EXTENDING SR-CORRNET TO LABEL-QUERIED TARGET SOUND EXTRACTION

Technical Report

Bon-Hyeok Ku *Woocheol Jeong* *Hyung-Min Park*

Sogang University
Intelligent Information Processing Lab
Seoul, Korea
{nine4409, charmingwc, hpark}@sogang.ac.kr

ABSTRACT

This paper describes our submission to DCASE 2026 Task 4. We extend SR-CorrNet, originally designed for blind source separation, into a label-queried target sound extraction and separation model. The system reformulates blind separation into target extraction by conditioning the separator with frame-level strong class labels via Feature-wise Linear Modulation (FiLM), steering each output slot to extract the queried class. To stabilize ambiguous regions, we supplement this with a time-invariant weak class label FiLM bias. The class labels are predicted by a front-end fusion tagger that combines two complementary AudioSet-pretrained Transformers (M2D and PaSST) via feature-axis concatenation. The extended model operates with block streaming inference, coupling the tagger and separator through a soft-query interface.

Index Terms— spatial semantic segmentation, sound source separation, FiLM, M2D, fPaSST

1. INTRODUCTION

Task 4 of the DCASE 2026 Challenge, *Spatial Semantic Segmentation of Sound Scenes* (S5), asks a system to take a multi-channel spatial recording of an acoustic scene and return, for each active foreground sound event, both its class label and an isolated waveform [1]. Each mixture contains between zero and three foreground events drawn from 18 predefined classes, overlaid with up to two interfering sources and an always-present background, and rendered into four-channel FOA through measured room impulse responses.

We adopt the established two-stage decomposition of the baseline [2] — AT to decide *what* is present, then label-conditioned separation to recover *each* present source — but our focus is the separation stage. SR-CorrNet[3] is a strong multi-channel separator, yet in its original form it performs *blind* separation: it splits a fixed number of sources without being told which classes to recover. We extend it into a label-queried extractor, so the same correlation-network backbone can be steered, per source slot and per frame, to extract a specific requested class.

2. PROPOSED SYSTEM

As illustrated in fig. 1, the deployed pipeline produces per-source strong labels via a frame-level fusion tagger, which are subsequently formulated as a query sequence to guide the extended SR-CorrNet separator.

2.1. Target sound extraction : Label-Queried SR-CorrNet

2.1.1. SR-CorrNet

SR-CorrNet [3] is a time–frequency (TF) domain, asymmetric encoder–decoder framework employing a separation–reconstruction strategy. Instead of direct mapping, it formulates separation as a structured correlation-to-filter problem. The model first extracts generalized spatio–spectro–temporal correlation features $\mathbf{z}_{tf} \in C^{M(2L+1)(2I+1)}$ from multi-channel observations $\tilde{\mathbf{x}}_{tf}$ via local frame (L) and frequency (I) taps against a reference channel.

The processing pipeline is divided into three key stages: (1) a *Separation TF-Encoder* (B_E blocks) utilizing dual-path self-attention with rotary positional encoding [4] for coarse separation; (2) an *Dynamic Split Module* that adaptively determines the speaker streams; and (3) a *Reconstruction TF-Decoder* (B_D blocks) featuring weight-shared TF blocks and a speaker interaction module for cross-speaker refinement. Finally, a parallel filter estimator predicts multi-channel multi-tap complex deep filters $\mathbf{w}_{k,tf}$ to explicitly reconstruct each target source via MISO filtering ($Y_{k,tf} = \mathbf{w}_{k,tf}\tilde{\mathbf{x}}_{tf}$).

2.1.2. Label-queried conditioning interface (Strong + Weak labels)

We extended the original speech separation model by adapting its *Dynamic Split Module* into a *label-queried Split Module*. This is achieved by conditioning the encoder features simultaneously with a *time-varying, per-source* strong class label and a *time-invariant* per-source weak label through Feature-wise Linear Modulation (FiLM)[5].

While the frame-level strong query is highly discriminative, it can be inherently noisy at frame granularity. Therefore, we complement it with a static weak bias—the utterance-level multi-hot identity of the slot—to provide each slot with a stable, global identity channel. Formally, for source slot s at frame t, the per-frame strong label produces a FiLM gain $\boldsymbol{\gamma}_s^{\text{str}}(t)$ and shift $\boldsymbol{\beta}_s^{\text{str}}(t)$, while the weak label introduces an additive bias $\boldsymbol{\beta}_s^{\text{wk}}$ broadcast over time and frequency. The combined conditioning interface modulates the encoder feature $\boldsymbol{x}(t)$ as follows:

$$\boldsymbol{x}_s(t) = \boldsymbol{x}(t) \odot \boldsymbol{\gamma}_s^{\text{str}}(t) + \boldsymbol{\beta}_s^{\text{str}}(t) + \boldsymbol{\beta}_s^{\text{wk}}, \tag{1}$$

where the slot is informed not only *which* class to extract but precisely *when* it is active. This unified formulation effectively converts SR-CorrNet from a blind "separate the K sources" framework into

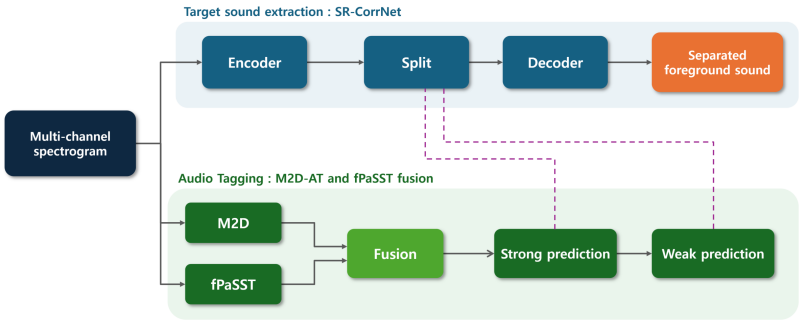

Figure 1: System architecture and overall pipeline of the proposed framework.

a targeted "extract the queried class on each slot" system, leveraging the local precision of strong labels alongside the global stability of weak biases.

2.1.3. *Block-wise streaming inference*

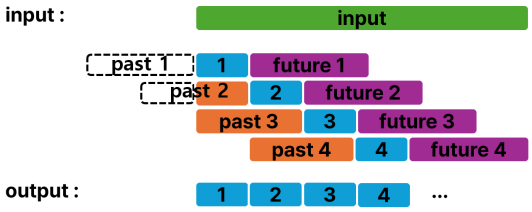

Figure 2: Block-wise streaming inference

The extended separator operates on the 4-channel STFT and is executed at test time through block streaming inference with key/value (KV) caching due to VRAM capacity limitations: the utterance is processed as a sequence of present blocks while attending to a cached past context and a future lookahead.

As illustrated in fig. 2, each inference block consists of three distinct segments: past context, present block, and future lookahead . The final output sequence is synthesized by stitching the processed present blocks together. Since the model's separation performance inherently depends on the empirical sizing of these past, present, and future windows, we utilize a tailored configuration to balance resource constraints and performance.This configuration deliberately uses a *long-context* window rather than the low-latency default.

2.2. Audio Tagging: M2D-AT and fPaSST fusion

2.2.1. *Embedding fusion*

The queries are produced by a frame-level (strong) tagger that fuses two complementary AudioSet-pretrained [6] Transformers: an M2D encoder [2], which we modified to enable frame-wise prediction, and a frame-level PaSST detector [7].

Each backbone is *shared* across the four FOA channels and applied per channel, and the resulting per-channel embeddings are concatenated along the feature axis, so spatial cues are carried implicitly in the inter-channel structure rather than through an explicit directional feature. The two channel-concatenated streams are each projected to a common width d_{fused} and fused by feature-axis concatenation to width $2\,d_{\text{fused}}$; no hand-designed cross-stream operator is used, and the cross-stream interaction is delegated to the first linear layer of the classification head. This "project-then-concatenate" fusion keeps each pretrained backbone intact in its own subspace while letting a lightweight head learn to mix semantic and temporal evidence.

2.2.2. *Strong and weak prediction label*

The two stages are joined by a confidence-gated bridge with *decoupled report and query thresholds*. Slot- and frame-level decisions used for *reporting* the predicted labels are gated at a higher confidence threshold, while the labels actually *handed to the extended SR-CorrNet* — as strong query and weak bias — are gated at a lower threshold. This soft-query scheme lets borderline slots and frames still reach the separator, giving it broader temporal context, without relaxing the labels against which the tagging metrics are scored. Slots below the weak gate are forced to silence and contribute no query.

3. EXPERIMENTAL SETUP

3.1. Data and synthesis

Training mixtures are synthesized on the fly from the DCASE 2026 S5 development set's foreground events, interference, background noise, and measured FOA room impulse response[8], with randomized event counts, SNRs, and spatial placements. Validation use the provided pre-rendered development split;

3.2. Audio tagging training

For Audio Tagging (AT) training, following the baseline configuration, the input consists of 4-channel First-Order Ambisonics (FOA) signals sampled at 32 kHz. We employ shared per-channel M2D and PaSST backbones to extract audio representations, which are subsequently integrated using a project-then-concatenate fusion mechanism. A per-source classification head is utilized to classify the inputs into 18 distinct target classes plus an additional silence class.

3.3. Target sound extraction training

For target sound extraction training, the model utilizes a 4-channel Short-Time Fourier Transform (STFT) front end. The condition-

Table 1: Results on the DCASE 2026 Task 4 development test set.

System	CAPI-SDRi	$\mathbf{Acc}_{\mathrm{mix}}$	$\mathbf{Acc}_{\mathrm{src}}$
Baseline	8.489	**60.714**	**70.394**
Submission 1	9.049	55.357	68.292
Submission 2	10.527	56.085	68.248
Submission 3	11.422	56.085	68.248
Submission 4	**11.434**	54.696	68.082

ing mechanism leverages a time-varying per-source strong-label FiLM query combined with a static per-source weak-label FiLM bias to guide the separation process. The network is optimized using a class-aware permutation-invariant loss function formulated with a Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [9] objective. Although the baseline configuration processes 10-second segments, limitations regarding model capacity and hardware computing power restricted our training input length to 4 seconds. During evaluation, long-context block streaming inference is deployed at test time to process extended audio inputs.

3.4. Inference and evaluation

The full pipeline is scored with the class-aware permutation-invariant SDR improvement (CAPI-SDRi) and label accuracy used by the challenge. The individual submissions vary based on their specific configurations of model checkpoints and block inference window sizes.

4. RESULTS

Table 1 reports the results of the four submitted systems on the development test set. All four submissions outperform the official baseline in terms of CAPI-SDRi. Submission 1 achieves 9.049, improving the baseline score of 8.489 by 0.560. The score further increases to 10.527 for Submission 2 and 11.422 for Submission 3.

Submissions 1–3 use the same model-selection criterion but different block-context configurations. The increase in CAPI-SDRi across these systems demonstrates that the inference context is an important factor for the proposed separator. In particular, Submissions 2 and 3 produce identical mixture- and source-level accuracy, while Submission 3 improves CAPI-SDRi by 0.895. Since their reported label accuracies are unchanged, this improvement is consistent with a gain in waveform separation arising from the block-context configuration rather than from a change in audio-tagging decisions.

Submission 4 uses the same block configuration as Submission 3 but changes the separator checkpoint-selection criterion from the best development metric to the lowest validation loss. It obtains the best CAPI-SDRi of 11.434, exceeding the official baseline by 2.945. However, the improvement over Submission 3 is only 0.012, indicating that checkpoint selection has a considerably smaller numerical effect than the block-context configuration in the present experiments.

The proposed systems show a different trend for the tagging metrics. Their mixture-level and source-level accuracies remain below those of the official baseline. Thus, the CAPI-SDRi improvement is primarily associated with the separation component and does not correspond to an improvement in standalone classification accuracy. This result also indicates that stronger calibration of the fusion tagger and its confidence thresholds may further improve the complete system, particularly for zero-target mixtures and low-confidence source slots.

Since CAPI-SDRi is the official ranking metric, Submission 4 is selected as our primary submitted system. Nevertheless, the accuracy results are retained to make the trade-off between sound event classification and waveform separation explicit.

5. CONCLUSION

We presented a two-stage system for DCASE 2026 Task 4 that combines an M2D–fPaSST fusion tagger with a label-queried SRCorrNet separator. Frame-level strong labels and clip-level weak labels condition the separator through FiLM, while block-wise long-context inference enables memory-efficient processing of the four-channel input. Our best submission achieves a development CAPI-SDRi of 11.434, exceeding the official baseline by 2.945. However, its mixture- and source-level classification accuracies remain below the baseline, suggesting that further improvements should focus on tagger calibration and the interface between class prediction and source separation.

6. 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 preprint arXiv:2604.00776*, 2026. [Online]. Available: https://arxiv.org/abs/2604.00776

[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] U.-H. Shin and H.-M. Park, "Asymmetric encoder-decoder based on time-frequency correlation for speech separation," arXiv preprint arXiv:2603.29097, 2026.

[4] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

[5] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[6] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "AudioSet: An ontology and human-labeled dataset for audio events," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017.

[7] F. Schmid, P. Primus, T. Morocutti, J. Greif, and G. Widmer, "Multi-iteration multi-stage fine-tuning of transformers for sound event detection with heterogeneous datasets," in *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2024.

[8] M. Yasuda, B. T. Nguyen, N. Harada, and D. Takeuchi, "DCASE2026Task4Dataset: The Dataset for Spatial Semantic Segmentation of Sound Scenes," Apr. 2026, version 1.0.

[9] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.