Probing Audio Understanding in Realtime Generative Music Models

Arjun Bahuguna
Universitat Pompeu Fabra
Barcelona, Spain
arjun.bahuguna01@estudiant.upf.edu

***Abstract*—Text-prompt probes of music foundation models often show high linear separability for semantic attributes, but it is unclear whether that signal reflects acoustic understanding or lexical regularities. We investigate one guiding question throughout this paper: when a MusicCoCa embedding, used in Magenta's Realtime generative music model, separates an attribute from text prompts, does that separation persist for real audio under artist-disjoint evaluation? To answer this, we compare matched text and audio probing conditions using 500 MagnaTagATune clips, six feasible tag families, and 3-fold group cross-validation by artist identity. We report both raw accuracy and margin over chance to normalize for different classes. The main result is a consistent text-audio gap for most semantic attributes, especially instrument and timbre, while loudness remains strongly separable in both modalities. This pattern suggests that text-side separability can overstate acoustic grounding if interpreted without modality-matched controls. We contribute a reproducible protocol for side-by-side text versus audio probing and a transparent analysis of what the probes do and do not justify.**

***Index Terms*—music representation learning, probing, evaluation, MagnaTagATune**

I. Introduction

Large text-audio music models are increasingly used as semantic interfaces for music generation and controllable editing. A common diagnostic is linear probing: if a lightweight classifier recovers a musical attribute from embeddings, we infer that the representation encodes that attribute. The central problem is that most reports emphasize text-side probing, where lexical tokens may make attributes look cleaner than they are acoustically.

We therefore follow a single research thread from start to finish: *when text prompts produce a separable attribute axis, does that axis remain separable for real audio under artist-disjoint control?* This is a meaningful question because without good separability in the audio domain, it is difficult to achieve robust semantic control.

The remainder of the paper is organized around this thread. Section II positions the question in prior work. Section III translates it into a reproducible probe protocol. Section IV reports what the probes recover in audio and where they diverge from text. Section V interprets what this gap implies for evaluation practice. Section VI closes with limitations and concrete next steps.

II. Related Work

A. Multimodal Embedding Models for Audio and Language

CoCa and related contrastive and generative paradigms show that language supervision can create semantically structured embedding spaces [1]. MusicCoCa extends this paradigm to music style embedding with shared text-audio representations and practical APIs for both text and waveform encoding [2]. These works motivate probing as a fast diagnostic, but they do not guarantee that text-derived axes align with acoustic structure in the same way across modalities.

B. Evaluation Practices in Generative and Representation Models

Directed readings in this project emphasized evaluation as a strong design goal, especially for generative music systems where there is a lack of downstream tasks with measurable performance to evaluate with. A recent unified framework perspective argues that multimodal evaluation should separate representational, behavioral, and confound-driven effects rather than treating scalar benchmark scores as sufficient [3]. In parallel, work on disentangled representation evaluation in music stresses that interpretability claims should be tested against controllable probes and robustness criteria [4].

C. Live and Real-Time Music Modeling Context

A practical thread in live music modeling highlights the tension between latency, controllability, and semantic reliability. In real-time settings, overconfident semantic control can produce brittle behavior when input acoustics deviate fromthe prompt, leading to prompt adherence issues [5]. This paper contributes to that discussion by quantifying a text-audio separability gap on concrete attributes.

III. Research Question and Study Design

A. Probe Question

Our probe question is: *for a fixed embedding model, which musical attributes remain linearly separable when moving from text prompts to real audio, and where does separability collapse?*

Given an attribute family a, define probe accuracy under condition $c \in \{\text{text}, \text{audio}\}$ as $\mathrm{Acc}_{a,c}$ and chance baseline as

TABLE I
AUDIO-GROUNDED PROBE RESULTS WITH ARTIST-DISJOINT 3-FOLD CV (500-CLIP RUN).

Attribute	Accuracy	Chance	Margin
Loudness proxy	0.930 ± 0.038	0.500	0.430
Structure (chorus) proxy	0.807 ± 0.111	0.500	0.307
Timbre proxy	0.718 ± 0.100	0.250	0.468
Octave proxy	0.689 ± 0.075	0.500	0.189
Tempo proxy	0.583 ± 0.064	0.333	0.250
Instrument (8-way)	0.536 ± 0.155	0.125	0.411

$\text{Chance}_{a,c} = 1/K_{a,c}$ where K is number of classes. We report margin over chance:

$$\text{Margin}_{a,c} = \text{Acc}_{a,c} - \text{Chance}_{a,c}. \quad (1)$$

This normalization helps compare tasks with different class cardinalities. We use this measure because the central question is comparative across modalities, not absolute within one task.

B. Dataset and Feasible Attributes

Audio comes from MagnaTagATune clip files and metadata. Because tag supervision is weak and incomplete, only feasible attributes were used:

- instrument (8-way proxy classes)
- tempo proxy (slow, fast, upbeat)
- timbre proxy (airy, dark, soft, hard)
- loudness proxy (loud vs quiet)
- octave proxy (low register vs not-low)
- structure chorus proxy (chorus present vs absent)

Unsupported targets were explicitly excluded: full section-level form taxonomy, chord choice, melody type, and duration as a musical-semantic attribute.

This pruning is part of the narrative logic of the study: if the question is whether text-side separability transfers to audio, then label definitions must be honest about what the dataset can actually supervise.

For each task, we train logistic regression probes with 3-fold artist-disjoint group cross-validation. Artist disjointness is enforced at split construction, not as a post-hoc check.

IV. RESULTS

Table I reports audio-side probe performance (500 clips total across tasks). All tasks are above chance; however, separability strength varies substantially.

Fig. 1 compares text and audio probe accuracies directly. Fig. 2 compares margins over chance, which better controls for class-count differences.

A. Interpretation

We interpret the results by returning to the same guiding question: where does text-side separability survive in audio, and where does it break?

1) Text generally outperforms audio for semantic category probes. The largest gap is instrument: text reaches near-perfect separability in the earlier prompt condition, while audio is substantially lower (accuracy 0.536, high fold variance).

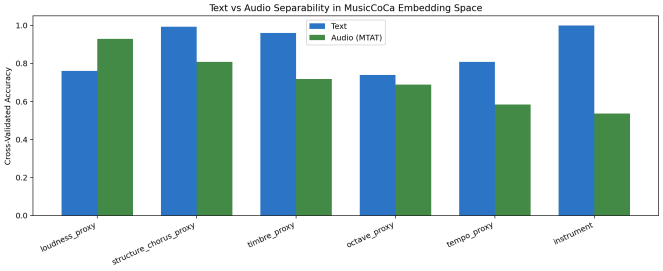

Fig. 1. Text versus audio separability by attribute. Audio uses artist-disjoint MTAT probes; text uses prompt-only probes from the same embedding model.

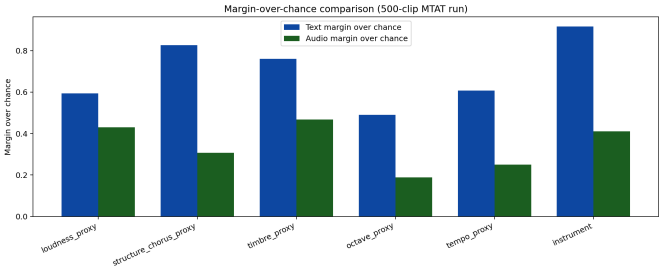

Fig. 2. Margin-over-chance comparison. Positive margin indicates performance above random baseline.

This suggests that lexical tokens for instruments are cleanly encoded, whereas acoustic instrument identity in weakly labeled clips remains harder and more confounded.

2) Loudness is robust across modalities. Loudness is strong in both settings and very strong in audio (0.930). This is plausible because loudness cues are directly encoded in waveform-level statistics and less dependent on symbolic semantics.

3) Some structure evidence may reflect proxy behavior. The chorus proxy is above chance in audio, but this should not be overinterpreted as full musical-form understanding. The task is binary and tag-based, not section-level structural parsing.

Together, these results support a conservative reading: text probe success demonstrates semantic organization in language-conditioned embedding space, but does not directly guarantee equivalent acoustic recoverability.

V. DISCUSSION

A. Why This Matters

Within this setup, the text-audio gap indicates that high text-side separability does not consistently transfer to audio-side probes. For model analysis, this supports reporting both modalities when possible. For deployment-oriented use cases, it suggests caution when using text-only probe scores to estimate behavior on real audio.

B. Limitations

This study has practical constraints that bound how far we can interpret the results.

- MTAT tags are imperfect and incomplete, so label noise is expected.
- Some attributes (tempo, timbre, structure) are represented through tag-based proxies rather than dense annotations.
- The 500-clip sample is sufficient for a comparative pilot, but not for fine-grained estimates across many subclasses.
- This draft reports quantitative visuals only and does not include listening examples.
- Residual class imbalance remains after filtering in several tasks.

VI. Conclusion and Future Directions

This paper followed one question: when text probes suggest semantic structure, does that structure remain recoverable from audio? In a 500-clip artist-disjoint setting, the answer is mixed. Text and audio agree for some attributes, but diverge substantially for others, with loudness as a robust exception and instrument/timbre showing large modality gaps.

In this dataset and probe configuration, text-only probe scores are best treated as provisional indicators rather than direct evidence of acoustic grounding. Future work can test the stability of this pattern with larger artist-disjoint runs, aligned class taxonomies across modalities, stronger annotations for harmony and form, and listening-based checks linked to each probe axis.

The main takeaway from this experiment is narrowly scoped: matched text-audio probing provides more reliable evidence than text-only or audio-only probing when the goal is to characterize audio-grounded behavior.

Acknowledgment

We thank the course instructors and directed-reading authors for framing this project around evaluation rigor, interpretability, and critical reflection rather than leaderboard performance.

References

[1] J. Yu, J. Wang, Z. Zhou, Y. Yang, Z. Chen, A. Ross, W. Wang, C. Schmid, and D. Le, "CoCa: Contrastive captioners are image-text foundation models," 2022, arXiv:2205.01917.

[2] Q. Huang, D. S. Park, T. Wang, T. Wang, K. Ma, C. Zhang, and Y. Wu, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," 2022, arXiv:2208.12415.

[3] C. Plachouras, J. Guinot, G. Fazekas, E. Quinton, E. Benetos, and J. Pauwels, "Towards a unified representation evaluation framework beyond downstream tasks," arXiv:2505.06224, 2025. doi: 10.48550/arXiv.2505.06224.

[4] L. Ibanez-Martinez, C. Nkama, A. Poltronieri, X. Serra, and M. Rocamora, "Evaluating disentangled representations for controllable music generation," arXiv:2602.10058, 2026. doi: 10.48550/arXiv.2602.10058.

[5] Lyria Team, A. Caillon, B. McWilliams, C. Tarakajian, I. Simon, I. Manco, J. Engel, N. Constant, Y. Li, et al., "Live music models," arXiv:2508.04651, 2025. doi: 10.48550/arXiv.2508.04651.