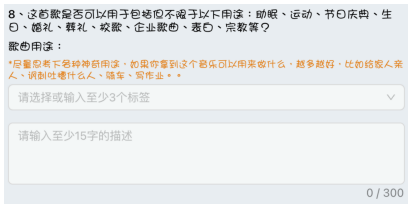


Figure 4: Supplementary actual screenshots from the main text. A screenshot of the 'Song Purpose' section during the Description Annotation Phase.

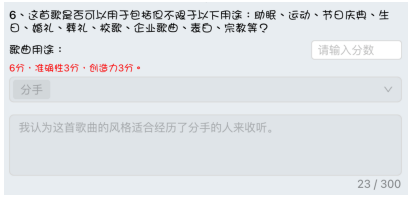


Figure 5: Supplementary actual screenshots from the main text. A screenshot of the 'Song Purpose' section during the Description Quality Assurance Phase.

Figure 6: Supplementary actual screenshots from the main text. A screenshot of the 'Instrumentation' section during the Description Annotation Phase.

D CaiMAP: Caichong Multitask Music Annotation Platform

In Appendix C, we have launched a comprehensive suite of annotation tasks alongside an advanced quality assurance system. To bring these complex designs to life, we developed the Caichong Multitask Music Annotation Platform (CaiMAP), which harmonizes this series of tasks and systems. This section will provide a brief overview of the platform.

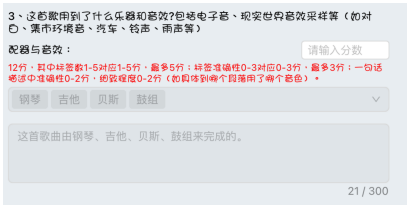


Figure 7: Supplementary actual screenshots from the main text. A screenshot of the 'Instrumentation' section during the Description Quality Assurance Phase.

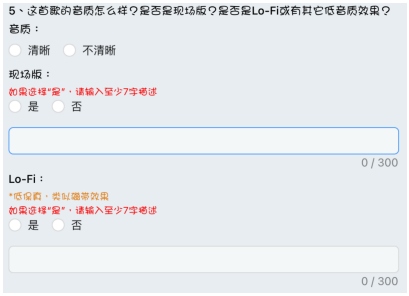


Figure 8: Supplementary actual screenshots from the main text. A screenshot of the 'Audio Effects' section during the Description Annotation Phase.

- **Account and Login.** The platform utilizes an access control system, assigning specific roles to each user account. Users can log into their accounts, review and complete assigned tasks, and submit their results.
- **Annotation Interface.** Upon logging in and selecting a specific piece of music, annotators are directed to a dedicated annotation interface designed for the task. This interface includes a media player and a specialized text box. Users have the ability to control the progress bar and playback speed of the media player. Furthermore, the music description annotation interface incorporates a comprehensive lexicon and search tool, enabling users to select suitable descriptive terms directly from the lexicon or to search for specific terms as needed.
- **Quality Assurance Interface.** Upon logging in and selecting a specific piece of music, quality assurance inspectors are taken to the quality assurance interface. For Type A tasks, inspectors are responsible for simultaneously evaluating the annotations provided by two users. The interface presents these annotations side-by-side, highlighting the differences for easy comparison. Inspectors can then decide which annotation is correct, make adjustments to either, or choose to re-annotate the piece. For Type B tasks, the interface displays a single, complete annotation for the inspector to verify and score. Inspectors simply review the annotation and submit their scores.
- **Administrator Interface.** Administrators have the access to view the submissions of any designated user, including annotators and quality assurance inspectors. Both the annotation and quality assurance interfaces in-

corporate a feedback button for reporting platform issues, enabling annotators and quality assurance inspectors to communicate with administrators for resolution.

We have provided screenshots of several platform pages as examples, as shown in Figures 4–8.

E Individual Grouping and Training

E.1 Grouping

During the structure annotation phase, which consists of Type A tasks, each piece of data requires two annotations. In contrast, the music description annotation phase, made up of Type B tasks, necessitates only one annotation. As a result, the latter phase involves fewer participants. The task of annotating the musical sections in the lyric annotation phase demands a basic knowledge of music theory. Consequently, only 104 professionals are engaged in this task. Out of these, 11 individuals, distinguished by their high level of expertise and conscientious approach, are chosen as quality assurance inspectors. This selection process involves screening their resumes and conducting further assessments. The remaining 93 individuals function as annotators.

During the music description annotation phase, the 109 amateurs form the amateur group, and the 93 professionals from the previous phase form the professional group. Additionally, the 11 inspectors from the previous phase continue to serve as inspectors in this phase. Beyond the roles of annotators and quality assurance inspectors, we also select a member from our research team who is adept at using the platform, with a high level of expertise, and with strong communication skills to act as the platform administrator.

E.2 Training

Next, we offer training for both the annotators and quality assurance inspectors, focusing on their specific roles. Initially, each annotator accesses CaiMAP to pre-annotate a compact dataset of around 20 entries, which encompasses tasks of both Type A and B. This phase allows annotators to acquaint themselves with the platform's features and learn the correct procedures for completing annotation tasks. Additionally, we provide specialized training to address common mistakes, such as the elimination of extraneous information from lyric texts and the accurate identification of each interjection.

On the other hand, training for inspectors entails a more intricate process. They must not only master the platform's use but also develop a set of consistent evaluation standards. We gather data annotated by the annotators during the pre-annotation phase and distribute the same dataset to all inspectors. For the lyric annotation phase, inspectors must choose the annotation they consider correct based on the guidelines outlined in **Section 2.2**, or provide an alternative correct annotation if they find the existing ones inaccurate. During the music description annotation phase, inspectors evaluate each annotation independently. Once the inspectors have completed their tasks, we compile all the scores for the music descriptions and organize a meeting with the inspectors. At this meeting, we identify instances where scores from different inspectors significantly vary, with a maximum discrepancy exceeding 10 points, and encourage inspectors to discuss and agree on a unified evaluation criterion. This training process is repeated until the inspectors' scores for the same dataset show substantial consistency.

(verse1)
晚上来临了
ccccR
游戏通关了
ccccR
我们的爱情早已结束了
ccccccccR
(verse2)
我的心碎了
ccccR
你也解脱了
ccccR
你就像只船开走了
cccccccR
(chorus1)
船要起航了
ccccR
你要出发了
ccccR
到处漂泊的你也许会累了
ccccR

Figure 9: A Fragment from an Illustrative Example of Structure Annotation

Main Question: 这首歌带给你的感受?
Main Question: How does this song make you feel?

Label Selection
Q1: 特色感受
Q1: Perception of Uniqueness
A1: "情歌","青春","积极面对","动听"
A1: "Love song," "Youth," "Face positively," "Melodious."
Q2: 快慢感受
Q2: Perception of Tempo
A2: "欢快","踩点","跟着哼唱"
A2: "Cheerful," "On beat," "Hum along."
Q3: 表现力感受（歌手）
Q3: Expressive Impact (Singer)
A3: "感悟","情深意切","余音袅袅","动情"
A3: "Insight," "Deep emotion," "Lingering sound," "Moving."
Q4: 情绪感受（歌词）
Q4: Emotional Impact (Lyrics)
A4: "成长","追忆","愉悦","释然"
A4: "Growth," "Reminiscence," "Joy," "Relief."

Compose Description
A: 这是一首正能量的歌曲，在成长中难免会遇到困难挫折，克服它们继续向前，向着人生的目标奔跑，要无所畏惧。
A: This is a positive song that acknowledges the inevitable difficulties and setbacks encountered during growth. It encourages overcoming these obstacles and continuing to move forward fearlessly towards the goals of life.

Figure 10: A Fragment from an Illustrative Example of Amateur Description Annotation

F Caichong Music Dataset

F.1 Annotated Data Processing

On one hand, we seamlessly incorporate annotations of musical sections into the lyrics by marking the start of each musical section with a section label, positioned before the lyrics

Main Question: 这首歌带给你的感受?
Main Question: How does this song make you feel?

Label Selection
Q1: 表现力感受（歌手和伴奏）
Q1: Expressive Impact (Singer & Accompaniment)
A1: "美妙的声音","温暖磁性的男声","完美","柔和的男声"
A1: "Exquisite voice," "Warm and magnetic male voice," "Perfection," "Gentle male voice."
Q2: 情绪感受（歌词和旋律）
Q2: Emotional Impact (Lyrics & Melody)
A2: "感性的","放松和安慰","浪漫情感"
A2: "Sentimental," "Relaxing and comforting," "Romantic emotions."

Compose Description
A: 这首歌整体表现了较为感性的情感主题，描述了两人的陪伴和旅行方面的故事。歌手的表现力很棒，情感充沛，技法高超，很快能把人带入到一个舒缓平静的氛围中。
A: The song overall conveys a sentimentally charged emotional theme, depicting the story of companionship and travel between two individuals. The singer's performance is impressive, brimming with emotion and technical proficiency, swiftly drawing the listener into a soothing and tranquil atmosphere.

Figure 11: A Fragment from an Illustrative Example of Professional Description Annotation

of that section begin. We denote rhyming information using strings that include 'c' and 'R' markers: an 'R' is added at the end of any sentence that rhymes with the one before it, while 'c' indicates words that do not rhyme. This method is used to compile all annotated lyric information—encompassing the lyrics' theme, musical sections, and rhyming details—into a JSON file.

On the other hand, during the phase dedicated to annotating music descriptions, we collect textual descriptions of each music piece from various perspectives. Each annotation consists of several descriptive terms along with a comprehensive descriptive text. To enhance the richness of these descriptions, we integrate these terms into the textual descriptions, which are then combined with the texts. Furthermore, we concatenate descriptions from different aspects to create a single, detailed annotation that captures the multifaceted nature of the music.

F.2 Overview

This section provides an overview of the descriptive tag distribution and song structure distribution in CaiMD, as illustrated in Figure 12–14. Song structure is the arrangement of musical sections.

F.3 Examples

This section presents a range of annotation examples, encompassing both professional and colloquial musical descriptions, along with the musical sections and rhymes featured in CaiMD, as depicted in Figures 9–11.

G Evaluation Metrics of Structured Lyric Generation

G.1 Formula

The similarity of the overall structure and musical section structure is calculated according to Equation 1, where K_m represents the number of matching characters in the longest

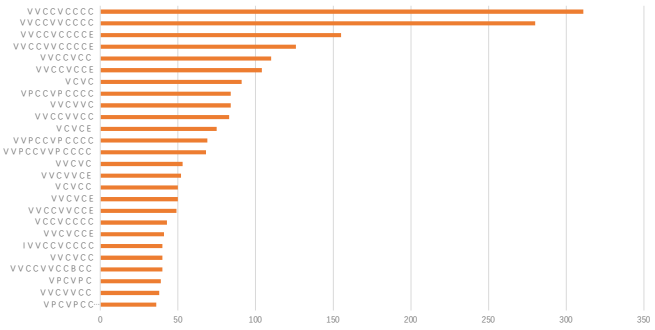


Figure 12: Distribution of Song Structures. The bin labels on the left side of the histogram represent the various musical sections of a song. Specifically, 'i' stands for "Introduction," 'v' corresponds to "Verse," 'c' denotes "Chorus," 'p' indicates "Pre-chorus," 'b' signifies "Bridge," and 'e' represents the "Ending."

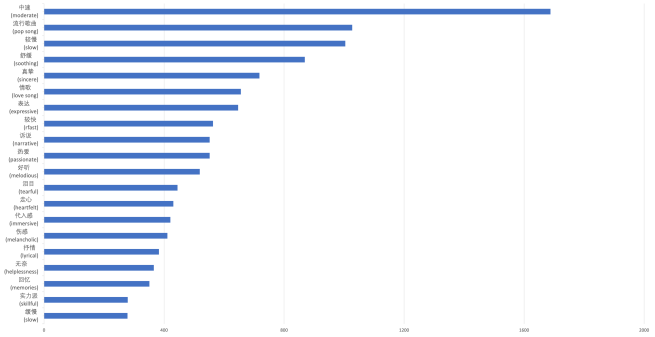


Figure 13: Distribution of Colloquial Descriptive Tags

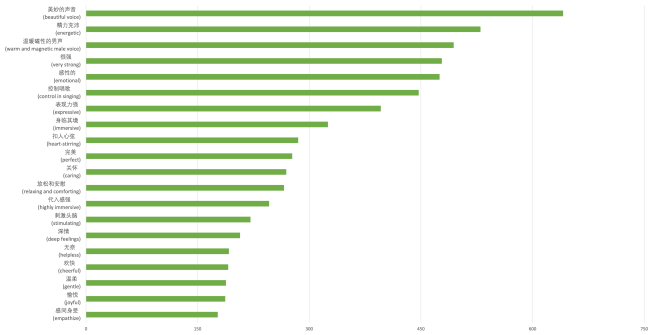


Figure 14: Distribution of Professional Descriptive Tags

common subsequence between strings A and B. L_A denotes the length of string A, and L_B denotes the length of string B. In the context of the overall structure, A and B represent the entire set of lyrics. In the context of musical section structure, A and B refer to the sequence of musical section labels.

$$p = \frac{2K_m}{L_A + L_B} \tag{1}$$

The within-section structure similarity is calculated according to Equation 2. In this equation, each element of $ListA$ and $ListB$ represents the number of sentences contained in each matching musical section of song A and B, respectively, e.g., [4, 8, 4] indicates that the three matching musical sections contain 4, 8, and 4 sentences, respectively.

Algorithm 1 Reward Score Algorithm

Input: max_equ_slc_sum, rc_ing, acmp_sr, rc_ino
Parameter: EXTRA_POINTS
Output: extscore

```
if max_equ_slc_sum == 0 then
    r_ratio = 0
else
    r_ratio = rc_ing / max_equ_slc_sum
end if
extscore = EXTRA_POINTS * acmp_sr
if 0.6 <= r_ratio and r_ratio <= 0.8 then
    extscore *= 1.0
else if rc_ino == rc_ing and rc_ino > 0 then
    extscore *= 0.7
else
    r_delta = |r_ratio - 0.7|
    if r_delta <= 0.3 then
        extscore *= 0.4 * (1 - r_delta)
    else
        extscore *= 0.0
    end if
end if
return extscore
```

$$p = \frac{2\sum \min(ListA, ListB)}{\sum ListA + \sum ListB} \tag{2}$$

Similarly, the within-sentence structure similarity can also be calculated using Equation 2. In this calculation, each element of $ListA$ and $ListB$ represents the number of words in each matching sentence of songs A and B.

The calculation of rhyming similarity follows Equation 1, where K_m represents the number of sentences that contain rhyming markers in the lyrics, and L_A and L_B respectively represent the total number of sentences in songs A and B.

Since each more detailed structure depends on the match of the preceding structure, cumulative similarity is used when calculating similarity, to take into account the influence of more macroscopic structures on the similarity of more microscopic structures. With the similarities of the overall structure, musical section structure, within-section structure, within-sentence structure, and rhyming structure calculated as p_1 to p_5 respectively, and their corresponding weights in the overall scoring as w_1 to w_5, the overall similarity can be calculated using Equation 3.

$$p = \sum_{i=1}^{5} w_i \prod_{j=1}^{i} p_j \tag{3}$$

Multiplying the overall similarity by 100 gives the overall score. Additionally, the extra reward score based on the proportion of rhyming sentences within the overall lyrics is also incorporated into the overall score.

G.2 Reward Score

The calculation method of the reward score is shown as Algorithm 1, by which generated lyrics are assigned a certain amount of reward points based on the proportion of rhyming. In this algorithm, max_equ_slc_sum denotes the maximum number of phrases that match; rc_ing denotes the number of rhyming phrases that match; acmp_sr denotes the cumulative product of similarities across the first 5 dimensions; rc_ino denotes the proportion of rhyming within the given rhyme scheme. And EXTRA_POINTS denotes the total score of the reward score.

H Details of Evaluating Music Understanding Models

H.1 Pipeline of MLP

To assess the effectiveness of music understanding models, we feed music audio into them and obtain their respective encoded sequences. Subsequently, for each model, we utilize an MLP comprising an average pooling layer and 5 linear layers to extract 10 sets of descriptive music tags corresponding to the dimensions of its output encoded sequences. The pipeline of this process can be found in Figure 15.

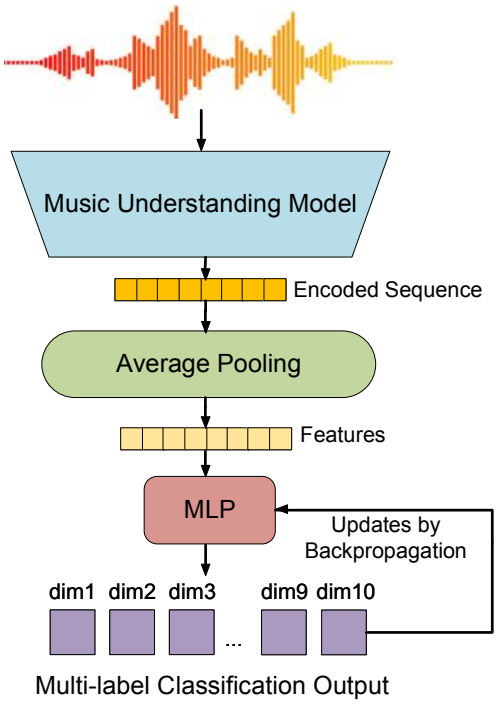
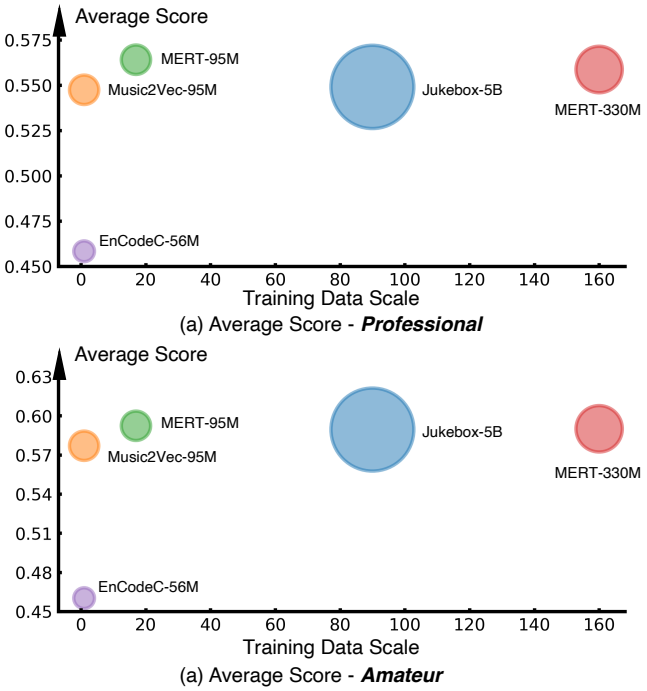


Figure 15: The pipeline of evaluating music understanding models

H.2 Result Analysis

Figure 16 shows, despite having fewer parameters and a smaller amount of training data, MERT-95M performs best overall in the task of professional and colloquial music description.



(a) Average Score - ***Professional***

(a) Average Score - ***Amateur***

Figure 16: Evaluation of selected music understanding models on the benchmark as represented in a scatter plot.