Cantus: Toward Pedagogical Feedback Generation for Sight-Singing Learning

Jiale Chen[1]*, Xuelian Dong[1]*, Yubing Chen[2], Yongxin Hu[3], Wenxiu Xie[4]†, Tianyong Hao[1]†,
[1]School of Computer Science, South China Normal University
[2]School of Music, South China Normal University
[3]Musicology Department, China Conservatory of Music
[4]School of Artificial Intelligence, Guangdong Polytechnic Normal University
{jlchen, xldong, 202410216257, haoty}@m.scnu.edu.cn, hyx@stu.ccmusic.edu.cn, wxxie@gpnu.edu.cn

Abstract

Sight-singing is a fundamental yet challenging task in music education, where timely and fine-grained feedback is crucial for helping learners identify mistakes and improve their performance efficiently. However, existing automatic methods mainly focus on isolated subtasks such as alignment or scoring, and are unable to provide end-to-end diagnostic feedback with explicit identification of pitch, rhythm, and musical fluency errors. In this paper, we propose Cantus, a novel multimodal large language model framework for sight-singing assessment and exercise recommendation generation. Cantus jointly models vocal audio and symbolic musical representations, and leverages a pedagogically grounded error representation module together with a large language model to diagnose performance errors, generate structured assessment comments, and provide personalized exercise recommendations. To support this task, we construct a new dataset *MIVO* (MIDI-Voice) containing 1,003 real-world sight-singing samples with aligned MIDI notes and vocal recordings collected from authentic teaching scenarios. Experimental results demonstrate the effectiveness of Cantus in sight-singing assessment, achieving strong performance across multiple metrics. Human evaluation of feedback quality shows the potential of Cantus in supporting student-centered sight-singing learning.

1 Introduction

Sight-singing learning is one of the core parts of music education (Payne, 2005; Gordon, 2003; Chen and Lin, 2026). It typically involves learners performing unaccompanied singing based on a given musical score within a limited preparation time. Through this exercise, learners can develop their musical perception of each note (Cleland and Dobrea-Grindahl, 2013). During sight-singing practice, learners rely heavily on teacher feedback. In particular, timely and targeted feedback plays a crucial role in the early stages of learning, as it supports the development of pitch sensitivity, fosters accurate interval perception, and enables learners to effectively identify and correct errors during practice (Hattie and Timperley, 2007).

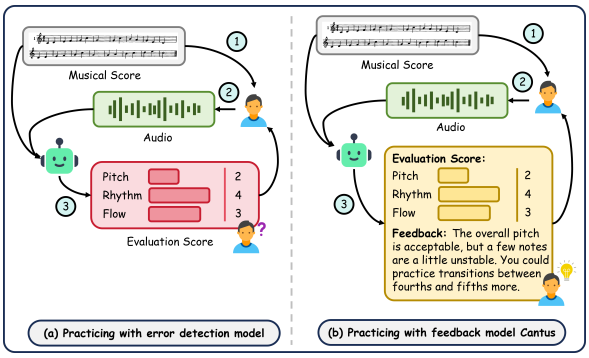


Figure 1: Illustration of the difference between our real-time feedback model Cantus and vanilla model in sight-singing education.

However, in real-world teaching, the effectiveness of sight-singing training is often constrained by practical limitations in educational settings. In large classroom environments, teachers may be unable to provide individualized attention to every student, particularly those with weaker sight-singing skills. Moreover, beginners often require consistent guidance to develop stable pitch and rhythm perception; however, during out-of-class self-practice, the lack of timely teacher feedback makes it difficult for them to identify and correct mistakes promptly (Henry, 2015). To alleviate these challenges, it is necessary to design an objective and reliable automatic sight-singing feedback system that can provide real-time assessment and exercise recommendations.

Most existing studies on sight-singing focus on one aspect of the process. For example, Yang et al. (2022) propose a CRNN-based method to align

*Equal Contribution.
†Corresponding author.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1825–1837
November 6-10, 2026 ©2026 Association for Computational Linguistics

the pitch of sight-singing with the musical score. Xu et al. (2022) propose SingMaster, which compares Optical Music Recognition (OMR) reference sequences (Calvo-Zaragoza et al., 2020) with transcribed sight-singing note sequences and performs evaluation to output a score. These approaches typically address isolated stages such as transcription, alignment, or musical score following, rather than providing step-by-step diagnostic feedback for learners. An illustration of existing methods is shown in Figure 1(a). This leaves a gap in the development of intuitive, accessible, and real-time feedback systems tailored for students and educators in practical environments.

To bridge this gap, an end-to-end architecture that integrates perception, evaluation, and feedback generation is required. Recent advances in Large Language Models (LLMs) offer strong capabilities in reasoning and natural language generation, making them a promising technology for supporting such human-centered feedback. However, developing such systems remains challenging for two main reasons. First, obtaining high-quality feedback annotations is costly and labor-intensive, as it requires domain expertise from educators. Second, effectively fusing vocal signals with musical score representations, while leveraging LLMs to generate meaningful, learner-oriented assessment comments and exercise recommendations, remains a non-trivial problem.

In this paper, we propose a new task named sight-singing feedback generation. For this task, we introduce Cantus, a novel multimodal LLM framework designed to support sight-singing assessment and exercise recommendation generation in real-world learning contexts, as illustrated in Figure 1(b). The framework integrates vocal audio signals with symbolic musical representations to enable end-to-end perception, evaluation, and generation of pedagogically meaningful feedback that is accessible and actionable for learners. Specifically, Cantus leverages a pedagogically grounded error representation module to align audio and musical score representations and capture performance discrepancies. It further utilizes LLM-based reasoning to diagnose performance errors and produce structured, interpretable feedback, as well as personalized exercise recommendations to support iterative learning. To facilitate this human-centered task, we introduce the first benchmark for sight-singing feedback generation, MIDI-Voice (MIVO), which contains 1,003 sight-singing samples collected from authentic teaching scenarios, along with aligned MIDI[1] and vocal recordings. Our main contributions are summarized as follows:

- We introduce a novel multimodal framework that moves beyond pipeline-based approaches to enable end-to-end fine-grained error diagnosis and pedagogical feedback generation for sight-singing learners.
- We release the MIVO, the first dataset for sight-singing feedback generation, which provides real-world sight-singing data to facilitate research on this task.
- Experiments demonstrate the effectiveness of Cantus and the utility of the MIVO dataset in supporting sight-singing feedback generation, highlighting their value for student-centered learning.

2 Related Work

2.1 AI for Music Education

In recent years, Artificial Intelligence (AI) has been increasingly applied in music education. Zhang et al. (2025a) propose LLaQo, a large language query-based music coaching system that leverages audio-language modeling to provide detailed and formative evaluations of instrumental performances. Zhao et al. (2025) introduce MuseAgent, a system for music understanding, which unifies heterogeneous modalities into structured textual representations (e.g., ABC notation, MusicXML, and JSON) that can be directly processed by large language models.

In addition, Li et al. (2025) explore the integration of AI with e-learning tools in vocal education, emphasizing feedback and reflective interventions to enhance pre-service teachers' metacognitive abilities and singing performance, and providing a systematic framework for practical implementation. Jin et al. (2025) propose an LLM-based teachable agent (Chat Melody) based on the Learning-by-Teaching paradigm, where students improve music theory understanding by teaching a simulated novice agent through structured dialogue and error-driven interaction.

While promising, these approaches lack mechanisms for grounding feedback in score-aligned errors, making them unsuitable for fine-grained diagnosis and actionable feedback in sight-singing.

[1]More details about MIDI can be found in Appendix A.

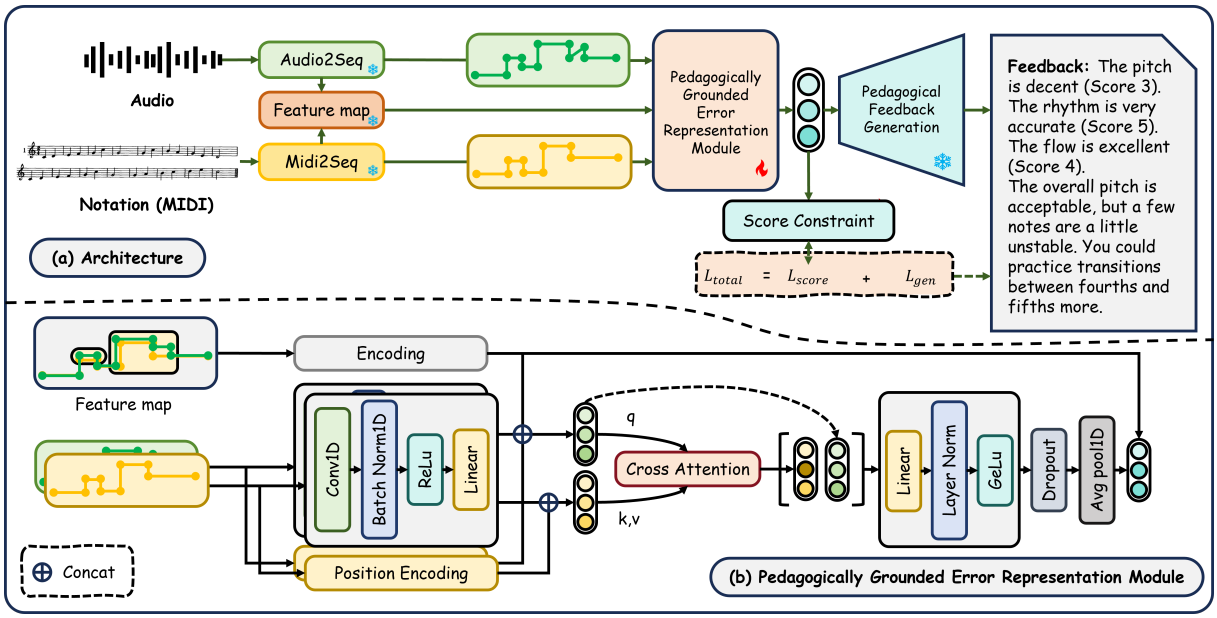


Figure 2: Overall architecture of Cantus.

2.2 Sight-Singing Assessment

Building upon advances in AI-driven music understanding and education, sight-singing assessment has recently emerged as an important research direction. Most existing approaches aim to evaluate singing quality by modeling acoustic signals, optionally in comparison with reference performances or musical scores. Early work utilizes handcrafted or perceptually motivated acoustic features to improve assessment accuracy (Tsai and Lee, 2011; Gupta et al., 2017), while more recent methods incorporate symbolic information. For example, Yang et al. (2022) combine note transcription and alignment to provide note-level feedback, and Zhang et al. (2025b) introduce the SPED dataset to support large-scale performance evaluation and knowledge tracing.

In addition, Kumar et al. (2026) formulate singing error detection as a frame-level audio event detection task based on student–teacher comparisons, focusing on pitch and amplitude deviations. Beyond reference-based approaches, some studies attempt to evaluate singing performance without explicit benchmarks by modeling audio signals directly. For instance, Nakano et al. (2006) assess singing quality using acoustic features such as interval accuracy and vibrato, while Gupta et al. (2019) propose ranking-based methods using pitch histograms and population-level statistics.

Despite these advances, existing methods focus on low-level scoring or error detection, and provide limited actionable, high-level feedback for learning. Different from prior work, our method not only produces quantitative assessments but also generates interpretable, practice-oriented feedback, forming a closed-loop learning process for sight-singing training.

3 Methodology

3.1 Problem Formulation

Let $\mathbf{x}^a$ denote the input audio waveform from a sight-singing learner and $\mathbf{x}^m$ denote the corresponding MIDI musical score. Unlike conventional sight-singing evaluation systems that only output scalar scores, our goal is to generate both a textual pedagogical assessment and a multi-dimensional score vector, enabling the model to provide feedback that is not only accurate but also actionable for learning. Formally, the task is defined as:

$$\hat{\mathbf{y}} = f(\mathbf{x}^a, \mathbf{x}^m), \tag{1}$$

where $\hat{\mathbf{y}}$ denotes the unified textual prediction generated by model, including the score description, assessment comments, and exercise recommendations. In our setting, the assessment covers three dimensions: pitch, rhythm, and musical fluency.

3.2 Overall Architecture

Figure 2 illustrates the overall architecture of Cantus. Given paired vocal audio and MIDI musical score inputs, the framework first converts both inputs into aligned sequential representations and

a corresponding two-dimensional feature map I, which is encoded by a CLIP encoder (Radford et al., 2021). These multimodal features are then processed by a pedagogically grounded error representation module, which captures fine-grained discrepancies between the learner's performance and the reference musical score. The resulting representation serves as a conditioning signal for the LLM, enabling unified generation of score descriptions, assessment comments, and exercise recommendations. During training, an auxiliary score constraint is introduced to encourage the model to learn pedagogically grounded error representations.

3.3 Signal Pre-processing

Before encoding, the raw audio and MIDI inputs are transformed into structured sequential representations using standard audio and symbolic music processing toolkits, including PrettyMIDI[2] and Librosa[3]. The resulting representations are denoted as $\mathbf{x}^{est}$ and $\mathbf{x}^{ref}$, corresponding to the performed audio and reference MIDI modalities, respectively. We then align the two modalities by compensating for temporal offsets and pitch discrepancies, including octave-level transpositions. Finally, the aligned sequences are visualized as a two-dimensional feature map, denoted as I, for downstream feedback generation.

3.4 Pedagogically Grounded Error Representation

Since raw sequence signals are difficult for an LLM to compare directly, we introduce a pedagogically grounded error representation module to align audio and score representations, model their discrepancies, and provide error-aware signals for feedback generation.

Local token embedding. To convert each signal sequence into a sequence of latent tokens, we apply an independent 1-D convolutional block to each channel:

$$\begin{aligned}\mathbf{F}^{ref} &= \phi\Big(\mathrm{BN}\Big(\mathrm{Conv1D}(\mathbf{x}^{ref})\Big)\Big),\\ \mathbf{F}^{est} &= \phi\big(\mathrm{BN}\big(\mathrm{Conv1D}(\mathbf{x}^{est})\big)\big),\end{aligned} \tag{2}$$

where the convolution uses a kernel size of 7, $\mathrm{BN}(\cdot)$ denotes batch normalization, and $\phi(\cdot)$ is ReLU. This operation maps each time step to a local context-aware feature vector, producing $\mathbf{F}^{ref}, \mathbf{F}^{est} \in \mathbb{R}^{d\times L_s}$. d is the feature dimension, and L_s is the length of the sequence. After transposing the channel and temporal dimensions, we obtain frame-level token sequences $\mathbf{H}^{ref}, \mathbf{H}^{est} \in \mathbb{R}^{L_s\times d}$. Each token thus encodes a short temporal neighborhood rather than a single scalar value.

Latent projection. We project the local features into a shared latent space of dimension d and add learnable positional embeddings to preserve temporal order:

$$\begin{aligned}\mathbf{U}^{*} &= \mathbf{H}^{*}\mathbf{W}_{*} + \mathbf{b}_{*},\\ \tilde{\mathbf{U}}^{*}_{t} &= \mathbf{U}^{*}_{t} + \mathbf{e}^{*}_{t}, \quad * \in \{ref, est\},\end{aligned} \tag{3}$$

where $\mathbf{W}_{*} \in \mathbb{R}^{d}$ and $\mathbf{e}^{*}_{t}$ denotes the learnable positional embedding at position t.

Error grounding and fusion. To support pedagogical diagnosis, the model must identify where the learner deviates from the reference musical score and retain this discrepancy information in a form that can be interpreted by the language model. We therefore model the interaction between the two sequences using cross-attention, where the estimated sequence serves as the query and the reference sequence serves as the key-value memory:

$$\mathbf{C} = \mathrm{MHA}\Big(\mathbf{Q} = \tilde{\mathbf{U}}^{est}, \mathbf{K} = \tilde{\mathbf{U}}^{ref}, \mathbf{V} = \tilde{\mathbf{U}}^{ref}\Big). \tag{4}$$

This design aligns each token in the estimated performance sequence with relevant reference tokens, while keeping the output organized according to the temporal order of the estimated sequence.

In addition to the cross-attended representation, we explicitly include the raw frame-wise difference between the two input sequences: $\boldsymbol{\delta}_t = \mathbf{x}^{ref}_t - \mathbf{x}^{est}_t$.

We then concatenate the estimated features, cross-attended context, and raw difference $\mathbf{G}_t = \Big[\tilde{\mathbf{U}}^{est}_t; \mathbf{C}_t; \boldsymbol{\delta}_t\Big]$, and map concatenated representation back to the model dimension using a feed-forward projection:

$$\mathbf{Z}_t = \mathrm{GELU}(\mathrm{LN}(\mathbf{W}_f\mathbf{G}_t + \mathbf{b}_f)). \tag{5}$$

Finally, we apply dropout and temporal average pooling with stride 2 to obtain a compressed multi-scale sequence representation:

$$\hat{\mathbf{Z}} = \mathrm{AvgPool1D}_{s=2}(\mathrm{Dropout}(\mathbf{Z})). \tag{6}$$

The output is a shorter latent sequence that preserves both local acoustic structure and cross-sequence alignment information, while emphasizing the discrepancy patterns needed for downstream feedback generation.

[2] https://github.com/craffel/pretty-midi
[3] https://librosa.org/

3.5 Pedagogical Feedback Generation

The pooled error-grounded representation $\hat{\mathbf{Z}}$ is provided to an LLM as an external conditioning signal. Concretely, $\hat{\mathbf{Z}}$ is injected into the language model through an adapter projection: $q = \mathbf{W}_p\hat{\mathbf{Z}} + \mathbf{b}_p$.

Given an instruction prompt $\mathbf{x}^{\text{instr}}$, a conditioning vector q, and a two-dimensional feature map I, we feed I into the visual encoder of the multimodal LLM and inject q through an adapter as additional soft conditioning tokens. The LLM then autoregressively generates a unified textual prediction sequence: $\hat{\mathbf{y}} = [\hat{\mathbf{y}}^{\text{score}}, \hat{\mathbf{y}}^{\text{assessment}}, \hat{\mathbf{y}}^{\text{feedback}}]$, which contains performance scoring, assessment comments, and subsequent training feedback in a concatenated form. The generation process is formulated as

$$P(\hat{\mathbf{y}} \mid q, I, \mathbf{x}^{\text{instr}}) = \prod_{\ell=1} P\big(\hat{\mathbf{y}}_\ell \mid \hat{\mathbf{y}}_{<\ell}, q, I, \mathbf{x}^{\text{instr}}\big)\,. \tag{7}$$

During training, the unified target sequence is optimized by a standard cross-entropy objective against reference annotations. In this way, the model is trained to generate not only fluent language, but also feedback that is grounded in the identified performance discrepancies and aligned with pedagogical intent.

3.6 Score Constraint

To provide additional supervision for pedagogically grounded error representation learning, we introduce an auxiliary score prediction constraint during training. Specifically, a lightweight scoring head operates on the pooled representation $\hat{\mathbf{Z}}$:

$$\hat{\mathbf{s}} = g(\hat{\mathbf{Z}}) = \sigma(\mathbf{W}_s\hat{\mathbf{Z}} + \mathbf{b}_s), \tag{8}$$

where $\sigma(\cdot)$ is a sigmoid function for normalized scores. The target scores are defined within the range $[0, 5]$, and normalized predictions can be rescaled accordingly. The scoring branch predicts multiple evaluation dimensions independently:

$$\hat{\mathbf{s}} = [\hat{s}_{\text{pitch}}, \hat{s}_{\text{rhythm}}, \hat{s}_{\text{fluency}}]^\top. \tag{9}$$

Importantly, this scoring branch is only used as an auxiliary training constraint to regularize the learned representation space and encourage agreement between quantitative assessment and textual feedback. During inference, the score prediction head is discarded to avoid additional inference overhead and prevent conflicts.

Category	Value
Total Samples	1,003
Total Musical Notes	29,507
Total Audio Duration (s)	20,321
Total Review Words	58,857
Modality	Audio + MIDI + Text
Annotation Types	Text + Multi-dimensional Scores
Scoring Dimensions	Pitch, Rhythm, Musical Fluency

Table 1: Summary of the MIVO dataset.

3.7 Training Objective

The full model is trained by combining the language generation loss and the scoring loss. The generation loss is the standard negative log-likelihood:

$$\mathcal{L}_{\text{gen}} = -\sum_{\ell=1}^{|\mathbf{y}|} \log P(\mathbf{y}_\ell \mid \mathbf{y}_{<\ell}, q, I, \mathbf{x}^{\text{instr}}), \tag{10}$$

where $\mathbf{y}$ is the reference text. The score loss is defined as the mean squared error:

$$\mathcal{L}_{\text{score}} = \frac{1}{K}\sum_{k=1}^{K} (\hat{s}_k - s_k)^2\,, \tag{11}$$

with s_k denoting the ground-truth score for the k-th dimension. With hyper-parameters λ_1 and λ_2, the final training objective is formally: $\mathcal{L} = \lambda_1\mathcal{L}_{\text{gen}} + \lambda_2\mathcal{L}_{\text{score}}$.

4 MIVO Dataset

4.1 Overview

We introduce MIVO, the first benchmark for sight-singing feedback generation. The dataset contains paired MIDI musical score and vocal performance recordings collected from authentic student practice scenarios, together with expert-annotated assessment results and pedagogical feedback. MIVO is designed to support learner-centered sight-singing evaluation, including score prediction, performance assessment, and exercise recommendation generation.

4.2 Data Collection and Annotation

MIVO consists of 1,003 sight-singing samples collected from real student practice sessions. Each sample contains a reference MIDI musical score, a corresponding vocal recording, and expert annotations of the learner's performance.

For each sample, annotators provide: (1) numerical ratings ranging from 0 to 5 for pitch accuracy, rhythmic precision, and musical fluency; (2) textual assessment comments describing performance

issues; and (3) exercise recommendations for subsequent practice. All annotations are produced by five experienced sight-singing experts. To ensure annotation consistency, annotations are cross-checked among annotators, and disagreements are resolved through discussion.

4.3 Statistics

MIVO contains 1,003 annotated audio-MIDI pairs spanning singing performances from learners at different proficiency levels. The dataset captures common sight-singing errors, including pitch instability and temporal deviations, reflecting realistic learning conditions in music education scenarios. Detailed statistics are summarized in Table 1.

5 Experiment

5.1 Setting

Baselines We compare Cantus with a diverse set of baselines covering pipeline systems, zero-shot methods, and few shot methods. All baselines are provided with comparable inputs whenever applicable to ensure fair comparison. For visual LLM baselines, aligned audio-MIDI sequences are rendered into piano-roll-style visual representations.

Data Split We divide the dataset into training, validation, and test sets with a ratio of 8:1:1. To avoid potential data leakage, samples from the same student are assigned exclusively to one split.

Evaluation To comprehensively assess all baselines and Cantus, we use three types of metrics to evaluate performance: language quality, scoring quality, and overall judgment. Specifically, we report BLEU (Papineni et al., 2002), chrF(Popović, 2015), BERTScore (Zhang et al., 2019), MAE, Point Accuracy (Point Acc.), Sample Accuracy (Sample Acc.), and Judge Accuracy (Judge Acc.).

Implementation For the generation backbone, we employ open-source Qwen-3.5 models. We fine-tune these models using Low-Rank Adaptation (Hu et al., 2022). The models are trained for 30 epochs with a learning rate of 5×10^{-5}. All experiments[4] are conducted on a single NVIDIA RTX 4090D GPU with 24 GB memory.

[4]Further details on the baseline methods, evaluation metrics, and implementation settings are provided in Appendix B, Appendix C, and Appendix D, respectively.

5.2 Quantitative Performance

Table 2 compares all methods on the MIVO dataset. The results show that traditional pipeline-based approaches, including rule-based, CNN-based, and transformer-based models, can achieve reasonable performance on isolated scoring tasks, with the transformer-based model obtaining a competitive MAE of 0.52. However, these methods cannot generate diagnostic feedback, which limits their usefulness in practical sight-singing education.

General-purpose visual LLMs can generate fluent natural language, but they still struggle with fine-grained musical understanding. Although few-shot prompting improves the semantic quality of generated feedback, the overall assessment accuracy remains limited, with the best overall Judge Accuracy reaching only 27.85%. While Qwen-3-VL-8B-Instruct$^{\square}$ obtains the highest Point Accuracy and Sample Accuracy among the baselines, this advantage does not translate into strong overall judge performance. In contrast, Cantus supports both assessment comment generation and exercise recommendation in a unified framework. The Qwen-3.5-2B$^{\diamond}$ variant achieves competitive performance across most metrics, obtaining the lowest MAE of 0.43 and the highest overall judge accuracy of 55.70%, while also showing strong language generation performance. These results suggest that grounded error representations can effectively guide LLM-based generation for end-to-end sight-singing pedagogical feedback.

5.3 Human Evaluation of Feedback Quality

To evaluate the pedagogical quality of the generated feedback, we conduct a human evaluation study with domain experts. Specifically, we recruit five evaluators with formal music training in sight-singing instruction. For each method, evaluators are presented with anonymized and randomly shuffled feedback samples to avoid bias introduced by model identity or presentation order.

Feedback is evaluated along three dimensions: (1) Correctness, (2) Usefulness, and (3) Specificity. Table 3 presents the human evaluation results. Although Qwen-3-VL-8B-Instruct performs well on quantitative metrics, its generated feedback tends to be generic or misaligned with specific singing errors, resulting in lower human evaluation scores. Cantus consistently achieves the highest scores across all three dimensions, outperforming all compared general-purpose multimodal LLMs. In par-

Method	Language Quality			Scoring Quality			Overall
	BLEU ↑	chrF ↑	BERTScore ↑	MAE ↓	Point Acc. ↑	Sample Acc. ↑	Judge Acc. ↑
Pipeline Methods							
Rule-based	-	-	-	0.94	62.03	45.57	-
CNN-based	-	-	-	0.58	66.24	44.31	-
Transformer-based	-	-	-	0.52	68.35	41.77	-
Zero-shot Methods							
ERNIE-4.6-Turbo-VL	2.76	5.17	38.79	1.05	51.05	20.25	10.13
GLM-4.6V	0.04	3.35	23.14	1.70	45.15	24.05	2.53
Gemini-3.1-Pro	3.12	5.73	57.28	1.31	46.41	26.58	10.13
GPT-5.4	3.46	6.24	60.81	1.06	46.84	21.52	15.19
Qwen-3-VL-Flash	0.98	3.52	35.77	0.96	58.23	29.11	3.80
Qwen-3-VL-Plus	0.86	3.01	23.10	1.08	46.84	13.92	1.27
Qwen-3-VL-8B-Instruct	2.54	7.16	64.33	1.20	32.07	1.27	1.27
Qwen-3-VL-8B-Instruct$^{\square}$	27.66	25.02	71.23	0.53	**86.92**	**82.28**	21.55
Few-shot Methods							
ERNIE-4.6-Turbo-VL	5.90	9.70	71.61	0.91	58.65	29.11	27.85
GLM-4.6V	1.40	3.95	26.85	2.50	22.36	7.59	6.33
Gemini-3.1-Pro	9.49	11.41	62.58	1.23	51.05	32.91	26.58
GPT-5.4	4.71	10.15	70.31	1.02	47.26	24.05	22.78
Qwen-3-VL-Flash	2.68	5.75	58.03	0.85	67.09	39.24	25.32
Qwen-3-VL-Plus	3.47	6.46	58.40	0.92	59.49	29.11	16.46
Cantus (Ours)							
Qwen-3.5-0.8B$^{\diamond}$	<u>31.12</u>	<u>27.46</u>	<u>74.81</u>	<u>0.46</u>	75.95	58.23	<u>54.43</u>
Qwen-3.5-0.8B$^{\square}$	19.92	17.57	63.81	0.61	78.48	60.76	48.10
Qwen-3.5-2B$^{\diamond}$	**33.57**	**29.08**	**75.80**	**0.43**	<u>80.59</u>	<u>62.03</u>	**55.70**
Qwen-3.5-2B$^{\square}$	17.15	15.87	67.67	0.58	75.53	56.96	45.57

Table 2: Quantitative performance comparison. **Bold** denotes the highest score and <u>Underlining</u> denotes the second-highest score. $^{\diamond}$ means using the frozen LLM as decoder. $^{\square}$ means LLM training using the Lora.

Method	Correctness	Usefulness	Specificity
Qwen-3-VL-Flash	2.8	3.0	3.1
Qwen-3-VL-8B-Instruct$^{\square}$	2.3	1.6	1.3
GLM-4.6V	1.8	2.3	2.2
Gemini-3.1-Pro	2.2	2.6	2.4
GPT-5.4	2.4	2.3	2.6
Cantus (Qwen-3.5-2B)	**3.5**	**3.3**	**3.4**

Table 3: Human evaluation of pedagogical feedback.

Method	Point Acc. ↑	BERTScore ↑	Judge Acc. ↑
Cantus	**80.59**	<u>75.80</u>	**55.70**
w/o SC	71.73	74.67	45.57
w/o PGER	74.68	**76.00**	44.30
w/o FM	75.53	67.67	51.90

Table 4: Ablation study.

ticular, Cantus obtains substantial improvements in correctness and usefulness, indicating that its multimodal modeling of vocal audio and musical score enables more accurate diagnosis and pedagogical guidance. Meanwhile, the better performance in specificity suggests that Cantus is capable of generating detailed and fine-grained feedback rather than generic comments.

6 Analysis

6.1 Ablation Study

Table 4 presents the ablation study of Cantus. We evaluate the contribution of three key components: the score constraint (SC), the pedagogically grounded error representation module (PGER), and the feature map input (FM).

Removing the SC causes a noticeable decline in Point Accuracy, dropping from 80.59 to 71.73. This result suggests that the additional score supervision helps the model better capture assessment-related information and improves the consistency between generated feedback and evaluation scores. Eliminating the PGER also reduces performance, particularly on Judge Accuracy. Although BERTScore remains comparable to the full model, the lower evaluation accuracy suggests that fluent feedback alone is insufficient without ac-

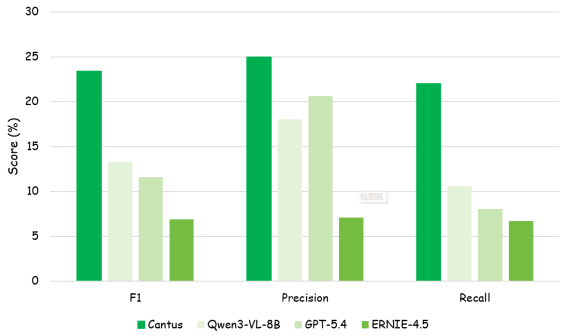


Figure 3: Comparison of fine-grained error diagnosis performance on the MIVO dataset. We evaluate Cantus against general-purpose LLMs (GPT-5.4, ERNIE-4.6). Metrics include F_1-score, precision, and recall.

curate note-level correspondence between singing audio and the musical score. Without the FM signal, the model shows consistent degradation on all metrics, especially BERTScore, indicating that global musical structure contributes to generating coherent and informative feedback.

Overall, the full Cantus model achieves better performance, demonstrating the complementary effects of our proposed modules in supporting accurate sight-singing feedback generation.

6.2 Fine-grained Error Diagnosis

We evaluate the fine-grained diagnostic capability of Cantus by comparing its predicted learner error labels against expert annotations in the MIVO dataset. As shown in Figure 3, despite the overall difficulty of fine-grained error diagnosis, Cantus achieves clear improvements over general-purpose LLMs. Specifically, Cantus obtains an F_1-score of 23.41, which outperforms general-purpose LLMs such as GPT-5.4 (11.59) and ERNIE-4.6 (6.89). While a performance gap still exists between the model and human practitioners, Cantus demonstrates a more balanced precision-recall profile compared to GPT-5.4, which suffers from high precision but low recall. This suggests that our pedagogically grounded error representation module allows the model to capture subtle vocal errors that general LLMs overlook.

6.3 User Study: Learning Effect

As illustrated in Table 5, to evaluate the effectiveness of fine-grained pedagogical feedback, we conducted a one-week controlled experiment with novice learners who had limited prior training in sight-singing. A total of 24 anonymous university student participants with basic singing ability were divided into an Experimental Group (EG) and a Control Group (CG). During the study, participants were required to practice sight-singing for at least 15 minutes per day using MIDI score provided by the system. Participants in the EG were provided with assessment comments and exercise recommendations, whereas those in the CG practiced without such assistance. Performance was evaluated on a 15-point scale consisting of three dimensions: pitch, rhythm, and musical fluency (each worth 5 points).

Group	Pre-test	Post-test	Δ Score	Improvement
CG	6.28	6.56	+0.28	+4.46%
EG	6.26	6.94	+0.68	+10.86%

Table 5: Student performance comparison between the EG and CG. EG shows greater improvement than CG (significance test $p < 0.05$).

While both groups showed improvement after one week of practice, the EG outperformed the CG. The average score of the EG increased from 6.26 to 6.94 (a 10.86% relative improvement), whereas the CG only showed a marginal gain of 4.46%. A t-test ($p < 0.05$) suggests that detailed diagnostic feedback may contribute to improved learning outcomes. These results suggest that end-to-end pedagogical feedback provides benefits over score-only evaluation in practical sight-singing education.

7 Conclusion

This paper introduces Cantus, a novel multimodal framework for automatic sight-singing feedback generation. It integrates vocal audio, musical score representations, assessment generation, and exercise recommendation generation into a unified end-to-end framework, making it more suitable for practical student-centered sight-singing learning. We also present MIVO, a new dataset consisting of 1,003 real-world sight-singing samples collected from authentic teaching scenarios. The dataset is intended to support research on automatic sight-singing understanding and pedagogical feedback generation. Experimental results show that Cantus achieves better performance in end-to-end feedback generation, particularly in language quality, judge accuracy, and human evaluation. These findings suggest the potential of the multimodal LLM-based framework for sight-singing pedagogy and student-centered music learning.

Acknowledgements

The work is supported by grants from National Natural Science Foundation of China (No. 62372189), Guangdong Special Funding for Science and Technology Innovation Strategy (2024A1111120004), SCNU Training Program of Innovation and Entrepreneurship for Undergraduates (No. 202610006) and SCNU Student Extracurricular Research Program (No. 25YYGA11).

Limitations

Despite the promising results, there remains room for further improvement. First, the current framework primarily focuses on pitch and rhythm accuracy, while higher-level musical factors such as expressive phrasing, emotional delivery, stylistic interpretation, and performance confidence remain underexplored. Second, the framework mainly provides feedback based on single-session performance analysis and does not yet model learners' long-term progress, individual learning habits, or personalized pedagogical needs. Future work will explore longitudinal learner modeling and teacher-in-the-loop refinement to provide more adaptive and human-centered feedback.

Ethics Statement

The MIVO dataset was collected from students' sight-singing practice sessions with informed consent. All recordings were anonymized, and personally identifiable information was removed before dataset construction. The dataset contains only audio recordings, MIDI files, scores, and textual annotations used for music education research.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Jorge Calvo-Zaragoza, Jan Hajič Jr, and Alexander Pacha. 2020. Understanding optical music recognition. *ACM Computing Surveys (CSUR)*, 53(4):1–35.

Zhao-ling Chen and He Lin. 2026. Ai and digital technologies in sight-singing and aural training: A systematic review of innovations in higher music education. *European Journal of Education*, 61(1):e70502.

Kent D Cleland and Mary Dobrea-Grindahl. 2013. *Developing Musicianship Through Aural Skills: A Holistic Approach to Sight Singing and Ear Training*. Routledge.

Edwin Gordon. 2003. *A music learning theory for newborn and young children*. Gia Publications.

Chitralekha Gupta, Haizhou Li, and Ye Wang. 2017. Perceptual evaluation of singing quality. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 577–586. IEEE.

Chitralekha Gupta, Haizhou Li, and Ye Wang. 2019. Automatic leaderboard: Evaluation of singing quality without a standard reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:13–26.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research*, 77(1):81–112.

Michele Henry. 2015. Vocal sight-reading assessment: Technological advances, student perceptions, and instructional implications. *Update: Applications of Research in Music Education*, 33(2):58–64.

Edward Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Lingxi Jin, Baicheng Lin, Mengze Hong, Hyo-Jeong So, and Kun Zhang. 2025. Learning by teaching: Enhancing music learning through llm-based teachable agents. In *International Conference on Artificial Intelligence in Education*, pages 148–155. Springer.

Sumit Kumar, Suraj Jaiswal, Parampreet Singh, and Vipul Arora. 2026. Automatic detection and analysis of singing mistakes for music pedagogy. *arXiv preprint arXiv:2602.06917*.

Wen Li, Xuerong Cui, Pravina Manoharan, Lu Dai, Ke Liu, and Li Huang. 2025. Ai-assisted feedback and reflection in vocal music training: effects on metacognition and singing performance. *Frontiers in Psychology*, 16:1598867.

Tomoyasu Nakano, Masataka Goto, and Yuzuru Hiraga. 2006. An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. In *Interspeech 2006*, pages paper–1854.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Dorothy Payne. 2005. Essential skills for promoting a lifelong love of music and music making. *The American Music Teacher*, 54(4):26.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Wei-Ho Tsai and Hsin-Chieh Lee. 2011. Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features. *IEEE transactions on audio, speech, and language processing*, 20(4):1233–1243.

Wei Xu, Bowen Tian, Lijie Luo, Weiming Yang, Xianke Wang, and Lei Wu. 2022. Singmaster: A sight-singing evaluation system of" shoot and sing" based on smartphone. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6944–6946.

Weiming Yang, Xianke Wang, Bowen Tian, Wei Xu, and Wenqing Cheng. 2022. A multi-stage automatic evaluation system for sight-singing. *IEEE transactions on Multimedia*, 25:3881–3893.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Huan Zhang, Vincent KM Cheung, Hayato Nishioka, Simon Dixon, and Shinichi Furuya. 2025a. Llaqo: Towards a query-based coach in expressive music performance assessment. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yan Zhang, Jie Luo, Tianrui Li, and Wei Xu. 2025b. Sped: A sight-singing dataset for performance evaluation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Qihao Zhao, Yunqi Cao, Yangyu Huang, Hui Yi Leong, Fan Zhang, and Wei Hu. 2025. Towards multimodal understanding of music scores and performance audio. *openreview.net*.

A Why We Choose MIDI

Musical Instrument Digital Interface (MIDI) is an American-Japanese technical standard that describes a communication protocol, digital interface, and electrical connectors that connect a wide variety of electronic musical instruments, computers, and related audio devices for playing, editing, and recording music[5]. A screenshot of the MIDI notaion is shown in Figure 4. We adopt MIDI as the symbolic representation of musical score for two key reasons. First, MIDI provides a structured representation of music, explicitly encoding pitch, onset time, duration, and velocity, which facilitates precise evaluation of pitch accuracy and rhythmic precision. Second, MIDI enables straightforward alignment between the reference musical score and the performed audio, supporting interpretable, fine-grained feedback generation. Its widespread use in music research ensures interoperability and reproducibility.

B Baseline

Pipeline methods. We implement a conventional pipeline that aligns the sung audio with the reference MIDI and computes pitch and rhythm deviations. These methods include rule-based method, CNN-based method and Transformer-based method. All these methods just can predict score, and cannot provide feedback for the sight-singing learner.

Visual LLM-based methods. These methods include ERNIE-4.6-Turbo-VL, GLM-4.6V (Zeng et al., 2025), Gemini-3.1-Pro, GPT-5.4 and Qwen-3-VL (Bai et al., 2025) series. To assess the contribution of LLM-based generation, we feed visual feature map (The original picture is illustrated in Figure 5) as input to an LLM to produce feedback. We evaluate both zero-shot and few-shot prompting, where the latter includes a small number of in-context examples.

C Metrics

To comprehensively evaluate the overall performance of our approach, we adopt a diverse set of automatic and semi-automatic metrics. These metrics can be divided into three kinds: language quality, score accuracy and overall judge, including BLEU, chrF, BERTScore, MAE, Score accuracy, Review Accuracy, and Judge Accuracy. These metrics capture different aspects of generation quality, ranging from surface-level similarity to semantic fidelity and task-specific correctness.

C.1 Language Quality Evaluation.

For language quality assessment, we use BLEU, chrF, and BERTScore.

BLEU measures the n-gram overlap between the generated text and golden text. It evaluates precision over matched n-grams and is widely used in text generation tasks. Higher BLEU scores indicate better lexical similarity.

chrF is a character-level F-score metric that computes the harmonic mean of character n-gram precision and recall. Compared to BLEU, chrF is more sensitive to morphological variations and is particularly suitable for evaluating fine-grained textual similarity.

BERTScore Given a generated text $x = \{x_1, \ldots, x_n\}$ and a reference (golden) text $y = \{y_1, \ldots, y_m\}$, let $\mathbf{e}(x_i)$ and $\mathbf{e}(y_j)$ denote their contextual embeddings. We compute the cosine similarity matrix:

$$s_{ij} = \cos\big(\mathbf{e}(x_i), \mathbf{e}(y_j)\big). \tag{12}$$

The precision and recall are defined as:

$$P = \frac{1}{n}\sum_{i=1}^{n} \max_j s_{ij}, \quad R = \frac{1}{m}\sum_{j=1}^{m} \max_i s_{ij}. \tag{13}$$

The BERTScore F1 is then computed as:

$$F_{\text{BERT}} = 2 \cdot \frac{P \cdot R}{P + R}. \tag{14}$$

BERTScore can capture semantic fidelity beyond surface-level match.

C.2 Score Quality.

For score quality evaluation, we use MAE, Point Accuracy, and Sample Accuracy. MAE measures the magnitude of scoring errors, Point Accuracy evaluates the correctness of each individual scoring dimension, and Sample Accuracy evaluates whether all scoring dimensions of a sample are correctly predicted.

MAE (Mean Absolute Error) measures the average absolute difference between predicted score and ground truth. Lower MAE indicates better performance.

[5]More detail about MIDI can be found in https://en.wikipedia.org/wiki/MIDI

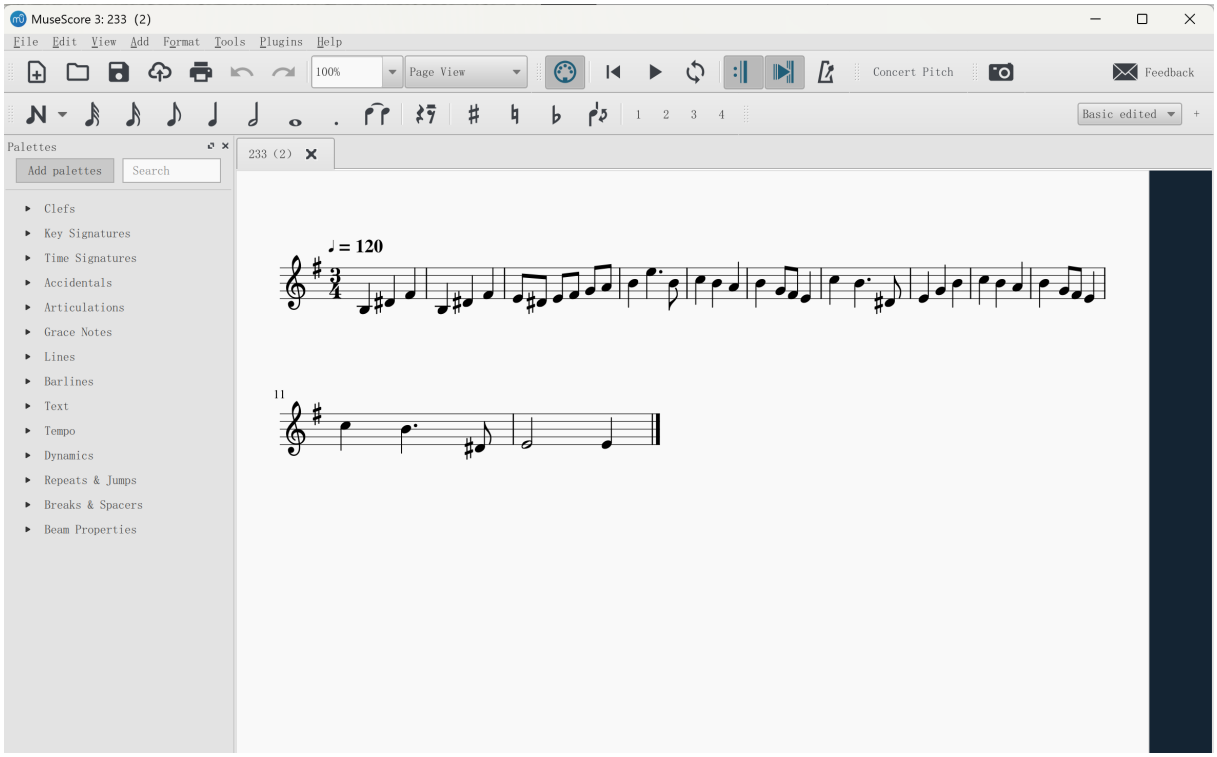


Figure 4: A screenshot of MIDI notaion.

Point Accuracy evaluates the accuracy of individual scoring dimensions. Given a dataset with N samples and K scoring dimensions, let $\hat{s}_{i,k}$ and $s_{i,k}$ denote the predicted and ground-truth scores of the k-th dimension for the i-th sample, respectively. A scoring point is considered correct if the absolute difference between the predicted score and the ground-truth score does not exceed 0.5:

$$\text{Acc}_{i,k}^{\text{point}} = \begin{cases} 1, & \text{if } |\hat{s}_{i,k} - s_{i,k}| \leq 0.5, \\ 0, & \text{otherwise.} \end{cases} \quad (15)$$

The final Point Accuracy is computed over all scoring points:

$$\text{Point Accuracy} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} \text{Acc}_{i,k}^{\text{point}}. \quad (16)$$

Sample Accuracy evaluates whether the complete set of predicted scores for a sample is consistent with the ground-truth scores. A sample is considered correct only when all scoring dimensions are predicted within the allowed error threshold:

$$\text{Acc}_{i}^{\text{sample}} = \begin{cases} 1, & \text{if } \max_k |\hat{s}_{i,k} - s_{i,k}| \leq 0.5, \\ 0, & \text{otherwise.} \end{cases} \quad (17)$$

The final Sample Accuracy is computed as:

$$\text{Sample Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \text{Acc}_{i}^{\text{sample}}. \quad (18)$$

C.3 Judge Accuracy

We define an indicator function for numerical consistency:

$$\mathbb{I}_{\text{score}} = \begin{cases} 1, & \text{if Acc}_{i}^{\text{sample}} = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (19)$$

Let F_{BERT} denote the BERTScore between the generated and reference texts. The Judge Accuracy for a single sample is defined as:

$$\text{Acc}_{\text{judge}} = \begin{cases} 1, & \text{if } \mathbb{I}_{\text{score}} = 1 \text{ and } F_{\text{BERT}} > 0.6, \\ 0, & \text{otherwise.} \end{cases} \quad (20)$$

Here, we empirically set the BERTScore threshold to 0.6. The final Judge Accuracy over the dataset is:

$$\text{Judge Accuracy} = (1 - \alpha) \frac{1}{N} \sum_{i=1}^{N} \text{Acc}_{\text{judge}}^{(i)}. \quad (21)$$

Here, α represents the redundancy ratio, defined as the percentage of duplicate feedback generated

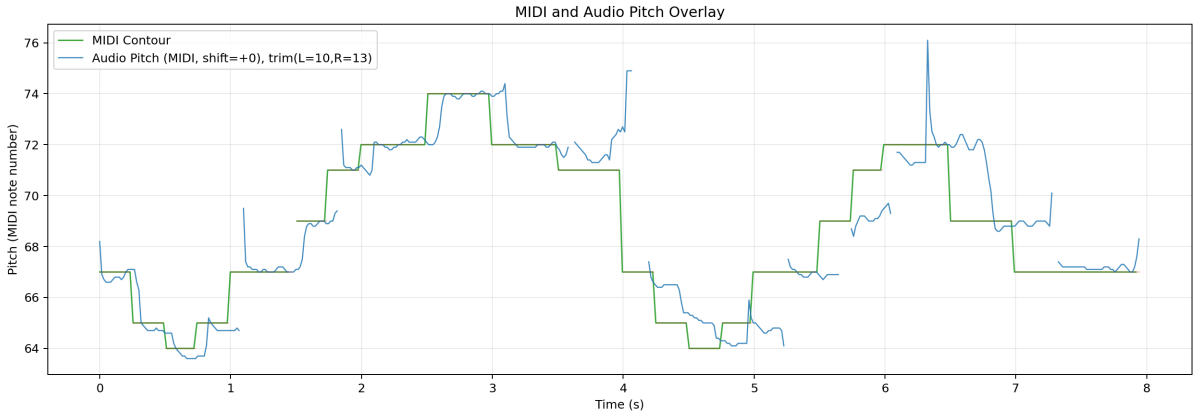


Figure 5: Example of the two-dimensional audio-MIDI feature map.

across the test set. By incorporating $(1 - \alpha)$, the metric penalizes models that favor repetitive, "safe" templates over diverse, case-specific diagnostics, thereby reflecting both accuracy and model generalization. A feedback response is considered redundant if its normalized textual form exactly matches another generated response in the test set.

D Implementation supplementation

The hyperparameters λ_1 and λ_2 are both set to 0.5, balancing the generation and score loss based on preliminary experiments.

E Background of Annotation Participants

The annotation team for the MIVO dataset consisted of five domain experts with professional backgrounds in music education. All annotators hold at least a Bachelor's degree in music or related disciplines, with the highest qualification being a Master's degree. Beyond their academic credentials, each participant possesses multiple years of professional experience in sight-singing pedagogy and performance evaluation. This combination of advanced academic training and extensive practical expertise ensures the high quality, consistency, and pedagogical reliability of the textual feedback and numerical ratings provided in the dataset.

F Instructions Given to Participants

The data collection and annotation process involved two distinct groups of participants: the students who provided the vocal performances and the domain experts who annotated the data.

Instructions to Data Providers You will be asked to perform unaccompanied sight-singing exercises based on the given musical score during your regular practice sessions. Your vocal performance will be recorded and paired with the corresponding MIDI musical score for research and dataset development purposes. Before participation, you will be asked to provide informed consent. All recordings will be anonymized, and any personally identifiable information (PII) will be removed. The dataset will focus only on non-sensitive audio features relevant to music education.

Instructions to Annotators You will be asked to evaluate each collected audio-MIDI pair by comparing the student's vocal performance with the reference MIDI musical score. For each sample, please complete the following tasks: (1) assign numerical scores from 0 to 5 for pitch accuracy, rhythmic precision, and musical fluency; (2) provide diagnostic comments explaining the reasons for the assigned scores; and (3) offer structured and actionable practice suggestions based on the diagnostic feedback. To ensure annotation quality and consistency, please cross-check your evaluations with other annotators and resolve subjective disagreements through collaborative discussion when necessary.