

SYMBOLIC AUXILIARY SUPERVISION FOR AUDIO MUSIC PREFERENCE MODELING

Rui Guo

ABSTRACT

We investigate whether symbolic defect supervision can improve audio music preference modeling. Three instances of a MIDI detector assign harmony and timing targets to controlled interventions, which are rendered to audio. A small shared model learns both these targets and corrections to frozen CMI-RM scores from human comparisons. Inference uses only audio features and learned score corrections. Against a capacity-matched preference-only control, auxiliary supervision yields a small improvement on reused calibration pairs. The experiment establishes a practical training bridge; independent evaluation is needed to determine its benefit.

1. INTRODUCTION

CMI-RewardBench evaluates audio preferences under text, lyric, and reference-audio conditioning [1]. Human comparisons do not explicitly constrain sensitivity to localized musical defects. Controlled MIDI interventions isolate such changes; rendering supplies targeted audio supervision, but teacher judgments may not transfer to listening preferences. We test a symbolic-to-audio training bridge against a capacity-matched preference-only control. Deployment requires only audio and its conditioning.

2. METHOD

2.1 Symbolic Teachers and Rendered Targets

The MIDI detector has three defect heads: local track-timing misalignment, relational harmonic clash, and global excessive repetition. Their scopes are individual track-bars, track pairs within a bar, and cross-bar recurrence, respectively. Only the first two supply this paper’s supervision; three random-seed replicas provide an ensemble.

Inputs are note events grouped by track and bar. POP909 [2] supplies three source parts: MELODY (main melody), BRIDGE (secondary melody), and PIANO (accompaniment). The prepared multicorpus pool retains two to four pitched tracks per piece, including variable Lakh parts [3]; model slots are anonymous and batch-padded, without fixed musical-role assignments. Pitch, pitch class,

onset, duration, and continuation embeddings are pooled per track-bar. A two-layer temporal Transformer contextualizes each track; symmetric track-pair features support the harmony head, while a global temporal/recurrence branch supports repetition.

Before audio training, the detector replicas are trained on controlled MIDI defects: harmony supervision uses POP909, while timing and repetition also use filtered Lakh. Harmony revoicing creates inter-track pitch conflicts, with variants preserving or disrupting interval flow. Timing perturbations alter onset spacing and durations while retaining pitches, order, chord attacks, and note counts; repetition copies short phrases across longer spans. These are detector-training defects, rather than human quality labels.

For the downstream audio auxiliary set, new eight-bar candidate comparisons use the harmony and timing operators: harmony draws from POP909 and timing from POP909/Lakh. Each MIDI receives mean predicted defect risk over occupied track-bars (timing) or track pairs (harmony). Targets are normalized using synthetic training statistics and averaged across seeds. We retain comparisons with unanimous seed ordering and a sufficiently large score gap.

This audio-training stage retains 1,200 MIDI content comparisons (600 per task). Two timbral views per comparison yield 2,400 rendered pair views, or 4,800 candidate clips. Settings match within each pair; candidates inherit continuous MIDI targets. Song identities are split before segment creation/rendering; auxiliary validation excludes teacher-training/test identities and uses separate timbres. Exact masks, cutoffs, and settings accompany the source.

2.2 Learning Audio Score Corrections

Frozen CMI-RM, using MuQ/MuQ-MuLan representations [4], extracts conditioned features x_i . A shared audio MLP predicts a utility correction $u_\theta(x_i)$ and auxiliary harmony/timing targets (Figure 1). The zero-initialized utility starts from the baseline.

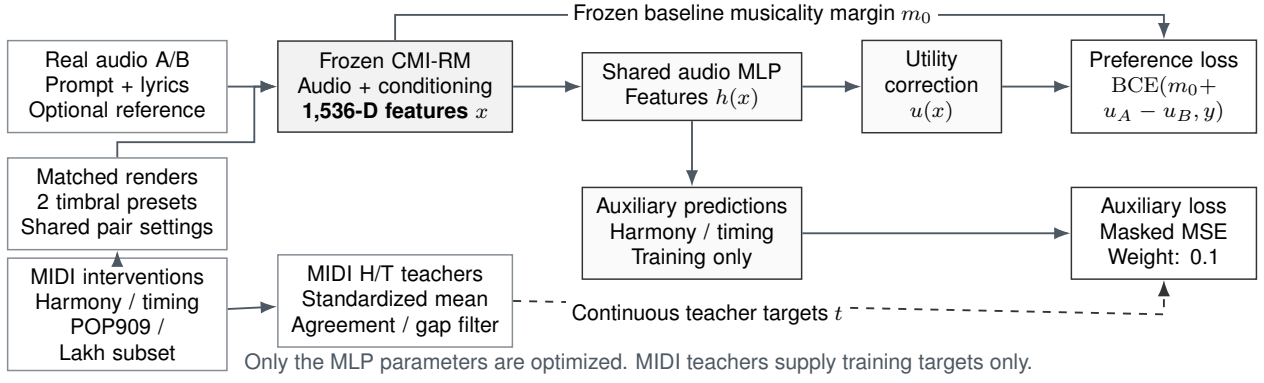
For a real pair, m_0 is the baseline score of A minus B; positive margins favor A. The utility difference adjusts this margin. With $y = 1$ for a human preference for A and $y = 0$ for B, the preference loss is

$$\mathcal{L}_{\text{pref}} = \text{BCEWithLogits}(m_0 + u_\theta(x_A) - u_\theta(x_B), y). \quad (1)$$

For rendered candidates, masked mean squared error $\mathcal{L}_{\text{aux}}$ compares audio predictions with the available MIDI-derived task targets. Variant C minimizes $\mathcal{L}_{\text{pref}} + 0.1\mathcal{L}_{\text{aux}}$;



(a) Training: one independently trained residual head per seed



(b) Inference: calibrated ensemble of three frozen residual heads

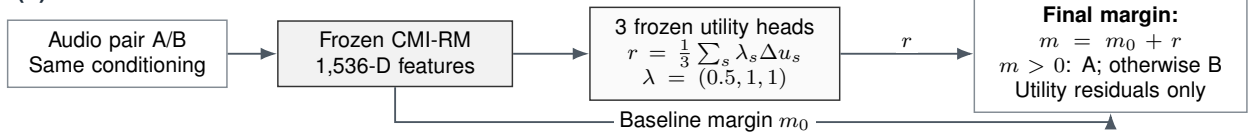


Figure 1. Symbolic supervision during training and audio-only deployment. Dashed arrows carry continuous MIDI targets. Real and rendered audio share the encoder and MLP; auxiliary predictions guide training, while only utility corrections enter the final preference margin.

control B uses preference loss alone. The shared representation lets auxiliary supervision influence utility, but auxiliary scores never directly enter the preference margin.

2.3 Audio-Only Deployment

At deployment, three trained audio models share frozen features. Calibration selects each residual scale λ_s ; their corrections are averaged:

$$m = m_0 + \frac{1}{3} \sum_{s=1}^3 \lambda_s [u_s(x_A) - u_s(x_B)]. \quad (2)$$

A positive margin selects A; otherwise B. Neither auxiliary predictions nor MIDI teachers are required.

3. EXPERIMENTS AND RESULTS

We use 2,220 real training pairs and 128 calibration pairs from eligible CMI-Pref annotations [1], grouped by shared audio identities and repeated prompts. This corpus informed prior experiments. B and C share architecture, three seeds, AdamW, and the real-update budget; C balances harmony/timing examples. Calibration accuracy selects checkpoints and residual scales, with BCE and smaller scales breaking ties. Table 1 compares the frozen baseline (A) with B and C.

C improves calibration accuracy over B by one pair (0.78 percentage points). Since the same calibration set selects checkpoints and residual scales, this difference is exploratory and requires independent confirmation. Official test accuracy has not been measured.

AI usage. An LLM assisted with writing and implementation-record analysis; the author is responsible for validation.

4. REFERENCES

- [1] Y. Ma et al., “CMI-RewardBench: Evaluating Music Reward Models with Compositional Multimodal Instruction,” arXiv:2603.00610, 2026.
- [2] Z. Wang et al., “POP909: A Pop-song Dataset for Music Arrangement Generation,” in *Proc. ISMIR*, 2020, arXiv:2008.07142.
- [3] C. Raffel, *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*, Ph.D. dissertation, 2016. [Online]. Available: <https://colinraffel.com/projects/lmd/>
- [4] H. Zhu et al., “MuQ: Self-Supervised Music Representation Learning with Mel Residual Vector Quantization,” arXiv:2501.01108, 2025.

Model	Correct / 128	Accuracy
A: frozen baseline	105	82.03%
B: preference only	107	83.59%
C: harmony/timing	108	84.38%

Table 1. Accuracy on the 128 calibration pairs.