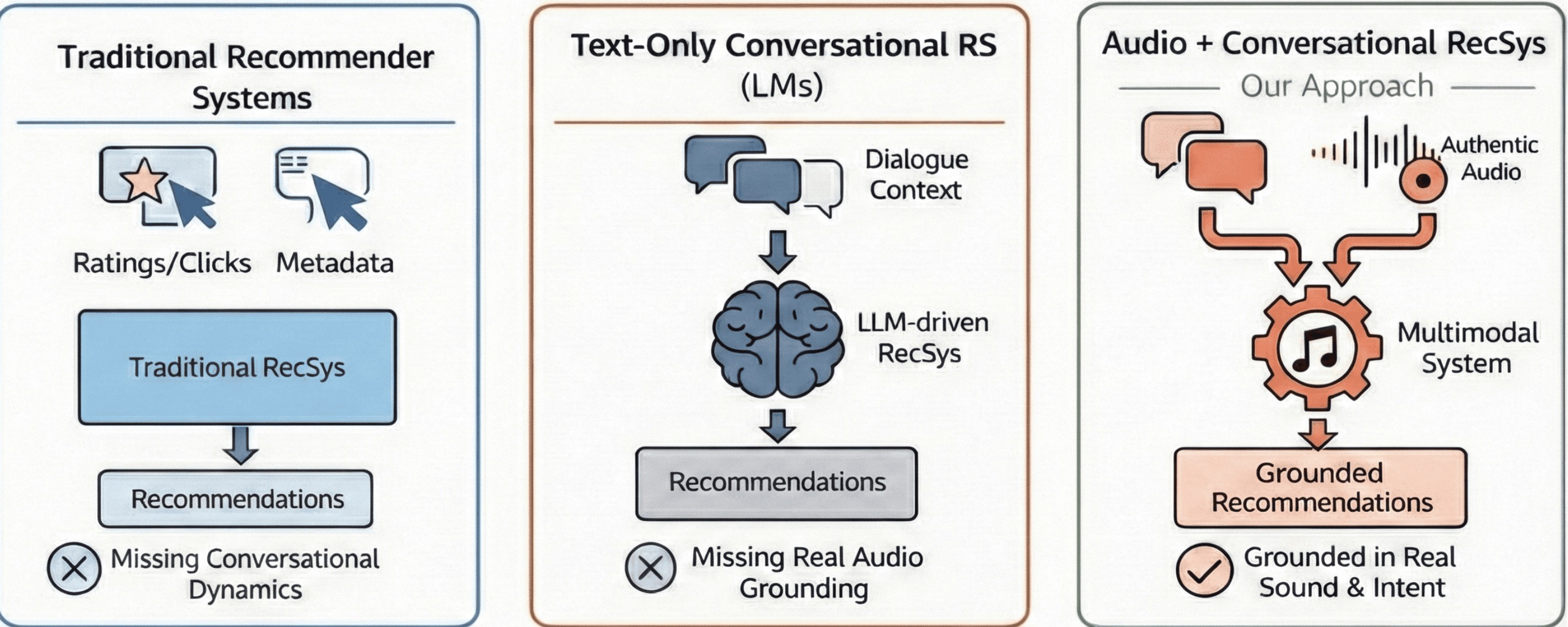


Motivation

- Conversational recommender systems have advanced rapidly with LLMs, but **music** remains uniquely challenging.
- Listener preferences depend on rhythm, timbre, instrumentation, and production style. These are hard to capture through text or metadata alone.
- Existing benchmarks focus on movies or use synthetic dialogue. None combine authentic conversations, real audio grounding, and recommendation evaluation.

Where MusiCRS Fits



Traditional RS ignore dialogue. Text-only CRS miss the audio. MusiCRS grounds conversations in real music.

Contributions

01 New Benchmark

First audio-centric CRS benchmark: 477 authentic Reddit discussions grounded in validated YouTube audio.

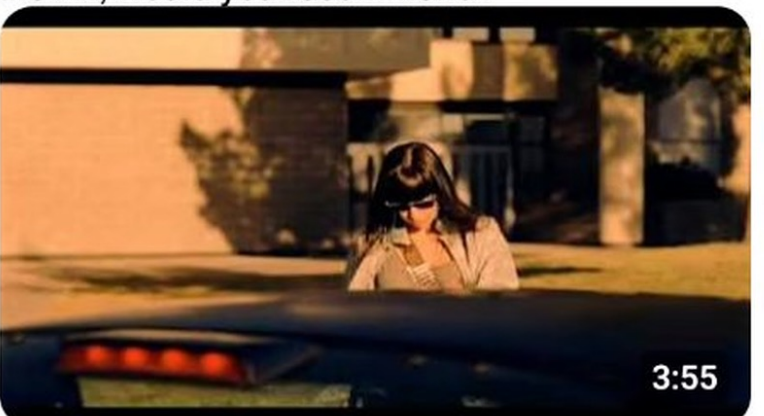
02 Rigorous Protocol

Three input modalities (audio, query, combined) evaluated across 11 generative, retrieval, and traditional systems.

03 Key Insight

Multimodal models frequently lose to single-modality baselines, exposing cross-modal integration gaps.

Genre Examples

Classical	Electronic Music
Reddit Post Title: What classical pieces would make good guitar solos? ... Query: What classical pieces, specifically Faure's Pavane, Paganini's Caprice No. 5, and CPE Bach's Solfeggietto, would be suitable for adaptation into compelling guitar solos?  Ground Truth: - Waldstein sonata 1st movement - Moonlight Sonata 3rd movement - chromatic fantasy and fugue - Adagio K 540 Candidates: - The Four Last Songs - Choral Symphony - Adagio K 540 - Je vous revois - Devil's trill sonata	Reddit Post Title: What's your favorite song with a huge build-up and an incredibly satisfying climax? Examples of this would be:[deadmau5 - Str.... Query:What are some songs, like deadmau5's Strobe or Uppermost's No Human Code, that feature a significant build-up leading to an incredibly satisfying climax  Ground Truth: - Rebound - Cott's Face - Right This Second - The Palace Candidates: - Right This Second - Scala - The Bay - Kensington Gardens - Out of Touch
Indie	Hip-Hop
Reddit Post Title: Looking for songs that have a nostalgic feel to them Songs like Wait by DIIV or Gold in the Air by Jesse Woods. Query: What songs, similar in nostalgic feel to DIIV's Wait and Jesse Woods' Gold in the Air, would you recommend?  Ground Truth: - Rill Rill - You get what you give - Cherub Rock - Hummer Candidates: - Malcyon (Beautiful Days) - For Free - Hummer - You get what you give - Comrade	Reddit Post Title: What is your favorite posse track? Like the title says, what is your favorite posse track? I actually have a tie for my favorite with [1 Train ... Query: What are some of your favorite posse tracks, particularly ones comparable to the collaborative tracks 1 Train featuring A\$AP Rocky, Kendrick Lamar, Joey...  Ground Truth: - Da Graveyard - Chesdown - Oldie - So Appalled Candidates: - The Magic Number - Hello Goodbye - Passing Me By - Knuck If You Buck - Chesdown

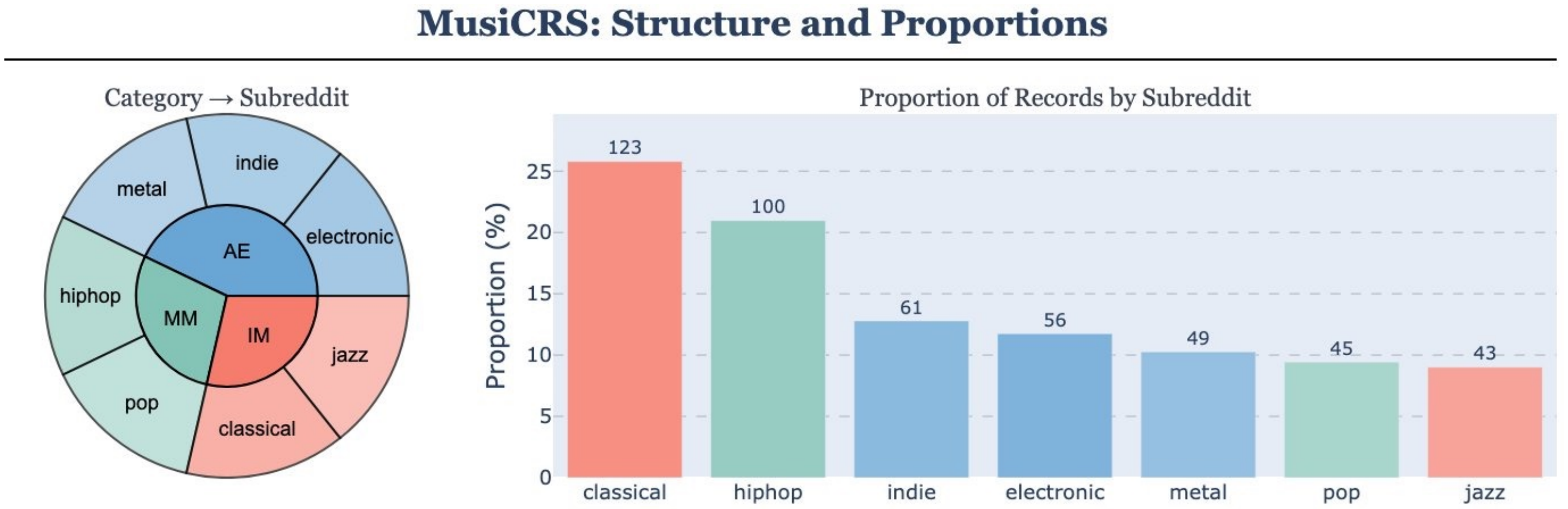
The MusiCRS Dataset

477 Conversations	3,589 Music Entities	7 Genres	100 Candidates / query
-----------------------------	--------------------------------	--------------------	----------------------------------

Construction Pipeline

1 Reddit Mining 2.7M posts + 28.5M comments from 7 music subreddits
2 Three-Stage Filter Valid YouTube links, multi-turn threads, substantive replies
3 Entity + Query Gen Qwen2.5-7B extracts entities, Gemma-3-12B generates queries
4 Human Validation 4 graduate annotators reviewed 1,623 threads, kept 477

Genre Coverage



Balanced across modern/mainstream (MM), instrumental/art (IM), and alternative/experimental (AE) macro-genres.

MusiCRS vs. Existing Datasets

Dataset	Real Conv.	Audio	Ground	Recom.	Multi
CPCD	✓	✗	✓	✓	✗
TalkTheWalk	✗	✗	✓	✓	✗
LP-MusicDialog	✗	✗	✓	✓	✗
Audio Dialogues	✗	✓	✓	✗	✓
LastFM	✗	✗	✗	✓	✗
MusiCRS	✓	✓	✓	✓	✓

MusiCRS is the only benchmark combining authentic conversations, real audio, ground truth, recommendation evaluation, and multimodality.

Three Evaluation Modalities

Audio-only Isolates musical content	Query-only Isolates dialogue text	Audio + Query Tests cross-modal fusion
---	---	--

Human Validation Impact

1,623 Threads reviewed	4 Graduate annotators	41.3% Invalid URLs removed	31.1% Recategorized
----------------------------------	---------------------------------	--------------------------------------	-------------------------------

Experimental Setup

11 Models Evaluated Generative (7) Qwen2-Audio, Qwen2.5-Omni, SALMONN-7B/13B, FUTGA, Audio Flamingo 3, Phi-4-MM Retrieval (2) CLAP, CoLLAP Traditional (2) Popularity, Neighborhood CF	Metrics & Protocol Metrics Recall@20, nDCG@20, MRR Ranking task 100 candidates per query (10 ground truth + 90 subreddit distractors) Audio budget 300s across up to 10 clips, resampled per model
--	--

Key Findings

RQ1 Does multimodal win?

Only 3 of 9 models perform best when combining audio + query. Current fusion mechanisms under-exploit complementary signals.

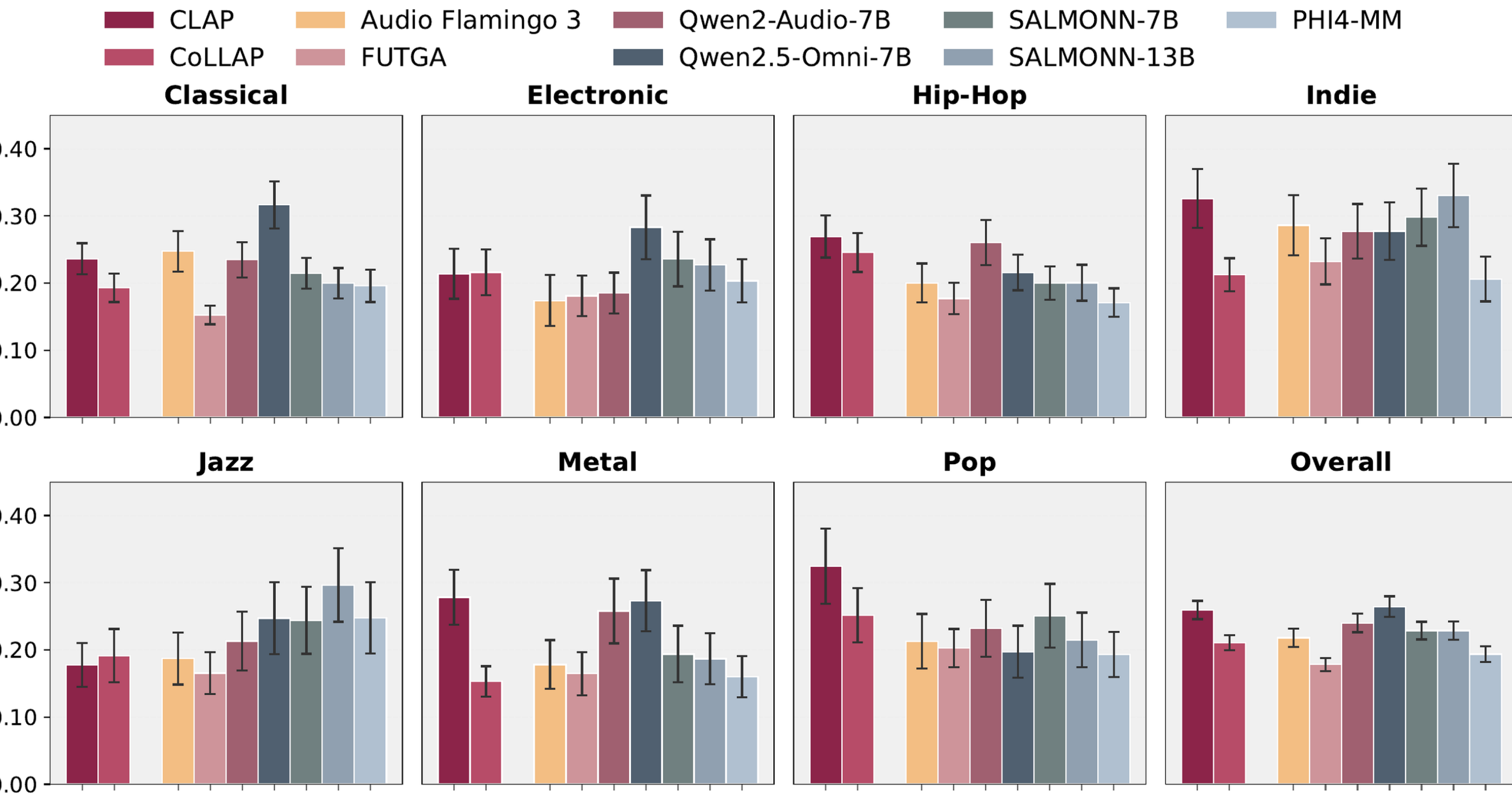
RQ2 Which paradigm is strongest?

Retrieval beats generation. CLAP reaches 22.71% Recall@20 vs. 21.93% for the best generative model (Qwen2.5-Omni).

RQ3 Any genre patterns?

Jazz consistently ranks highest (up to 28% Recall@20). Classical and Hip-Hop remain persistently difficult.

MRR Across Genres



Retrieval models (CLAP, CoLLAP) remain robust across genres. Generative models show higher variance and struggle most with Classical and Hip-Hop.

Takeaway

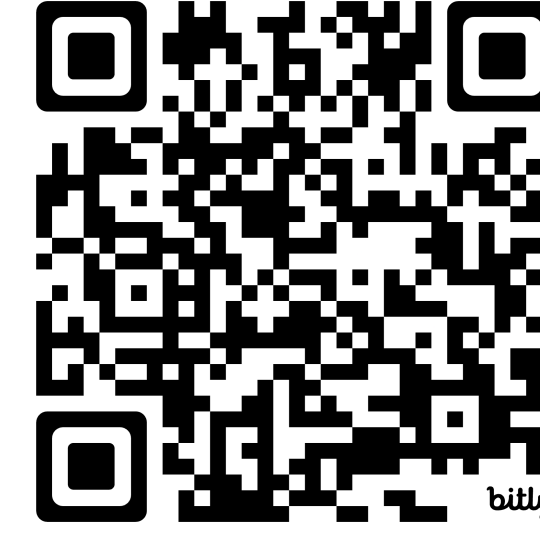
- Current multimodal models fail to ground **abstract musical concepts** in actual audio.
- MusiCRS exposes this gap and provides the benchmark to close it.

Acknowledgments

This work was partially supported by the U.S. National Science Foundation under Grant IIS-2432486.



Paper



Code + Data