# TAGGING-DRIVEN INFERENCE REFINEMENT AND DUAL-SEPARATOR SELECTION FOR SPATIAL SEMANTIC SEGMENTATION OF SOUND SCENES

Technical Report

*Seunggyu Jeong*[1,2]*, Seong-Eun Kim*[1,2]

[1] Medisensing, Seoul, Korea
[2] Seoul National University of Science and Technology, Seoul, Korea
wa3229433@gmail.com, sekim@seoultech.ac.kr

## ABSTRACT

We describe our submission to DCASE 2026 Challenge Task 4, spatial semantic segmentation of sound scenes. A system has to detect which sound classes are active in a four-channel mixture and to separate each detected source, and is scored by class-aware permutation-invariant signal-to-distortion ratio improvement (CA-SDRi). We build on the official baseline, which couples two Masked Modeling Duo (M2D) audio taggers with a FiLM-conditioned ResUNet label-queried separator. An oracle study, in which ground-truth labels are fed to the unchanged baseline separator, reaches 9.52 dB, so for the given separator the label decision is the limiting factor up to about 9.5 dB. Our system follows this finding and concentrates on the labelling and decision stages, almost all of it at inference time. We fuse the single- and four-channel taggers, fine-tune the four-channel tagger with a curriculum that oversamples silent-target clips, re-tag the separated stems to verify and clean the queries, gate residual false positives using the reward structure of CA-SDRi, and select stems from two separators by re-tagging agreement. These steps raise development-set CA-SDRi from 8.49 dB to 9.06 dB and mixture tagging accuracy from 60.7% to 63.8%. We submit three systems that add these components in turn.

***Index Terms*—** sound separation, audio tagging, spatial audio, label-queried separation, source separation, DCASE

## 1. INTRODUCTION

DCASE 2026 Task 4 asks for both detection and separation of sound events in a four-channel spatial mixture. Each output stem carries a class label, and the score is the class-aware permutation-invariant SDR improvement (CA-SDRi) [1, 2], which matches estimated and reference sources of the same class, takes the permutation that maximises SDR improvement, and averages over sources. Two properties of the metric shape system design. A clip whose reference contains no target source contributes nothing to the mean when the system also predicts silence, but any false positive on such a clip scores zero and pulls the average down. New for 2026, a class may appear more than once in a mixture [1], so the label decision is a multiset rather than a set, and source indices are needed to keep stem filenames unique.

The official baseline [2] reaches 8.49 dB CA-SDRi on the development test set. To find where the error lies, we ran an oracle study: we fed the ground-truth labels into the baseline separation pipeline while leaving the separator weights unchanged, and obtained 9.52 dB. The gap from 8.49 to 9.52 dB is therefore recoverable by better labelling alone, and exceeding roughly 9.5 dB would require a stronger separator. We confirmed the second half of this statement separately: fine-tuning the separator and adding a trainable second stage did not move the full-pipeline score (Section 4). We therefore direct our effort at the tagging and decision stages, which can be improved quickly and mostly at inference time.

Section 2 gives the system overview. Section 3 describes the taggers and their fine-tuning, Section 4 the separation refinements, and Section 5 the decision thresholds. Section 6 lists the three submitted systems and Section 7 reports development-set results.

## 2. SYSTEM OVERVIEW

All models are trained on the DCASE 2026 Task 4 dataset [1], whose four-channel mixtures are synthesised on the fly from foreground events and interference drawn from FSD50K [7], EARS [8], ESC-50 [10], and the DISCO noise set [11], together with the Semantic Hearing interference set [12], and spatialised with recorded room impulse responses [9]. The pipeline keeps the two baseline taggers, a single-channel (1ch) and a four-channel (4ch) M2D ViT-base model [3, 4], and the baseline FiLM-conditioned ResUNet separator [5, 6]. Inference proceeds in four stages. The two taggers are run on the mixture and their per-track label multisets are fused into a query list. The separator produces one stem per query. Each stem is re-tagged by the 1ch tagger to verify the query, after which unsure queries are dropped and the mixture is separated again. Finally, energy and probability gates relabel weak stems to silence, and where a second separator is available the better of the two stems is kept per slot. All thresholds in these stages are tuned offline against the exact CA-SDRi objective on cached inference outputs, so the search does not require re-running the networks.

## 3. AUDIO TAGGING

**Dual-tagger fusion.** The 1ch and 4ch taggers make different errors, so we combine them rather than trust either alone. For each track we take the most probable non-silence class from each tagger and merge the two multisets by class. When both taggers propose a class, its fused confidence is a weighted sum of the two probabilities; when only one tagger proposes a class, it is kept only if its probability passes a solo-keep threshold that is higher for the weaker 1ch tagger. The merged list is truncated to the three strongest classes, which become the separation queries.

**Zero-target fine-tuning.** The dominant labelling error under CA-SDRi is a false positive on a clip that should be silent, because such a clip scores zero instead of being excluded from the mean.

We fine-tune the head of the 4ch tagger with a synthesis curriculum that raises the fraction of zero-target (silent) clips to 40%, so the model sees far more of the case the metric punishes. This single change raised the four-channel tagging proxy from 8.49 to 8.62 dB and mixture tagging accuracy from 60.7% to 63%. We also tried unfreezing the last two transformer blocks of the M2D backbone and training longer, but the four-channel proxy did not improve beyond the head fine-tune, which indicates the head is already close to the practical tagging limit on this data.

## 4. SEPARATION REFINEMENT

**Verify and refine.** Separation uses the baseline ResUNet from its 492-epoch checkpoint. After the first separation pass, each stem is re-tagged by the 1ch tagger. A stem is relabelled to silence when its query class is not confirmed by the re-tagging and its fused query confidence is also low, which removes queries that the separator could not realise as a clean source. The mixture is then separated a second time with the cleaned query list, so the remaining stems no longer compete with spurious queries for the same energy.

**Dual-separator selection.** We additionally keep a second separator obtained by a light fine-tune of the baseline. For each output slot we run both separators and keep the stem whose 1ch re-tagging gives the higher probability for the query class. This selection is decided per slot and per clip, so the system can take the cleaner stem from either model on a case-by-case basis. The gain is small but consistent on the development set.

**Separator capacity.** For completeness we tried to lift the 9.5 dB ceiling by improving the separator itself. A direct fine-tune of the ResUNet and a two-stage cascade, in which a frozen baseline is followed by a trainable residual refiner conditioned on the mixture and the first-stage stems, both converged to the quality of the baseline without a net gain. This is consistent with the oracle analysis: with the current architecture the separator is near its representational limit, and meaningful gains above 9.5 dB would need a larger model and substantially more training than a fine-tune provides.

## 5. DECISION THRESHOLDS

The zero-target structure of CA-SDRi makes false-positive suppression almost free on silent clips, so aggressive gating is favourable as long as it rarely removes a true source. We apply two gates after refinement. An energy gate relabels a stem to silence when its power relative to the mixture is below a threshold, and a probability gate removes a stem whose fused confidence is too low. Both thresholds, together with the fusion weights and the verification thresholds, are searched offline. We cache the taggers' probabilities, the per-stem re-tagging probabilities, the stem energies, and the pairwise SDR and SDR-improvement matrices for every development clip, and replay the exact CA-SDRi scoring under any threshold setting without re-running the networks. The search is staged, first the fusion weights, then the verification thresholds, then the gates, followed by a short coordinate-descent refinement.

**Per-class thresholds.** Classes differ in how reliably the taggers and the re-tagging behave, so a single global threshold is not optimal. We extend the verification, drop, and probability-gate thresholds to per-class values and tune them one class at a time on the same cached objective. To limit overfitting to the development set, a per-class change is accepted only when it improves the offline score by more than a fixed margin; twelve of the classes received an adjustment under this rule. The per-class thresholds give the largest single improvement after the tagger fine-tune.

Table 1: Development-set CA-SDRi (dB) and mixture tagging accuracy. The last row is the oracle labelling bound with the separator held fixed.

| System | CA-SDRi | Tag. acc. |
|---|---|---|
| Baseline (4ch) | 8.49 | 60.7 |
| Inference, baseline taggers | 8.92 | 60.7 |
| System 1: FT tagger + global thr. | 8.98 | 63.1 |
| System 2: per-class thresholds | 9.05 | 64.0 |
| System 3: dual-separator select | 9.06 | 63.8 |
| Oracle labelling (fixed sep.) | 9.52 | 100 |

## 6. SUBMITTED SYSTEMS

We submit three systems that add the components above in turn, so that at least one is robust if the evaluation set behaves differently from development. System 1 uses dual-tagger fusion, the zero-target fine-tuned 4ch tagger, verify and refine, and a single global threshold set. System 2 adds the per-class thresholds. System 3 adds the dual-separator stem selection and is our primary system. Systems 1 and 2 use a single separator (215.9M parameters); System 3 runs two separators at inference and totals 245.8M parameters. All three share the same taggers and the same refinement logic.

## 7. RESULTS

Table 1 reports development-set CA-SDRi and mixture tagging accuracy. Each component adds to the score over the 8.49 dB baseline. The inference pipeline without the fine-tuned tagger already reaches 8.92 dB through fusion, refinement, and gating. Adding the zero-target fine-tuned tagger and the global threshold search gives System 1 at 8.98 dB, the per-class thresholds give System 2 at 9.05 dB, and the dual-separator selection gives System 3 at 9.06 dB. Mixture tagging accuracy rises in step from 60.7% to 63.8%, and the gap to the oracle labelling bound of 9.52 dB narrows from 1.03 to 0.46 dB. The remaining gap is almost entirely the separator ceiling rather than the labelling, which matches the oracle analysis of Section 1.

## 8. CONCLUSION

Our DCASE 2026 Task 4 system treats labelling as the limiting stage, as indicated by an oracle study that places the separator-bound ceiling near 9.5 dB. Working mostly at inference time, dual-tagger fusion, a zero-target fine-tune of the four-channel tagger, verify-and-refine re-tagging, CA-SDRi-aware gating, per-class thresholds, and dual-separator selection together raise the baseline from 8.49 to 9.06 dB. Attempts to push past the ceiling by fine-tuning or cascading the separator did not help, which points to separator capacity as the next direction for this task.

## 9. REFERENCES

[1] B. T. Nguyen, M. Yasuda, N. Harada, R. Serizel, M. Mishra, M. Delcroix, C. Hernandez-Olivan, S. Araki, D. Takeuchi, T. Nakatani, and N. Ono, "Description and Discussion on

DCASE 2026 Challenge Task 4: Spatial Semantic Segmentation of Sound Scenes," arXiv:2604.00776, 2026.

[2] B. T. Nguyen, M. Yasuda, D. Takeuchi, D. Niizumi, and N. Harada, "Class-Aware Permutation-Invariant Signal-to-Distortion Ratio for Semantic Segmentation of Sound Scene with Same-Class Sources," in *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

[3] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked Modeling Duo: Learning Representations by Encouraging Both Networks to Model the Input," in *Proc. IEEE ICASSP*, 2023.

[4] 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 *Proc. IEEE ICASSP*, 2017.

[5] Q. Kong, Y. Wang, X. Song, Y. Cao, W. Wang, and M. D. Plumbley, "Source Separation with Weakly Labelled Data: An Approach to Computational Auditory Scene Analysis," in *Proc. IEEE ICASSP*, 2020.

[6] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual Reasoning with a General Conditioning Layer," in *Proc. AAAI*, 2018.

[7] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An Open Dataset of Human-Labeled Sound Events," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 829–852, 2022.

[8] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation," in *Proc. Interspeech*, 2024.

[9] M. Yasuda, Y. Ohishi, and S. Saito, "Echo-aware Adaptation of Sound Event Localization and Detection in Unknown Environments," in *Proc. IEEE ICASSP*, 2022.

[10] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 1015–1018.

[11] N. Furnon, "Noise files for the DISCO dataset," 2020. [Online]. Available: `https://github.com/nfurnon/disco`

[12] B. Veluri, M. Itani, J. Chan, T. Yoshioka, and S. Gollakota, "Semantic Hearing: Programming Acoustic Scenes with Binaural Hearables," in *Proc. ACM Symp. User Interface Software and Technology (UIST)*, 2023.