DialBGM: A Benchmark for Background Music Recommendation from Everyday Multi-Turn Dialogues

Anonymous ACL submission

Abstract

Background music (BGM) selection that fits natural human conversation is common in media production, yet automating it remains largely unaddressed. We introduce dialogue-conditioned BGM recommendation, the task of selecting non-intrusive music that fits a multi-turn conversation, even when the conversation contains no explicit music descriptors. We present DialBGM, a benchmark of 1,200 open-domain daily dialogues, each paired with four candidate music clips and human preference rankings under three suitability criteria: relevance, non-intrusiveness, and consistency. In our primary zero-shot candidate-ranking benchmark, no audio–language or multimodal LLM exceeds 35% Hit@1, and inter-model ordinal agreement stays low (Kendall's $\tau_b < 0.4$). We then train a lightweight supervised score head on three multimodal LLM backbones; all improve over zero-shot, and two reach 0.49–0.52 Hit@1 under a dialogue-level split. These results suggest that DialBGM provides learnable supervision for dialogue-conditioned BGM preference, while track-disjoint generalization remains future work. DialBGM provides a standardized benchmark for discourse-aware BGM recommendation, supporting both retrieval-based and generative model evaluation.

1 Introduction

Background music (BGM) is widely used in movies, games, and interactive systems; it shapes atmosphere, guides attention, and strengthens emotional impact (Ansani et al., 2020). Despite its practical importance, automatically selecting suitable BGM from textual dialogues remains largely unaddressed. In particular, to the best of our knowledge, there is no prior work that directly addresses dialogue-conditioned BGM recommendation, especially when the dialogue itself contains no explicit music descriptors.

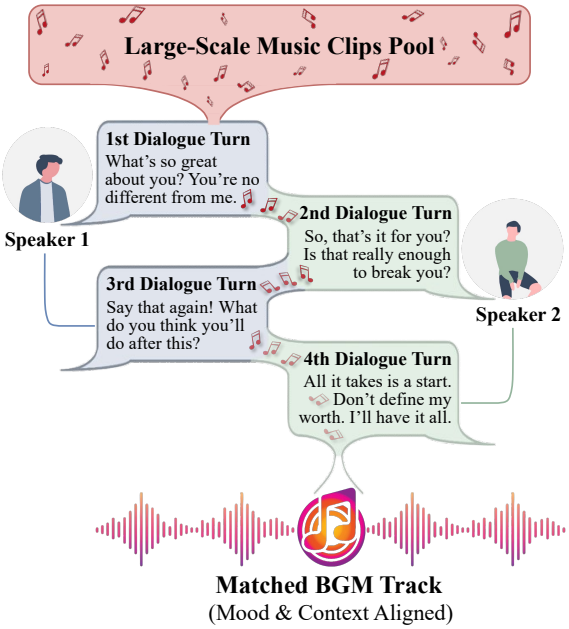

Figure 1: **Dialogue-conditioned BGM recommendation**. Given a multi-turn dialogue and a large-scale music clip database, the system uses the dialogue as a contextual filter to rank candidates and selects the one that best matches the dialogue as background music (BGM).

Automation of BGM recommendation is intrinsically difficult because it requires bridging linguistic context and musical attributes. Prior efforts in audio-text alignment have tackled music captioning and text-based music retrieval (e.g., CLAP (Elizalde et al., 2023), MuseChat (Dong et al., 2024)), but they often assume that the textual context contains sufficient clues like music descriptors (e.g., tags, artists, captions). In contrast, selecting a suitable BGM for multi-turn daily dialogues is significantly more challenging. Unlike general text-to-music retrieval, our input is conversational text that often contains no music descriptors, and our objective is background suitability (e.g., fit and non-intrusiveness), not simply seman-

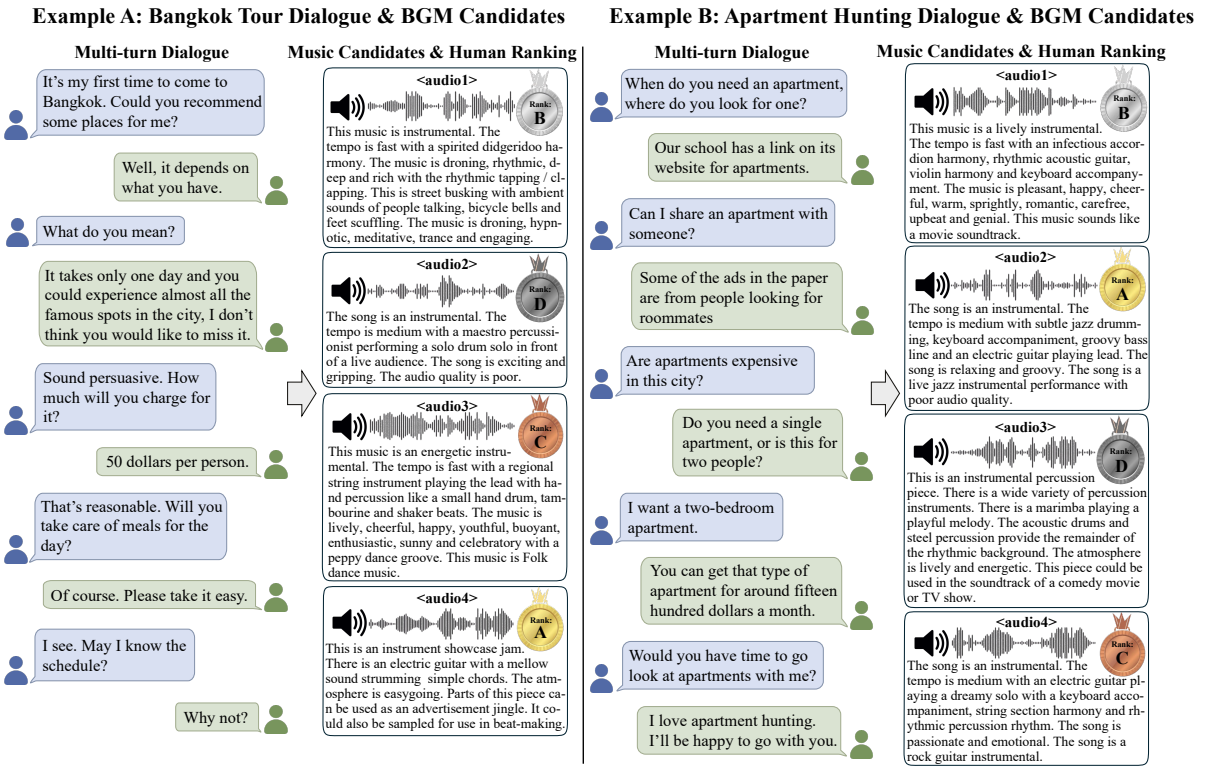

Figure 2: **DialBGM examples.** Each dataset instance consists of a multi-turn dialogue paired with four candidate background music (BGM) clips, along with human preference rankings indicating which clip best matches the conversational atmosphere. Each music clip is presented with its corresponding caption and human-annotated rank.

tic correspondence. Unlike prior work on BGM selection for films, trailers, or video summaries, dialogue-conditioned BGM recommendation cannot rely on rich audiovisual signals or scene-level visual context. Instead, the model must infer latent affect and intent from conversational text alone. The conversation may begin with tension and resolve warmly; it may convey sarcasm, embarrassment, or excitement through interaction patterns rather than explicit sentiment or musical words. Notably, this application has practical significance for AI-driven media creation (e.g., adding soundtracks to dialogues in games, virtual reality, or storytelling applications) and for conversational agents that aim to enhance user experience with adaptive background ambience.

In this paper, we introduce a new task, **dialogue-conditioned BGM recommendation**, which requires selecting suitable background music for multi-turn conversations without explicit music descriptors. To support this task, we present the **DialBGM** benchmark (see Figure 1). DialBGM pairs open-domain multi-turn dialogues with candidate music clips, utilizing human preference rankings to evaluate nuanced matching. Specifically, each dialogue is paired with four candidate BGM clips, and the task is to rank the four music clips by how well they fit the dialogue (see Figure 2). We systematically curate the candidate pool and employ human annotators to rank clips. The resulting dataset provides a ground truth ranking for each dialogue's quartet of music options, which supports retrieval/ranking-based evaluation and serves as a testbed for generative models.

Our experiments show that standard audio–text retrieval pipelines, such as summarizing the dialogue and retrieving music by caption similarity in a shared embedding space, remain inadequate for dialogue-conditioned BGM recommendation. Across a wide range of open-source and proprietary models, no method achieves a Hit@1 score above 35% in our primary zero-shot setting, highlighting a considerable gap relative to the aggregated human rankings. These findings motivate a benchmark that explicitly targets complex dialogue understanding, rather than generic audio–text semantic alignment.

Our main contributions are as follows:

- **Dataset and Construction Pipeline**: We present DialBGM, the first dataset that pairs multi-turn dialogues with background music,

containing 1,200 daily dialogues, four music candidates per dialogue, and human preference rankings. We provide a reproducible four-stage construction pipeline (rule-based BGM filtering, LLM-based dialogue captioning, embedding-based candidate retrieval, human ranking) with explicit annotation criteria (relevance, non-intrusiveness, consistency), and we will release DialBGM publicly.

- **Baseline and Analysis**: We evaluate a wide range of models, including embedding-based retrieval baselines (e.g., CLAP (Elizalde et al., 2023)) and recent multimodal LLMs (e.g., Qwen2.5-Omni (Xu et al., 2025), Gemini 2.5 (Comanici et al., 2025), Music Flamingo (Ghosh et al., 2025)), and show that no system exceeds Hit@1 $= 0.35$ with inter-model agreement below $\tau_b = 0.4$, establishing that DialBGM is hard.

- **Supervised Score Head**: We train a lightweight DoRA-adapted score head on three multimodal LLM backbones (Qwen2.5-Omni, Audio-Flamingo-3, Phi-4-Multimodal) using a 1,000/200 dialogue-level split. All three improve over zero-shot, suggesting that DialBGM provides learnable supervision for dialogue-conditioned BGM ranking.

2 Related Work

2.1 Audio-Language Models and Cross-Modal Retrieval

Cross-modal retrieval between audio and text has advanced through contrastive learning approaches. CLAP (Elizalde et al., 2023) learns shared embeddings through contrastive learning, enabling zero-shot audio-text retrieval and classification. LAION-CLAP (Wu et al., 2023) scales this approach with a large-scale dataset and feature fusion, achieving strong performance on audio-text retrieval benchmarks. ParaCLAP (Jing et al., 2024) extends this framework to paralinguistic attributes such as emotion and speaking style. While these models capture acoustic characteristics, they are trained on short audio clips with brief captions, and their ability to align music with multi-turn dialogue context remains unexplored.

Beyond retrieval, large audio-language models (LALMs) have emerged for audio reasoning tasks. Audio Flamingo (Kong et al., 2024) supports few-shot learning and multi-turn dialogue, and its successor Audio Flamingo 3 (Ghosh et al., 2026) unifies speech, sound, and music understanding within a single open model, while Music Flamingo (Ghosh et al., 2025) specializes in music understanding via chain-of-thought reasoning. Qwen2-Audio (Chu et al., 2024) provides both voice chat and audio analysis modes, trained on diverse audio, including speech, music, and environmental sounds. More recently, omni-modal models such as GPT-4o (Hurst et al., 2024), Gemini 2.5 (Comanici et al., 2025), Qwen2.5-Omni (Xu et al., 2025), and Phi-4-Multimodal (Abouelenin et al., 2025) integrate text, image, audio, and video understanding into unified architectures. Despite recent progress, music understanding itself remains challenging, with models struggling to capture nuanced content (Kang and Herremans, 2025), and dedicated benchmarks such as MuChoMusic (Weck et al., 2024) show that audio–language models often over-rely on the language modality rather than the audio itself.

2.2 Conversational Music Recommendation

Traditional music recommendation relies on collaborative filtering or content-based approaches that focus on user listening history. Early work, such as MusicRoBot (Zhou et al., 2018), combined knowledge graphs with chatbots for context-aware music recommendation. Talk the Walk (Leszczynski et al., 2023) generates synthetic conversational data by performing biased random walks on playlist graphs, addressing data scarcity in conversational recommendation. MuseChat (Dong et al., 2024) combines video understanding with dialogue-based refinement, explaining why specific tracks match visual content. TalkPlay (Doh et al., 2025a,b) formulates recommendations as an LLM-driven agentic task, invoking external tools such as SQL queries and dense retrieval based on user profiles and dialogue history. More recently, MusiCRS (Surana et al., 2026) introduced a benchmark for audio-centric conversational recommendation, linking music-seeking user conversations to candidate tracks.

These systems target explicit user preferences, playlist continuation, or music-seeking conversations. In contrast, DialBGM focuses on implicit affective alignment: selecting BGM that matches the emotional trajectory of a conversation without explicit user requests, requiring models to interpret subtle mood cues rather than executing database queries.

2.3 Datasets for Audio-Text Research

High-quality paired datasets and benchmarks have driven progress in audio-language research (Sakshi et al., 2024; Kumar et al., 2026). For general audio captioning, AudioCaps (Kim et al., 2019) provides 46k human-annotated captions, while Clotho (Drossos et al., 2020) offers crowdsourced descriptions for 6k environmental sound clips. WavCaps (Mei et al., 2024) further scales this effort to 400k clips using weakly labeled web data. In the music domain, MusicCaps (Agostinelli et al., 2023) provides 5.5k clips with professionally written captions. LP-MusicCaps (Doh et al., 2023) extends this to 2.2M samples using LLM-generated pseudo-captions. MTG-Jamendo (Bogdanov et al., 2019) offers large-scale music tagging annotations, while Song Describer (Manco et al., 2023) provides free-form textual descriptions for music retrieval.

For emotion-aware dialogue, DailyDialog (Li et al., 2017) contains 13k conversations with emotion and dialogue act labels. EmpatheticDialogues (Rashkin et al., 2019) provides 25k dialogues grounded in emotional situations, and MELD (Poria et al., 2019) offers multimodal emotion annotations from TV drama conversations.

Despite the availability of these resources, no existing dataset bridges multi-turn dialogue understanding with music selection. DialBGM addresses this gap by pairing dialogues with human-ranked BGM candidates, enabling evaluation of affective alignment between conversational context and musical accompaniment.

3 DialBGM Dataset

DialBGM is constructed via a reproducible semi-automatic pipeline (Figure 3): an automatic construction phase (Stages 1–3) and a human annotation phase (Stage 4).

3.1 Task Definition and Evaluation Metrics

Task Definition. Given a multi-turn dialogue d and four music tracks $\mathcal{M} = \{m_1, m_2, m_3, m_4\}$ with varying semantic relevance, the model predicts a ranking permutation R. The ground truth is a human-annotated ranking G from 1 (best) to 4 (worst). This formulation tests whether a model identifies relative BGM preferences rather than retrieving a single matching item.

Evaluation Metrics. We adopt four metrics measuring distinct aspects of ranking alignment: Hit@1 (top-1 accuracy), MRR, nDCG (graded relevance with 3-2-1-0), and Kendall's τ_b (ordinal consistency). Formulas are listed in Table 1. Multimodal models frequently produce tied scores; we adopt tie-aware variants throughout: Hit@1 assigns $1/k$ for k-way ties, MRR uses expected reciprocal rank, nDCG averages over tie-permutations, and τ_b accounts for tied pairs (T_x, T_y).

Table 1: Metric formulas (idealized, no-tie). Tie-aware variants are described in Section 3.

Metric	Formula
Hit@1	$\mathbb{I}(\text{top}(R) = \text{top}(G))$
MRR	$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i}$
nDCG	$\frac{\text{DCG}}{\text{IDCG}}, \quad \text{DCG} = \sum_{i=1}^{4}\frac{rel_i}{\log_2(i+1)}$
Kendall's τ_b	$\frac{P-Q}{\sqrt{(P+Q+T_x)(P+Q+T_y)}}$

Table 2: **Statistics of the DialBGM.**

Dialogue Statistics	
Total Dialogues	1,200
Avg. Turns per Dialogue	7.85
Avg. Words per Dialogue	85.98
Audio Statistics	
Unique Audio Tracks	1,020
Avg. Duration (sec)	10.0
Avg. Instrument Families per Clip	2.63

3.2 Automatic Dataset Construction Pipeline

Source Data. We leverage DailyDialog (Li et al., 2017) (∼13,000 multi-turn dialogues with emotion/dialogue-act labels) and MusicCaps (Agostinelli et al., 2023) (5,500 ten-second music clips with professional captions).

Stage 1: BGM Suitability Filtering. Not all MusicCaps clips are appropriate as background music. We apply a rule-based exclusion filter on the MusicCaps `aspect_list` field, removing vocal-related terms ("vocal", "speech"), noise descriptors ("static", "hiss"), and SFX ("pop", "knock"). This yields ∼2,000 instrumental clips.

Stage 2: Dialogue Caption Generation. We next process the dialogue side. Multi-turn dialogues are verbose yet semantically sparse, so directly embedding raw text yields poor retrieval. We employ GPT-4o (Hurst et al., 2024) to distill each dialogue

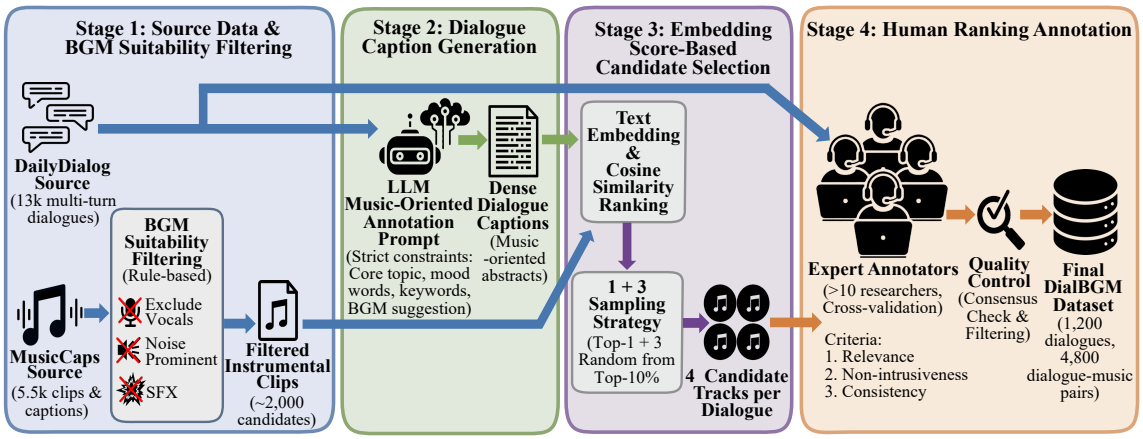

Figure 3: **Dataset construction.** The DialBGM dataset is constructed through a four-stage pipeline: (1) source data collection and rule-based BGM suitability filtering, (2) dialogue caption generation via an LLM, (3) embedding score-based candidate selection, and (4) expert human ranking annotation with quality control.

into a single-sentence music-oriented caption, conditioned on a BGM suggestion (style, instrument, tempo, energy). This Stage-2 caption is used only as the retrieval query for candidate construction; it is never provided as model input at training or evaluation time. The summary-input baselines in Section 4 instead use a separate generic dialogue summary, not this construction caption. The full system prompt is in Appendix C.

Stage 3: Embedding Score-Based Candidate Selection. Using the generated captions, we treat BGM selection as text-to-text semantic retrieval over the 2,000 BGM-filtered captions, computing cosine similarity with OpenAI's `text-embedding-3-large`[1]. We select four candidates per dialogue (six pairwise comparisons, manageable cognitive load); rather than top-4 by similarity (overly homogeneous), we adopt a "1+3 Sampling Strategy": top-1 plus three uniformly sampled from the top-10%, intended to balance relevance with acoustic contrast.

3.3 Human Ranking Annotation

Stage 3 retrieval captures surface semantics but cannot assess background-music suitability. We therefore collect human rankings on each retrieved quartet as ground truth beyond textual similarity.

Annotation Protocol. We recruit 12 researchers with speech/audio experience as annotators. Each annotator independently views a dialogue with four BGM candidates, listens via a Gradio interface (Appendix B), and ranks 1 (best) to 4 (worst). We prefer ranking over absolute scoring because it reduces calibration issues and aggregates more reliably for subjective judgments.

Annotation Criteria. Three criteria: (1) *Relevance* (mood/energy alignment with dialogue); (2) *Non-intrusiveness* (background suitability; vocals/effects penalized); (3) *Consistency* (stable atmosphere without abrupt changes).

Quality Control. We assess inter-annotator agreement via Kendall's W, with all 12 annotators per dialogue. Samples with $W < 0.25$ or in the bottom 10% of agreement are excluded; this removes $\sim$15% of annotations, yielding 1,200 dialogues with mean $W = 0.79$, median $W = 0.90$, and 92.16% achieving $W \geq 0.50$. Consensus rankings are aggregated via Borda count (Rank-1$\rightarrow$3, . . . , Rank-4$\rightarrow$0); ties break hierarchically by Rank-1 votes, Rank-2 votes, mean rank, and candidate ID.

3.4 Dataset Statistics

Beyond the totals in Table 2, dialogues span Daily Life & Lifestyle (68%), Work (19%), Relationships (9%), and others (4%); emotional tones split into Neutral (41%), Positive (36%), Negative (14%), and Mixed (9%). The 1,020 unique audio tracks span Classical/Orchestral (15%), Electronic (11%), Jazz/Blues (10%), and Rock/Metal (10%), with Bass, Drums, and Guitar appearing most frequently.

Supervised split. For the supervised score-head experiments in Section 5, we use a dialogue-level 1,000/200 train/test split. Unless otherwise speci-

[1] `https://openai.com/index/new-embedding-models-and-api-updates/`

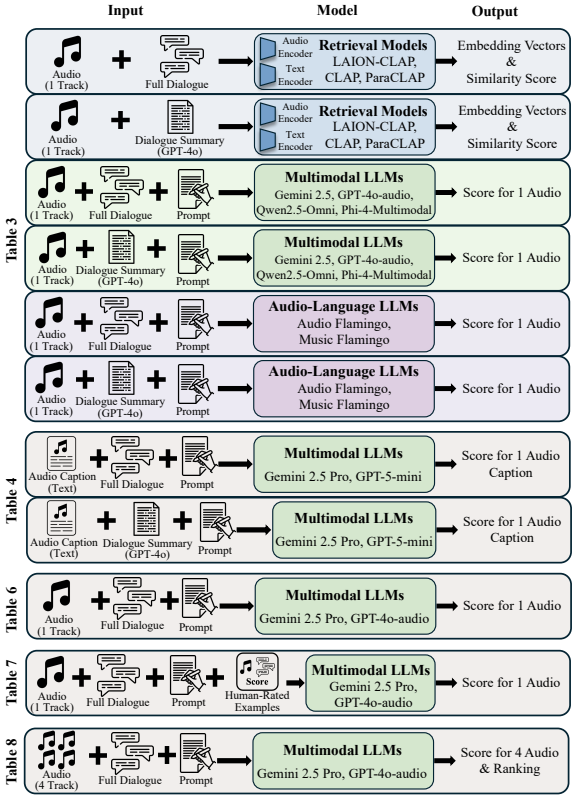

Figure 4: **Tested input-model-output configurations.** Overview of the experimental settings (Tables 3, 4, 6, 7, and 8): input compositions (audio or audio caption; full dialogue or GPT-4o summary), model families (retrieval models, multimodal LLMs, audio-language LLMs), and output formats. Tables 6–8 correspond to the advanced-prompting settings (without CoT, few-shot, and joint reranking).

fied, zero-shot benchmark results in Section 4 are reported on all 1,200 dialogues. Audio-track overlap across splits is discussed under Limitations.

4 Benchmark Results

We evaluate contemporary multimodal LLMs, audio-language models, and audio-text retrieval models on all 1,200 DialBGM dialogues across two input configurations: an audio-text setting (full dialogues with audio candidates) and a text-only setting (dialogue captions with music captions). Figure 4 illustrates the configurations; pairwise model agreement is analyzed in Section 4.4 (Figure 5).

4.1 Baseline Performance on DialBGM

The evaluated models exhibit consistently low performance (Table 3): the best Hit@1 is $\approx$ 0.33–0.35, MRR stays below 0.60, and Kendall's τ_b remains low (-0.04 to 0.18) despite relatively high nDCG (0.74–0.80); nDCG is less discriminative in this

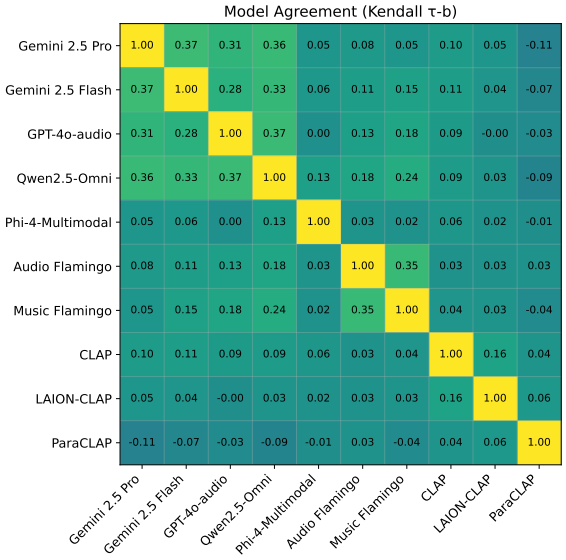

Figure 5: **Pairwise Kendall's τ_b across model rankings** on full-dialogue audio input. Intra-family clustering is visible; cross-family agreement is examined in Section 4.4.

four-candidate setting because relevance scores are concentrated. Replacing the full dialogue with a one-line summary yields only marginal changes. Multiple evaluation runs yield stable rankings, with Hit@1 standard deviations ranging from ± 0.002 to ± 0.038. One recurring failure mode illustrates this gap. When a speaker refuses a "party" because of academic failure, the models still rank "spirited" BGM highest: they latch onto a locally salient token ("party") and miss the affective turn (the refusal and its cause) that sets the intended mood, consistent with surface-lexical matching rather than affect-grounded selection. Resolving such cases requires reading the dialogue as an emotional trajectory rather than a set of isolated cues, the competence our ranking task is designed to isolate and that caption-similarity retrieval cannot supply.

4.2 Text-Only Does Not Close the Gap

The lexical-overfitting pattern hints at a language-side bottleneck. To isolate it from audio perception, we conduct a text-only ablation with Gemini 2.5 Pro and GPT-5-mini (OpenAI, 2025), pairing music captions with full dialogues or summaries.

The text-only condition yields the highest scores yet remains low relative to the aggregated human rankings (Table 4). This pattern is consistent with a limitation in affective reasoning rather than audio perception alone: models struggle to bridge dialogue narrative with the descriptive "atmosphere"

Table 3: Performance of baseline models on the **DialBGM** Ranking task. †Accessed via the `gpt-audio` API.

		Full Dialogue				One-line Summary			
Model Category	**Model**	**Hit@1**	**MRR**	**nDCG**	**Kendall's τ_b**	**Hit@1**	**MRR**	**nDCG**	**Kendall's τ_b**
Multimodal LLMs	Gemini 2.5 Pro	0.3233	0.5851	0.7904	0.1745	0.3474	0.5958	0.7965	0.1808
	Gemini 2.5 Flash	0.3293	0.5872	0.7905	0.1657	0.3228	0.5842	0.7912	0.1765
	GPT-4o-audio†	0.3051	0.5629	0.7822	0.1515	0.3192	0.5732	0.7856	0.1625
	Qwen2.5-Omni	0.3137	0.5742	0.7853	0.1739	0.3153	0.5740	0.7830	0.1674
	Phi-4-Multimodal	0.2506	0.5206	0.7490	-0.0092	0.2622	0.5311	0.7550	0.0241
Audio-Language Models	Audio Flamingo	0.2651	0.5359	0.7657	0.0945	0.2799	0.5441	0.7694	0.1004
	Music Flamingo	0.2596	0.5306	0.7582	0.0598	0.2877	0.5518	0.7697	0.0965
Retrieval Models	CLAP	0.2392	0.5194	0.7529	0.0397	0.3033	0.5643	0.7795	0.1289
	LAION-CLAP	0.2708	0.5378	0.7612	0.0531	0.2933	0.5570	0.7748	0.1092
	ParaCLAP	0.2100	0.4924	0.7362	-0.0264	0.2158	0.4957	0.7356	-0.0442

Table 4: Text-only performance with full multi-turn dialogue or LLM-generated summary as input.

Model	**Hit@1**	**MRR**	**nDCG**	τ_b
Gemini 2.5 Pro (full)	0.3491	0.5978	0.8000	0.2003
GPT-5-mini (full)	0.3474	0.5945	0.7991	0.2016
Gemini 2.5 Pro (summary)	0.3422	0.5954	0.8003	0.2106
GPT-5-mini (summary)	0.3321	0.5866	0.7968	0.1991

of music.

4.3 Limitations of Advanced Prompting

If reasoning is the bottleneck, prompting should help. Yet across multimodal LLMs (GPT-4o-audio, Gemini 2.5), chain-of-thought (Wei et al., 2022), few-shot demonstrations with explicit guidelines (Brown et al., 2020), and four-track joint reranking all fail to consistently improve over zero-shot; full results in Appendix A.

4.4 Inter-Model Preference Alignment

Even if individual models fall short, they might still share a notion of BGM suitability. We compute pairwise Kendall's τ_b correlations across rankings produced by all evaluated models under the full dialogue and audio input setting. Figure 5 visualizes the resulting correlation matrix.

Architectural Clustering. Models with similar architectures exhibit higher agreement, forming distinct preference clusters. Multimodal LLMs (Gemini 2.5 Pro, Gemini 2.5 Flash, GPT-4o-audio, and Qwen2.5-Omni) show moderate positive correlations with one another ($\tau_b \approx 0.28$–0.37), suggesting shared inductive biases derived from comparable pretraining objectives and model scale. In contrast, the Flamingo-based models (Audio Flamingo and Music Flamingo) show strong intra-family agreement ($\tau_b \approx 0.35$) but weak alignment with the multimodal LLM cluster ($\tau_b < 0.2$).

Divergence of Retrieval Models. CLAP and its variants (LAION-CLAP, ParaCLAP) exhibit near-zero or negative correlation with the multimodal LLM cluster ($\tau_b < 0.1$), consistent with their reliance on static audio-text embedding similarity rather than affective reasoning.

Lack of Consensus. Despite the observed clustering, overall agreement remains low ($\tau_b < 0.4$ even for the most similar pairs). This absence of a shared preference structure across models is consistent with the difficulty of dialogue-conditioned BGM recommendation and suggests that current systems do not converge on a stable notion of BGM suitability for conversational context. Together with the gap to the annotated consensus, this motivates the supervised score head in Section 5 rather than further zero-shot tuning.

5 Supervised Score Head on DialBGM

To probe whether DialBGM is learnable from training data alone, we train a lightweight supervised score head on three multimodal LLM backbones and evaluate on the 200-dialogue test split defined in Section 3. We compare each trained head against its zero-shot baseline, which is evaluated under the same generative-ranking protocol as the benchmark in Section 4.

Recipe. We evaluate Qwen2.5-Omni (Xu et al., 2025), Audio-Flamingo-3 (Ghosh et al., 2026), and Phi-4-Multimodal (Abouelenin et al., 2025). Qwen2.5-Omni and Audio-Flamingo-3 share the Qwen2.5 language module but differ in their audio

Backbone	Method	Hit@1	Kendall's τ_b
Qwen2.5-Omni	Zero-shot	0.3188	0.2098
	+ DoRA	**0.5200**	**0.3979**
Audio-Flamingo-3	Zero-shot	0.2487	-0.0059
	+ DoRA	**0.4875**	**0.3699**
Phi-4-Multimodal	Zero-shot	0.2387	-0.0407
	+ DoRA	0.2850	0.1650

Table 5: **Supervised score head on the 200-dialogue test split.** Zero-shot is evaluated under the same generative-ranking protocol as the benchmark in Section 4.

encoder, while Phi-4-Multimodal uses a different LLM family, letting us probe both encoder- and family-level sensitivity. Each backbone produces a scalar score $s_i = \mathbf{w}^\top \mathbf{h}(d, a_i)$ via a linear head on the last-token hidden state; the four scores induce the predicted ranking. We adapt the LLM's attention and feed-forward (MLP) projections with DoRA (Liu et al., 2024), an extension of LoRA (Hu et al., 2022), at rank $r = 16$, $\alpha = 32$, leaving the audio encoder and projector frozen. With gold permutation π and graded reward $r_i \in \{3, 2, 1, 0\}$, the head is trained with a pairwise Bradley–Terry loss (Bradley and Terry, 1952) plus a listwise ListMLE term (Xia et al., 2008),

$$\mathcal{L} = -\sum_{r_i > r_j} \log \sigma(s_i - s_j) - \lambda \sum_{t=1}^{3} \Big[s_{\pi_t} - \log \sum_{j \in \mathcal{R}_t} e^{s_j} \Big], \tag{1}$$

with $\mathcal{R}_t = \{\pi_t, \ldots, \pi_4\}$ and $\lambda = 0.3$. We train for 3 epochs at learning rate 1×10^{-4}, batch size 1 with gradient accumulation 4, cosine warmup 0.03, in bfloat16 on 3×A100 (40 GB).

Results. All three backbones improve over their zero-shot baselines (Table 5). Qwen2.5-Omni reaches Hit@1 = 0.5200 and Kendall's τ_b = 0.3979 (+0.2012 Hit@1, +0.1881 τ_b), and Audio-Flamingo-3 reaches 0.4875 and 0.3699 (+0.2388, +0.3758); both exceed the best zero-shot Hit@1 of any model in our benchmark (Section 4, ≈ 0.33–0.35). Phi-4-Multimodal improves clearly in τ_b (+0.2057, to 0.1650), with a smaller Hit@1 change (+0.0463, to 0.2850) that stays below this benchmark ceiling. The gain therefore holds across both audio encoders and LLM families; together, these results suggest that DialBGM provides learnable supervision for dialogue-conditioned BGM ranking.

What the score head learns. Two patterns in Table 5 clarify what the supervision contributes. First, the largest relative gains are in ordinal consistency for the backbones whose zero-shot rankings are essentially uncorrelated with the consensus: Audio-Flamingo-3 and Phi-4-Multimodal move from near-zero or negative Kendall's τ_b to clearly positive values, indicating that the linear head induces a consistent candidate ordering that the frozen encoders did not expose zero-shot. This holds for Phi-4-Multimodal even though its Hit@1 stays below this benchmark ceiling, so the supervision signal remains informative for the full ranking even when top-1 selection is still hard. Second, the two backbones that exceed this ceiling share the Qwen2.5 language module, whereas Phi-4-Multimodal, built on a different LLM family, improves but does not; the gain is thus not tied to a single audio encoder, though our three backbones are too few to isolate which component drives it.

6 Conclusion and Future Work

Summary. DialBGM reframes BGM recommendation as a dialogue-understanding problem: systems must infer the latent affective trajectory of everyday conversation and judge which music can support it without becoming intrusive. We contribute the benchmark itself—1,200 dialogues with four ranked BGM candidates—together with a reproducible construction pipeline and a ranking-based evaluation protocol over Hit@1, MRR, nDCG, and Kendall's τ_b.

Findings and Future Directions. Across audio-input, text-only, and prompting-based settings, current models remain weakly aligned with the human consensus rankings, and advanced prompting (chain-of-thought, few-shot) does not close the gap. That even the strongest text-only variant stays low suggests this capability does not emerge zero-shot from existing pretraining, but must instead be learned from task-specific supervision. As a first step, a lightweight supervised score head improves performance under a dialogue-level split, suggesting that DialBGM provides useful training signal while leaving track-disjoint generalization as an important next step. Future work will use DialBGM to train specialized audio-language adapters and to develop pretraining objectives for affective alignment between dialogue and music.

Limitations

Despite the systematic construction and evaluation of the DialBGM benchmark, several limitations remain that future work should address.

Scale and Diversity of Data. While DialBGM provides high-quality, human-verified rankings, the dataset size (1,200 dialogues) is relatively small compared to large-scale pretraining corpora. Consequently, the dataset is more suitable as an evaluation benchmark or a resource for fine-tuning, rather than for training LALMs from scratch. In addition, music candidates are restricted to the MusicCaps dataset, which implies that DialBGM may reflect biases present in MusicCaps.

Subjectivity of Annotations. Although strict evaluation criteria are applied to achieve a high level of inter-annotator agreement, the suitability of BGM remains inherently subjective. For a given dialogue, a single or objectively "correct" music track rarely exists, and the rankings in DialBGM therefore reflect annotator consensus, which may not comprehensively capture the full spectrum of valid artistic interpretations. Our annotators are speech/audio researchers rather than professional music supervisors; incorporating professional supervisors, along with absolute-quality scoring and contrastive fitting/non-fitting labels, is left to future work.

Audio-Track Overlap Across Splits. The 1,020 unique audio tracks are shared across the 1,200 dialogues, and our train/test split is defined at the dialogue level rather than the track level. Some tracks therefore appear in both splits, so the supervised results in Section 5 should be read as dialogue-level rather than track-disjoint generalization.

Ethics Statements

This work introduces DialBGM, a dataset for dialogue-conditioned BGM recommendation built on publicly available datasets (DailyDialog, MusicCaps); it contains no personally identifying or sensitive information. Annotation was performed by 12 adult researchers in speech and audio who participated voluntarily; we did not recruit through a paid crowdsourcing platform and collected no demographic data, so the task involves only minimal-risk ranking judgments. We will release DialBGM annotations and metadata with source-specific license notices for DailyDialog and MusicCaps. We do not redistribute raw audio; downstream users should comply with the original dataset licenses and relevant copyright terms. Overall, DialBGM is provided as a research resource that does not raise direct societal concerns and is constructed to facilitate progress in context-aware multimedia AI.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, and 1 others. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Alessandro Ansani, Marco Marini, Francesca D'Errico, and Isabella Poggi. 2020. How soundtracks shape what we see: Analyzing the influence of music on visual scenes through self-assessment, eye tracking, and pupillometry. *Frontiers in Psychology*, 11.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The mtg-jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Seungheon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. Lp-musiccaps: Llm-based pseudo music captioning. In *Ismir 2023 Hybrid Conference*.

Seungheon Doh, Keunwoo Choi, and Juhan Nam. 2025a. Talkplay: Multimodal music recommendation with large language models. *arXiv preprint arXiv:2502.13713*.

Seungheon Doh, Keunwoo Choi, and Juhan Nam. 2025b. Talkplay-tools: Conversational music recommendation with llm tool calling. *arXiv preprint arXiv:2510.01698*.

Zhikang Dong, Xiulong Liu, Bin Chen, Pawel Polak, and Peng Zhang. 2024. Musechat: A conversational music recommendation system for videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12775–12785.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and 1 others. 2026. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Advances in Neural Information Processing Systems*, 38:41819–41886.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sanggil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, and 1 others. 2025. Music flamingo: Scaling music understanding in audio language models. *arXiv preprint arXiv:2511.10289*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Xin Jing, Andreas Triantafyllopoulos, and Björn Schuller. 2024. Paraclap–towards a general language-audio model for computational paralinguistic tasks. In *Proc. Interspeech 2024*, pages 1155–1159.

Jaeyong Kang and Dorien Herremans. 2025. Are we there yet? a brief survey of music emotion prediction datasets, models and outstanding challenges. *IEEE Transactions on Affective Computing*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *International Conference on Machine Learning*, pages 25125–25148. PMLR.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, and 1 others. 2026. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 22688–22697.

Megan Leszczynski, Shu Zhang, Ravi Ganti, Krisztian Balog, Filip Radlinski, Fernando Pereira, and Arun Tejasvi Chaganty. 2023. Talk the walk: Synthetic data generation for conversational music recommendation. *arXiv preprint arXiv:2301.11489*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*.

Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, and 1 others. 2023. The song describer dataset: a corpus of audio captions for music-and-language evaluation. In *Workshop on Machine Learning for Audio, Neural Information Processing Systems (NeurIPS)*. Neural Information Processing Systems.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354.

OpenAI. 2025. GPT-5 System Card. https://openai.com/index/gpt-5-system-card/. Accessed: 2025-01-06.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5370–5381.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

Rohan Surana, Amit Namburi, Gagan Mundada, Abhay Lal, Zachary Novack, Julian McAuley, and Junda Wu. 2026. Musicrs: Benchmarking audio-centric conversational recommendation. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12412–12416. IEEE.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. Muchomusic: Evaluating music understanding in multimodal audio-language models. *arXiv preprint arXiv:2408.01337*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Chunyi Zhou, Yuanyuan Jin, Kai Zhang, Jiahao Yuan, Shengyuan Li, and Xiaoling Wang. 2018. Musicrobot: Towards conversational context-aware music recommender system. In *International Conference on Database Systems for Advanced Applications*, pages 817–820. Springer.

A Additional Experimental Results for Advanced Prompting

Chain-of-Thought (CoT). Table 6 reports results obtained without CoT prompts under the same evaluation setting. When compared to the CoT-based results in Table 3, performance differences remain minor across all metrics, indicating that explicit reasoning decomposition does not provide consistent gains.

Model	Hit@1	MRR	nDCG	Kendall's τ_b
Gemini 2.5 Pro	0.3712	0.6193	0.8103	0.2544
GPT-4o-audio	0.3142	0.5695	0.7837	0.1578

Table 6: Results without Chain-of-Thought prompting.

Few-Shot Examples and Guidelines. Table 7 presents results using few-shot demonstrations with human-labeled scores and explicit guidelines (Relevance, Non-intrusiveness, Consistency). Relative to the corresponding baseline in Table 3, the differences remain small, suggesting that this form of example-based instruction does not reliably improve performance.

Model	Hit@1	MRR	nDCG	Kendall's τ_b
Gemini 2.5 Pro	0.3443	0.5976	0.7941	0.1758
GPT-4o-audio	0.2949	0.5553	0.7769	0.1365

Table 7: Results with few-shot demonstrations and explicit guidelines.

Joint Ranking (Four-Track Comparison). Table 8 reports results for joint reranking, where the model receives all four candidate clips simultaneously and outputs a single global ordering. Consistent with Section 4, joint reranking does not yield reliable improvements across models.

Model	Hit@1	MRR	nDCG	Kendall's τ_b
Gemini 2.5 Pro	0.3204	0.5848	0.7924	0.1783
GPT-4o-audio	0.3058	0.5695	0.7849	0.1690

Table 8: Joint reranking results where all four audio candidates are provided simultaneously.

B Data Collection Interface

To facilitate efficient data collection, we implement a lightweight web-based annotation tool using Gradio. For each dialogue-music candidate set, the interface presents (i) the full multi-turn dialogue

context and (ii) four candidate audio tracks. Annotators listen to the candidates and assign a complete ranking from 1 (best) to 4 (worst) according to the task definition (selecting background music that best supports the dialogue).

C LLM Prompt for Dialogue Caption Generation

To bridge the modality gap between multi-turn dialogues and concise music descriptions, we prompt an LLM to produce a single-sentence caption that encodes (i) the core topic of the dialogue, (ii) the implied conversational mood, (iii) salient keywords, and (iv) a brief background-music (BGM) suggestion.

System Prompt: Dialogue-to-Caption for BGM Retrieval

```
You are an expert annotator for background
music selection. Given a two-speaker
multi-turn dialogue, write exactly ONE
natural sentence in English that covers:
the core topic, 3–5 mood words, 3–5
concise keywords, and a short BGM suggestion
(style/instrument/tempo/energy). Do not use
lists, line breaks, quotes, or brackets.
Keep the sentence between 20 and 35 words.
```

User Prompt

```
Read the dialogue below and produce exactly
ONE sentence in English.
Rules: - Include the core summary, and
weave 3–5 mood words naturally (no
brackets). - Include 3–5 concise keywords,
comma-separated, while keeping the sentence
natural. - Provide a brief BGM suggestion
(style/instrument/tempo/energy) within the
same sentence. - No line breaks, no lists,
and no quotes/brackets/braces. - Output one
sentence only.
Dialogue: {DIALOGUE}
```

D LLM Evaluation Prompt

For data construction, we also employ a structured prompt that elicits a continuous $[0.0, 10.0]$ suitability score for how well a candidate track functions as background music for a given dialogue.

The prompt explicitly discourages central-tendency scoring, prioritizes technical interference checks (e.g., masking around speech frequencies), and enforces a strict JSON-only output format for reliable downstream parsing. It further specifies an internal (non-disclosed) scoring rubric that combines (A) acoustic technicality, (B) emotional/contextual fit, and (C) pacing/rhythm using a weighted formula, with additional constraints for prominent vocals/lyrics and poor technical compatibility. The full system prompt used in our experiments is presented in the box below.

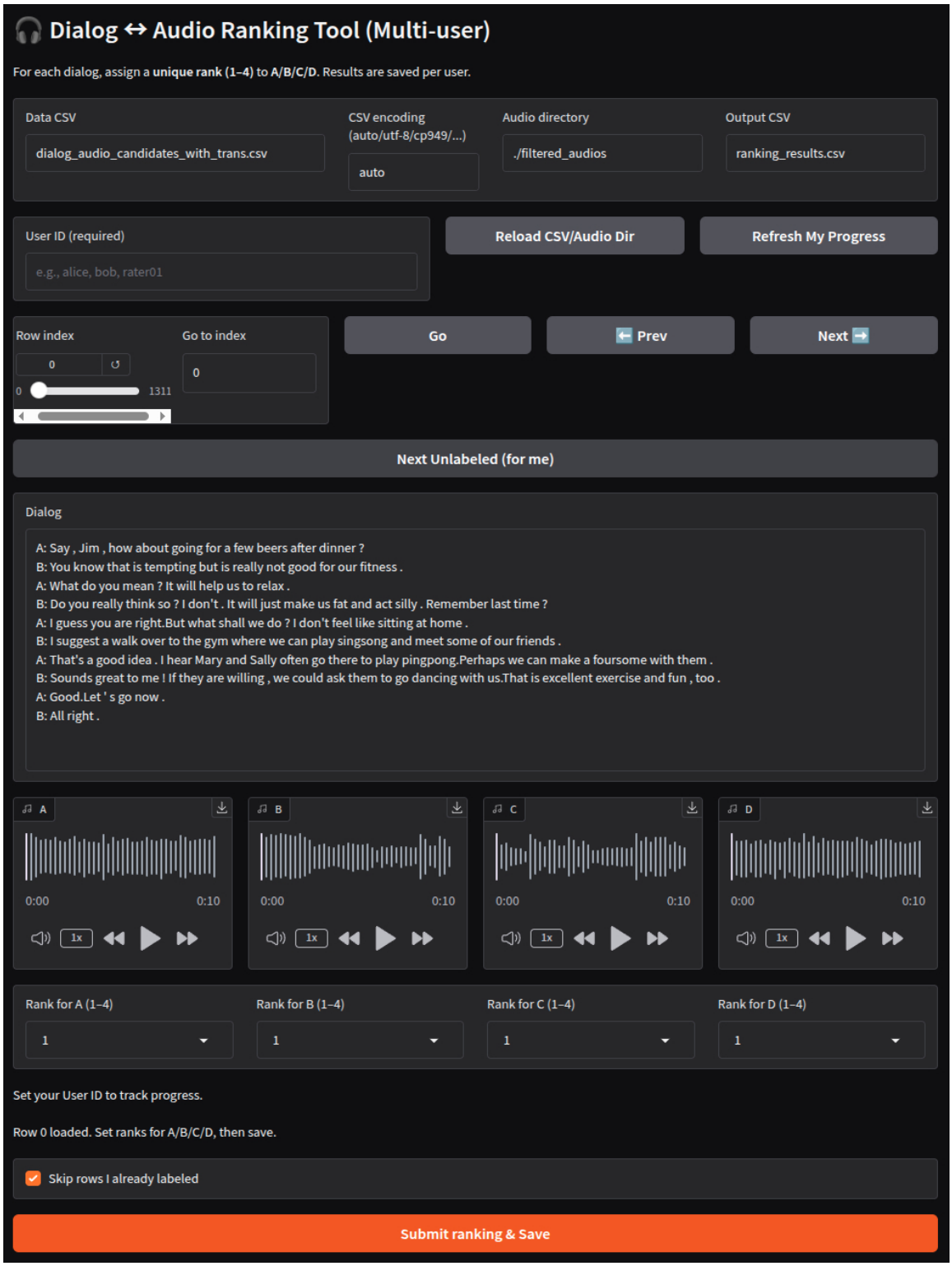

Figure 6: Screenshot of the Gradio-based data collection interface. The tool displays the dialogue and allows annotators to rank four BGM candidates (Candidates A–D) before proceeding to the next turn.

System Prompt for BGM Evaluation

"You are a legendary Chief Audio Director and film music critic with 30+ years of experience in Hollywood. You are strict, detail-obsessed, and you do not compromise on quality.
Your task: judge how well the given candidate track works specifically as BACKGROUND MUSIC (BGM) for the dialogue context.
IMPORTANT RULES: 1) Reject central tendency: 'okay' is a lazy answer. Use the full 0–10 scale. Great matches deserve high scores; distracting or contradictory choices deserve low scores. 2) Technical interference first: before artistic taste, check for physical clashes. - Frequency masking risk: would the music likely intrude into speech intelligibility (roughly 1–4 kHz)? - Level balance risk: would the music likely overpower or compete with the dialogue? - If the track contains prominent vocals/lyrics, treat it as a Fatal Error for BGM under dialogue. 3) Narrative synchronization: the music must support the dialogue's emotional arc and setting. 4) Think step-by-step internally, but DO NOT reveal your reasoning.
Return ONLY a JSON object and nothing else.
Score how well this track fits as BACKGROUND MUSIC (BGM) for the given dialogue context.
OUTPUT REQUIREMENTS (STRICT): - Return ONLY one JSON object with exactly one key: score. - score must be a NUMBER in [0.0, 10.0] with EXACTLY ONE decimal place. - Do NOT output an integer. Always include one decimal place. - Use 0.1 granularity; avoid repeating the same score unless truly indistinguishable.
To avoid score clustering, compute the final score using this internal procedure (do NOT output the steps): Step A — Acoustic Technicality (0.0–10.0, one decimal): Evaluate masking risk (busy midrange, sharp leads, dense percussion), perceived loudness competition, and how well it can sit under speech without needing aggressive ducking. Step B — Emotional & Contextual Fit (0.0–10.0, one decimal): Match mood, tension, warmth, setting, and emotional arc implied by the dialogue. Step C — Pacing & Rhythm (0.0–10.0, one decimal): Match energy/tempo to implied speech rate and scene pacing; penalize off-beat urgency/sleepiness.
Final score rule (compute numerically, then round to ONE decimal): base = 0.2*A + 0.5*B + 0.3*C If prominent vocals/lyrics are present (Fatal Error), then score = min(base, 2.0). If A < 3.0, then score must not exceed 3.0. subtle fit differences (e.g., slightly too busy -> -0.2; exceptionally well-undercored -> +0.2). Then clamp to [0.0, 10.0] and round to 1 decimal."

Figure 7: The full system prompt used for LLM-based BGM evaluation. The prompt incorporates a weighted scoring mechanism ($0.2 \times$ Technicality $+ 0.5 \times$ Fit $+ 0.3 \times$ Pacing) and explicit penalty rules for vocal interference.