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

Anonymous ACL submission

Abstract

Selecting an appropriate background music (BGM) that supports natural human conversation is a common production step in media and interactive systems. In this paper, we introduce dialogue-conditioned BGM recommendation, where a model should select non-intrusive, fitting music for a multi-turn conversation that often contains no music descriptors. To study this novel problem, we present DialBGM, a benchmark of 1,200 open-domain daily dialogues, each paired with four candidate music clips and annotated with human preference rankings. Rankings are determined by background suitability criteria, including contextual relevance, non-intrusiveness, and consistency. We evaluate a wide range of open-source and proprietary models, including audio–language models and multimodal LLMs, and show that current models fall far short of human judgments; no model exceeds 35% Hit@1 when selecting the top-ranked clip. DialBGM provides a standardized benchmark for developing discourse-aware methods for BGM selection and for evaluating both retrieval-based and generative models.

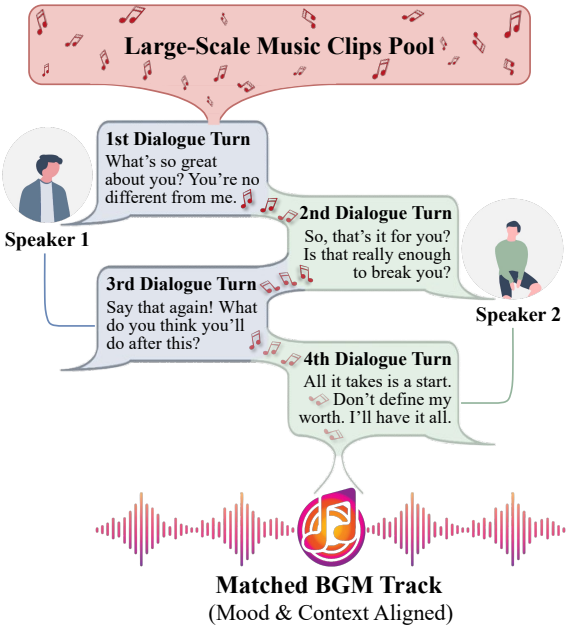

Figure 1: **Dialogue-conditioned BGM recommendation**. Given a multi-turn dialogue and a large-scale music clip database, the system uses the dialogue as a contextual filter to rank candidates and selects the one that best matches the dialogue as background music (BGM).

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 textual content is *not related to the music* itself.

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 musical characteristics (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 semantic correspondence. The speakers may casually talk about various topics, such as vacation plans, shopping recommendations, or opinions about food. The conversation may begin with ten-

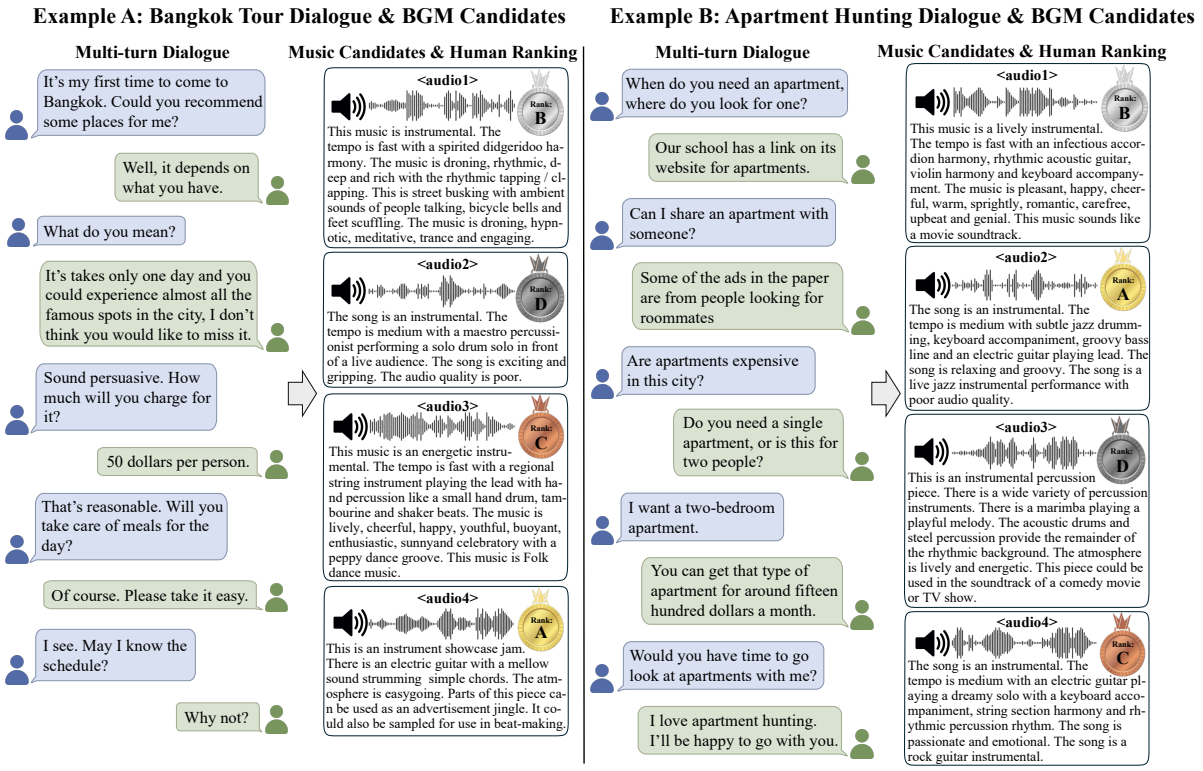

Figure 2: **DialBGM examples.** Each dataset instance consists of a multi-turn dialogue paired with four candidate background music (BGM) clips, along with human preference rankings indicating which clip best matches the conversational atmosphere. Each music clip is presented with its corresponding caption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

sion 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%, highlighting a considerable gap relative to human judgments. These findings motivate a benchmark that explicitly targets complex dialogue understanding, rather than generic audio–text semantic alignment.

Our main contributions are as follows:

- **New Benchmark Dataset**: We present DialBGM, the first dataset that focuses on matching multi-turn dialogues to background music. The dataset contains 1,200 daily dialogues, each paired with four music candidates and human preference rankings. We will release DialBGM publicly for research.
- **Dataset Construction Pipeline**: We provide a scalable pipeline that combines music fil-

tering, query rewriting, and embedding-based candidate retrieval, followed by human ranking. We define clear annotation criteria (relevance, non-intrusiveness, and consistency) to improve annotation reliability.

- **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 all exhibit a substantial gap to human preference.

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, 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).

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.

These systems target explicit user preferences or playlist continuation. 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 have driven progress in audio-language research (Sakshi et al., 2024). 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

We present the DialBGM dataset, constructed through a reproducible semi-automatic pipeline. The pipeline consists of two high-level phases: an automatic construction phase (Stages 1–3) and a human annotation phase (Stage 4), as illustrated in Figure 3. To bridge the modality gap between dialogue and music, our pipeline incorporates multiple refinement steps, as described below.

3.1 Task Definition and Evaluation Metrics

Task Definition. We formulate background music selection as a conditional ranking task. Given a multi-turn dialogue context d and a candidate set of music tracks $\mathcal{M} = \{m_1, m_2, m_3, m_4\}$, the model must predict a ranking permutation R. The candidate set $\mathcal{M}$ is carefully designed to contain tracks with varying semantic relevance to the dialogue. The ground truth is a human-annotated ranking G from 1 (best) to 4 (worst). This formulation allows us to evaluate whether the model can distinguish relative preferences rather than simply retrieving a single item.

Evaluation Metrics. We adopt four metrics to quantify alignment between model predictions and human judgments. Hit@1 measures strict top-1 accuracy by checking if the model's top prediction matches the human-annotated Rank-1. MRR (Mean Reciprocal Rank) evaluates the rank position of the ground-truth top item in the predicted ranking. To assess overall ranking quality with graded relevance, the nDCG@4 metric maps human rankings to relevance scores via a 3-2-1-0 scheme. Finally, Kendall's τ_b captures ordinal consistency between model and human rankings. Table 1 summarizes the formulas.

Handling Tied Predictions. Empirically, LLMs and multimodal models frequently produce tied scores. To mitigate this ambiguity, we adopt tie-aware variants: Hit@1 assigns $1/k$ for k-way ties, while MRR, nDCG@4, and Kendall's τ_b incorporate probabilistic expectations or tie-corrected denominators.

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 1: Evaluation metrics for DialBGM benchmark. We employ tie-aware variants to handle equal scores. N denotes the number of samples. For Kendall's τ_b, P and Q denote concordant and discordant pairs, and T_x, T_y denote ties.

3.2 Automatic Dataset Construction Pipeline

Source Data. We leverage DailyDialog (Li et al., 2017), which provides ∼13,000 multi-turn dialogues with emotion and dialogue act labels, and MusicCaps (Agostinelli et al., 2023), which offers 5,500 ten-second music clips with professionally written captions. To the best of our knowledge, DialBGM is the first dataset that pairs multi-turn dialogues with human-ranked background music candidates.

Stage 1: BGM Suitability Filtering. While MusicCaps provides high-quality audio-text pairs, not all clips are appropriate as background music. Background music should complement dialogue without drawing excessive attention or conflicting with speech. Therefore, clips containing vocals, noise, or prominent sound effects are unsuitable. We apply a rule-based exclusion filter based on keywords in the MusicCaps `aspect_list` field, which contains descriptive tags for each clip. Specifically, we exclude clips containing vocal-related terms (e.g., "vocal", "speech"), noise descriptors (e.g., "static", "hiss"), and prominent sound effects (e.g., "pop", "knock"). This process yields ∼2,000 instrumental music clips suitable for background music use.

Stage 2: Dialogue Caption Generation. Multi-turn dialogues are typically verbose yet semantically sparse, creating a mismatch with concise music captions. As a result, directly embedding raw dialogue text leads to suboptimal retrieval quality. To bridge this gap, we generate dense captions that capture both narrative content and emotional tone.

We employ GPT-4o (Hurst et al., 2024) to distill each dialogue into a single-sentence caption. Rather than standard summarization, we adopt a music-oriented annotation strategy that aligns di-

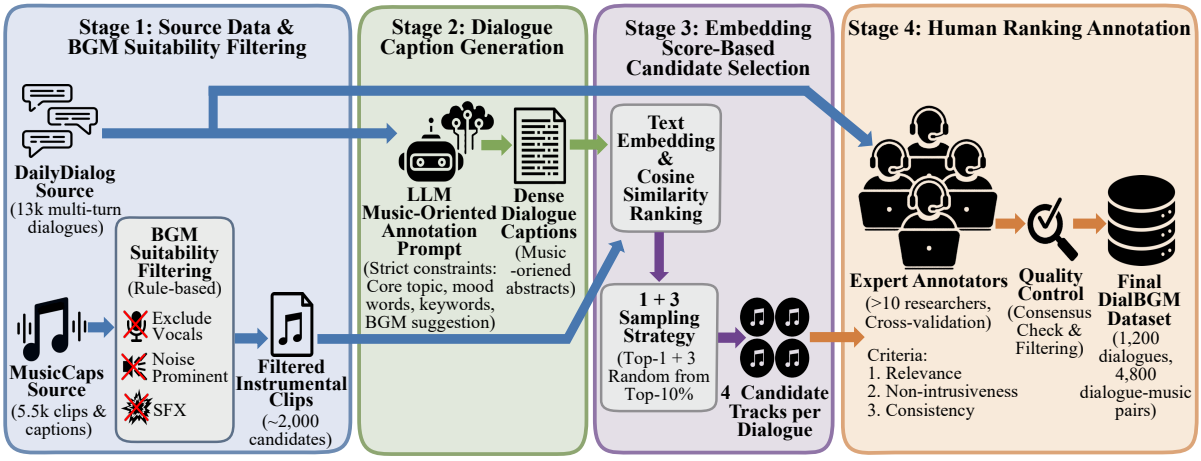

Figure 3: **Dataset construction.** The DialBGM dataset is constructed through a four-stage pipeline, consisting of (1) source data collection and rule-based BGM suitability filtering, (2) dialogue caption generation via an LLM, (3) embedding score-based candidate selection, and (4) expert human ranking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

alogue semantics with music descriptions. We instruct the model to act as a background music selector, conditioning output on a BGM suggestion (style, instrument, tempo, energy). This helps the generated caption capture implicit mood cues that are often omitted in factual summaries. The full system prompt is provided in Appendix C.

Stage 3: Embedding Score-Based Candidate Selection. Using the generated captions, we formulate BGM selection as a text-to-text semantic retrieval task, where the ∼13,000 dialogue captions serve as queries against the corpus of 2,000 BGM-filtered music captions. We employ OpenAI's `text-embedding-3-large`[1] to map both dialogue and music captions into a shared high-dimensional vector space. Music clips are ranked for each dialogue based on cosine similarity between their respective embeddings.

To construct the final candidate set for human evaluation, we adopt a "1+3 Sampling" strategy, selecting four BGM candidates per dialogue: (1) the top-1 clip with the highest cosine similarity, and (2) three clips randomly sampled from the top-10% of the similarity ranking. This strategy allows all four candidates to remain relevant to the dialogue while providing sufficient acoustic contrast for meaningful annotation. In preliminary experiments, selecting the top-4 by similarity alone yielded overly homogeneous candidate sets.

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

3.3 Human Ranking Annotation

Although embedding-based retrieval captures surface-level semantic overlap, it commonly fails to assess whether a track is appropriate for a given conversation. Human annotation provides ground-truth rankings that reflect judgment beyond textual similarity.

Annotation Protocol. We recruit 12 researchers with experience in speech and audio research as annotators. Annotators are presented with dialogue text and four BGM candidates, and instructed to rank clips from 1 (best) to 4 (worst) based on suitability as background music.

Annotation Criteria. Rankings are guided by three criteria: (1) *Relevance*: alignment between music mood/energy and dialogue semantics/emotion; (2) *Non-intrusiveness*: suitability as background audio, where clips containing vocals, prominent sound effects, or distracting elements are penalized; and (3) *Consistency*: a stable atmosphere throughout the clip without abrupt changes that disrupt dialogue flow.

Quality Control. Given the subjective nature of music recommendation, disagreement among annotators is expected for some samples. We assess inter-annotator agreement using Kendall's coefficient of concordance (W). Samples where W falls below 0.25 or ranks in the bottom 10% of agreement scores are excluded as low-consistency cases. This filtering removes ∼15% of initial annotations, yielding the final dataset of 1,200 dialogues.

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

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

To aggregate rankings into a single consensus label, we apply Borda count scoring (Rank-1→3, Rank-2→2, Rank-3→1, Rank-4→0). Ties are resolved hierarchically: (i) number of Rank-1 votes, (ii) number of Rank-2 votes, (iii) mean rank, and (iv) candidate ID for reproducibility.

3.4 Dataset Statistics

The DialBGM dataset consists of 1,200 dialogues, each paired with four candidate music clips and human-annotated rankings, yielding 4,800 dialogue-music pairs in total.

Dialogue Characteristics. On average, each dialogue contains 7.85 turns and 85.98 words, providing richer context compared to single-sentence captions typically used in music retrieval tasks. Topics center on Daily Life & Lifestyle (68%), followed by Work & Career (19%), Relationships & Love (9%), and others (4%). Emotional tones are distributed across Neutral (41%), Positive (36%), Negative (14%), and Mixed (9%).

Music Characteristics. The number of unique clips is 1,020, because audio clips can be assigned to multiple conversations as candidates. The genre distribution is diverse, led by Classical/Orchestral (15%), Electronic (11%), Jazz/Blues (10%), and Rock/Metal (10%), while a portion of tracks remain unclassified or fall into minor categories. The instrument distribution is similarly varied, with Bass, Drums, and Guitar appearing most frequently across the clips.

4 Experimental Results

To validate the necessity of DialBGM and assess the capabilities of current AI systems, we evaluate a wide range of models, including contemporary multimodal LLMs, specialized audio-language models, and audio-text retrieval models (see Figure 4).

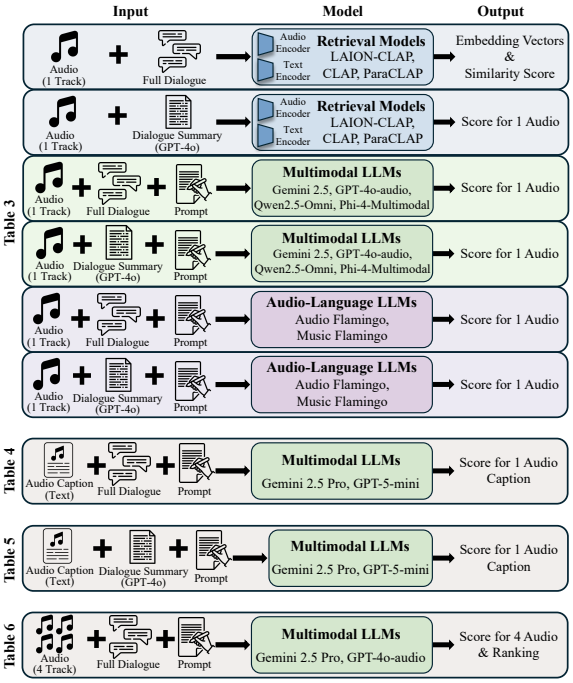

Figure 4: **Tested input-model-output configurations.** Overview of the experimental settings (Tables 3–6), illustrating the input compositions (audio track or audio caption, full dialogue or GPT-4o dialogue summary, and optional prompts), the model families (retrieval models, multimodal LLMs, and audio-language LLMs), and the corresponding outputs (embedding-based similarity or scalar scores for each candidate).

4.1 Baseline Performance on DialBGM

Experimental results show a significant discrepancy between model predictions and human annotations, as the evaluated AI models exhibit consistently low performance (Table 3).

Performance in Metrics. The best models select the human top-ranked clip in only about one-third of cases (Hit@1 $\approx$ 0.33-0.35), and MRR remains below 0.60 (0.49-0.60) across settings. Although nDCG is relatively high (0.74-0.80), Kendall's τ_b stays low (-0.04 to 0.18), indicating limited rank-order agreement with human preferences. Moreover, using a one-line summary instead of the full dialogue yields only marginal changes in these metrics, suggesting no substantial difference between the two dialogue representations.

Qualitative Analysis. Models tend to overfit to surface-level keywords, neglecting pragmatic context. For instance, despite a speaker refusing a "party" due to academic failure, models retrieve "spirited" BGM, focusing on the word rather than

Model Category	Model	Full Dialogue				One-line Summary			
		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 3: Performance of baseline models on the **DialBGM** Ranking task. †Accessed via the `gpt-audio` API.

Model	Hit@1	MRR	nDCG	Kendall's τ_b
Gemini 2.5 Pro	0.3491	0.5978	0.8000	0.2003
GPT-5-mini	0.3474	0.5945	0.7991	0.2016

Table 4: Text-only performance using full multi-turn dialogue as input.

Model	Hit@1	MRR	nDCG	Kendall's τ_b
Gemini 2.5 Pro	0.3422	0.5954	0.8003	0.2106
GPT-5-mini	0.3321	0.5866	0.7968	0.1991

Table 5: Text-only performance using the generated summary caption as input.

the speaker's distress. This highlights the persistent challenge of distinguishing superficial lexical cues from the true emotional narrative.

4.2 Text-Only Modality Failure

To investigate whether the bottleneck stems from audio processing or a lack of affective reasoning, we conducted a text-only ablation using Gemini 2.5 Pro and GPT-5-mini (OpenAI, 2025). We paired music text captions with either full dialogues or LLM-generated summaries, asking models to rank candidates based solely on textual descriptions.

As shown in Tables 4 and 5, performance in the text-only setting is similarly poor across both input variations. Even when the dialogue was condensed into a summary caption, which aligns the textual modalities, proprietary LLMs still fail to accurately match dialogue representations to the corresponding music descriptions. This failure suggests that the challenge of DialBGM is not solely an audio perception issue. It reveals a deeper deficiency in affective reasoning: models struggle to bridge the discrepancy between the narrative content of a dialogue and the descriptive "atmosphere" of music.

4.3 Limitations of Advanced Prompting

In experiments with multimodal LLMs (e.g., GPT-4o-audio, Gemini 2.5), we examine whether advanced prompting techniques improve the models' performance. Contrary to expectations, these complex prompting strategies often degrade performance compared to simple zero-shot baselines. Please see Appendix A for more details.

Chain-of-Thought (CoT). Step-by-step reasoning prompts (Wei et al., 2022) (e.g., "analyze the dialogue emotion, analyze the audio mood, then compare") frequently induce hallucinations, where the model fabricates audio attributes to justify its decision. This behavior suggests that affective dialogue-conditioned BGM matching does not reliably benefit from explicit logical decomposition.

Few-Shot Example and Guideline. Few-shot demonstrations (Brown et al., 2020) with human-labeled scores and explicit guidelines (Relevance, Non-intrusiveness, Consistency) do not yield consistent improvements. Models often overfit to superficial cues from the demonstrations rather than learning the intended criteria, indicating limited robustness without task-specific adaptation.

Joint Ranking (Four-Track Comparison). We test joint reranking to mitigate noise from scoring candidates independently, where the model receives all four clips and returns a single global ordering. However, this formulation does not provide reliable gains. For GPT-4o-audio, improvements over single-candidate scoring are marginal,

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

whereas the performance of Gemini 2.5 Pro degrades under the same setting despite identical dialogue context and candidate sets.

Overall, advanced prompting techniques, including CoT, few-shot instruction, and joint reranking, do not reliably improve performance and may even degrade it for some models.

4.4 Inter-Model Preference Alignment

While the previous results demonstrate a gap between model predictions and human judgments, a natural follow-up question is whether different models share similar internal ranking criteria. To investigate this, we compute pairwise Kendall's τ_b correlation 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 LMM cluster ($\tau_b < 0.2$).

Divergence of Retrieval Models. CLAP and its variants (LAION-CLAP, ParaCLAP) exhibit near-zero or negative correlation with the LMM 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 underscores the difficulty of dialogue-conditioned BGM recommendation and suggests that current systems

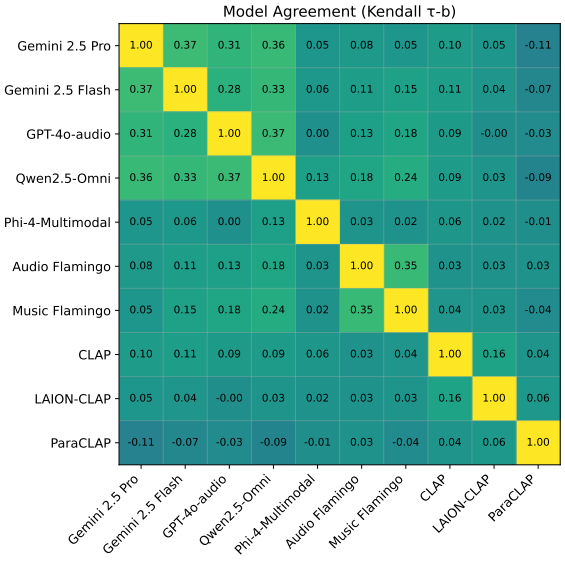

Figure 5: Pairwise Kendall's τ_b correlation matrix of ranking predictions from different models. Brighter boxes indicate the higher agreement in ranking.

do not converge on a stable notion of BGM suitability for conversational context.

5 Conclusion and Future Work

Summary of Contributions. In this work, we introduced DialBGM, a novel benchmark dataset for dialogue-conditioned BGM recommendation. We established a systematic construction pipeline combining LLM summarization, rule-based filtering, and embedding-based candidate selection, followed by human annotation. We also defined a comprehensive evaluation protocol using standard ranking metrics (Hit@1, MRR, nDCG, τ_b).

Current Limitations and Future Directions. Our experiments confirmed that selecting appropriate background music for a conversation is inherently difficult. The results demonstrate that even the most advanced multimodal systems remain far from human preferences in this task. The failure of advanced prompting strategies (CoT, few-shot) and the poor performance in text-only settings suggest that this capability does not emerge zero-shot from existing pretraining, necessitating specialized training data and tasks. Future work will leverage DialBGM to train specialized adapters for audio-language models or to develop new pretraining objectives that explicitly model the alignment of "mood" 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.

Ethics Statements

This work introduces DialBGM, a dataset for dialogue-conditioned BGM recommendation. Since DialBGM does not involve human subjects or contain personal or sensitive information, we believe there exist minimal ethical concerns. However, the utilization of music-related resources may require consideration of intellectual property rights in downstream applications, making it important to comply with copyright and relevant licensing 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)*.

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, 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*.

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.

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.

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*.

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.

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 7 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 7: Results without Chain-of-Thought prompting.

Few-Shot Examples and Guidelines. Table 8 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 8: Results with few-shot demonstrations and explicit guidelines.

Joint Ranking (Four-Track Comparison). Table 6 reports results for joint reranking, where the model receives all four candidate clips simultaneously and outputs a single global ordering. Consistent with Section 4.3, joint reranking does not yield reliable improvements across models.

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.

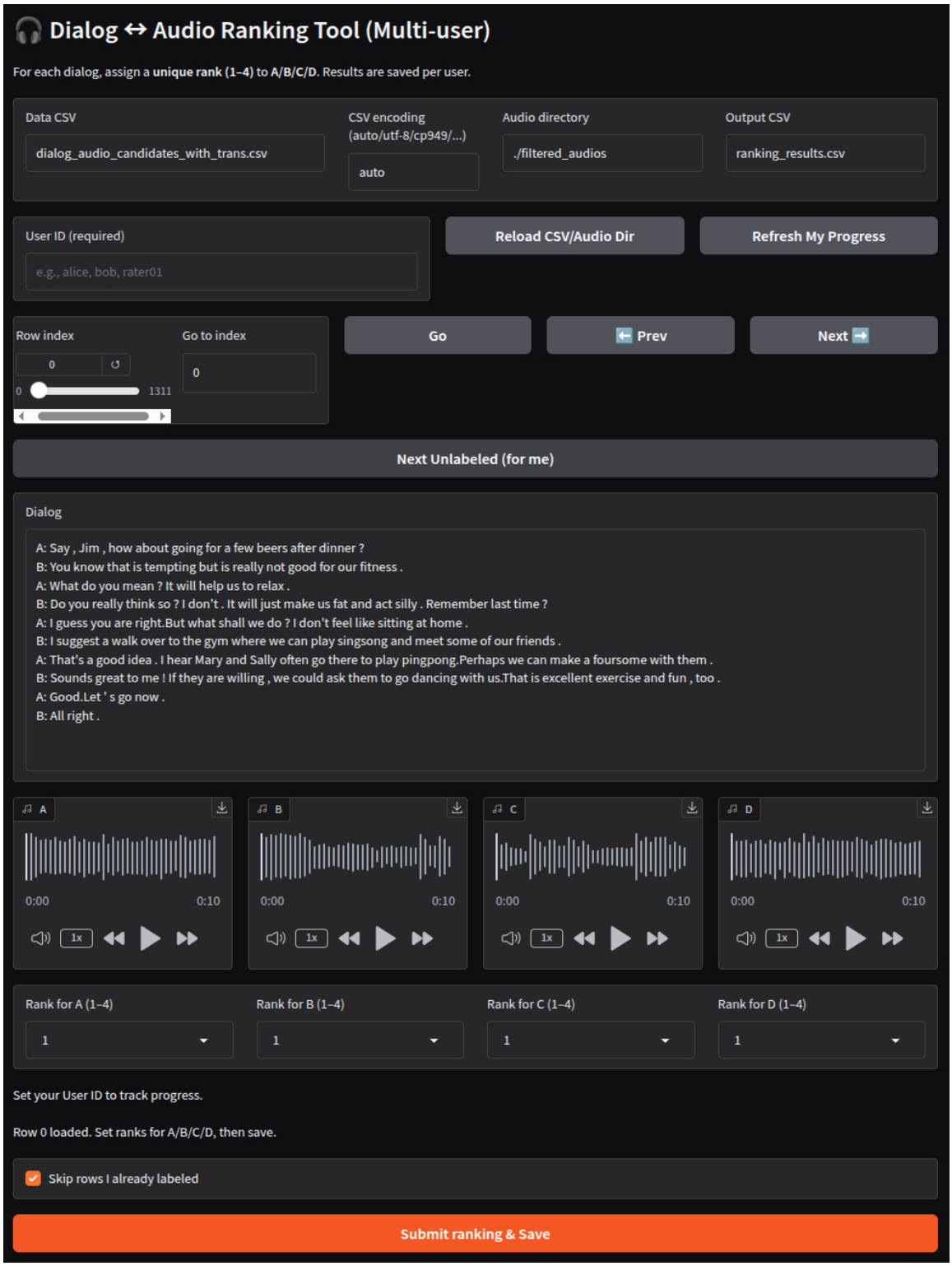

Figure 6: Screenshot of the Gradio-based data collection interface. The tool displays the dialogue and allows annotators to rank four BGM candidates (Candidates A–D) before proceeding to the next turn.

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.