Do Joint Language-Audio Embeddings Encode Perceptual Timbre Semantics?

Qixin Deng, Bryan Pardo, Thrasyvoulos N Pappas

qixindeng2026@u.northwestern.edu

Northwestern | McCORMICK SCHOOL OF ENGINEERING

NeurIPS 2025 Workshop on AI for Music — Creativity Meets Computation

Overview

- Joint language–audio models are widely used for retrieval, captioning, and text-guided audio generation.
- Unclear whether these models encode **perceptual timbre semantics** (e.g., bright, rough).
- We evaluate three embedding models using both **human-rated instrument timbre** and **DSP-controlled timbre manipulations.**

Background

- Timbre descriptors (bright, dark, warm) are central in music production, effects control, and sound design
- Joint language-audio embeddings map audio and text into a shared space.
- Excel at identifying sound sources/events
- Their representation of **subtle perceptual sound qualities**, especially timbre, has not been systematically studied.

Research Questions

- **Do joint language–audio embeddings reflect human-perceived timbre semantics?**
- How well do different models align with:
  - **Human timbre ratings** across diverse instruments?
  - **Controlled timbre changes** produced by EQ and reverb?
- Which model provides the most perceptually grounded timbre representation?

Experiment 1: Instrumental Timbre Semantics

Dataset: Jiang et al.'s **CCMusic-Database-Instrument-Timbre** datase (**37** Chinese, **24** Western instruments, **16** descriptors rated by trained listeners).

instrument_name	slim	bright	dim	sharp	thick	thin
violin	5.2	5.3	3.4	4.1	4.1	3.7
viola	3.5	4.4	4.4	3.4	6.2	2.9

human-rated data sample from CCMusic, each descriptor is rated out of a scale of 9(maximum)

Data: **61** Audio clips of different instruments
16 human-rated timbre descriptors.

Audio Encoder →Audio Embeddings (N x D)

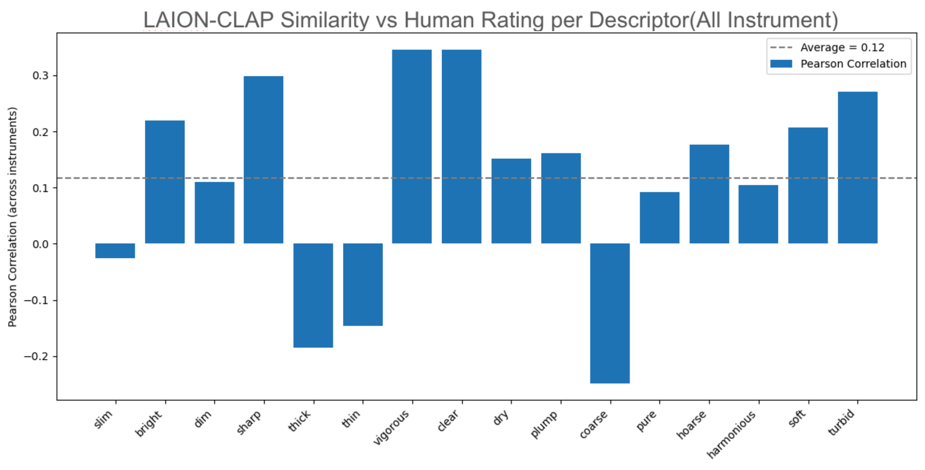
Text Encoder → Text Embeddings (16 x D)

Cosine Similarity Computation
Instrument i, audio embedding a_i,
Descriptor d, text embedding t_d
$s_i = \text{cosine similarity}(a_i, t_d),\ \ s_i = [s_{i,d}]_{d \in D}$

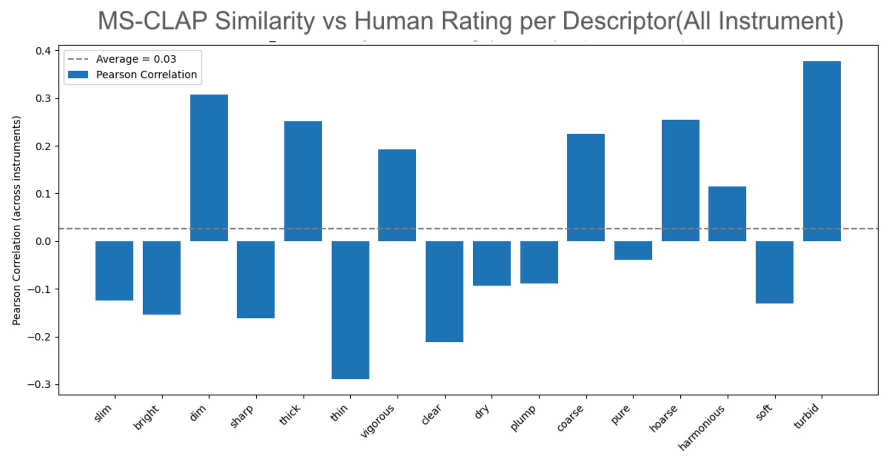
Correlation Computation
calculated similarities: $s_{i,d}$ human-rated scores: $h_{i,d}$
Per Descriptor: $r_d^{\text{Pearson}} = \text{corr}_{\text{Pearson}}(h_{i,d}, s_{i,d})$
Per Instrument: $r_i^{\text{Pearson}} = \text{corr}_{\text{Pearson}}(h_i, s_i)$

Results

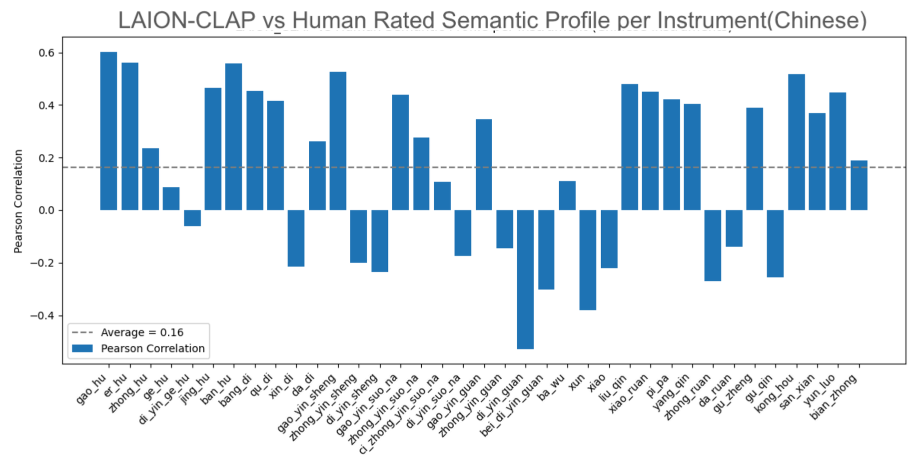
LAION-CLAP Similarity vs Human Rating per Descriptor(All Instrument)



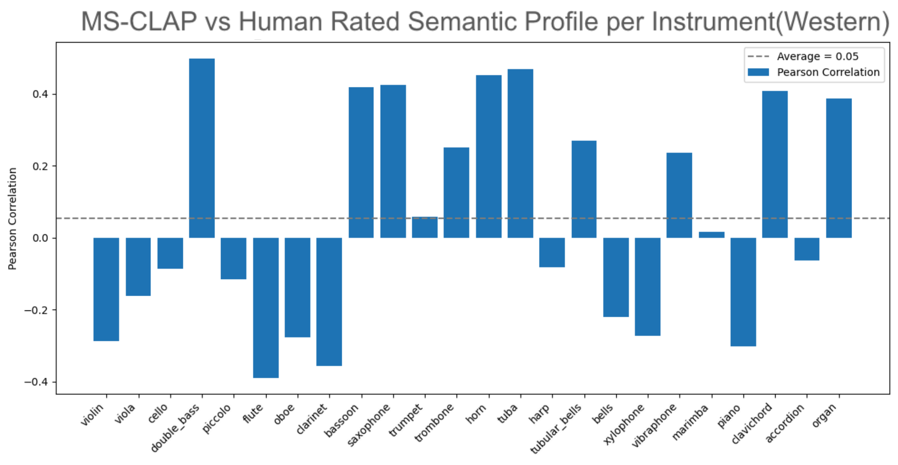
MS-CLAP Similarity vs Human Rating per Descriptor(All Instrument)



LAION-CLAP vs Human Rated Semantic Profile per Instrument(Chinese)



MS-CLAP vs Human Rated Semantic Profile per Instrument(Western)



Model	Descriptor Level (#Positive /16)	Chinese Instrument (# Positive /37)	Western Instrument (# Positive /24)
Laion-CLAP	12	24	10
MS-CLAP	7	24	12
MUQ-MULAN	7	16	10

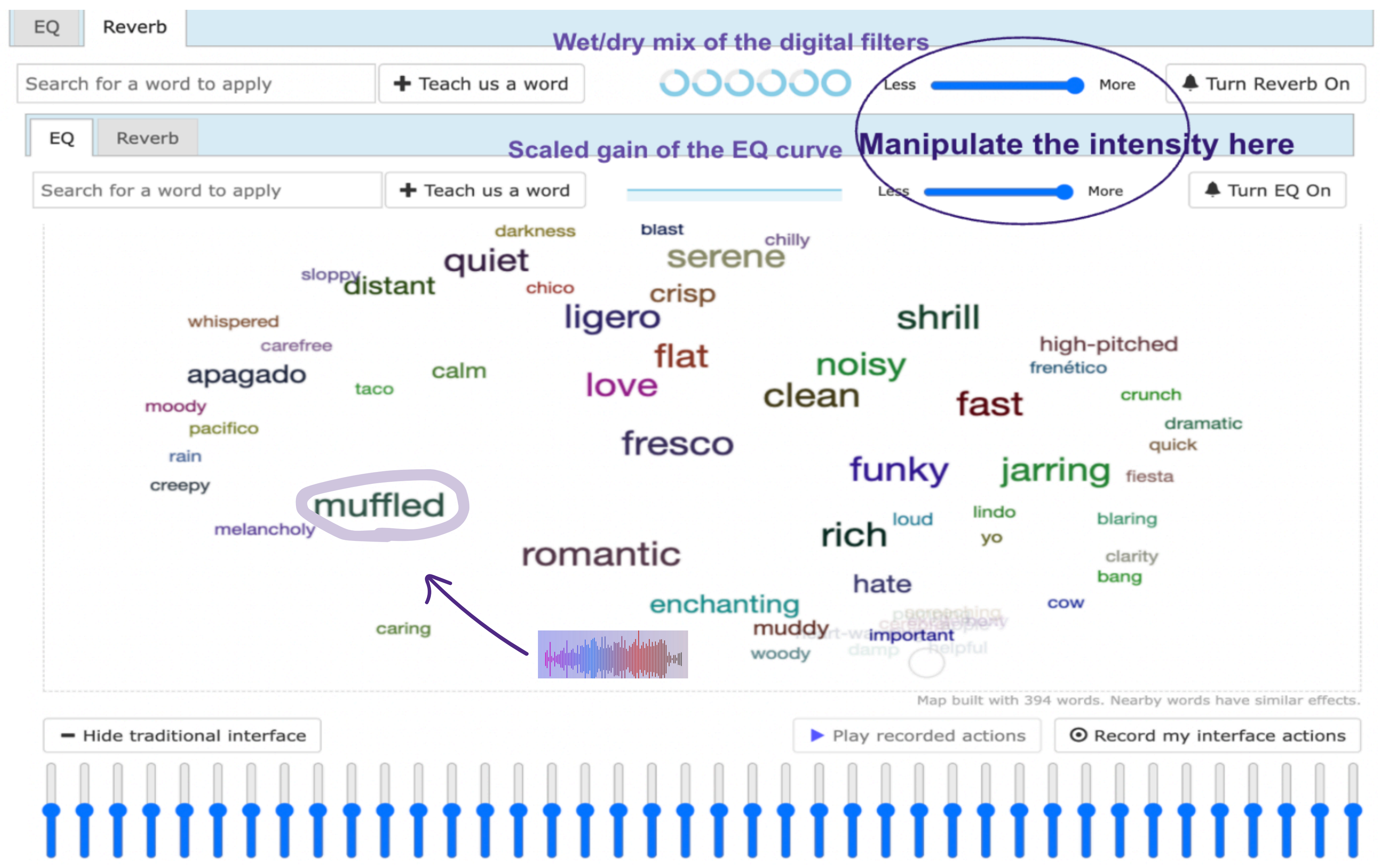
Conclusion

- Across both instrumental timbre and audio-effect manipulations, all three embedding spaces show **limited alignment** with human timbre perception.
- **LAION-CLAP is relatively stronger,** but significant gaps remain.
- MS-CLAP and MuQ-MuLan show weak or inconsistent alignment.

Experiment 2: Audio Effect Timbre Semantics

Dataset: Top 20 EQ and Top 20 Reverb descriptors and their corresponding parameters from **Audealize**.

For each descriptor: Original reference + audio EQ or reverb at three intensity levels (0.3, 0.6, 1.0) → **7 audio files per descriptor**



text embeddings for descriptor d, audio embedding (original + manipulated)
Compute similarity change:

$$\Delta(a) = Sim(\text{audio}_{\text{manip}}(d, a), \text{text(d)}) - Sim(\text{audio}_{\text{orig}}), \text{text(d)})$$

Δ↑ monotonic: strong semantic alignment
Δ flat: inconsistent encoding
Δ↓ monotonic: opposite perceptual meaning

Results

LAION-CLAP = Best performance, 14 / 20 monotonic ↑ trend for EQ and **12/20** monotonic ↑ for reverb
MS-CLAP: No prominent trend of alignment for both EQ and reverb
MUQ-MULAN: 8/20 monotonic ↑ trend for EQ, no prominent trend of alignment for reverb

Descriptor	MS-CLAP	LAION-CLAP	MuQ-MuLan
bass	-	-	↓
big	-	-	↓
church	-	↑	↓
clear	↓	↓	↓
deep	↓	↑	↓
distant	↓	↑	↓
distorted	↑	-	↓
echo	↓	↑	↑
hall	-	↑	↓
haunting	↑	↑	↑
hollow	↓	↑	-
loud	-	-	↓
low	-	↑	-
muffled	↓	-	↑
sad	↑	-	-
soft	↓	↑	↓
spacious	-	↑	↓
strong	-	↑	↓
tinny	↓	↑	↓
warm	↓	↓	↓

Reverb

Descriptor	MS-CLAP	LAION-CLAP	MuQ-MuLan
bright	-	↑	-
calm	↓	↑	↓
clear	-	↑	-
cold	↓	-	↓
cool	-	↓	↓
crisp	↓	↑	-
dark	-	↑	↑
gentle	-	↓	↓
hard	↓	↑	↑
harsh	-	-	↓
heavy	↓	↑	↑
loud	-	↑	↑
mellow	-	↑	↑
peaceful	↓	-	↑
sharp	↓	↑	↓
smooth	-	↑	↑
soft	-	↑	↓
soothing	↓	↑	↓
tinny	-	↑	↓
warm	↓	↓	↑

EQ

Reference

Wei Jiang, Jingyu Liu, Zijin Li, Jiaxing Zhu, Xiaoyi Zhang, and Shuang Wang. Analysis and modeling of timbre perception features of chinese musical instruments. In 2019 IEEE/ACIS18th International Conference on Computer and Information Science (ICIS), pages 191–195,2019

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 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, 2023

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

Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Yi Luo, Wei Tan,and Xie Chen. Muq: Self-supervised music representation learning with mel residual vectorquantization. arXiv preprint arXiv:2501.01108, 2025.

Prem Seetharaman and Bryan Pardo. Audealize: Crowdsourced audio production tools. Journal of the Audio Engineering Society, 64(9):683–695, 2016.