CMI-RewardBench: Evaluating Music Reward Models with Compositional Multimodal Instruction

Anonymous Authors[1]

Abstract

While music generation models have evolved to handle complex multimodal inputs mixing text, lyrics, and reference audio, evaluation mechanisms have lagged behind, remaining fragmented and narrowly focused. In this paper, we bridge this critical gap by establishing a comprehensive ecosystem for **Compositional Music Instruction (CMI)** reward modeling, where the generated music may be conditioned on text descriptions, lyrics, and/or audio prompts. We first introduce **CMI-Pref-Pseudo**, a large-scale preference dataset comprising 110k pseudo-labeled samples, and **CMI-Pref**, a high-quality, human-annotated corpus tailored for fine-grained alignment tasks. To unify the evaluation landscape, we propose CMI-RewardBench, a unified benchmark that evaluates music reward models on heterogeneous samples across musicality, text–music alignment, and compositional instruction alignment. Leveraging these resources, we develop **CMI reward models (CMI-RMs)**, a parameter-efficient reward model family capable of processing heterogeneous inputs. We evaluate their correlation with human judgments scores on Music Arena and CMI-Pref test set, as well as preference agreement on Music Arena and CMI-Pref. Additional analyses examine performance variation across factors such as annotators, annotation timing and confidence, music generation models, and audio length. Experiments demonstrate that CMI-RM not only correlates strongly with human judgments, but also enables effective **inference-time scaling** via top-k filtering. Our work provides the necessary data, benchmarks, and models to advance aligned music generation.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

The rapid advancement of Artificial Intelligence Generated Content (AIGC) has significantly impacted the creative industries, with music generation emerging as one of the cornerstones for numerous commercial applications in music, movie and entertainment industries (Cao et al., 2025; Ren et al., 2025; Ma et al., 2024). Despite the proliferation of sophisticated generative models, evaluating their outputs remains a fundamental challenge. Unlike text or image generation which sometimes focus on functional or purpose, music evaluation is typically aesthetically subjective and multi-dimensional, requiring assessments of musicality and adherence to complex, often multimodal, user instructions.

Developing these evaluation models is hindered by a critical data scarcity. Although large-scale user interaction data exists in music recommendation (e.g., Spotify Million Playlist (Papreja et al., 2019)), it fundamentally captures user-item affinity—a global preference for genre styles or playlists—rather than generative alignment, which demands assessment of perceptual quality and precise instruction-following. Such recommendation datasets lack the fine-grained, comparative rankings of generated samples against complex, multimodal instructions (such as interwoven lyrics, text descriptions, and reference audio) required to train alignment models (Deshmukh et al., 2024).

Consequently, evaluation methodologies have struggled to bridge this gap. Traditional metrics like Fréchet Audio Distance (FAD) (Kilgour et al., 2019) operate at the distribution level, failing to provide the sample-level signals necessary for post-training or filtering. More recent approaches, such as SongEval (Yao et al., 2025), PAM (Deshmukh et al., 2024), and various MOS predictors (Jeong et al., 2025), have advanced the field by offering sample-level scoring. However, these efforts remain fragmented and narrowly specialized. They typically focus on isolated attributes (e.g., only caption alignment) and rely on rigid input assumptions, whereas state-of-the-art music generation models already support flexible input combinations, ranging from simple text prompts to interwoven lyrics and audio references, as illustrated in Figure 1. This highlights a growing mismatch between model capabilities and existing evaluation methodologies.

We argue that effective evaluation requires Compositional Alignment, defined here not merely as adherence to simultaneous constraints, but as the capability of a unified model to adaptively agree with human preferences across these optional and varying input conditions - human must determine which of two audio samples is superior in terms of musicality or better aligned with the provided compositional instructions. A framework capable of judging this versatility—handling cases where inputs may be text-only, lyric-guided, or audio-referenced—is currently missing. To bridge this gap, our proposed **CMI-RewardBench** integrates diverse task-specific datasets to vigorously evaluate whether a single reward model can judge generation quality against the heterogeneous instruction sets inherent to modern AIGC flows.

This paper introduces a comprehensive ecosystem for music reward modeling. Our contributions are three-fold:

1. We construct **CMI-Pref-Pseudo**, containing 110k samples labeled via a robust pipeline using Qwen3-Omni (Xu et al., 2025a) with consistency filtering. Complementing this, we introduce **CMI-Pref**, a high-quality corpus of 4,027 pairs annotated by 31 human experts. These annotations capture fine-grained preferences for musicality, alignment, and confidence levels across diverse genres, instruments, and multimodal prompts (including lyrics and audio-to-audio conditioning).

2. We propose **CMI-RewardBench**, a unified benchmark for music reward models. By integrating existing resources (PAM, MusicEval, Music Arena) with our CMI-Pref test split, this benchmark evaluates models on five distinct tasks ranging from absolute musicality scoring to complex compositional alignment. This unified approach serves as a rigorous testbed for model **versatility** across optional input settings. Our baseline evaluations on this benchmark expose a significant capability gap, revealing that even state-of-the-art multimodal LLMs (e.g., Gemini-2.5-Pro) struggle to exceed 80% agreement with human preferences.

3. We develop **CMI-RM**, a family of music reward models supporting compositional conditioning over text, lyrics, and audio. Uniquely supporting all evaluation settings in CMI-RewardBench via a single, parameter-efficient architecture (∼30M), CMI-RM achieves performance comparable to or better than specialized open-source baselines like SongEval. Furthermore, we demonstrate that CMI-RM provides measurable benefits when used for top-k filtering, enabling "inference-time scaling" for music generation.

We will release our dataset, benchmark, and model weights to facilitate future research in aligned music generation.

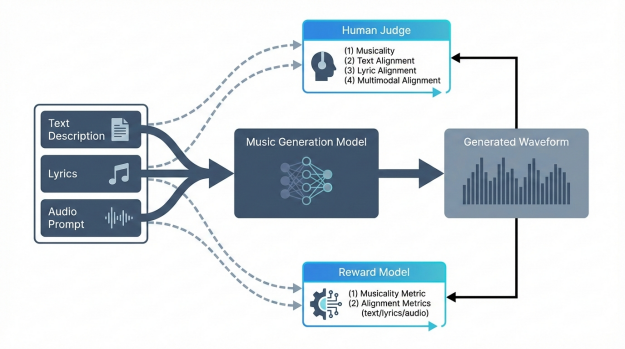


Figure 1. Comparison between traditional music evaluation and the human judge based on compositional multimodal instruction (CMI), which inspire our reward modeling benchmark framework.

2. Related Work

2.1. RLHF for LLMs and MLLMs

Reinforcement Learning from Human Feedback (RLHF) has successfully aligned generative models with human intent by replacing heuristic proxies (e.g., ROUGE) with learned reward models (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022). This paradigm has successfully extended to multimodal domains, including text-to-image generation (Kirstain et al., 2023; Xu et al., 2023) and video synthesis (Ahn et al., 2024; Liu et al., 2025b), to address aesthetic quality and semantic alignment. Recent research has also applied preference optimization to speech synthesis to improve naturalness and bridge inference-time distribution gaps (Zhang et al., 2024a; 2025b). Concurrently, the use of LLMs as scalable evaluators (Zheng et al., 2023; Gu et al., 2024) has emerged as a robust alternative to human experts.

2.2. Evaluation Metrics of Music

Music evaluation typically bifurcated into distribution-level quality assessment and sample-level alignment metrics. However, existing approaches struggle to address the complexity of compositional instructions.

Quality and Subjective Metrics. Distributional metrics like the Fréchet Audio Distance (FAD) (Kilgour et al., 2019) are the standard for global corpus assessment, with recent variants like MAD (Huang et al., 2025b) and KAD (Chung et al., 2025) improving correlation with human perception. For sample-level musicality, MOS predictors such as PAM (Deshmukh et al., 2024), Audiobox (Jeong et al., 2025) and SongEval (Yao et al., 2025) evaluate aesthetic quality. While effective for general audio, high-performing systems like MusicRL (Cideron et al., 2024), WhisQ (Emon et al., 2025), QAMRO (Wang et al., 2025a), and DRAGON (Bai et al., 2025) remain closed-source.

Alignment and LLM Judging. Alignment is primarily measured via contrastive scores like CLAP (Wu et al., 2023), CLaMP3 (Sharma et al., 2024), MuQ-Mulan (Zhu et al., 2025). While music-specific checkpoints improve human-preference alignment (Grötschla et al., 2025), these metrics are largely restricted to text-to-audio pairs and neglect lyrics or audio prompts. Besides, emerging "LLM-as-a-judge" frameworks like AutoMV (Tang et al., 2025) and music recommendation AutoRaters (Chen et al.) offer scalable, multimodality evaluation with complex instructions. However, these rely on proprietary models and lack an open-source framework for evaluating CMIs.

2.3. Preference Dataset and Platform

The development of robust reward models relies on standardized preference datasets and evaluation platforms. Early efforts such as MusicEval (Liu et al., 2025a), **SongEval** (Yao et al., 2025), and AudioEval (Wang et al., 2025b) provided the first expert-annotated corpora for absolute quality prediction, forming the basis for major community data challenge and benchmarks (Huang et al., 2025a; Ma et al., 2026; Zhang et al., 2025a). To better capture subjective nuances and reduce the cost of data collection, recent work has pivoted toward pairwise preference datasets, such as the AIME dataset (Grötschla et al., 2025). which benchmarks models via large-scale human comparisons .

Inspired by the crowdsourcing success of Chatbot Arena (Chiang et al.), GenAI Arena (Jiang et al., 2024), and Copilot Arena (Chi et al.) etc., the recently introduced Music Arena (Kim et al., 2025) provides a live platform for comparative text-to-music evaluation. While these resources represent significant progress, they primarily focus on text-to-music alignment. Our proposed **CMI-Pref** dataset fills a critical gap by providing the first large-scale preference corpus specifically targeting compositional instructions, including lyrics and audio-to-audio conditioning.

3. Method

3.1. Datasets for CMI-RewardBench

To develop and evaluate music reward models under compositional multimodal music instructions, we introduce a comprehensive dataset ecosystem. We move beyond simple text-to-music evaluation by incorporating existing datasets and a new, large-scale multimodal preference corpus.

3.1.1. Evaluating Musicality and Text-to-Music Alignment

To establish a foundation for single-modality evaluation, we curate a subset from three established resources: **PAM:** We include 500 audio clips from the music subset. Each clip is associated with MOS for musicality and text-music alignment with a text description. **MusicEval:** We utilize 413 samples from the test split. These provide expert-validated MOS for musicality (musical impression). Due to the mismatch of audio, filename and text prompt, we omit its MOS for text-music alignment. **Music Arena:** We process 2,800 historical interaction datafrom the Arena platform until Dec 2025. To ensure high-quality preference pairs, we remove instances where generation failed and filter out "tie" or "both bad" labels given different user has different tolerance margin when two given audio are similar, yielding 1,340 preference pairs.

3.1.2. CMI-Pref: A Large-Scale Compositional Multimodal Preference Dataset

We propose a new dataset **CMI-Pref** in our benchmark to capture human preferences when music is conditioned on CMIs including text descriptions, lyrics, and audio prompts.

Data Collection We distilled audio from a diverse set of 12 models and 11 commercial APIs to ensure a broad distribution of quality and style. For commercial APIs, we generated samples using Suno (v3.5, v4, v4.5, v4.5+, v5), Stable Audio 2.0[1], Minimax-Music-2.0[2], Mureka (v7.5, o2)[3], and Loudly[4]. These include equal splits of instrumental and vocal tracks (with lyrics) if the model supports lyrics as input, and equal splits of input with and without audio prompts if applicable. For open-source models, we generated audio from MusicGen (Copet et al., 2023), Stable Audio Open (Evans et al., 2025), YUE (Yuan et al., 2025), SongGen (Liu et al.), AudioLDM (Liu et al., 2023), AudioLDM 2 (Liu et al., 2024), DiffRhythm(Ning et al., 2025), Levo(Lei et al., 2025), Magenta Lyria-RealTime(Team et al., 2025), Jamify(Liu et al., 2025c), MusicLDM(Chen et al., 2024), and ACE-step(Gong et al., 2025). 35.6% of samples are conditioned on audio prompts for style transfer or continuation in addition to text and lyrics. Audio caption of the audio prompt is provided by Qwen3-Omni as additional text condition if model input cannot support audio prompt.

Annotation and Statistics We first utilized Qwen3-Omni to generate 130k pseudo-labels, retaining 110k pairs after consistency checks. This **CMI-Pref-Pseudo** dataset spans 47,546 generations (797.34h) across audio+lyrics (18.3%), audio-only (17.0%), lyrics-only (19.8%), and text-only (44.8%) settings. Following the annotation protocol described in C, 31 annotators constructed **CMI-Pref**, comprising 4,027 preference samples from generations produced by 23 models, with a total duration of 133.80 hours. The modality distribution remains balanced: audio+lyrics

[1] https://platform.stability.ai/
[2] https://platform.minimaxi.com/
[3] https://platform.mureka.ai/
[4] https://www.loudly.com/

Table 1. Statistics of CMI-Pref compared with previous datasets.

DATA SET	PAM	MUSICEVAL	SONGEVAL	MUSIC ARENA	CMI-PREF-PSEUDO	CMI-PREF
TEXT CONDITION	✔	UNAVAILABLE	✗	✔	✔	✔
LYRICS CONDITION	✗	✗	UNAVAILABLE	✔	✔	✔
AUDIO CONDITION	✗	✗	✗	✗	✔	✔
#SAMPLES	500	2748	2,399	2,800	110K	4,027
#TESTING SAMPLES	500	413	-	1,340	-	500
#DURATION OF AUDIO(HOURS)	0.83	16.62	140.54	88.30	797.34	133.80
#DURATION OF REFERENCE AUDIO(HOURS)	-	-	-	-	123.95	48.56
#UNIQUE PROMPTS	100	-	-	883	10,213	2,632
#MODELS & APIS	5	31	7	17	23	23

(15.2%), audio-only (20.4%), lyrics-only (14.9%), and text-only (49.5%). We reserved a balanced 500-pair test set and collected 2,574 rationale feedbacks (68% Chinese, 32% English). Table 1 highlights our distinct advantages in scale, duration, and modality diversity over existing benchmarks.

3.2. CMI-RewardBench

We introduce CMI-RewardBench, a unified benchmark designed to evaluate the capability of music reward models in capturing human aesthetic and instructional preferences.

3.2.1. EVALUATION PROTOCOL

To ensure robust evaluation across heterogeneous models and datasets, we employ two primary protocols: regression-based correlation and preference-based accuracy. For the PAM and MusicEval datasets, we evaluate reward models on absolute musicality and alignment scoring. Given that different models and datasets utilize varying score ranges, Mean Squared Error (MSE) is insufficient for measuring generalization. Instead, we prioritize relative trend alignment using Linear Correlation Coefficient (LCC) , Spearman Rank Correlation (SRCC) , and Kendall-Tau (K-Tau). For Music Arena and the CMI-Pref test split, we evaluate models on pairwise preference accuracy. Models must determine which of two audio samples is superior in terms of musicality or better aligned with the provided compositional instructions. Accuracy is calculated by comparing model predictions against experts' annotation.

3.2.2. BASELINE MODELS

We benchmark a diverse set of current state-of-the-art models, categorized by their primary training objective. The musicality baselines includes PAM (Deshmukh et al., 2024), audiobox (Jeong et al., 2025) and SongEval(Yao et al., 2025). PAM is a reference-free metric leveraging Audio-Language Models (ALMs) for general-purpose quality assessment. Audiobox-Aesthetic is a unified assessment framework providing four specialized metrics: Production Quality (PQ), Production Complexity (PC), Content Enjoyment (CE), and Content Usefulness (CU). SongEval is a model specifically trained on full-length songs using five aesthetic dimensions: coherence, memorability, naturalness, clarity, and overall musicality. For text-music alignment, we exam CLAP-Score in both default (Wu et al., 2023) and suggested music-specialized checkpoints (Grötschla et al., 2025), as well as CLAMP3 (Sharma et al., 2024) is a state-of-the-art multimodal alignment model for ABC symbolic notation, waveform and text. MuQ-MuLan(Zhu et al., 2025) is a joint music-text embedding framework utilizing self-supervised music representation learning. Furthermore, we evaluate the zero-shot capabilities of frontier AudioLLMs for all three tasks—musicality, text-music alignment, and compositional instruction alignment. The evaluated suite includes: Qwen: Qwen2-Audio (Chu et al., 2024), Qwen2.5-Omni (Xu et al., 2025b), and Qwen3-Omni. Gemini series including 2.5 Flash, 2.5 Pro (Comanici et al., 2025), and 3 Pro.

3.3. Compositional Music Reward Modeling

We develop a reward model architecture to handle multi-modal conditioning and predict fine-grained human scores.

3.3.1. MODEL ARCHITECTURE

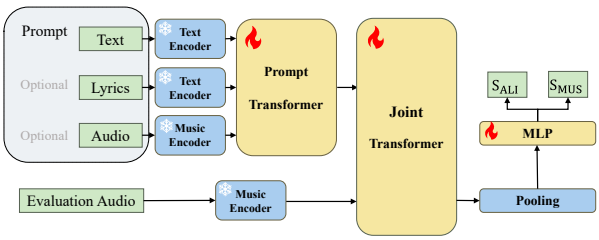


Figure 2. Model architecture of compositional music instruction reward model (CMI-RM).

Task Formulation. Given a CMI prompt

$$\mathcal{P} = (t, l, a_{\text{ref}}),$$

where t denotes optional text description, l optional lyrics, and a_{ref} optional reference audio, together with an evaluation audio a_{eval}, our reward model predicts human-aligned

preferences over generated music along the two complementary dimensions musicality (MUS) and alignment (ALI). The model outputs two scalar scores $(s_{\text{MUS}}, s_{\text{ALI}}) \in \mathbb{R}^2$.

Architecture. We adopt a two-tower multimodal architecture following Zhang et al. (2024b). One tower encodes the multimodal prompt, while the other processes the target audio. All encoders are frozen and instantiated from MuQ-MuLan. Text descriptions t and lyrics l are encoded separately using the text encoder, while reference audio a_{ref} and evaluation audio a_{eval} are encoded using the audio encoder. For each optional modality, we introduce a learnable null embedding when it is absent, enabling a unified architecture across different compositional instructions. The encoded prompt components are concatenated and fused using a 4-layer Prompt Transformer:

$$\mathbf{h}_{\text{prompt}} = \text{PromptTF}([\mathbf{E}_t; \mathbf{E}_l; \mathbf{E}_{a_{\text{ref}}}]), \tag{1}$$

where $[\cdot;\cdot]$ denotes sequence concatenation.

To model interactions between the prompt and the generated music, the fused prompt embedding and the evaluation audio embedding are concatenated and processed by a single-layer self-attention Joint Transformer:

$$\mathbf{h}_{\text{prompt}}; \mathbf{h}_{\text{eval}} = \text{JointTF}([\mathbf{h}_{\text{prompt}}; \mathbf{E}_{a_{\text{eval}}}]). \tag{2}$$

We extract the hidden states corresponding to the evaluation audio tokens, apply temporal pooling, and project them through a lightweight MLP to obtain the final scores:

$$(s_{\text{ALI}}, s_{\text{MUS}}) = \text{MLP}(\text{Pool}(\mathbf{h}_{\text{eval}})). \tag{3}$$

3.3.2. Training Strategy

We train the reward model using a two-stage pipeline that leverages both large-scale pseudo-labeled data and high-quality human annotations. Both MUS and ALI heads are optimized jointly throughout both stages. Given a training sample, we compute the musicality-related loss $\mathcal{L}_{\text{MUS}}$ and the alignment-related loss $\mathcal{L}_{\text{ALI}}$ when applicable. The overall training objective is a combination of the two:

$$\mathcal{L}_{\text{total}} = 0.5 \cdot \mathcal{L}_{\text{MUS}} + 0.5 \cdot \mathcal{L}_{\text{ALI}}. \tag{4}$$

Stage 1: Preference Pre-training. We first pre-train the model on preference pairs from CMI-Pref-Pseudo. Training is conducted for 2k steps with a batch size of 48. Pairwise preferences are modeled using Bradley–Terry (Bradley & Terry, 1952) formulation. Given a prompt $\mathcal{P}$ and two candidate audio files A and B,

$$P(A > B) = \sigma\big(s_\theta(\mathcal{P}, A) - s_\theta(\mathcal{P}, B)\big), \tag{5}$$

where s_θ denotes the predicted MUS or ALI score depending on the annotation type. The model is optimized using cross-entropy loss, and tied preferences are excluded from training. To mitigate over-confident decision boundaries induced by noisy pseudo labels, we apply label smoothing with a ratio of 0.2 during this stage.

Stage 2: Expert Fine-tuning. We then fine-tune the model on a mixture of high-quality human annotations, combining the training split of CMI-Pref and MusicEval, resulting in a total of 6,647 training samples. We use a batch size of 48 and perform early stopping based on validation performance. The selected checkpoint is obtained after 250 optimization steps selected with early stopping. Human annotations appear in two formats: (1) pairwise preferences $(\mathcal{P}, A, B)$, trained using Bradley–Terry loss as in Stage 1; and (2) scalar ratings $(\mathcal{P}, A, y)$, where $y \in [1, 5]$. For scalar ratings, we regress the predicted scores using

$$\mathcal{L}_{\text{reg}} = \text{MSE}\big(2\tanh(as + b) + 3,\ y\big), \tag{6}$$

where s denotes the predicted MUS or ALI score. The scaling parameters are initialized to $a = 0.2$ and $b = 0$ during fine-tuning. We drop these constants during inference.

3.3.3. Test-time Scaling

To evaluate the efficacy of test-time scaling (Jin et al., 2025), we conduct experiments using two backbone models: **MusicGen-small** and **Stable-Audio-Open-small**. For each of the 2,183 text prompts from the MusicCaps (Agostinelli et al., 2023) dataset eval-split, we generate 10 audio samples (10 sec each) per model. Our reward model serves as a "best-of-N" filter to select the top-performing sample, where $N \in \{1, 3, 10\}$. We evaluate the effectiveness via subjective A/B testing: Validating whether reward-model selection consistently aligns with human preferences for superior musical quality. The selection criterion is the average of musicality and alignment scores.

4. Discussion

4.1. Benchmark Results

4.1.1. Comprehensive Evaluation on Musicality

Superior Generalization on Musicality. Table 2 presents a quantitative comparison of our proposed **CMI-RM** against all baselines. Our method demonstrates superior generalization capabilities across regression tasks. Specifically, the model finetuned on CMI-Pref achieves state-of-the-art performance on the PAM music subset. Furthermore, incorporating external data (w/ f.t.: CMI + MusicEval) pushes the Music Arena accuracy to **73.21%** and MusicEval LCC to **0.8222**, indicating that our preference alignment strategy yields representations that are linearly correlated with granular human ratings.

Deficiency of General-Purpose MLLMs. Comparisons

Table 2. Musicality results on CMI-RewardBench. For each metric, the best performance is marked in boldface and second with underline.

Musicality	PAM (Music Subset)			MusicEval (Test Split)			Music Arena	CMI-Pref
Method&Model	LCC	SRCC	K-Tau	LCC	SRCC	K-Tau	ACC	ACC
PAM score	0.5873	0.6099	0.4367	0.6466	0.6724	0.4874	63.13%	65.40%
audiobox-CE	0.5283	0.5204	0.3665	0.6393	0.6599	0.4830	64.25%	71.80%
audiobox-CU	0.4645	0.4704	0.3279	0.6272	0.6764	0.4950	67.76%	71.40%
audiobox-PC	0.2505	0.2230	0.1552	0.1225	0.0768	0.0514	58.73%	59.00%
audiobox-PQ	0.4636	0.4513	0.3166	0.6016	0.6335	0.4620	67.54%	73.80%
SongEval-RM	<u>0.6987</u>	<u>0.6977</u>	<u>0.4997</u>	0.7140	0.6949	0.5185	**73.88%**	72.40%
Qwen2-audio	0.1468	0.1523	0.1120	0.1455	0.2196	0.1585	5.99%	8.60%
Qwen2.5-omni	0.2776	0.2837	0.2144	0.1655	0.1454	0.1145	36.05%	17.40%
Qwen3-omni	0.4155	0.4113	0.3146	0.3693	0.3101	0.2205	59.63%	60.40%
Gemini2.5-flash	0.3813	0.3693	0.2571	0.4188	0.3886	0.2694	64.12%	64.20%
Gemini2.5-pro	0.4463	0.4355	0.3068	0.4966	0.4902	0.3454	69.75%	70.00%
Gemini3-pro	0.5972	0.5967	0.4283	0.6044	0.6018	0.4400	68.85%	65.80%
- w/o f.t.: Distill	0.4277	0.4219	0.2919	0.5309	0.5189	0.3758	65.37%	70.40%
- w/ f.t.: CMI-Pref	**0.7003**	**0.6988**	**0.5106**	<u>0.7381</u>	<u>0.7435</u>	<u>0.5579</u>	71.57%	**78.60%**
- w/ f.t.: CMI + MusicEval	0.6417	0.6646	0.4806	**0.8222**	**0.8294**	**0.6486**	<u>73.21%</u>	<u>78.20%</u>

with general-purpose MLLMs reveal their critical lack of domain-specific sensitivity. Despite the massive scale of models like Gemini 3 Pro and Qwen3-omni, they lack the domain-specific sensitivity required for fine-grained musical assessment. For instance, on the internal CMI-Pref test set, Gemini 3 Pro and Qwen3-omni achieve accuracies of 65.80% and 60.40%, respectively. In sharp contrast, our CMI-RM reaches 78.60%, demonstrating a significant improvement against the baselines. This empirical evidence validates our hypothesis (see Introduction) that general instruction tuning is insufficient for aesthetic critique in the audio domain. Besides, the low accuracy rate of some AudioLLMs such as Qwen2-audio is due to their low instruction following rate, typically generating music captions or lyrics, making results significantly lower than random guess.

4.1.2. Compositional MultiModal Instruction Alignment Evaluation

Table 3 display the compositional alignment setting of evaluation. Unlike standard benchmarks that focus solely on text-to-music consistency, CMI-RewardBench utilizes our **CMI-Pref** dataset to test the capability to follow complex instructions involving text, lyrics, and reference audio.

Comparison with Objective Metrics and General MLLMs. Standard embedding-based metrics and general MLLMs struggle to handle the complexity of compositional instructions. For instance, CLAP scores generally yield moderate accuracy (∼62-64% on CMI-Pref Subsets) but fail to process audio prompts, resulting in missing data for audio-conditioned tasks. Similarly, while Qwen3-Omni achieves a high correlation on the standard PAM (Music Subset) with an LCC of 0.5841, its performance drops significantly on our comprehensive CMI-Pref benchmark (e.g., 64.0% ACC on w/ Audio). This discrepancy suggests that standard metrics often oversimplify the alignment problem by ignoring lyrics and reference audio constraints. In contrast, our CMI-RM (w/ f.t.: CMI-Pref) demonstrates superior alignment capabilities, achieving 73.6% accuracy on text-centric tasks (w/o Audio) and 78.4% on audio-centric tasks (w/ Audio), substantially outperforming Gemini 2.5 Pro (67.2% and 72.8%, respectively). This validates that our CMI-RM effectively bridges the gap between varying modalities, offering a unified solution for diverse generation controls.

Detailed Breakdown by Modality. The breakdown of CMI-Pref Subsets highlights the unique contributions of our dataset in teaching specific alignment skills. Regarding reference audio alignment, general models like Gemini 3 Pro perform ordinarily with 66.8% accuracy, indicating a lack of understanding of acoustic similarity. In contrast, our model finetuned on CMI-Pref achieves a 78.4% accuracy, proving that CMI-Pref provides essential supervision signals for audio-to-music alignment. For lyrics and text adherence, the variant w/ f.t.: CMI + MusicEval reaches 80.0% accuracy. This suggests that while CMI-Pref establishes a strong foundation for multimodal alignment, the inclusion of MusicEval further refines the model's sensitivity to semantic and lyrical content.

4.2. The effectiveness of different training sets

We ablate how training data sources affect CMI-RM under a fixed architecture and identical trainable parameters. All variants share the same two-head setup (MUS/ALI) and differ only in (i) initialization and (ii) fine-tuning data.

Training variants.

Table 3. Benchmark Results on Compositional Multimodal Instruction Alignment.

	PAM (Music Subset)			CMI-Pref (Subsets)					
Text-Music	✔	✔	✔	✔	✔	✔	✔	CMI-Pref w/o Audio	CMI-Pref w/ Audio
Lyrics-Music	✗	✗	✗	✗	✔	✗	✔		
Audio-Music	✗	✗	✗	✗	✗	✔	✔		
Metrics	LCC	SRCC	K-Tau	ACC	ACC	ACC	ACC	ACC	ACC
CLAP score (default)	0.4692	0.4517	0.3171	60.8%	64.0%	-	-	62.4%	-
CLAP score (music)	0.3192	0.2881	0.1978	67.2%	73.6%	-	-	70.4%	-
MuQ-Mulan	0.4984	0.4741	0.3341	64.8%	68.0%	-	-	66.4%	-
CLAMP3 score	0.2998	0.3013	0.2068	63.2%	62.4%	-	-	62.8%	-
Qwen2-audio	-0.024	-0.025	-0.020	0.8%	2.4%	8.8%	14.4%	1.6%	11.6%
Qwen2.5-Omni	0.1529	0.2084	0.1696	31.2%	37.6%	25.6%	32.0%	34.4%	28.8%
Qwen3-Omni	**0.5841**	**0.5907**	**0.4714**	67.2%	60.0%	64.8%	63.2%	63.6%	64.0%
Gemini2.5-Flash	0.3686	0.2454	0.1851	65.6%	56.0%	69.6%	54.4%	60.8%	62.0%
Gemini2.5-Pro	0.4562	0.4179	0.3192	**71.2%**	63.2%	73.6%	72.0%	67.2%	72.8%
Gemini3-Pro	0.5201	0.5373	0.4047	67.2%	60.8%	68.8%	64.8%	64.0%	66.8%
- w/o f.t.: Distill	0.3531	0.3463	0.2395	61.6%	74.4%	70.4%	75.2%	68.0%	72.8%
- w/ f.t.: CMI-Pref	0.5236	0.5280	0.3745	68.8%	78.4%	**76.8%**	80.0%	73.6%	**78.4%**
- w/ f.t.: CMI + MusicEval	0.4411	0.4328	0.3002	68.0%	**80.0%**	72.8%	**80.8%**	**74.0%**	76.8%

Distill Pre-trained on CMI-Pref-Pseudo only (pairwise preference, 1.5 epochs).

Distill+CMI Initialized from **Distill**, then fine-tuned on CMI-Pref (train split).

Distill+MusicEval Initialized from **Distill**, then fine-tuned on MusicEval (train split).

Distill+Both Initialized from **Distill**, then jointly fine-tuned on CMI-Pref + MusicEval.

Scratch+Both Random initialization, trained on CMI-Pref + MusicEval.

Table 4. Aggregated ablation results with a fixed CMI-RM architecture. Mean SRCC on PAM averages the *musicality* and *text–music alignment* regression tasks. Music Arena accuracy uses the combined score $s_{\text{MUS}}+s_{\text{ALI}}$ for pairwise prediction. Mean Acc. on CMI-Pref averages musicality and alignment preference accuracy.

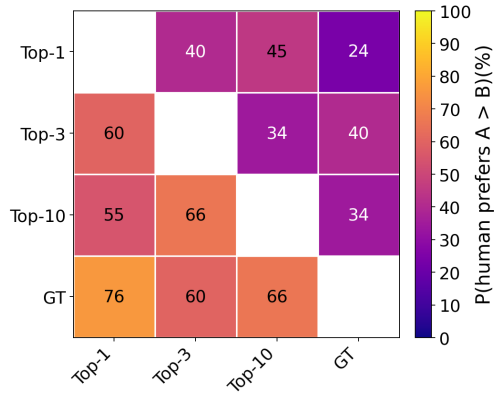
Training variant / Metric ↑	PAM Mean SRCC	MusicEval SRCC	Music Arena Acc.	CMI-Pref Mean Acc.
Distill (Pseudo only)	0.3841	0.5189	65.37%	70.4%
Distill+CMI (CMI-Pref)	**0.6134**	0.7435	71.57%	**77.3%**
Distill+MusicEval	0.4338	0.3460	64.78%	69.0%
Distill+Both (CMI-Pref+MusicEval)	0.5487	**0.8294**	73.21%	76.8%
Scratch+Both (CMI-Pref+MusicEval)	0.3844	0.8281	**75.0%**	75.4%
Qwen3-Omni (Judge)	0.5010	0.3101	59.63%	62.1%

Aggregated metric. To summarize results across heterogeneous benchmarks, we report one aggregated score per dataset in Table 4. For PAM, we average SRCC over the *musicality* and *text–music alignment* regression tasks. For MusicEval, we report SRCC on its musicality MOS. For Music Arena, we report preference accuracy, where the model prediction uses the combined score $s_{\text{MUS}} + s_{\text{ALI}}$ to match the platform's overall preference signal. For CMI-Pref, we report mean accuracy averaged over musicality and alignment preferences. Detailed per-task numbers are provided in Table 2, Table 3, and Table 13.

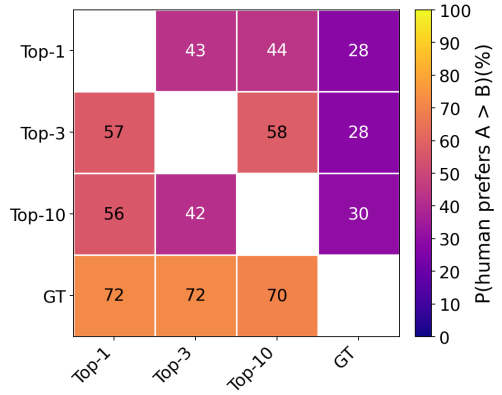
Findings. (1) CMI-Pref is the key driver for cross-benchmark generalization. Fine-tuning on CMI-Pref consistently improves all four aggregated metrics: **Distill+CMI** outperforms **Distill** on PAM (0.613 vs. 0.384), MusicEval (0.744 vs. 0.519), Music Arena (71.6% vs. 65.4%), and CMI-Pref (77.3% vs. 70.4%). This suggests that high-quality human preferences with compositional conditions provide transferable supervision signals.

(2) MusicEval provides complementary signal, but is insufficient alone. Distill+MusicEval does not yield consistent gains on the other benchmarks, while **Distill+Both** substantially improves MusicEval correlation (0.829) and maintains strong performance elsewhere. A mild trade-off is observed on PAM compared to **Distill+CMI**, indicating a distribution/objective mismatch between PAM and MusicEval that joint training partially balances.

(3) Distillation initialization helps most metrics, with a small trade-off on Music Arena. Comparing **Distill+Both**



(a) MusicGen-small



(b) Stable-Audio-Open-small

Figure 3. Pairwise preference matrices for test-time scaling with RM reranking. Each cell reports the percentage of trials in which annotators preferred system A (row) over system B (column).

vs. **Scratch+Both**, distillation initialization yields clear gains on PAM and CMI-Pref, while **Scratch+Both** performs slightly better on Music Arena. This is consistent with Music Arena exhibiting time-varying and platform-specific preference patterns (see Appendix D for a temporal drift analysis), which may favor a from-scratch fit to that distribution. Overall, **Distill+Both** remains the most balanced choice across benchmarks.

4.3. Test-Time Scaling

Table 5 shows that RM-based best-of-N reranking provides consistent test-time scaling gains across both backbones. The improvements are more pronounced for **MusicGen-small**, where MUQ-MULAN and AudioBox/SongEval metrics increase monotonically from N=1 to N=10. For **Stable-Audio-Open-small**, the gains are smaller and begin to saturate, suggesting diminishing returns when candidate quality is concentrated.

The human preference matrices in Fig. 3 provide two takeaways. First, **Ground Truth (GT) is consistently preferred over all reranked outputs** for both backbones, indicating that reranking improves generation quality but does not close the gap to real data. Second, preferences among Top-k selections are *not strictly monotonic* with N, and the gains can saturate (e.g., Stable Audio shows limited differences between N=3 and N=10). Overall, best-of-N reranking is a simple and effective test-time scaling strategy, with diminishing returns beyond moderate N.

Table 5. Test-time scaling results on objective metrics, AudioBox, and SongEval.

Model	Alignment	AudioBox		SongEval
	MuQ-MuLan↑	CE↑	CU↑	Mus↑
MusicGen	0.298	6.046	6.989	2.143
MusicGen (N=3)	0.323	6.405	7.255	2.213
MusicGen (N=10)	0.339	6.647	7.416	2.273
Stable Audio	0.293	5.567	7.170	2.055
Stable Audio (N=3)	0.301	5.732	7.245	2.078
Stable Audio (N=10)	0.307	5.799	7.290	2.090

5. Conclusion

We introduced **CMI-RewardBench**, a unified benchmark for evaluating music reward models under *Compositional Multimodal Instruction* (CMI), where the models handle optional and heterogeneous conditions (text-only, lyric-guided, and audio-referenced) and aligns with human preferences on both musicality and instruction following. To support this evaluation setting, we built **CMI-Pref**, combining (i) a large-scale pseudo-labeled corpus (**CMI-Pref-Pseudo**, 110k pairs) constructed via a robust Qwen3-Omni based pipeline with consistency filtering, and (ii) a high-quality expert-annotated set (**CMI-Pref**, 4k pairs) covering diverse genres, instruments, and multimodal prompts with confidence annotations. By integrating CMI-Pref with existing resources (e.g., PAM, MusicEval, Music Arena), CMI-RewardBench provides a rigorous testbed spanning five tasks from absolute scoring to pairwise preference, revealing a clear capability gap that even frontier multimodal LLM judges struggle to reach strong agreement with expert preferences in this setting.

We further developed **CMI-RM**, a parameter-efficient reward model that supports compositional conditioning over text, lyrics, and audio within a single architecture, achieving performance competitive with or exceeding specialized open-source baselines, and providing measurable gains when used for best-of-N reranking as a simple inference-time scaling strategy. We will release the dataset, benchmark, and model weights to facilitate future research.

Acknowledgement

Impact Statement

This work advances the evaluation and alignment of music generation systems by providing a unified benchmark and lightweight reward models that better reflect human preference under compositional multimodal conditions. Our pipeline includes audio generated via commercial APIs; therefore we follow a TOS-aware release policy: only components that are explicitly redistributable are made public, and any restricted parts are shared (if at all) via an application-based access process. We also acknowledge potential copyright concerns from style/melody similarity in generated music and provide a takedown/correction mechanism. Human preference data are collected via a MusicArena-style platform under informed consent and data minimization (no personal identifiers; limited metadata), with an opt-out/withdrawal mechanism. Where permitted, the dataset is released under **CC BY-NC-SA** and accompanied by a datasheet/data card documenting sources, licenses/TOS compatibility, and the public release fields.

References

Agostinelli, A., Musil, T. I., Roblek, D., Han, M., Bavishi, R., Zeghidour, N., Li, C., Ritter, M., Rolland, C., Piergiovanni, A., et al. Musiclm: Generating music from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16225–16236, 2023.

Ahn, D., Choi, Y., Yu, Y., Kang, D., and Choi, J. Tuning large multimodal models for videos using reinforcement learning from AI feedback. In Ku, L., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 923–940. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.52. URL https://doi.org/10.18653/v1/2024.acl-long.52.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Bai, Y., Casebeer, J., Sojoudi, S., and Bryan, N. J. Dragon: Distributional rewards optimize diffusion generative models. *arXiv preprint arXiv:2504.15217*, 2025.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.2307/2334029.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P., and Sun, L. A survey of ai-generated content (aigc). *ACM Computing Surveys*, 57(5):1–38, 2025.

Chen, K., Wu, Y., Liu, H., Nezhurina, M., Berg-Kirkpatrick, T., and Dubnov, S. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1206–1210. IEEE, 2024.

Chen, S., Vasilevski, Y., Lampinen, A. K., Ahmad, A., Ndebele, N., Goldman, S., Mozer, M. C., and Ren, J. Comparing human and llm ratings of music-recommendation quality with user context. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Chi, W., Chen, V., Angelopoulos, A. N., Chiang, W.-L., Mittal, A., Jain, N., Zhang, T., Stoica, I., Donahue, C., and Talwalkar, A. Copilot arena: A platform for code llm evaluation in the wild. In *Forty-second International Conference on Machine Learning*.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J. E., et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Chung, Y., Eu, P., Lee, J., Choi, K., Nam, J., and Chon, B. S. Kad: No more fad! an effective and efficient evaluation metric for audio generation. *arXiv preprint arXiv:2502.15602*, 2025.

Cideron, G., Girgin, S., Verzetti, M., Vincent, D., Kastelic, M., Borsos, Z., Mcwilliams, B., Ungureanu, V., Bachem, O., Pietquin, O., et al. Musicrl: Aligning music generation to human preferences. In *International Conference on Machine Learning*, pp. 8968–8984. PMLR, 2024.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Défossez, A. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36:47704–47720, 2023.

Deshmukh, S., Alharthi, D., Elizalde, B., Gamper, H., Al Ismail, M., Singh, R., Raj, B., and Wang, H. Pam: Prompting audio-language models for audio quality assessment. In *Proc. Interspeech 2024*, pp. 3320–3324, 2024.

Emon, J. I., Alam, K. T., and Salek, M. A. Whisq: Cross-modal representation learning for text-to-music mos prediction. *arXiv preprint arXiv:2506.05899*, 2025.

Evans, Z., Parker, J. D., Carr, C., Zukowski, Z., Taylor, J., and Pons, J. Stable audio open. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pp. 1–5. IEEE, 2025. doi: 10.1109/ICASSP49660.2025.10888461. URL https://doi.org/10.1109/ICASSP49660.2025.10888461.

Gong, J., Zhao, S., Wang, S., Xu, S., and Guo, J. Ace-step: A step towards music generation foundation model. *arXiv preprint arXiv:2506.00045*, 2025.

Grötschla, F., Solak, A., Lanzendörfer, L. A., and Wattenhofer, R. Benchmarking music generation models and metrics via human preference studies. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

Huang, W.-C., Wang, H., Liu, C., Wu, Y.-C., Tjandra, A., Hsu, W.-N., Cooper, E., Qin, Y., and Toda, T. The audiomos challenge 2025. *arXiv preprint arXiv:2509.01336*, 2025a.

Huang, Y., Novack, Z., Saito, K., Shi, J., Watanabe, S., Mitsufuji, Y., Thickstun, J., and Donahue, C. Aligning text-to-music evaluation with human preferences. *arXiv preprint arXiv:2503.16669*, 2025b.

Jeong, Y., Kim, Y., Chun, S., and Lee, J. Read, watch and scream! sound generation from text and video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17590–17598, 2025.

Jiang, D., Ku, M., Li, T., Ni, Y., Sun, S., Fan, R., and Chen, W. Genai arena: An open evaluation platform for generative models. *Advances in Neural Information Processing Systems*, 37:79889–79908, 2024.

Jin, Y., Ye, Z., Tian, Z., Liu, H., Kong, Q., Guo, Y., and Xue, W. Inference-time scaling for diffusion-based audio super-resolution. *arXiv preprint arXiv:2508.02391*, 2025.

Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proc. Interspeech 2019*, pp. 2350–2354, 2019.

Kim, Y., Chi, W., Angelopoulos, A. N., Chiang, W.-L., Saito, K., Watanabe, S., Mitsufuji, Y., and Donahue, C. Music arena: Live evaluation for text-to-music. *arXiv preprint arXiv:2507.20900*, 2025.

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., and Levy, O. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Lei, S., Xu, Y., Lin, Z., Zhang, H., Tan, W., Chen, H., Yu, J., Zhang, Y., Yang, C., Zhu, H., et al. Levo: High-quality song generation with multi-preference alignment. *arXiv preprint arXiv:2506.07520*, 2025.

Liu, C., Wang, H., Zhao, J., Zhao, S., Bu, H., Xu, X., Zhou, J., Sun, H., and Qin, Y. Musiceval: A generative music dataset with expert ratings for automatic text-to-music evaluation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025a.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., and Plumbley, M. D. Audioldm: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning*, pp. 21450–21474. PMLR, 2023.

Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Wang, Y., Wang, W., Wang, Y., and Plumbley, M. D. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024.

Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Xia, M., Wang, X., et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025b.

Liu, R., Hung, C.-Y., Majumder, N., Gautreaux, T., Bagherzadeh, A. A., Li, C., Herremans, D., and Poria, S. Jam: A tiny flow-based song generator with fine-grained controllability and aesthetic alignment. *arXiv preprint arXiv:2507.20880*, 2025c.

Liu, Z., Ding, S., Zhang, Z., Dong, X., Zhang, P., Zang, Y., Cao, Y., Lin, D., and Wang, J. Songgen: A single stage auto-regressive transformer for text-to-song generation. In *Forty-second International Conference on Machine Learning*.

Ma, G., Xia, Y., Yao, J., Xue, H., Liu, H., Wang, S., Liu, H., and Xie, L. The icassp 2026 automatic song aesthetics evaluation challenge. *arXiv preprint arXiv:2601.07237*, 2026.

Ma, Y., Øland, A., Ragni, A., Del Sette, B. M., Saitis, C., Donahue, C., Lin, C., Plachouras, C., Benetos, E., Shatri, E., et al. Foundation models for music: A survey. *arXiv preprint arXiv:2408.14340*, 2024.

Ning, Z., Chen, H., Jiang, Y., Hao, C., Ma, G., Wang, S., Yao, J., and Xie, L. Diffrhythm: Blazingly fast and embarrassingly simple end-to-end full-length song generation with latent diffusion. *arXiv preprint arXiv:2503.01183*, 2025.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Papreja, P., Venkateswara, H., and Panchanathan, S. Representation, exploration and recommendation of playlists. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 543–550. Springer, 2019.

Ren, L., Wang, H., Li, J., Tang, Y., and Yang, C. Aigc for industrial time series: From deep-generative models to large-generative models. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2025.

Sharma, Y. et al. Clamp3: Universal music information retrieval with contrastive language-audio pretraining. *arXiv preprint arXiv:2410.00001*, 2024.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Tang, X., Lei, X., Zhu, C., Chen, S., Yuan, R., Li, Y., Oh, C., Zhang, G., Huang, W., Benetos, E., et al. Automv: An automatic multi-agent system for music video generation. *arXiv preprint arXiv:2512.12196*, 2025.

Team, L., Caillon, A., McWilliams, B., Tarakajian, C., Simon, I., Manco, I., Engel, J., Constant, N., Li, Y., Denk, T. I., et al. Live music models. *arXiv preprint arXiv:2508.04651*, 2025.

Wang, C.-C., Huang, K.-T., Yang, C.-Y., Lee, H.-S., Wang, H.-M., and Chen, B. Qamro: Quality-aware adaptive margin ranking optimization for human-aligned assessment of audio generation systems. *arXiv preprint arXiv:2508.08957*, 2025a.

Wang, H., Zhao, J., Liu, C., Jia, Y., Sun, H., Zhou, J., and Qin, Y. Audioeval: Automatic dual-perspective and multi-dimensional evaluation of text-to-audio-generation. *arXiv preprint arXiv:2510.14570*, 2025b.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 15903–15935, 2023.

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

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

Yao, J., Ma, G., Xue, H., Chen, H., Hao, C., Jiang, Y., Liu, H., Yuan, R., Xu, J., Xue, W., et al. Songeval: A benchmark dataset for song aesthetics evaluation. *arXiv preprint arXiv:2505.10793*, 2025.

Yuan, R., Lin, H., Guo, S., Zhang, G., Pan, J., Zang, Y., Liu, H., Liang, Y., Ma, W., Du, X., Du, X., Ye, Z., Zheng, T., Ma, Y., Liu, M., Tian, Z., Zhou, Z., Xue, L., Qu, X., Li, Y., Wu, S., Shen, T., Ma, Z., Zhan, J., Wang, C., Wang, Y., Chi, X., Zhang, X., Yang, Z., Wang, X., Liu, S., Mei, L., Li, P., Wang, J., Yu, J., Pang, G., Li, X., Wang, Z., Zhou, X., Yu, L., Benetos, E., Chen, Y., Lin, C., Chen, X., Xia, G., Zhang, Z., Zhang, C., Chen, W., Zhou, X., Qiu, X., Dannenberg, R. B., Liu, Z., Yang, J., Huang, W., Xue, W., Tan, X., and Guo, Y. Yue: Scaling open foundation models for long-form music generation. *CoRR*, abs/2503.08638, 2025. doi: 10.48550/ARXIV.2503.08638. URL https://doi.org/10.48550/arXiv.2503.08638.

Zhang, D., Li, Z., Li, S., Zhang, X., Wang, P., Zhou, Y., and Qiu, X. Speechalign: Aligning speech generation to human preferences. *Advances in Neural Information Processing Systems*, 37:50343–50360, 2024a.

Zhang, H., Liang, J., Phan, H., Wang, W., and Benetos, E. From aesthetics to human preferences: Comparative perspectives of evaluating text-to-music systems. *arXiv preprint arXiv:2504.21815*, 2025a.

Zhang, S., Yang, X., Feng, Y., Qin, C., Chen, C.-C., Yu, N., Chen, Z., Wang, H., Savarese, S., Ermon, S., Xiong, C., and Xu, R. Hive: Harnessing human feedback for instructional visual editing, 2024b. URL https://arxiv.org/abs/2303.09618.

Zhang, X., Wang, C., Liao, H., Li, Z., Wang, Y., Wang, L., Jia, D., Chen, Y., Li, X., Chen, Z., et al. Speechjudge: Towards human-level judgment for speech naturalness. *arXiv preprint arXiv:2511.07931*, 2025b.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623, 2023.

Zhu, H., Zhou, Y., Chen, H., Yu, J., Ma, Z., Gu, R., Luo, Y., Tan, W., and Chen, X. Muq: Self-supervised music representation learning with mel residual vector quantization, 2025. URL https://arxiv.org/abs/2501.01108.

A. Detailed Metadata of Datasets

A.1. Diversity of Prompts

CMI-Pref contains a highly diverse collection of prompts and lyrics that reflect realistic and heterogeneous user instructions for music generation. Across the full dataset, we collect 10,213 unique prompts, while the human-annotated split contains 2,788 unique prompts. Prompt lengths follow a long-tailed distribution: most prompts are concise style or intent descriptors consisting of a few words, while a non-trivial subset includes longer compositional instructions specifying multiple musical attributes. This range captures both minimal user inputs and more detailed requests that require structured reasoning from reward models.

Semantically, the prompts span a wide variety of musical attributes. They cover a broad spectrum of genres, including popular, electronic, rock, jazz, classical, ambient, folk, and orchestral styles, with many prompts combining multiple genre cues. Beyond genre, prompts frequently specify mood, tempo, instrumentation, and production characteristics, introducing orthogonal dimensions of variation. Such compositional prompts prevent reward models from relying on single-keyword correlations and instead encourage holistic assessment of instruction adherence.

The prompt distribution is also linguistically diverse. While English prompts dominate, the dataset includes non-English and mixed-language prompts, reducing reliance on a single linguistic prior and better reflecting global usage patterns. Importantly, prompts are paired with generations produced by a wide range of music generation models and commercial APIs, ensuring that instruction diversity is not confounded with a specific synthesis pipeline.

In addition to text prompts, CMI-Pref includes lyric-conditioned instructions at a meaningful scale. The human-annotated split contains 840 examples with non-empty lyrics, and the full dataset contains 3,896 such examples. Lyrics vary substantially in structure and length, ranging from short repetitive hooks to multi-stanza verses with clear narrative progression. They include both vocal-focused instructions and lyrics intended to be adapted to different musical styles, posing additional alignment challenges beyond text-to-music generation.

Overall, the diversity of prompts and lyrics in CMI-Pref provides a realistic and challenging testbed for reward modeling. By combining short and long instructions, multiple semantic control dimensions, diverse musical styles, and lyric-conditioned inputs, the dataset supports robust evaluation of reward models under compositional and multimodal instruction settings.

Table 6. Inter-annotator agreement on overlapping votes, computed over 644 vote pairs from 492 comparisons with multiple annotations.

Metric	Instruction Following	Music Quality
Agreement Count	445	466
Disagreement Count	199	178
Agreement Rate	0.691	0.724
Krippendorff's α	0.382	0.447

A.2. Annotators Agreement of Overlapping Votes

We analyze inter-annotator consistency for comparisons that received multiple independent votes. A total of 492 comparison pairs were annotated by two or more annotators, producing 1,056 votes. For each comparison, we form all pairwise combinations of its votes, resulting in 644 vote pairs.

We measure agreement using *agreement rate* and *Krippendorff's alpha*, computed separately for **instruction following** and **music quality**; we adopt Krippendorff's alpha as it supports agreement estimation when annotator identities are not fixed across vote pairs. As shown in Table 6, both dimensions exhibit relatively high agreement rate. Music quality judgments exhibit slightly higher consistency than instruction following, as they are more intuitive and place lower demands on specialized musical knowledge. After correcting for chance agreement, Krippendorff's alpha indicates a *moderate* level of true agreement, which we consider reasonable given the inherent subjectivity of music evaluation.

A.3. Confidence Scores of Human Annotation

We analyze the distribution of annotators' confidence scores (Fig. 4). Overall, most votes are associated with relatively high confidence. The average confidence for music quality is slightly higher than that for instruction following, consistent with earlier observations that music quality judgments are more intuitive.

Table 7. Agreement between instruction-following and music-quality preferences, and the corresponding average confidence.

Metric	Value
#Votes with same preference	3309
#Votes with different preference	718
Agreement rate	0.822
Avg. confidence (same Pref.)	IF: 3.674, MQ: 3.734
Avg. confidence (diff. Pref.)	IF: 3.171, MQ: 3.311

We further examine the relationship between agreement across instruction following and music quality and annotator confidence (Table 7). Overall, the two dimensions show a high agreement rate, and votes have higher average confidence when they agree. This suggests that disagreements between instruction following and music quality often correspond to more ambiguous or trade-off cases, where annotators are less certain about the overall preference.

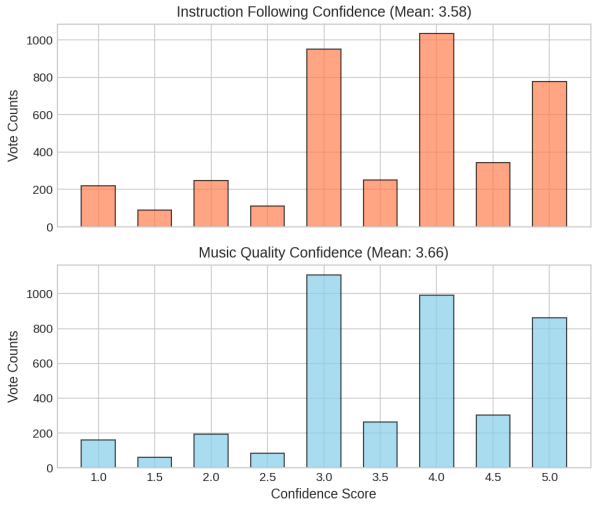


Figure 4. Distribution of annotators' confidence scores for instruction following and music quality.

Alignment among heads: We observed a high consistency between musicality and instruction-music alignment preferences. In pseudo-labels, the agreement is 91% between musicality and instruction alignment. In human data, this agreement is lower (81%), but conflicts typically occur when annotators report high confidence in instruction alignment despite lower musicality, highlighting the necessity of evaluating these dimensions separately.

B. Pseudo-label Generation for CMI-Pseudo

B.1. Label Acquisition Protocol

To ensure the reliability of our pseudo-labels, we address the well-documented phenomenon of positional bias in Large Language Models (LLMs), where the model's preference is influenced by the presentation order of the options rather than the content quality alone.

Our protocol for collecting pseudo-labels adopts a *Position-Consistency* strategy. For a given pair of audio samples (A, B) generated from the same prompt P, we conduct a bidirectional assessment:

1. **Forward Pass:** We query the model with the sequence (A, B) to obtain preference L_{fwd}.
2. **Reverse Pass:** We swap the positions to (B, A) and query the model again to obtain preference L_{rev}.

A pseudo-label is considered valid and retained only if the judgment is invariant to position—that is, the model prefers

the same underlying audio clip in both the forward and reverse passes ($L_{fwd} = L_{rev}$). Comparisons yielding conflicting results or inconsistent ties are discarded as hallucinations or uncertain boundaries.

B.2. Bias Analysis and Dataset Statistics

We initially sampled 129,545 pairwise comparisons from our generated audio pool. As demonstrated in Table 8, we observe significant distributional shifts when the presentation order is swapped, confirming the presence of positional bias. For instance, in the Musicality metric, the win rate for Candidate A fluctuates from 51.96% in the original configuration to 59.27% in the reversed configuration.

By applying our consistency filter, we retain only the high-confidence labels, resulting in 114,694 valid musicality labels and 117,828 valid alignment labels. As shown in the "Agreed" rows of Table 8, the resulting distribution stabilizes, effectively mitigating the variance introduced by the model's sensitivity to input order. For the final dataset construction, we retained the **intersection** of these valid subsets to ensure quality across all dimensions, yielding approximately 110k pairs.

Configuration	Win A (%)	Win B (%)	Tie (%)
Musicality			
Original (A, B)	51.96	40.96	7.08
Reversed (B, A)	59.27	33.48	7.25
Agreed (Filtered)	**57.80**	**36.97**	**5.22**
Alignment			
Original (A, B)	33.65	43.71	22.64
Reversed (B, A)	38.75	38.84	22.77
Agreed (Filtered)	**36.05**	**41.23**	**22.73**

Table 8. Distribution of pseudo-labels across varying query configurations. "Original" and "Reversed" denote the presentation order of the audio pair. "Agreed" represents the subset of labels where the model's preference remained consistent across both permutations. The discrepancy between Original and Reversed highlights positional bias, which is rectified in the Agreed set.

The "Original" and "Reversed" distributions should theoretically be identical if the evaluator were perfectly unbiased. The observed divergence necessitates our rigorous filtering approach. We did not shuffle the comparison pairs during analysis, therefore its normal that the labels A, B are uneven.

C. Human Annotation Details

C.1. Annotation Protocol for CMI-Pref

C.1.1. Core Objectives

The annotation process is decomposed into three distinct components to ensure a multi-dimensional evaluation of the generated music:

1. **Preference Label (A/B):** A forced-choice selection between two candidates.
2. **Confidence Score (1–5):** A quantitative measure of the annotator's certainty, grounded in constraint satisfaction (for alignment) or quality delta (for musicality).
3. **Free-text Feedback:** Qualitative justifications focusing on fine-grained details that discrete labels cannot capture.

C.1.2. General Principles

- **Instruction-First:** Annotators must strictly evaluate the *instruction/prompt* before listening to avoid post-hoc rationalization.
- **Holistic and Granular Review:** Each sample is evaluated for overall coherence as well as specific details (instrumentation, emotion, structure, and audio fidelity).
- **Dimensional Isolation:** Annotators are instructed to decouple **Alignment** (adherence to the prompt) from **Musicality** (aesthetic quality and production value). A sample may win in alignment while losing in musicality.

C.1.3. Q1: Textual Music Alignment Preference

Annotators identify which sample better follows the elements specified in the instruction, regardless of aesthetic appeal. Key alignment dimensions include:

- **Instrumentation:** Specific instruments (e.g., piano solo, guitar riff).
- **Mood/Atmosphere:** Emotional valence (e.g., melancholy, upbeat, tense).
- **Genre/Style/Era:** Stylistic markers (e.g., Lo-fi, Baroque, 80s synth).
- **Rhythm/Tempo:** Temporal characteristics (e.g., driving drum beat, groovy, energetic).

Confidence Calibration (1–5):

- **5 (Very Certain):** Clear binary distinction; one satisfies all key constraints while the other fails or collapses.
- **3 (Moderate):** Default choice; a perceptible lean toward one candidate without overwhelming dominance.
- **1 (Uncertain):** Highly ambiguous prompts or both samples are indistinguishable in their failure/success.

C.1.4. Q2: Musicality Preference

Annotators evaluate which sample sounds more like a "finished, natural, and professional" musical work, independent of the prompt.

- **Key Criteria:** Melodic memorability, structural progression, rhythmic stability, and production clarity (lack of distortion/artifacts).
- **Confidence:** Reflects the perceived "quality gap" between candidates.

C.1.5. Q3: Feedback Guidelines

Feedback should consist of 1–3 concise sentences focusing on "audible evidence." Annotators are encouraged to use specific timestamps and avoid vague descriptors.

- **Positive Justification:** "Sample A aligns better due to the presence of the specified saxophone; the melody is more distinct."
- **Negative Evidence:** "Sample B suffers from rhythmic instability at 0:20 and harsh high-frequency distortion."

C.1.6. Glossary for Annotators

Table 9. Taxonomy of musical attributes used in the annotation process.

Dimension	Positive Descriptors	Negative Descriptors
Melody	Catchy, Memorable, Distinct	Repetitive, Generic, Wandering
Structure	Coherent, Progression, Build-up	Disjointed, Abrupt, Random Loops
Rhythm	Groovy, Steady, Driving	Off-beat, Unstable, Chaotic
Audio Quality	Clean mix, Balanced, High-fidelity	Harsh, Muddy, Distorted/Clipping
Vocals	Natural, Clear articulation	Robotic, Slurred, Artifacts

C.2. Annotation Platform

Our annotation platform is illustrated in Figure 5. It provides a unified interface for pairwise audio comparison, confidence scoring, and free-text feedback collection. Detailed usage instructions are provided in the README included in the submitted supplementary materials.

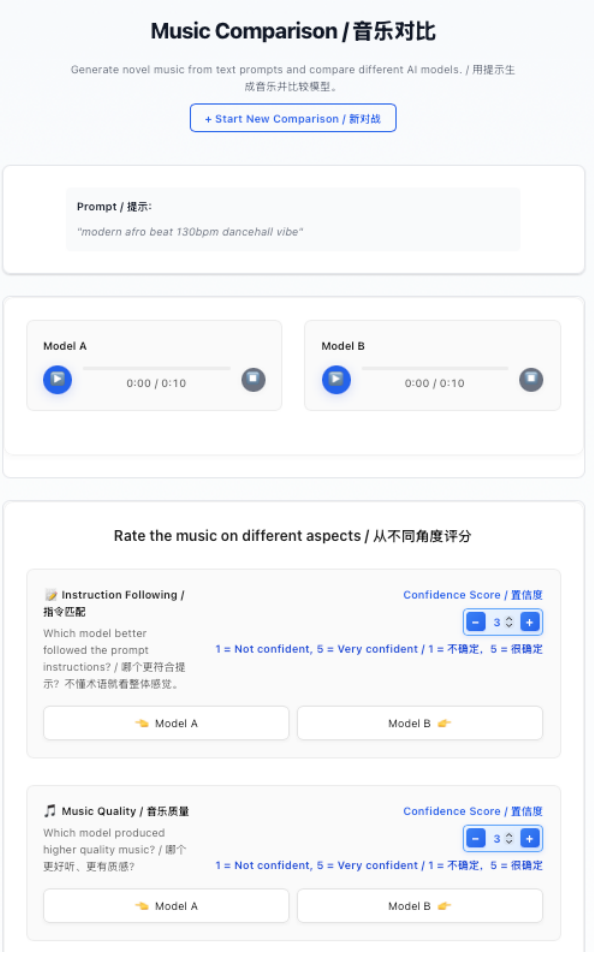


Figure 5. Platform of human annotation.

D. Detailed Analysis of Results

D.1. Performance on Music Arena Subsets

Tables 10 and 11 reveal two dominant factors that influence reward model performance: temporal distribution shift and annotation confidence.

Table 10 shows a clear temporal drift in Music Arena. Several reference-free metrics, particularly the Audiobox variants, exhibit a noticeable decline in accuracy from early to later months, indicating sensitivity to evolving generation quality and user preference distributions. SongEval-RM demonstrates relatively stronger stability across time, while general-purpose multimodal LLMs show substantial month-to-month variance, suggesting limited robustness for fine-grained musicality judgment. In contrast, our models maintain more consistent performance over time and achieve improved robustness on vocal music, especially when jointly fine-tuned with MusicEval, highlighting the benefit of music-specific supervision.

D.2. Performance on CMI-Pref Subsets

Table 11 analyzes performance on the CMI-Pref test set by stratifying samples according to annotator-reported confidence, which in our data collection protocol explicitly measures the perceived *preference margin* between two candidates. Unlike traditional pairwise annotation schemes that only record a binary choice, annotators are required to select a preferred sample and additionally indicate how strongly one sample is preferred over the other.

Table 10. Benchmark Musicality ACC Results on Music Arena (Time & Data Type Analysis).

Model	Total (%)	Time Period (Month)					Data Type	
		Jul-Aug	Sep	Oct	Nov	Dec	Instru	Vocal
PAM score	63.13	69.13	75.41	58.56	53.14	55.27	56.84	68.35
audiobox-CE	64.25	67.79	73.77	63.36	60.14	55.27	64.25	64.25
audiobox-CU	67.76	71.36	77.0	66.78	60.14	60.73	68.04	67.53
audiobox-PC	58.73	63.75	70.49	55.82	54.54	48.00	52.88	63.57
audiobox-PQ	67.54	68.01	74.86	67.47	65.04	63.27	68.04	67.12
SongEval-RM	**73.88**	73.60	**78.69**	71.92	64.34	**78.18**	**77.27**	71.08
Qwen2-audio	35.99	38.26	37.91	31.23	38.24	34.91	30.66	39.63
Qwen2.5-Omni	36.05	40.04	39.89	38.36	24.48	30.55	29.98	41.23
Qwen3-Omni	59.63	59.06	58.47	59.93	60.84	60.36	62.93	56.89
Gemini2.5-Flash	64.12	60.54	62.84	69.52	61.27	66.55	65.84	62.70
Gemini2.5-Pro	69.75	65.99	64.48	71.88	73.24	75.27	75.78	64.69
Gemini3-Pro	68.85	63.59	60.28	74.31	**78.63**	72.22	76.03	62.46
- w/o f.t.: Distill	65.37	67.56	68.3	67.8	68.53	55.64	63.26	67.12
- w/ f.t.: CMI-Pref	71.57	71.36	75.41	72.95	65.04	71.27	73.81	69.71
- w/ f.t.: CMI + MusicEval	73.21	**74.27**	75.96	**74.66**	65.73	72.00	74.14	**72.44**

Table 11. Benchmark Musicality ACC Results on CMI-Pref (Ablation Analysis).

Musicality	Total (500)	Confidence Level			Data Type	
		Conf <3 (66)	Conf $=3$ (128)	Conf >3 (306)	Instru (250)	Vocal (250)
PAM score	65.40%	63.64%	64.84%	66.01%	67.60%	63.20%
audiobox-CE	71.80%	63.64%	72.66%	73.20%	70.80%	72.80%
audiobox-CU	71.40%	66.67%	71.88%	72.22%	68.80%	74.00%
audiobox-PC	59.00%	46.97%	54.69%	63.40%	60.00%	58.00%
audiobox-PQ	73.80%	**74.24%**	**78.13%**	71.90%	73.20%	74.40%
SongEval-RM	72.40%	63.64%	71.09%	74.84%	70.80%	74.00%
Qwen2-audio	8.60%	4.54%	7.81%	9.80%	5.20%	12.00%
Qwen2.5-omni	17.40%	15.15%	16.41%	18.40%	15.20%	19.60%
Qwen3-omni	60.40%	54.55%	53.12%	64.70%	63.20%	57.60%
Gemini2.5-flash	64.20%	56.57%	57.94%	71.94%	74.40%	54.00%
Gemini2.5-pro	70.00%	48.49%	69.29%	75.96%	**78.80%**	61.20%
Gemini3-pro	65.80%	57.94%	66.67%	72.67%	73.20%	58.40%
- w/o f.t.: Distill	70.40%	54.55%	70.31%	73.86%	64.40%	76.40%
C- w/ f.t.: CMI-Pref	**78.60%**	62.12%	77.34%	**85.71%**	74.80%	82.00%
- w/ f.t.: CMI + MusicEval	78.20%	65.15%	76.56%	81.70%	73.20%	**83.20%**

We observe a clear monotonic relationship between confidence and model accuracy across all methods. Higher-confidence comparisons, corresponding to larger perceptual gaps in musical quality or instruction adherence, are consistently easier to predict. Lower-confidence comparisons reflect fine-grained distinctions with smaller margins, which naturally impose a more challenging learning problem.

Importantly, our proposed reward models exhibit the largest performance gains in the high-confidence regime. The CMI-Pref fine-tuned model substantially outperforms all baselines when confidence is high, indicating strong alignment with clear and decisive human preferences. At low confidence levels, performance differences across methods narrow, suggesting that improvements on near-tie comparisons are fundamentally limited by the small preference margins rather than model capacity.

Overall, these results demonstrate that explicitly modeling preference strength during data collection provides a principled way to analyze reward model behavior beyond binary accuracy, and highlights the effectiveness of our approach in capturing dominant human preference signals relevant for downstream reranking and selection.

E. Reward Model

E.1. Ablation Study on Mapping Functions

We investigate the impact of different mapping functions used to transform the predicted preference scores for MOS regression. We compare four settings:

- **None**: No transformation is applied.
- **Tanh**: The method adopted in our main experiments, using a Tanh activation to bound the output.
- **Linear**: A linear projection without Tanh activation.
- **Ordinal**: We employ an ordinal regression objective with learnable classification margins during training. Note that for evaluation, we use the model's raw scalar output to compute metrics, bypassing the learned margins.

Table 12. Benchmark Text-Music Alignment ACC Results on CMI-Pref(Ablation Analysis).

Text-Music Alignment	Total (250)	Confidence Level			Data Type	
		Conf < 3 (37)	Conf $= 3$ (66)	Conf > 3 (147)	Instru (125)	Vocal (125)
audiobox-CE	60.00%	45.95%	63.64%	61.90%	56.80%	63.20%
audiobox-CU	59.60%	54.05%	63.64%	59.18%	53.60%	65.60%
audiobox-PC	56.00%	35.13%	59.09%	59.86%	55.20%	56.80%
audiobox-PQ	59.60%	54.05%	65.15%	58.50%	55.20%	64.00%
CLAP score	62.40%	51.35%	59.09%	66.67%	60.80%	64.00%
CLAP music score	70.40%	75.68%	60.61%	73.47%	67.20%	73.60%
MuQ-Mulan	66.40%	67.56%	60.00%	69.39%	64.80%	68.00%
CLAMP3 score	62.80%	64.86%	60.61%	63.27%	63.20%	62.40%
Qwen2-audio	1.60%	2.70%	1.51%	1.36%	0.80%	2.40%
Qwen2.5-Omni	34.40%	27.03%	33.33%	36.73%	31.20%	37.60%
Qwen3-Omni	63.60%	56.76%	63.64%	66.67%	67.20%	60.00%
Gemini2.5-Flash	60.80%	48.65%	60.61%	63.95%	65.60%	56.00%
Gemini2.5-Pro	67.20%	54.05%	66.67%	70.75%	71.20%	63.20%
Gemini3-Pro	64.00%	54.05%	66.67%	65.31%	67.20%	60.80%
- w/o f.t.: Distill	68.00%	51.35%	71.21%	70.75%	61.60%	74.40%
C- w/ f.t.: CMI-Pref	73.60%	54.05%	75.76%	74.07%	64.00%	81.60%
- w/ f.t.: CMI + MusicEval	74.00%	62.16%	72.73%	77.55%	68.00%	80.00%

Table 13. Benchmark Audio-Music Alignment ACC Results on CMI-Pref (Ablation Analysis).

Audio-Music Alignment	Total (250)	Confidence Level			Data Type	
		Conf < 3 (42)	Conf $= 3$ (62)	Conf > 3 (146)	Instru (125)	Vocal (125)
Qwen2-audio	11.60%	11.90%	4.84%	14.38%	8.80%	14.40%
Qwen2.5-Omni	28.80%	26.19%	25.81%	30.82%	25.60%	32.00%
Qwen3-Omni	64.00%	54.76%	61.29%	67.81%	64.80%	63.20%
Gemini2.5-Flash	62.00%	50.00%	61.29%	65.75%	69.60%	54.40%
Gemini2.5-Pro	72.80%	52.38%	72.58%	78.77%	73.60%	72.00%
Gemini3-Pro	66.80%	61.90%	53.23%	73.97%	68.80%	64.80%
- w/o f.t.: Distill	72.80%	58.62%	62.9%	79.25%	70.40%	75.20%
C- w/ f.t.: CMI-Pref	78.40%	72.41%	69.35%	83.02%	76.80%	80.00%
- w/ f.t.: CMI + MusicEval	76.80%	65.52%	67.74%	82.39%	72.80%	80.80%

To ensure a fair comparison, all models are initialized from the CMI-Pref-Pseudo pretraining checkpoint (2,000 steps) and finetuned on CMI-Pref and MusicEval. As shown in Table 14, we observe that the choice of mapping function does not yield significant performance differences.

Datasets Metrics ↑	PAM Mean SRCC	MusicEval SRCC	Music Arena ACC	CMI-Pref Mean ACC
None	0.5069	0.4659	73.73%	75.3%
Tanh	0.6024	0.4702	73.8%	74.8%
Linear	0.4595	0.4691	73.58%	74.40%
Oridinal	0.6079	0.4665	72.31%	75.80%

Table 14. Aggregated results on mapping ablations