TimbreCLAP

Adapting CLAP for Timbre-Oriented Sound Retrieval

Jiayi Hu[1]* Jiangnan Zhang[2]* Yifeng Huang[3]* Yuxuan Shi[4] Jinyue Guo[5] Shimeng Li[6]

[1] Twiddle AI, Shanghai, China [2] Shanghai Conservatory of Music, Shanghai, China [3] Fudan University, Shanghai, China
[4] University of Pennsylvania, Philadelphia, USA [6] Shanghai Jiao Tong University, Shanghai, China
[5] RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion / Department of Musicology, University of Oslo, Oslo, Norway
* These authors contributed equally

Motivation

“a warm pad with a slow attack”

Source identity alone does not describe the onset, texture and evolution that guide sound selection. CLAP only partly aligns with perceptual timbre semantics [1,4].

Caption supervision

708,606 short sounds

Instrument notes and synthesizer presets

Audio, metadata and DSP
A: source family
B: brightness, attack, envelope, character and cleanliness

MOSS-Music [3]
11 experts assessed 300 clips
84.82% acceptance / 57.91% best

Long / Medium / Short [2]
L: family, detail and evolution
M: family and salient attributes
S: compact phrases

Expert assessment guides captioner selection. Metadata and DSP provide annotation references, rather than independent perceptual ground truth.

Adapting CLAP

Retain the dual encoder. Update both backbones and projections, with the audio front end and input normalization fixed.

Audio encoder
HTS-AT-base
512-D projection

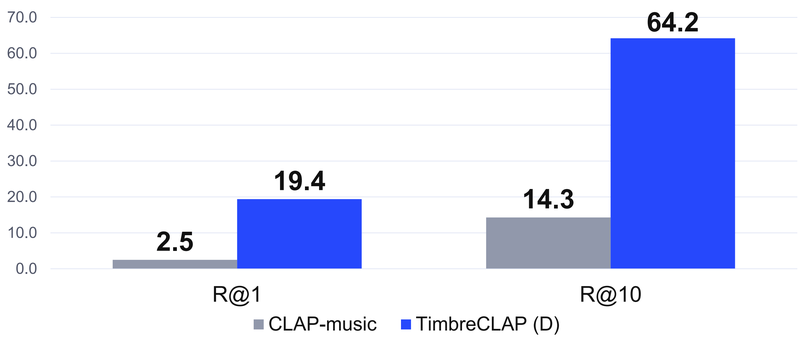
Text encoder
RoBERTa-base
512-D projection

Bidirectional InfoNCE

Rotate L / M / S captions across epochs.

610,090 training clips / 11,094 validation clips

Retrieval on Core-1K



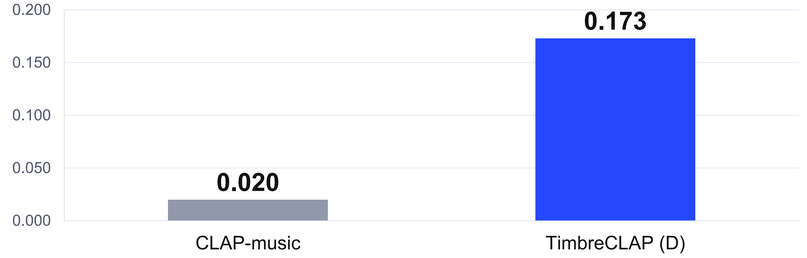
1,000 held-out candidates. Mean over T2A/A2T × L/M/S. Recall values are percentages.

L/M/S rotation raises full fine-tuning R@10 from 58.1% to 64.2%. Additional objectives bring no consistent gain.

Single-seed comparisons do not establish statistical significance.

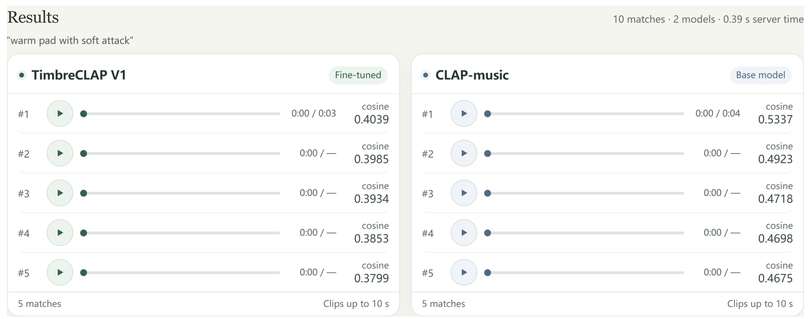
External timbre alignment

Mean descriptor Pearson r



CCMusic [5]: published ratings for 61 instruments and 16 descriptors, following [4]. Average alignment improves, but correlations remain modest and gains uneven.

Search and listen



3,786 demo sounds.

Try the demo

timbreclap-demo.twiddle-ai.com.cn

Audition both rankings. Cosine scales differ across models.

A baseline for finer sound retrieval, with potential uses in sample-library organization, recommendation and sound design.

[1] Wu et al., CLAP, ICASSP 2023. [2] Xu et al., arXiv:2602.18010, 2026. [3] OpenMOSS Team, MOSS-Music, 2026.
[4] Deng et al., arXiv:2510.14249v2. [5] Jiang et al., Applied Sciences 10(3), 789, 2020.

jiayifdzl@gmail.com