Multi-Backbone LoRA for Audio Instrument Recognition

MIREX 2026 Evaluation and Analysis

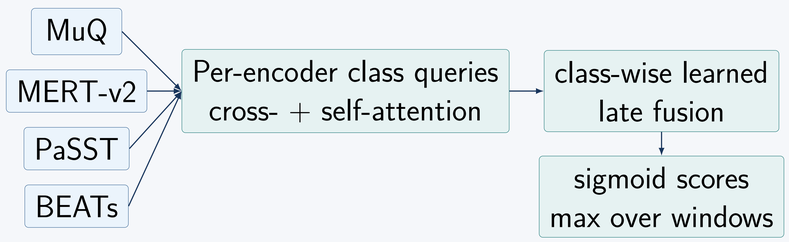
Shuangrun Shao – Central Conservatory of Music, Beijing, China

Task and motivation

MIREX Audio Instrument Recognition predicts the presence of 20 OpenMIC instrument classes in ten-second music excerpts. The organizer uses a fixed 0.5 threshold and ranks systems by clip-level macro-F1. We combine music-specialized and general-audio representations while respecting partial labels: an unobserved tag is unknown, not negative.

20 labels | **Fixed threshold: 0.5**

System overview



Each encoder has its own query head; the four sets of logits are fused separately for each instrument.

- Frozen pretrained backbones with rank-16 LoRA adapters.
- Auxiliary frame-activity and attention-MIL objectives.
- OpenMIC plus seven external datasets with dataset-balanced sampling.
- Long audio: overlapping 10-s windows and per-class maximum pooling.

Training and inference

$$z_{t,c} = \sum_b \mathrm{softmax}(a_{bc}) z_{t,c}^{(b)},$$
$$\hat{y}_c = \mathbf{1}\left[\max_t \sigma(z_{t,c}) \geq 0.5\right].$$

All class losses use observation masks. The final checkpoint uses 10 epochs, AdamW, 65,536 with-replacement draws per epoch, a 10% OpenMIC quota, and equal relative sampling weight for the seven external datasets.

Label policy: OpenMIC uses observed positives and negatives. IRMAS, FSD50K and MagnaTagATune supply positive-only supervision; Slakh, MusicNet, GOOD-SOUNDS and RWC also supply negative evidence.

Official MIREX 2026 results

+6.0 pp / +5.6 pp

mAP gain over OpenMIC-RF: public / hidden

System	Macro-F1	Micro-F1	mAP
OpenMIC public test			
OpenMIC-RF	0.707	0.787	0.794
SSR (ours)	**0.801**	**0.841**	**0.854**
MedleyDB hidden set			
OpenMIC-RF	**0.497**	**0.607**	0.632
SSR (ours)	0.488	0.451	**0.688**

The hidden set contains 316 ten-second excerpts from 115 MedleyDB recordings. Automatically derived labels average 3.43 present instruments per excerpt; only 12 classes meet the support criterion.

What the metrics reveal

Hidden-set mAP improves over the baseline, while both F1 measures decrease. At the fixed threshold of 0.5, SSR predicts 11.18 labels per clip, compared with 3.43 annotated present labels, indicating substantial overprediction. A likely contributor is that several weakly annotated datasets mask unobserved classes instead of treating them as negatives. Although this is appropriate for partial labels, it provides less negative supervision against overprediction. Combined with domain differences between the training data and MedleyDB, this may shift the decision boundary. This is a post-hoc hypothesis, not a causal finding.

Overprediction on the hidden set

11.18 vs. 3.43

Predicted labels / clip vs. annotated present labels

The baseline predicts 3.16 labels per clip. Our prediction count is about 3.26 times the annotated count. This is a cardinality diagnostic, not a precision estimate: evaluation excludes some classes and ambiguous labels.

Interpretation and limitations

- **Quantity is not coverage.** Easy negatives may not cover confusable sources in real mixtures.
- **Diversity is not invariance.** Labels can correlate with recording style and dataset source.
- **Sampling changes supervision.** OpenMIC receives 10% of draws; external datasets receive 90%. Class-wise loss weights also depend on masks.
- **Baseline is not immune.** Its hidden F1 is lower than its public F1, but its hidden threshold decisions outperform SSR's.

Future directions

1. Calibrate scores on an independent validation domain; retain the required 0.5 output threshold and never tune on the hidden set.
2. Audit per-class positive/negative supervision by source after sampling.
3. Test verified hard negatives in real mixtures; retain unknown-label masks.
4. Compare sampling and supervision under controlled training on held-out domains.

Conclusion

Strong instrument recognition performance.

Our model outperforms the baseline across all three public-test metrics and consistently achieves higher mAP on both datasets. However, hidden-set overprediction shows that translating this ranking advantage into reliable cross-domain decisions remains a challenge. Future work will focus on calibration and negative supervision.

Contact and resources

Shuangrun Shao
25tw20@mail.ccom.edu.cn

Official results: MIREX 2026 Audio Instrument Recognition