Do Joint Language-Audio Embeddings Encode Perceptual Timbre Semantics?

Qixin Deng **Bryan Pardo** **Thrasyvoulos N. Pappas**
Northwestern University, Evanston, IL, USA

Abstract

Understanding and modeling the relationship between language and sound is critical for applications such as music information retrieval, text-guided music generation, and audio captioning. Central to these tasks is the use of joint language–audio embedding spaces, which map textual descriptions and auditory content into a shared embedding space. While multimodal models such as MS-CLAP, LAION-CLAP and MuQ-MuLan have shown strong performance in aligning language and audio, their correspondence to human perception of timbre, a multifaceted attribute encompassing qualities such as brightness, roughness, and warmth, remains underexplored. In this paper, we evaluate the above three joint language–audio embedding models on their ability to capture perceptual dimensions of timbre. Our findings show that LAION-CLAP consistently provides the most reliable alignment with human-perceived timbre semantics across both instrumental sounds and audio effects.

1 Introduction

Joint language–audio embedding spaces align textual descriptions and auditory content in a shared semantic representation. These models learn to project audio and language into a common embedding space, where semantically related pairs are close together, enabling tasks such as cross-modal retrieval, audio captioning, and text-guided audio effects[1] and music generation[5]. Recent models such as MS-CLAP [2, 3], LAION-CLAP [8] and MuQ-MuLan [10]have demonstrated strong performance in identifying audio content, for example, recognizing that a clip contains a saxophone solo or footsteps on gravel. What remains less clear, however, is whether these models also capture how the sound is perceived, particularly its timbral qualities: a saxophone may be described as warm, bright, or raspy, while footsteps might sound light, crunchy, or heavy. Such timbral attributes are often more subtle and may not be well represented in training metadata.

Studies on timbral semantics largely examine how humans perceive and describe timbre, following two main approaches. The first focuses on single instruments. Jiang et al. [4] distilled 329 descriptors into 16 core terms (e.g., bright–dark, raspy–mellow) through listening tests on 72 instruments. Similarly, Roche et al. [6] collected 784 verbal expressions of synthetic sounds from 101 French-speaking listeners and clustered them into eight perceptual dimensions. The second approach, alternatively, links timbre descriptors to audio effect parameters. SocialFX [9] crowdsourced descriptions of equalization, reverberation, and compression from over 480 participants, yielding hundreds of unique terms. Across studies, adjectives such as warm, bright, sharp, and clear consistently emerged, suggesting that perceptual patterns generalize across sources and effects.

To our knowledge, no prior work thoroughly investigates how well descriptors of timbre are encoded in joint language–audio embedding spaces. While Text2FX [1] explores whether MS-CLAP encodes timbral semantics at all, it does not evaluate the extent of this encoding. In this work, we assess the perceptual validity of joint language–audio embeddings with respect to timbral semantics. Using human-annotated datasets from Jiang et al. [4] and SocialFX [9], we evaluate how popular embedding

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

spaces preserve timbral characteristics and how this preservation varies across models. These insights can inform the development of perceptually grounded systems for automatic timbre analysis and improve downstream applications such as timbre-based music retrieval, sound design, and interactive audio tools. The contributions of our work are:1. A methodology for assessing language-audio embedding model alignment with human perception of timbre. 2. An evaluation and comparison between popular language-audio embeddings using this methodology.

2 Experiments

We performed two experiments to evaluate the alignment between three popular audio-text embedding models (MS-CLAP, LAION-CLAP, and MuQ-MuLan) and human perception of timbre. In the first experiment, we assessed whether language-audio embedding models capture human-perceived timbre semantics of instruments. In the second experiment, we investigated how these three embedding models capture perceptual timbre descriptors in relation to audio effects control trends, specifically equalization (EQ) and reverberation.

2.1 The Models

Although all of these models use contrastive learning to align audio clips with their corresponding textual descriptions, they differ in training data and domain coverage. MS-CLAP and LAION-CLAP target general audio understanding, meaning that they are trained to represent a broad spectrum of sounds, including music, speech, environmental sounds(e.g., dogs barking, doors closing, waves crashing) to abstract auditory events (e.g., alarms, sirens). MS-CLAP is trained on a combination of FSD50k, Clotho V2, AudioCaps, and MACS, spanning music, speech, natural sounds, and abstract auditory events paired with human-written captions; LAION-CLAP uses their own curated large-scale LAION-Audio-630k dataset, which contains environmental and human-related audio clips labeled via keyword-to-caption augumentation. While the original MuLan model is not open-sourced, we use the open-source MuQ-MuLan, which focuses specifically on music and is trained on video soundtracks paired with rich metadata such as tags, titles, descriptions, and user comments.

2.2 Experiment 1: Instrumental Timbre Semantics

In the first experiment, we assessed whether language-audio embedding models capture human-perceived timbral semantics at both the descriptor and instrument level, using Jiang's CCMusic-Database-Instrument-Timbre dataset[4]. The dataset contains short audio clips for 37 Chinese and 24 Western instruments, each annotated with ratings for 16 semantic descriptors (e.g., bright, dark, raspy). Ratings were collected from 34 Chinese-speaking participants with musical training, who judged on a nine-point scale the degree to which each descriptor applied to each instrument. This dataset is openly available and was obtained through a controlled listening test, providing a reliable ground truth for perceptual timbre semantics.

In our experiment, for each pair of instrument audio recording and timbre descriptor, we encode them using the embedding model to obtain an audio embedding $\mathbf{a}_i$ and a text embedding $\mathbf{t}_d$. Cosine similarity was then computed between the audio embedding and text embedding, yielding a 16-dimensional similarity profile $\mathbf{s}_i = [s_{i,d}]_{d\in\mathcal{D}}$ for instrument i. Each entry $s_{i,d}$ reflects the strength of association, in the joint embedding space, between the instrument's sound and descriptor d. The underlying hypothesis is that if the embedding space encodes timbral semantics, for the instruments with higher human ratings for descriptor d (e.g., *bright*), its audio embedding should be positioned near to the text embeddings of d, resulting in higher cosine similarity value $s_{i,d}$. Two complementary correlation analyses were performed:

1. Descriptor-level correlation: For each descriptor d, Pearson correlations were computed between human ratings $\{h_{i,d}\}_i$ and embedding similarities $\{s_{i,d}\}_i$ across all instruments. A high positive correlation for a descriptor indicates that instruments judged by listeners as strongly expressing d also appear closer to $\mathbf{t}_d$ in the embedding space, reflecting semantic alignment for that perceptual quality. A low or near-zero correlation suggests weak or no alignment between the embedding space and human perception for that descriptor. A negative correlation indicates a mismatch, where instruments rated highly on descriptor d by humans are placed *farther away* from $\mathbf{t}_d$ in the embedding space, suggesting the model encodes an opposite or contradictory association.

2. Instrument-level semantic profile correlation: For each instrument i, its 16-dimensional human rating vector $\mathbf{h}_i$ was correlated with its 16-dimensional similarity profile $\mathbf{s}_i$. A high correlation indicates that the embedding captures the overall timbre profile of the instrument i across descriptors. A low correlation implies that the embedding fails to reproduce the joint configuration of timbral attributes as perceived by listeners. A negative correlation indicates a systematic inversion, where descriptors that listeners strongly associate with an instrument are those that the embedding places far away, suggesting the model misrepresents the instrument's perceptual timbre profile.

2.2.1 Results for Experiment 1

At the descriptor level, LAION-CLAP demonstrated the strongest alignment with human ratings, with 12 out of 16 descriptors showing positive correlations. Its highest correlation was observed for the descriptor *vigorous* ($r = 0.35$), while the lowest was for *coarse* ($r = -0.25$). In contrast, MS-CLAP achieved positive correlations for only 7 descriptors. Its strongest alignment was for *turbid* ($r = 0.40$), but it also exhibited notable negative correlations, particularly for *thin* ($r = -0.28$). Similarly, MuQ-MuLan yielded 7 positive and 9 negative correlations. Although it reached a relatively high positive value for *slim* ($r = 0.41$), it failed dramatically for *vigorous* ($r = -0.48$). These results suggest that LAION-CLAP provides the most consistent descriptor-level alignment with human perception than MS-CLAP and MuQ-MuLan. **The detailed visualization of the descriptor-level correlations can be found in Figure 1, Figure 2 and Figure 3 in Appendix**.

At the instrument level, LAION-CLAP demonstrated the strongest overall alignment with human semantic profiles, producing 24 positive correlations out of 37 Chinese instruments (mean $r = 0.162$). MS-CLAP showed the same number of positives correlations but with a weaker average correlation ($r = 0.058$). MuQ-MuLan was less consistent, with only 16 positive and 21 negative correlations with nearly zero mean Pearson correlations which indicates very little perceptual validity. For Western instruments, MS-CLAP results in the best alignment(12 positives, 12 negatives, mean $r = 0.0528$). LAION-CLAP has weaker alignment (10 positives, 14 negatives, mean $r = 0.027$). MuQ-MuLan archived a same result as LAION-CLAP (10 positives, 14 negatives) but slipped into a slightly negative mean ($r = -0.027$). **The detailed visualization of instrument-level correlation can be found in Figure 4 to Figure 9 in Appendix**.

2.3 Experiment 2: Audio Effect Timbre Semantics

While Experiment 1 evaluated embeddings using naturally occurring timbral variation across instruments, real-world recordings also differ in pitch, dynamics, and recording conditions, making it difficult to isolate timbre. To address this, Experiment 2 systematically manipulated timbre through digital signal processing (DSP), allowing precise control over the type and magnitude of change. This design builds on SocialFX[9], which is a large crowdsourced collection linking 4,297 unique vocabulary terms to precise and quantified audio effect parameter settings. These mappings provide a perceptually grounded reference for how layperson descriptors (e.g., *warm*, *harsh*) correspond to measurable timbral changes. Two effect types were considered:

1. Equalization (EQ): Implemented using a 40-band parametric equalizer, where each band is defined by a center frequency, bandwidth, and gain. For each descriptor, the SocialFX parameters specify the gain adjustments across bands. An amount scaling factor was applied to linearly scale all band gains, producing three discrete effect intensities (0.3 = low, 0.6 = medium, 1.0 = high).

2. Reverberation: Implemented with a digital reverberator combining parallel comb filters, all-pass filters, and low-pass filters. Parameters included decay time, feedback gain, modulation, low-pass cutoff frequency, and overall effect gain. The wet/dry ratio controlled effect intensity at three discrete levels((0.3 = low, 0.6 = medium, 1.0 = high).

From the full SocialFX vocabulary list, the 20 most frequently used descriptors were selected separately for EQ and reverb. For each descriptor, timbre-manipulated audio was generated from a common reference file at each intensity level. This reference file corresponds to the original audio track used during the SocialFX listening tests, ensuring consistency with the dataset's perceptual annotations, since all descriptor judgments were made relative to this track. The EQ and reverb implementations were reproduced from Audealize online demo [7].

For each embedding model, the following steps were performed:

1. Text embeddings. A text embedding was computed for each descriptor from SocialFX d.

2. Audio embeddings. Audio embeddings were computed for both the original reference file and files resulting from applying a single effect (EQ or reverb) at one of 3 levels: (low, medium, high), resulting in 7 files per descriptor.

3. Similarity computation. The cosine similarity was calculated between the audio embeddings of the manipulated audio files and the text embeddings of the corresponding timbre descriptors.

$$\mathrm{sim}_{\mathrm{manip}}(d, a) = \mathrm{sim}\big(\mathrm{audio}_{\mathrm{manip}}(d, a), \mathrm{text}(d)\big).$$

Here, a denotes the intensity level of the manipulation (e.g., $a = 0.3$ for low, $a = 0.6$ for medium, $a = 1.0$ for high). The similarity between the unprocessed reference audio file and the timbral manipulation descriptors descriptor was:

$$\mathrm{sim}_{\mathrm{orig}}(d) = \mathrm{sim}\big(\mathrm{audio}_{\mathrm{orig}}, \mathrm{text}(d)\big).$$

The change in similarity due to manipulation was then defined as:

$$\Delta_d(a) = \mathrm{sim}_{\mathrm{manip}}(d, a) - \mathrm{sim}_{\mathrm{orig}}(d).$$

A positive $\Delta_d(a)$ indicates that the manipulation moved the audio embedding closer to the descriptor d in the joint embedding space.

For each descriptor-effect pair, $\Delta_d(a)$ values were examined across intensity levels to classify the trend as monotonic increase, monotonic decrease, or peaking at a specific intensity. A *monotonic increase* suggests that the model's similarity space consistently aligns with the intended timbral change, indicating strong semantic encoding for that descriptor. A *flat or inconsistent* pattern implies weak or no alignment between the DSP-induced timbral changes and the descriptor's semantic representation in the model. A *monotonic decrease* indicates that increasing the manipulation intensity moves the audio embedding *away* from the descriptor's text embedding. This implies that the model associates the descriptor with the opposite perceptual timbral quality.

2.3.1 Results for Experiment 2

For EQ, LAION-CLAP demonstrates the strongest alignment, with 14 out of 20 descriptors exhibiting monotonic up trends, indicating a consistent mapping between timbral qualities and spectral changes. By contrast, MuQ-MuLan shows a mixed performance: while 9 descriptors follow monotonic up trends, several others display monotonic down or localized peak patterns, suggesting less reliable encoding. MS-CLAP performs the weakest, with most descriptors producing monotonic down trends or narrow localized peaks, indicating poor and inconsistent alignment with spectral manipulations. Overall, these results, shown in **Table 1 in Appendix**, suggest that LAION-CLAP provides the most robust text-timbre alignment, while MuQ-MuLan is partially effective, and MS-CLAP fails to encode descriptor-relevant EQ trends in a consistent manner. The reverb analysis reveals weaker and less consistent alignment across all three models, as shown in **Table 2 in Appendix**. LAION-CLAP still shows the strongest alignment, with 12 descriptors following monotonic up mappings as the intensity increases. MS-CLAP and MuQ-MuLan show much weaker performances, with most descriptors showing monotonic down trends or local peaks.

3 Conclusion and Future Work

We conducted a systematic evaluation of three prominent joint language–audio embedding spaces: MS-CLAP, LAION-CLAP, and MuQ-MuLan. Our results show that LAION-CLAP consistently provides the most reliable alignment with human-perceived timbre semantics across both instrumental sounds and audio effects, outperforming MS-CLAP and MuQ-MuLan.Future research may proceed along two directions. First, examining whether LAION-CLAP encodes interpretable timbral axes, such as a perceptual continuum from "bright" to "dark", could yield deeper insights into the structure of the embedding space and its correspondence with perceptual dimensions of timbre. Second, fine-tuning LAION-CLAP with timbre-specific objectives may improve its capacity to capture subtle qualities, thereby enhancing timbre-based retrieval, manipulation, and generative applications.

References

[1] Annie Chu, Patrick O'Reilly, Julia Barnett, and Bryan Pardo. Text2fx: Harnessing clap embeddings for text-guided audio effects. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025.

[2] 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.

[3] Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. Natural language supervision for general-purpose audio representations, 2023.

[4] 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/ACIS 18th International Conference on Computer and Information Science (ICIS)*, pages 191–195, 2019.

[5] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21450–21474. PMLR, 23–29 Jul 2023.

[6] Fanny Roche, Thomas Hueber, and Limier Samuel Garnier, Maëva, and Laurent Girin. Make that sound more metallic: Towards a perceptually relevant control of the timbre of synthesizer sounds using a variational autoencoder. *Transactions of the International Society for Music Information Retrieval*, 4(1):115–131, 2021.

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

[8] 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.

[9] Tianran Zheng, Prem Seetharaman, and Bryan Pardo. Socialfx: Studying a crowdsourced folksonomy of audio effects terms. In *Proceedings of the 24th ACM International Conference on Multimedia (ACM MM)*, pages 182–186. ACM, 2016.

[10] 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 vector quantization. *arXiv preprint arXiv:2501.01108*, 2025.

A Figures and Tables

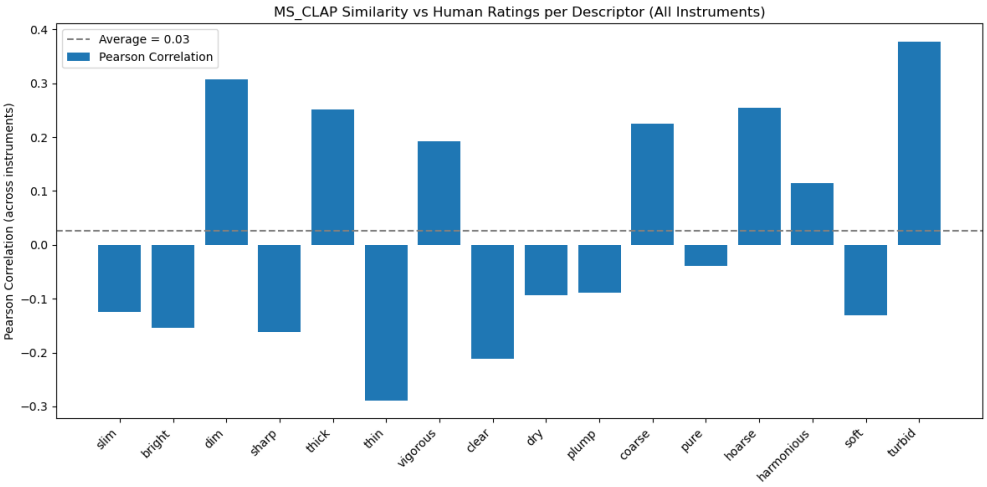

Figure 1: MS-CLAP Similarity vs Human Ratings per Descriptor

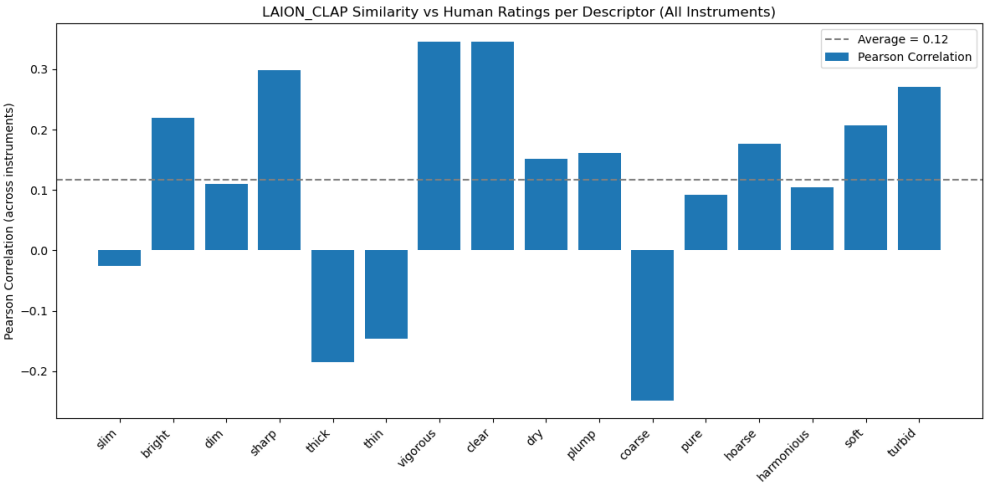

Figure 2: LAION-CLAP Similarity vs Human Ratings per Descriptor

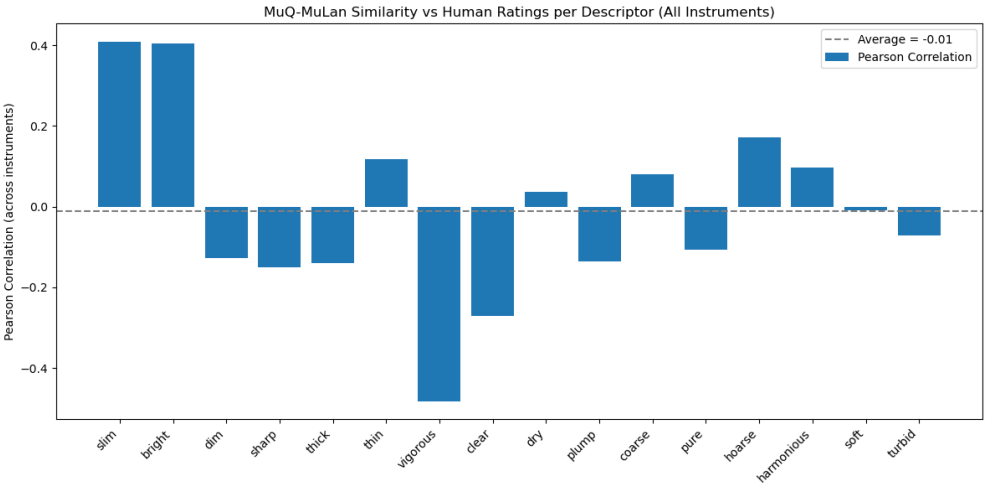

Figure 3: MuQ-MuLan Similarity vs Human Ratings per Descriptor

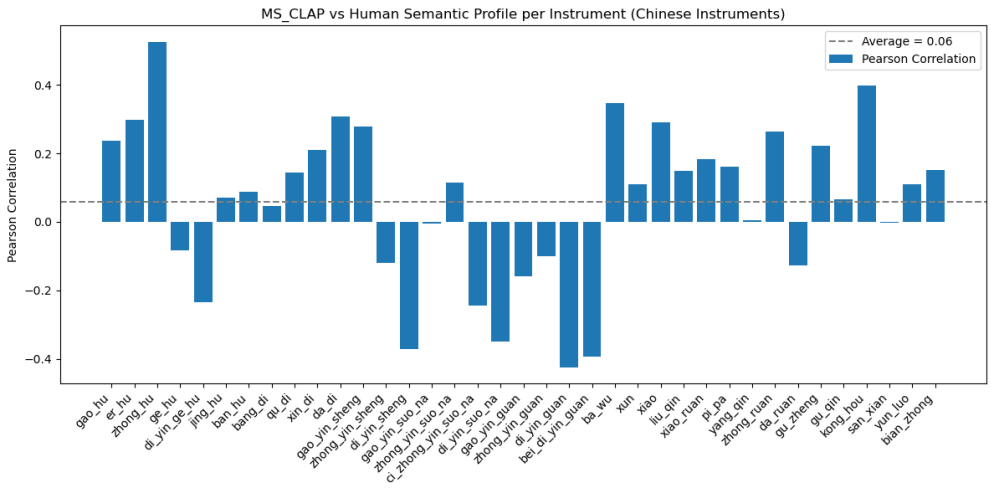

Figure 4: MS-CLAP vs Human-rated Timbre Semantic Profile for Chinese Instruments

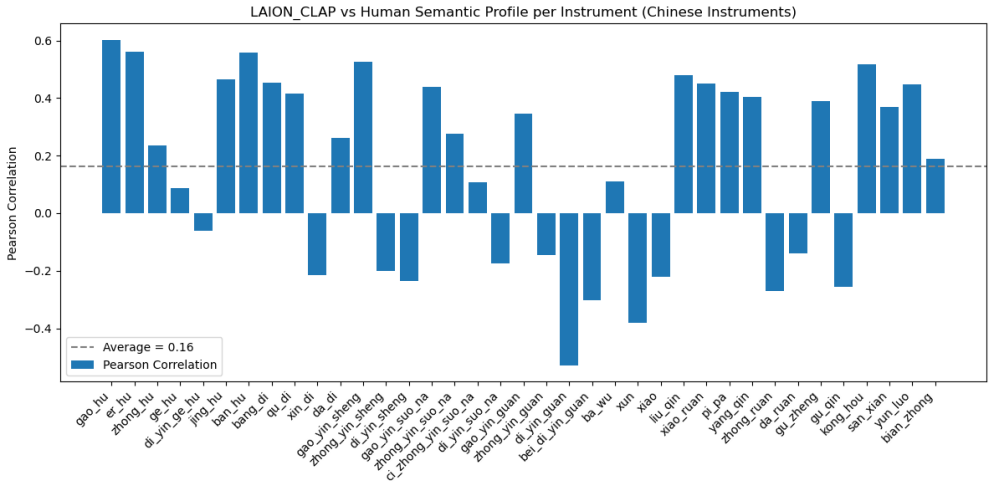

Figure 5: LAION-CLAP vs Human-rated Timbre Semantic Profile for Chinese Instruments

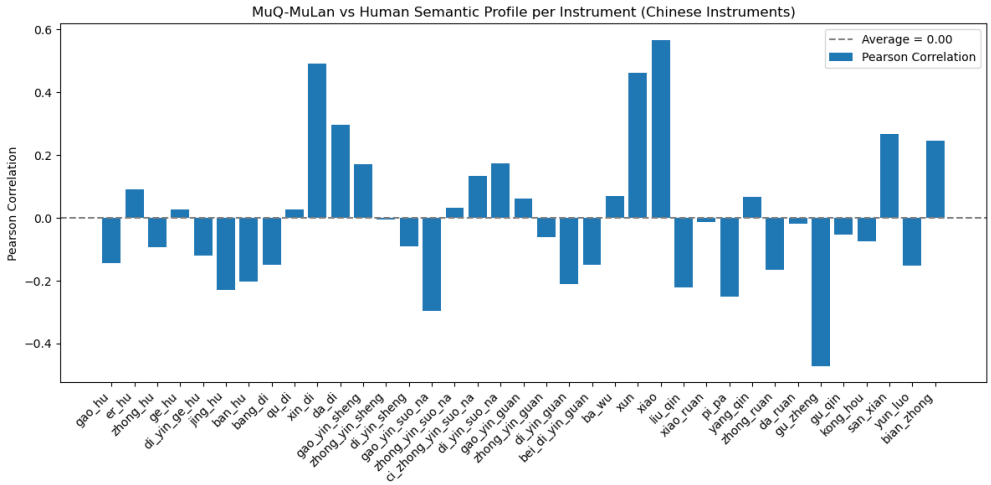

Figure 6: MuQ-MuLan vs Human-rated Timbre Semantic Profile for Chinese Instruments

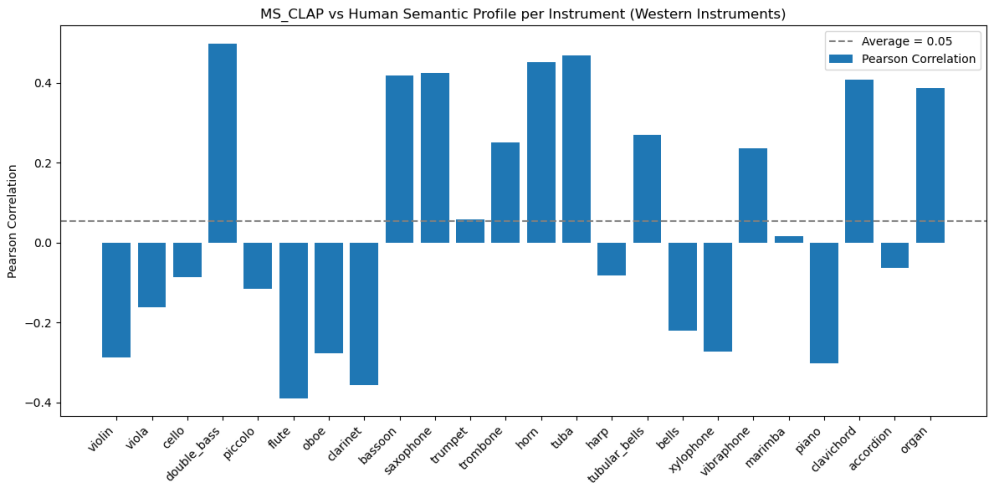

Figure 7: MS-CLAP vs Human-rated Timbre Semantic Profile for Western Instruments

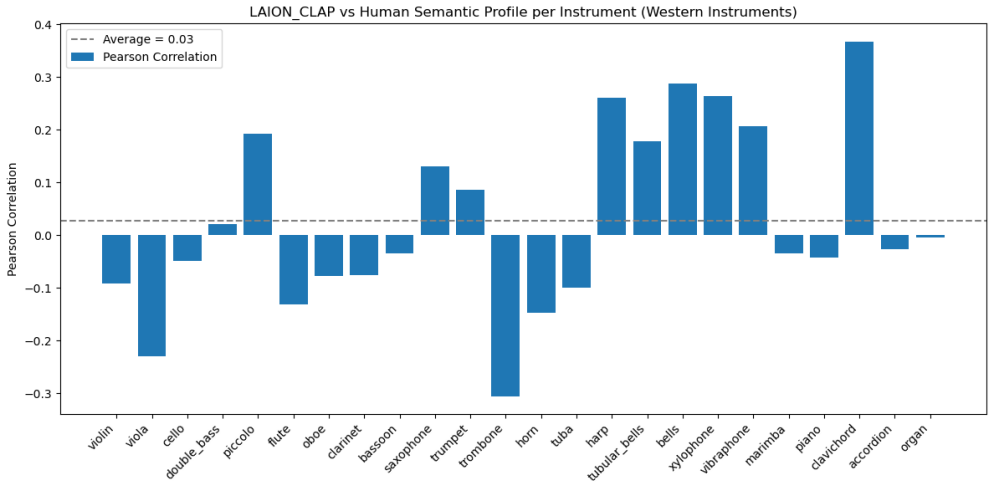

Figure 8: LAION-CLAP vs Human-rated Timbre Semantic Profile for Western Instruments

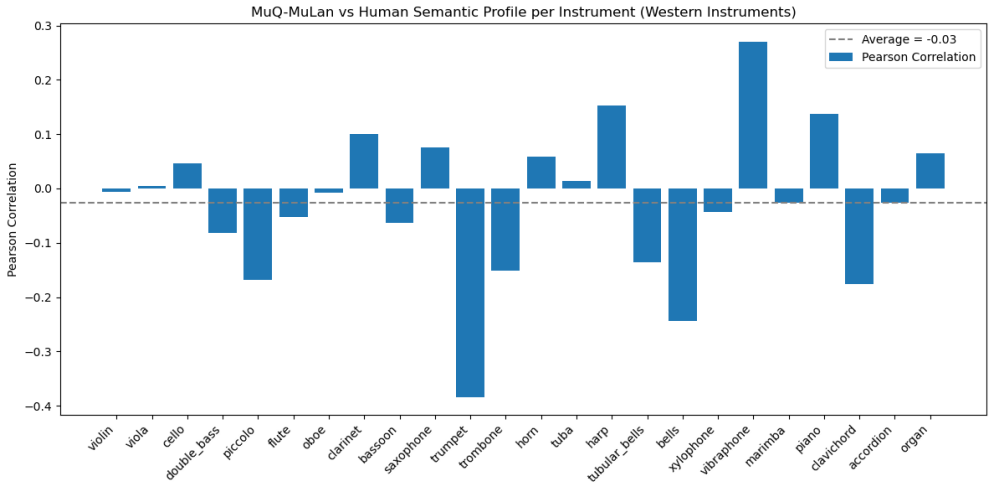

Figure 9: MuQ-MuLan vs Human-rated Timbre Semantic Profile for Western Instruments

Table 1: EQ trend types across MS-CLAP, LAION-CLAP, and MuQ-MuLan for the 20 most timbre descriptors from SocialFX.

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	↓	↓	↑

Legend: ↑ = Monotonic up, ↓ = Monotonic down, - = flat or inconsistent

Table 2: Reverb trend types across MS-CLAP, LAION-CLAP, and MuQ-MuLan for the 20 most timbre descriptors from SocialFX.

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	↓	↓	↓

Legend: ↑ = Monotonic up, ↓ = Monotonic down, - = flat or inconsistent