HETEROGENEOUS AUDIO CLASSIFICATION WITH FROZEN AUDIO–LANGUAGE EMBEDDINGS

Technical Report

Pao Lin

Student at Johannes Kepler Universität Linz, Linz, Austria
k12440897@students.jku.at

ABSTRACT

We describe our submission to DCASE 2026 Task 1, Heterogeneous Audio Classification, where each sound is assigned to one of 23 second-level Broad Sound Taxonomy classes and systems are ranked by macro hierarchical-F1 (h_F). Like the official baseline, we keep the CLAP audio–text encoder frozen and train small heads on its embeddings. Under matched 5-fold cross-validation on BSD10k-v1.2, our best single model, a gated multimodal head, reaches $78.8 \pm 1.1\%$ h_F, matching the published multimodal baseline ($78.8 \pm 0.8\%$; its folds differ from ours, as we quote its README figures[1]). Several of our additions add little once evaluated carefully. Agreement pseudo-labels mined from the noisy BSD35k-CS set give no gain that clears seed noise, so we treat them as a neutral data addition. A four-member ensemble, with weights tuned on our validation split, reaches $79.4 \pm 0.7\%$ h_F over three seeds. Of everything we tried, only adding the free-text description to the metadata helped cleanly ($+2.2$ h_F). Finally, the residual error appears strongly tied to label ambiguity: the classes our system confuses most are the ones the human annotators were least confident labeling, and most of the lost credit lies in cross-family confusions (such as crowd speech vs. urban soundscape) concentrated on those low-confidence labels. We view this label-ambiguity analysis as the main finding of our submission. We submit four systems with different complexity and robustness profiles.

***Index Terms*—** heterogeneous audio classification, Broad Sound Taxonomy, CLAP, multimodal fusion, pseudo-labeling

1. INTRODUCTION

Heterogeneous Audio Classification asks one model to place any sound (music, an instrument sample, speech, a sound effect, or a soundscape) into one of 23 second-level classes of the Broad Sound Taxonomy (BST) [1], which group into five top-level families [2, 3]. Systems are ranked by a macro-averaged hierarchical-F1 (h_F, $\lambda_{\text{prf}} = 0.75$, top-1) [4]. A prediction earns full credit for the exact class, partial credit when only the family is right, and nothing when both levels are wrong; the per-class scores are then averaged equally. Two properties follow. The hierarchy makes a near miss inside the right family cheap. Macro-averaging counts every class equally regardless of size, so a few weak classes can hold the whole score down.

The organizers highlight four questions [3]: how models cope with such heterogeneous categories, how the two-level taxonomy can be used to avoid errors that miss both levels, how noisy data can be turned into a useful training signal, and how audio-only and multimodal models compare. The official baseline trains a multilayer perceptron (MLP) on *frozen* CLAP audio and text embeddings [5, 6] and scores 78.76% h_F. Our aim was to improve on this baseline. Under matched cross-validation our single models only tie it, and the one clean gain we find, adding the free-text description, is likely to shrink on the held-out set. The more useful result, then, is less the score than an account of why it is hard to move. Our gated multimodal head performs on par with simple concatenation and matches the baseline; encoder fine-tuning does not help at this data scale and under the expected distribution shift; explicit hierarchy-aware mechanisms add almost nothing once CLAP embeddings are in use; and among noisy-data strategies, agreement-filtered pseudo-labels give no gain that clears seed noise.

[1]The official BSD10k baseline's published GitHub README numbers; we run our own five folds, so the protocol matches but the train/validation splits are not identical.

2. RELATED WORK

The BST and BSD10k were introduced to study classification across acoustically disparate but semantically related sounds [2, 1], where CLAP audio embeddings were found to outperform purely audio ones [7]. Hierarchical classification assigns partial credit through ancestor-aware metrics [4] and hierarchy-aware losses [8]. CLAP [6, 5] learns a shared audio–text space; PaSST [9] and RoBERTa [10] are strong single-modality encoders. One result shapes our design: full fine-tuning can distort pretrained features and underperform a linear probe under distribution shift [11], so we keep the encoder frozen. For label noise and imbalance we draw on confident learning [12], logit adjustment [13], and agreement-based pseudo-labeling [14].

3. SYSTEM DESCRIPTION

Figure 1 gives an overview; we describe each component below.

3.1. Frozen embeddings and classification heads

All audio–language members are MLP heads on *frozen* 512-d LAION-CLAP (`630k-audioset-fusion`) embeddings [5]; we verified that embeddings re-extracted with the public model match the challenge-provided ones (audio cosine ≈ 1.0). Each head is $\text{LN} \to \text{Linear}(d, 512) \to \text{GELU} \to \text{Dropout}(0.3) \to \text{Linear}(512, 23)$, trained with label smoothing 0.1 [15], AdamW at 10^{-3}, cosine schedule, best-epoch selection on validation h_F. We train audio-only, text-only (title, tags, description), and plain multimodal (concatenated ℓ_2-normalized audio+text) heads, fine-tune a

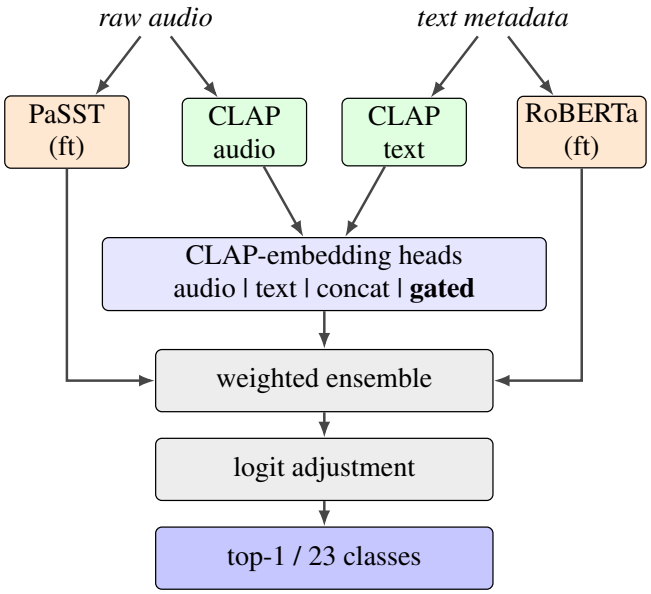

Figure 1: System overview: frozen CLAP heads (audio, text, concatenation, gated fusion) plus fine-tuned RoBERTa and PaSST, combined by weighted averaging with optional logit adjustment.

RoBERTa text classifier [10] on the metadata, and keep a PaSST [9] audio model as a complementary voice.

3.2. Gated multimodal fusion

Concatenation forces every clip to combine audio and text in the same fixed way. Reliability varies from clip to clip, so we instead learn a per-dimension gate, a gated multimodal unit [16]. From layer-normalized audio/text embeddings $\hat{a}, \hat{t}$,

$$g = \sigma\big(\mathrm{MLP}([\hat{a}; \hat{t}])\big), \quad z = g \odot W_a \hat{a} + (1 - g) \odot W_t \hat{t}, \quad (1)$$

and z feeds the shared head. Concatenation is just the special case of a constant gate that ignores content. Conditioning g on both embeddings lets the model lean on whichever modality is informative: it can down-weight audio when the sound is generic but the tags are specific, or down-weight text when the metadata is sparse or off-topic. A fixed head cannot make that per-clip judgment. This is our best single model (Sec. 5), though the margin over concatenation is small (78.81 vs. 78.50% h_F), which fits the picture of both modalities already sharing a CLAP space where even a linear combination is hard to beat.

3.3. Agreement-cleaned pseudo-labels

To exploit the large but noisy BSD35k-CS corpus without its label noise, we take, for each clip, the majority vote of three independent models (multimodal CLAP, RoBERTa, PaSST) and keep only clips where at least two agree, giving 27,551 extra training examples. This agreement filter is a practical instance of confident-learning-style noise removal [12] combined with self-training [14].

3.4. Ensemble and class rebalancing

Members are combined by weighted probability averaging, weights tuned on validation. Since the scored set is class-balanced while our training prior π is skewed ($\approx 7\times$), we apply logit adjustment [13], $s_c = \log p_c - \tau \log \pi_c$ ($\tau = 0.5$, untuned). Its effect depends on the test prior: it costs 1.5 h_F on the imbalanced validation split but ≈ 0 on a class-balanced subset, so we submit both the adjusted and unadjusted ensembles.

4. EXPERIMENTAL SETUP

We train on BSD10k-v1.2 [2, 7] (8,551 train / 2,405 validation; 23 classes; class prior skewed $\approx 7\times$, 124–964 clips/class). Our train/validation split is our own and is *uploader-disjoint* (no uploader appears in both). We report macro h_F, the official ranking metric, computed with the organizers' own evaluation code (we verified a six-decimal match), and add hierarchical accuracy h_A as a secondary diagnostic, as the baseline does.

We use BSD35k-CS [17] only for training, and only indirectly: we ignore its noisy crowd labels and form hard pseudo-labels by 2-of-3 model agreement (Sec. 3.3), keeping 27,551 clips.

Final scoring, and the official challenge ranking, uses ≈2,000 manually-labeled clips that are generally balanced across the 23 classes (the released inference set is larger). They were uploaded after 2025-04-01 and do not overlap the development data, so they differ from BSD10k in both label prior (balanced vs. our $\approx 7\times$ skew) and era; how much performance drops as a result is uncertain. We use no evaluation labels or sounds for training at any point.

5. RESULTS

We report h_F (the ranking metric) and h_A as percentages. We compare single models against the published baseline under matched 5-fold cross-validation on BSD10k (Table 1): we run our own five folds and quote the baseline's README figures, so the protocol matches but the folds differ. We then estimate the four submitted systems on a single fixed validation split, averaged over three seeds (Table 2). Because those systems' ensemble weights are tuned on that same split, their numbers are optimistic; the evaluation set is unlabeled, so there is no held-out labeled test set on which to tune and report separately.

Single models tie the baseline. In Table 1 our gated multimodal head reaches 78.81% h_F against the multimodal baseline's 78.76%, and on par with plain concatenation (78.50%); the audio-only head sits about three points lower (75.81%), a gap that clears the error bars. Adding text is thus the largest single lever, and the gate is on par with concatenation rather than better, consistent with both modalities already living in a shared CLAP space.

Pseudo-labels are a neutral addition. Agreement-filtered pseudo-labels mined from BSD35k-CS give no gain that clears seed noise, so we use them as a data addition in systems 1–3 but do not claim they improve on the baseline.

Submitted-system performance. Table 2 reports the four submitted systems on the fixed split, averaged over three seeds. The four-member ensemble (system 1) averages $79.4 \pm 0.7\%$ h_F; a single run once peaked at 81.5, confirming that one number was optimistic. It barely improves on the single gated model (system 3, 79.0), so in-domain ensembling adds little, and the clean concat probe (system 4) is the lowest-variance of the four ($78.1 \pm 0.1\%$).

Other design choices. Fine-tuning the encoder, hierarchy-aware losses, confidence and uploader filtering, and external ESC-50 data gave no clean gain; only adding the free-text description helped ($+2.2$ h_F).

Table 1: **5-fold cross-validation on BSD10k** (h_A and h_F, %; mean±std over five folds). Our rows use 5-fold stratified CV; the baseline rows are the official BSD10k baseline's README figures [2, 7], i.e. the same protocol on different folds.

System (5-fold CV)	h_A	h_F
Baseline, audio [7]	77.36±0.71	76.11±0.45
Baseline, multimodal [7]	79.71±0.82	78.76±0.79
Ours, audio (CLAP)	76.80±0.55	75.81±0.60
Ours, multimodal (concat)	79.46±0.95	78.50±1.03
Ours, multimodal (gated)	**79.56±0.96**	**78.81±1.08**

Table 2: **Submitted systems, fixed split, mean±std over 3 seeds** (not cross-validated, not directly comparable to Table 1; weights tuned on the same split, so optimistic). A single best run of system 1 reached 81.5 h_F, but averaged over seeds it is 79.4.

System (fixed split, 3 seeds)	h_A	h_F
1: gated+pseudo + ensemble	81.64±0.65	79.43±0.68
2: + logit adjustment (τ=0.5)	81.62±0.77	77.95±0.96
3: gated+pseudo single	81.20±0.20	79.02±0.18
4: clean concat probe (no pseudo)	79.72±0.08	78.08±0.12

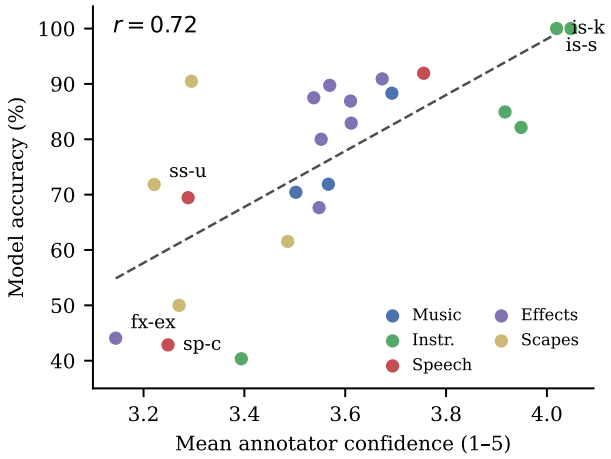

Figure 2: Per-class model accuracy vs. expert annotation confidence (r=0.72); each point is one of the 23 classes (validation split).

The ceiling appears to be a labeling limit, not a modeling one. Our flat results might suggest a stronger model would break through. But across the 23 classes, per-class accuracy tracks the dataset's expert annotation confidence (Fig. 2, $r = 0.72$): the model is near-perfect on the classes annotators labeled confidently and fails on the ones they were unsure of. The clips it gets wrong carry systematically lower annotation confidence than the ones it gets right (3.26 vs. 3.84 mean on the dataset's 1–5 rating), and this holds even for the costly cross-family confusions, the large majority of which sit below the median confidence. The metric concentrates the cost on exactly these cases: under h_F a cross-family error earns zero credit while a within-family one keeps the 0.375 floor, and 68% of the lost credit is cross-family, concentrated on the lowest-confidence labels—above all conversation/crowd speech (sp-c) confused with urban soundscape (ss-u), two of the three classes annotators were least confident labeling. The residual gap therefore appears strongly tied to label ambiguity on a few confusable pairs rather than to model capacity. (Confidence here is the dataset's per-clip annotation rating, a proxy for ambiguity, not a measured inter-annotator rate.)

6. SUBMITTED SYSTEMS

We submit four systems (Table 3). The pseudo-label generators (multimodal CLAP, RoBERTa, PaSST) were trained on clean BSD10k-train, and the submitted RoBERTa and PaSST members were then retrained on BSD10k-train plus the retained pseudo-labeled clips. Systems 1–3 use these pseudo-label-trained members with probability-average weights tuned on the fixed validation split, so system 3 is the pseudo-label-trained gated head; system 4 is a single frozen concatenation probe trained on 10k-train only, with no pseudo-labels and no validation-tuned weights. Each system trains three seeds, and the submitted `output.csv` averages the per-seed class probabilities, taking each clip's argmax as the prediction and the maximum probability as `prediction_score`. System 1 has the highest mean, but its validation-tuned weights are the most fragile under the eval shift, while systems 2–4 are progressively simpler, lower-variance alternatives—which is why we submit all four. All external resources are pretrained without Freesound data after 2025-04-01, as the task rule requires: CLAP (LAION-Audio-630K) [5], PaSST (AudioSet) [9], and `roberta-base` [10]; ESC-50 is used only in one ablation, not in any submitted system. Each system's `meta.yaml` lists its resources, and we used no evaluation labels or sounds at any point.

Table 3: Submitted systems. Members: F=gated fusion, A=CLAP audio, R=RoBERTa, P=PaSST, M=plain multimodal (concat) probe. Weights in member order. LA=logit adjustment; PL=uses 35k pseudo-labels.

#	Members / weights	PL	LA	Intended profile
1	F,A,R,P = 0.55,0.25,1.0,0.5	yes	no	primary ensemble
2	F,A,R,P = 0.55,0.25,1.0,0.5	yes	τ=0.5	balanced variant
3	F only (gated fusion)	yes	no	single PL model
4	M only (10k probe)	no	no	clean probe (low-var)

7. CONCLUSION

We separate the reliable matched-CV findings from the exploratory fixed-split ones. Under matched CV, multimodal CLAP models beat audio-only heads and our gated fusion performs on par with concatenation and ties the official multimodal baseline (78.8% h_F). Agreement-filtered pseudo-labels gave no gain that clears seed noise, and fixed-split ensembling adds little once seed-averaged (79.4 ± 0.7% h_F, barely above the single gated model). Of the data strategies we tried, only adding the free-text description was a clear win. The main result is structural: careful protocol control removes one apparent win and several additional modifications gave no clear gain under stricter evaluation, and the residual error appears strongly tied to label ambiguity—our worst classes are the ones annotators were least sure of, and most of the lost credit lies in

cross-family confusions concentrated on those low-confidence labels. For a first-year task, we consider this the more useful finding. Within the frozen-embedding regime, further refinements may still recover small gains, but larger improvements may require cleaner labels for the most confusable classes.

8. REFERENCES

[1] P. Anastasopoulou, X. Serra, and F. Font, "A general-purpose sound taxonomy for the classification of heterogeneous sound collections," *Research Square (in press)*, 2025, https://www.researchsquare.com/article/rs-7206795/v1.

[2] 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.

[3] DCASE Community, "DCASE 2026 Challenge Task 1: Heterogeneous Audio Classification," https://dcase.community/challenge2026/, 2026.

[4] S. Kiritchenko, S. Matwin, and A. F. Famili, "Functional annotation of genes using hierarchical text categorization," in *Proc. ACL Workshop on Linking Biological Literature, Ontologies and Databases (BioLINK)*, 2005.

[5] 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 *Proc. IEEE ICASSP*, 2023.

[6] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proc. IEEE ICASSP*, 2023.

[7] 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.

[8] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[9] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Proc. INTERSPEECH*, 2022.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[11] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," in *Proc. ICLR*, 2022.

[12] C. G. Northcutt, L. Jiang, and I. L. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *Journal of Artificial Intelligence Research*, vol. 70, pp. 1373–1411, 2021.

[13] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," in *Proc. ICLR*, 2021.

[14] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves ImageNet classification," in *Proc. IEEE CVPR*, 2020.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE CVPR*, 2016.

[16] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," in *Proc. ICLR Workshop Track*, 2017.

[17] P. Anastasopoulou and F. Font Corbera, "BSD35k-CS (broad sound dataset 35k – crowd sourced)," Zenodo, 2026.