NOISY-LABEL-AWARE MULTIMODAL ENSEMBLING WITH INFERENCE-SAFE CANDIDATE RERANKING FOR HETEROGENEOUS AUDIO CLASSIFICATION

Technical Report

Peihong Zhang, Shengchen Li

School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China

ABSTRACT

We present a noisy-label-aware multimodal system for DCASE 2026 Challenge Task 1, Heterogeneous Audio Classification. The task requires second-level Broad Sound Taxonomy prediction from audio and metadata and is evaluated using macro-averaged hierarchical F-score. Our system addresses the central tension between exploiting large but noisy crowd-sourced supervision and preserving inference safety on the unlabeled evaluation set. It combines LAION-CLAP audio/text embeddings, teacher-weighted soft and hard supervision from BSD35k-CS, a fresh long-run ensemble of 31 checkpointed model families, calibrated probability fusion, and an inference-safe candidate reranker with conservative parent-aware override. All submitted systems are generated from deployable checkpoints rather than replayed development-set artifacts. On a label-blind development-as-evaluation protocol, the primary system obtains 83.5214 hF, improving over our local multimodal baseline by 4.57 absolute hF points.

***Index Terms*—** heterogeneous audio classification, multimodal learning, noisy-label learning, hierarchical F-score

1. INTRODUCTION

DCASE 2026 Task 1 studies heterogeneous audio classification under the Broad Sound Taxonomy (BST), a two-level taxonomy with 5 top-level and 23 second-level sound categories [1, 2]. The development data include BSD10k-v1.2, a curated Freesound collection, and BSD35k-CS, a substantially larger crowd-sourced collection with noisier user-provided labels [3, 4, 5, 6]. The evaluation package provides audio and metadata but no released labels, and the challenge rules prohibit manual annotation of evaluation sounds or using evaluation predictions as additional training labels [1, 7].

The task is challenging because the input is heterogeneous in both acoustic and semantic form. Sounds differ in duration, recording condition, foreground-background structure, and category granularity. Metadata can provide useful semantic cues, but it may also be noisy, incomplete, or shortcut-prone. Moreover, the official metric is macro-averaged hierarchical F-score (hF), which makes cross-parent errors more harmful than within-parent confusions under the BST hierarchy [8]. A competitive system must therefore combine multimodal representations, robust use of noisy supervision, calibrated decision fusion, and hierarchy-aware correction.

Our submission is based on the observation that large noisy supervision is useful only when it is treated as auxiliary evidence rather than as a replacement for curated BSD10k labels. We use LAION-CLAP audio/text embeddings as the common representation, train diverse classifier families with teacher-weighted supervision from BSD35k-CS and external-data variants, and combine their predictions through calibrated ensembling. To further improve low-margin decisions, we build a candidate-level reranker that operates only on inference-available features and applies conservative parent-aware overrides. This design avoids development-only routing signals and ensures that the same pipeline can be executed on the unlabeled evaluation set.

The main practical contributions of our submission are threefold. First, we construct a fresh submit-grade ensemble of 31 model families and 155 fold checkpoints, avoiding reliance on replayed out-of-fold artifacts. Second, we show that calibrated model diversity provides most of the gain over the official-style multimodal baseline, while candidate reranking adds a smaller but useful hierarchy-aware correction layer. Third, we provide four deployable systems with different risk profiles: an aggressive ranker, a balanced ranker/base system, a no-ranker ensemble, and an external-diversity low-risk system.

2. METHOD

Our system is designed around a practical constraint of the challenge: all development-time model selection must be converted into an inference-safe pipeline that can run on unlabeled evaluation audio and metadata. We therefore distinguish between development-side supervision, model selection, and deployable evaluation-time features. The final submission consists of checkpointed multimodal classifiers, calibrated ensemble fusion, and a conservative candidate-level reranker.

2.1. Multimodal CLAP representation

We use the LAION-CLAP representation distributed with the Task 1 baseline, including the `630k-audioset-fusion-best.pt` checkpoint [9, 10, 11, 12, 13]. For each sound, an audio embedding is extracted from the waveform, while a text embedding is extracted from the official metadata fields: title, tags, and description. The base classifier follows the hierarchical audio-text representation family used by the official baseline [4, 9]. Audio and text branches are fused before predicting the 23 BST second-level classes.

The CLAP representation is useful for this task because it gives a shared audio-text space for acoustically diverse recordings and heterogeneous metadata. However, a single CLAP classifier is not sufficient for robust hierarchy-aware classification. We therefore train multiple classifier families that share the same inference interface but differ in supervision, calibration, metadata handling, and external-data exposure. These include audio-only variants, metadata-dropout variants, class-word-masked text variants, prior-corrected variants, external-data variants, and hierarchy-aware variants.

All submitted models are trained under the canonical five-fold BSD10k development protocol. Each fold stores `checkpoint.pt`, `args.json`, `config_snapshot.json`, held-out logits and probabilities, calibration parameters, and an evaluation inference registry. This checkpointed design is central to the final system. Earlier development systems based only on stored out-of-fold arrays can overstate deployability, whereas the submitted system must run from audio and metadata on the unlabeled evaluation set. The final P33 pool contains 31 freshly trained submit-grade model families and 155 new fold checkpoints. Two unstable confidence-weighted variants were excluded after abnormal validation behavior.

2.2. Teacher-weighted BSD35k-CS supervision

BSD35k-CS provides substantially broader acoustic and metadata coverage than BSD10k, but its labels are crowd-sourced and noisier than the curated BSD10k labels [6]. We therefore treat BSD35k-CS as weak auxiliary evidence rather than as clean supervision. Let (x_i, y_i) denote a clean BSD10k example, let $\tilde{p}_j$ denote a teacher distribution for a BSD35k-CS example, and let $\hat{y}_j$ denote a high-confidence relabeled target. The main mixed-supervision objective is

$$\mathcal{L} = \mathcal{L}_{\text{clean}} + \lambda_s \mathcal{L}_{\text{soft}} + \lambda_h \mathcal{L}_{\text{hard}} + \lambda_{\text{hier}} \mathcal{L}_{\text{hier}}, \tag{1}$$

where $\mathcal{L}_{\text{clean}}$ is cross-entropy on BSD10k, $\mathcal{L}_{\text{soft}}$ is soft cross-entropy to teacher probabilities, $\mathcal{L}_{\text{hard}}$ is cross-entropy for high-confidence teacher relabels, and $\mathcal{L}_{\text{hier}}$ is used only by hierarchy-aware variants. In the main v2mix family, the BSD35k-CS soft and hard terms are assigned low relative weights, while the clean BSD10k loss remains the dominant term.

This design is related to knowledge distillation and noisy-label learning [14, 15, 16], but it is constrained by the challenge setting. Teacher targets are constructed only from development-side models and allowed training resources. Evaluation sounds are never manually annotated, never pseudo-labeled for training, and never used to update the model. In practice, the BSD35k-CS teacher acts as a soft regularizer: it helps allocate probability mass among plausible BST leaves without replacing the curated fold labels. Broad audio-language relabeling was studied during development, but it was not selected as a final global training mechanism because the resulting corrections were not stable enough under strict inference-safety checks.

2.3. External resources and compliance

Table 1 summarizes the development resources, external data, and pretrained models used in the system design. We separate training resources, development diagnostics, and evaluation-time inputs to make the compliance boundary explicit. The official evaluation package is used only for inference. No evaluation sound is manually annotated, and no evaluation prediction is used as a training target.

External datasets are included for coverage rather than under the assumption that more external data always improves the primary classifier. ESC-50 and UrbanSound8K provide clean environmental and urban-event examples, while FSD50K provides broader Freesound-derived sound-event coverage [18, 19, 20, 21]. Because these labels must be mapped to BST second-level categories, they introduce mapping uncertainty. Consequently, external-heavy systems are used mainly as diversity components and low-risk alternatives, not as the primary predictor.

Table 1: Development resources, external data, and pretrained models.

Resource	Use stage	Role
LAION-CLAP	train/eval	Audio and metadata embedding backbone from the official baseline checkpoint.
BSD10k-v1.2	train/dev	Curated clean supervision and five-fold development protocol.
BSD35k-CS	train	Low-weight soft and hard teacher supervision; not treated as clean labels.
ESC-50, UrbanSound8K	train	Mapped external diversity variants with small ensemble weights.
FSD50K	train	Tail and diversity variants, retained mainly as low-risk diversity sources.
Qwen2.5-Omni	dev diagnostic	Development-side diagnostic and teacher-behavior evidence; not used to annotate evaluation data or as an evaluation-time relabeling model [17].
Evaluation package	eval only	Audio and metadata used only to produce final predictions.

2.4. Calibrated long-run ensemble

The final model pool consists of 10 core v2mix families, 9 prior, confidence, and hierarchy families, and 12 external, FSD, and text-diversity families. Each family is trained as a complete five-fold submit-grade system. Given M calibrated model predictions $p_m(y|x)$, the ensemble distribution is

$$p_{\text{ens}}(y|x) = \sum_{m=1}^{M} w_m p_m(y|x), \qquad w_m \geq 0, \quad \sum_{m=1}^{M} w_m = 1. \tag{2}$$

The ensemble weights are selected on development folds with family caps to reduce seed-level overfitting. This selection is still a development-set procedure, so it is converted into a frozen evaluation policy before submission. The best individual fresh model reaches 82.2893 hF, while the best fresh ensemble reaches 83.4690 hF. This shows that calibrated model diversity is the main source of improvement over the baseline and over any single classifier.

The ensemble also reduces the risk of overusing metadata shortcuts. Some variants remove or mask metadata, while others emphasize audio-only evidence or external-data diversity. The final ensemble is therefore not a simple seed average; it combines models that make different errors under the BST hierarchy.

2.5. Inference-safe candidate reranking

Candidate reranking is applied only after the ensemble has produced a calibrated 23-class distribution. Candidate sets are formed from top-ranked ensemble predictions and selected diversity alternatives. The ranker operates on candidate-level features, including model probabilities, ranks, family aggregates, entropy and margin, parent information, and metadata-semantic consistency. Teacher behavior from earlier development systems is used only as a training target. It is not used as an evaluation-time feature and never encodes whether an example was correctly classified by an earlier model.

The conservative override rule is

$$\hat{y} = \begin{cases} y_r, & s_r - s_b > \tau \text{ and } \text{parentSafe}(y_r, y_b), \\ y_b, & \text{otherwise}, \end{cases} \tag{3}$$

where y_b is the base ensemble prediction, y_r is the ranker-selected candidate, and $s_r - s_b$ is the ranker margin over the base choice. The parent-safe condition prevents aggressive corrections that cross top-level BST parents unless the learned margin and confidence thresholds support the change. This is important because an unconstrained correction can improve flat accuracy while increasing the cross-parent errors that are strongly penalized by hierarchical F-score.

The best nested ranker obtains 83.5543 hF. The deployable development-as-evaluation System 1 obtains 83.5214 hF. We report both values because nested model selection and label-blind inference answer different questions. The former estimates candidate-ranking capacity under development folds, while the latter evaluates the full submission pipeline in the same way that it is run on the unlabeled evaluation set.

Forbidden features include sample identifiers, fold identifiers, full-development confusion statistics, and any flag derived from whether a model was wrong on the same held-out example. This restriction is central to the system. Development-side teachers and diagnostics can shape the training objective, but the deployed ranker only receives features that are available for an unlabeled evaluation item.

3. EXPERIMENTS

3.1. Development protocol

All development scores are computed on the BSD10k five-fold protocol using the official hierarchical metric implementation. In addition to ordinary nested validation, we use a development-as-evaluation check: development audio and metadata are passed through the frozen inference chain used for the official evaluation package, and labels are read only after prediction to compute metrics. This is not an independent test set, but it verifies that the final system can run without label access, development-only error flags, or replayed out-of-fold artifacts.

The official evaluation archive was re-downloaded from the Task 1 page and verified against the official Zenodo record [22]; the full checksum is retained in the reproducibility artifacts. The released package contains 3246 anonymous identifiers and 3246 audio files; the organizers state that only a subset is used for final scoring [1]. Our verified index has no development overlap and one zero-byte audio file. The zero-byte item is handled by the same fallback policy used in evaluation inference, and all four submitted CSV files contain predictions for all 3246 identifiers.

Nested validation compares rankers and thresholds without tuning on their held-out fold. Development-as-evaluation instead tests the deployable chain: feature extraction, checkpoint inference, probability ensembling, candidate construction, ranker inference, parent-safe override, and CSV writing are all executed without labels. The score still reflects development-set model selection, but it rules out a common challenge failure mode: a high-scoring artifact that cannot be regenerated for hidden evaluation inputs.

3.2. Submitted systems

Table 2 summarizes the four submitted systems. They are intentionally related but not identical: System 1 is the primary ranker system, System 2 lowers correction pressure, System 3 removes the ranker, and System 4 keeps a lower-risk external-diversity path. This follows the challenge allowance of multiple systems and provides robustness against hidden-domain mismatch.

3.3. Main results and ablations

Table 4 shows the main development progression. The local baselines reproduce the expected trend of the official baseline family: metadata substantially improves hF over the audio-only system. The best individual fresh model reaches 82.2893 hF, but the top-12 calibrated ensemble reaches 83.4690 hF. This indicates that the main gain comes from calibrated diversity across supervision weights, priors, external-data exposure, and metadata handling, rather than from a single dominant classifier.

Compared with the local multimodal baseline, System 1 improves hF by 4.5714 absolute points, or about 5.79% relative. The gain from the top-12 ensemble to the deployable System 1 is smaller, about 0.0525 hF, but this matches the ranker's intended role: it corrects selected low-margin decisions when candidate evidence and parent-safe thresholds support the change.

The hierarchy-aware metrics in Table 3 support the same interpretation. System 1 has the lowest cross-parent error, but the margin is modest, suggesting that parent-safe reranking acts mainly as a risk-control layer rather than a replacement for calibrated probability fusion.

Metadata handling is another source of diversity. Titles and tags often help, but they can also encode recording context or ambiguous keywords. The pool therefore includes audio-text, audio-only, metadata-dropout, and class-word-masked variants, balancing useful semantic cues against shortcut risk.

External data is useful mainly as a diversity source. ESC-50 and UrbanSound8K are clean but narrow, while FSD50K is broader and noisier; mapping all of them to 23 BST leaves introduces uncertainty. We therefore use external and FSD variants with limited weights, especially in System 4.

The component-level findings are consistent with the design. Core v2mix models and prior or hierarchy variants dominate the strongest individual families; the large ensemble gain shows that the pool captures complementary errors. Candidate reranking gives a smaller gain, mainly when the correct leaf remains in the candidate set. Broad audio-language relabeling, direct BSD35k ranker pretraining, and unrestricted correction rules were not selected because they did not improve deployable development-as-evaluation hF under inference-safety constraints.

3.4. Submission validation

The final validation checks both inference safety and file-format requirements. Each system produces one CSV with identifier, predicted BST second-level class, and prediction score. Every anonymous evaluation identifier is present exactly once; all labels are among the 23 legal BST classes; all scores lie in $[0, 1]$; and the zero-byte evaluation file receives a fallback prediction. The metadata YAML files list the resources and pretrained models used by each system.

The validation is not only a formatting step. It also checks that final predictions are generated by the same label-free chain used in development-as-evaluation: no evaluation labels are read, no manual annotation is used, and no evaluation prediction updates the model.

4. DISCUSSION AND CONCLUSION

The results show that robust probability estimation is more important than an increasingly complex correction stack. LAION-CLAP

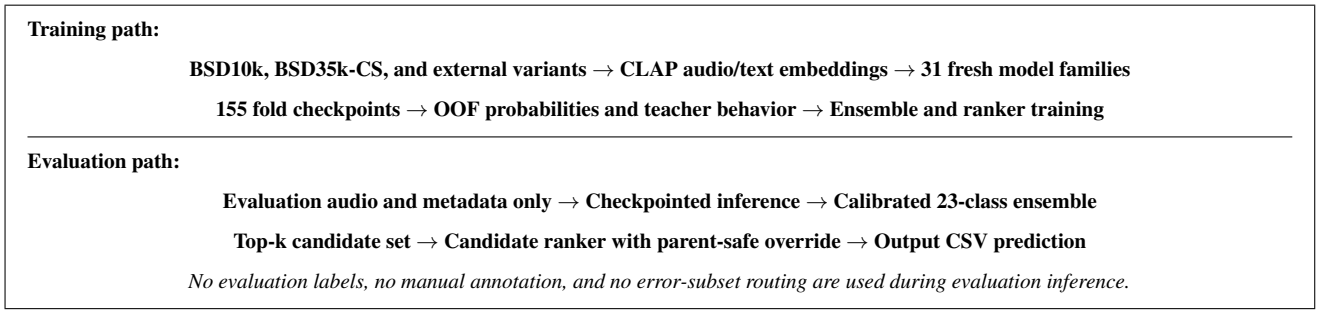

Figure 1: Training and evaluation pipeline. Development labels are used only to train checkpointed classifiers, ensembles, and rankers. Evaluation examples provide only audio and metadata, from which the frozen pipeline produces a BST second-level prediction and confidence score.

Table 2: Submitted systems. Scores are development-as-evaluation hF.

System	Main components	Ranker	External variants	hF	Role
1	Top-12 weighted fresh ensemble with candidate ranker and parent-safe override	yes	limited	83.5214	primary system
2	Balanced ranker and ensemble with lower correction pressure	limited	limited	83.4090	balanced alternative
3	Top-5 weighted fresh ensemble without candidate reranking	no	limited	83.2803	non-ranker diversity
4	Low-risk ensemble with light external, FSD, and text-diversity components	no	light	83.1567	diversity system

Table 3: Development-as-evaluation metrics for the four submitted systems.

System	hF	Flat	Top-level	Cross-parent
System 1	83.5214	84.6112	91.6941	8.3059
System 2	83.4090	84.4195	91.6028	8.3972
System 3	83.2803	84.3009	91.5207	8.4793
System 4	83.1567	84.2552	91.4477	8.5523

Table 4: Development progression.

Configuration	Protocol	hF	Comment
Audio-only baseline	5-fold	75.33	local run
Multimodal baseline	5-fold	78.95	local run
Best individual fresh model	5-fold	82.2893	single family
Top-12 fresh ensemble	5-fold	83.4690	calibrated fusion
Best nested ranker	nested CV	83.5543	ranker estimate
System 1	dev-as-eval	83.5214	deployable path

provides a strong multimodal foundation, but a single compact classifier is not sufficient for the heterogeneous BST taxonomy. The gain from the best individual model to the top-12 ensemble indicates that diversity across noisy-label weights, prior correction, hierarchy-aware variants, metadata handling, and external-data exposure captures complementary errors. Candidate reranking adds a modest correction layer on top of this estimate.

Several negative findings also shaped the final system. Broad audio-language relabeling was not stable enough to become global supervision. True segment-level CLAP evidence and pairwise Qwen-style diagnostics helped analyze hard cases, but did not outperform the final fresh ensemble and ranker when converted into a deployable system. BSD35k-CS was more useful for regularizing base classifiers than for directly pretraining the candidate ranker. Thus, teacher models can shape training targets or diagnostics, but final evaluation features must be available for an unlabeled example.

The main limitations are development-set selection bias, computational cost, and metadata noise. Development-as-evaluation verifies that labels are not read during inference, but the same development data still guide architecture, ensemble, and threshold selection. Hidden evaluation labels are unavailable, so domain shift cannot be measured before submission. The four submitted systems mitigate this risk by providing aggressive, balanced, no-ranker, and external-diversity alternatives.

The final limitation is reproducibility burden. The model pool is larger than a typical baseline submission and requires many fold checkpoints, registry files, and inference configurations. This overhead is intentional: earlier high-scoring prototypes based only on local out-of-fold artifacts were not sufficient for final submission. The main lesson is that development hF and deployability must be optimized together; for this challenge, a fully checkpointed system with conservative correction is preferable to a higher-scoring artifact that cannot be executed on the official evaluation package.

5. REFERENCES

[1] DCASE Community, "DCASE 2026 Challenge Task 1: Heterogeneous Audio Classification," https://dcase.community/challenge2026/task-heterogeneous-audio-classification, 2026, accessed: 2026-06-10.

[2] P. Anastasopoulou, X. Serra, and F. Font, "A general-purpose sound taxonomy for the classification of heterogeneous sound collections," In press. [Online]. Available: https://www.researchsquare.com/article/rs-7206795/v1

[3] P. Anastasopoulou, J. Torrey, X. Serra, and F. Font, "Heterogeneous sound classification with the Broad Sound Taxonomy and Dataset," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2024.

[4] P. Anastasopoulou, F. A. Dal Rí, X. Serra, and F. Font, "Hierarchical and multimodal learning for heterogeneous sound classification," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2025.

[5] P. Anastasopoulou, F. Font, D. Bogdanov, and L. Wyse, "BSD10k: Broad Sound Dataset 10k, version 1.2," Zenodo. doi:10.5281/zenodo.17233904, 2026. [Online]. Available: https://doi.org/10.5281/zenodo.17233904

[6] P. Anastasopoulou and F. Font Corbera, "BSD35k-CS (Broad Sound Dataset 35k - Crowd Sourced)," March 2026. [Online]. Available: https://doi.org/10.5281/zenodo.19187100

[7] DCASE Community, "DCASE 2026 Challenge Submission Instructions," https://dcase.community/challenge2026/submission, 2026, accessed: 2026-06-10.

[8] S. Kiritchenko, S. Matwin, and A. F. Famili, "Functional annotation of genes using hierarchical text categorization," in *Proc. of the ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, 2005.

[9] Music Technology Group, "DCASE 2026 Task 1 Baseline System," https://github.com/MTG/dcase2026_task1_baseline, 2026, accessed: 2026-06-10.

[10] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[11] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 8748–8763.

[13] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 776–780.

[14] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[15] J. Li, R. Socher, and S. C. H. Hoi, "DivideMix: Learning with noisy labels as semi-supervised learning," in *International Conference on Learning Representations*, 2020.

[16] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in Neural Information Processing Systems*, 2018.

[17] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, B. Zhang, X. Wang, Y. Chu, and J. Lin, "Qwen2.5-Omni technical report," *arXiv preprint arXiv:2503.20215*, 2025. [Online]. Available: https://arxiv.org/abs/2503.20215

[18] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proceedings of the ACM International Conference on Multimedia*, 2015, pp. 1015–1018.

[19] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the ACM International Conference on Multimedia*, 2014, pp. 1041–1044.

[20] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[21] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: A platform for the creation of open audio datasets," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2017, pp. 486–493.

[22] DCASE Task 1 Organizers, "DCASE 2026 Task 1 Evaluation Set," Zenodo. doi:10.5281/zenodo.20442928, 2026. [Online]. Available: https://doi.org/10.5281/zenodo.20442928