SEMANTIC DISTILLATION FOR SPATIAL SEMANTIC SEGMENTATION OF SOUND SCENES

Sen Wang, Chengyao Tang
Zhicheng Zhang, Jianqin Yin

Beijing University of Posts and Telecommunications, China
senwang@bupt.edu.cn, 2025111419@bupt.cn,
zczhang@bupt.edu.cn, jqyin@bupt.edu.cn

June 16, 2026

Abstract

This report describes our system for DCASE 2026 Task 4, spatial semantic segmentation of sound scenes. The system follows a two-stage pipeline: an M2D-based audio tagger first predicts up to three event labels, and a label-queried ResUNetK then separates the corresponding dry monaural sources from a four-channel spatial mixture. We improve audio tagging along two complementary directions. The first uses permutation-invariant deep supervision and an exponential-moving-average teacher. The second transfers semantic representations from a frozen Qwen2-Audio model to M2D using centered-kernel-alignment and cosine objectives. Both methods are developed for single-channel (1c) and four-channel (4c) tagging. On the development test set, the submitted 1c system obtains 8.557 dB CAPI-SDRi, while the submitted 4c system obtains 8.807 dB CAPI-SDRi with 61.442% mixture-level label accuracy and 72.535% source-level label accuracy.

Keywords: *spatial sound scene analysis, M2D, knowledge distillation, Qwen2-Audio, ResUNetK, source separation*

1 Introduction

Spatial semantic segmentation of sound scenes (S5) jointly addresses sound event recognition and source separation [1]. Given a first-order Ambisonics (FOA) mixture, a system must identify the active event classes and recover a dry monaural waveform for each predicted event. Unlike conventional source separation, the estimated sources are explicitly associated with semantic labels; therefore, both label errors and waveform distortion affect the final score.

Our system is based on the two-stage DCASE Task 4 pipeline and official baseline design [2]. The first stage uses Masked Modeling Duo (M2D) [3], pre-trained on AudioSet [4], to predict three event tracks including silence. The second stage uses ResUNetK to estimate three label-conditioned sources. We investigate two input configurations for the tagger: the reference channel alone (1c) and all four FOA channels (4c). The 4c model processes channel-wise M2D features and aggregates them with a channel-aware tagging head.

Recent S5 and target-sound-extraction systems have explored temporal guidance, iterative refinement, and label-query-conditioned separation architectures [5, 6, 7]. In contrast, our submitted systems keep the separator fixed and focus on improving semantic label prediction.

The main contribution to the tagging stage is a comparison of two distillation strategies. Masked self-distillation (Masked SD) combines intermediate supervision, token masking, and an exponential-moving-average (EMA) teacher, whereas Qwen semantic distillation transfers high-level representations from the audio encoder of Qwen2-Audio [8]. The separator is kept fixed to the released ResUNetK checkpoint, so the reported differences mainly reflect the tagger.

2 System

2.1 Two-stage inference

Let $\mathbf{x} \in \mathbb{R}^{C\times L}$ denote an FOA mixture, where $C = 4$. The tagger predicts $K = 3$ label distributions $\mathbf{p} \in \mathbb{R}^{K\times(N+1)}$ over $N = 18$ event classes and a silence class. The selected labels are concatenated into a query vector and embedded into a 512-dimensional condition. ResUNetK receives the original four-channel mixture and this condition, and jointly estimates K monaural waveforms. Tracks predicted as silence are suppressed by a source-presence mask.

2.2 Permutation-invariant masked self-distillation

The Masked SD system fine-tunes the complete M2D tagger and attaches auxiliary heads to Transformer blocks 9 and 10. In addition to the main permutation-invariant classification loss $\mathcal{L}_{\text{main}}$, the auxiliary heads are directly supervised. A teacher network is maintained as an EMA of the student parameters,

$$\theta_{\text{ema}} \leftarrow m\theta_{\text{ema}} + (1-m)\theta_{\text{student}}, \tag{1}$$

where $m = 0.999$. Since event tracks have no fixed order, distillation uses the minimum KL divergence over all track permutations:

$$\mathcal{L}_{\text{KD}} = \min_{\pi\in\mathcal{S}_K} T^2 D_{\text{KL}}\left(\sigma(\mathbf{z}^{\pi}_{\text{ema}}/T) \,\|\, \sigma(\mathbf{z}/T)\right), \tag{2}$$

where $T = 1.5$. The total objective is

$$\mathcal{L}_{\text{SD}} = \mathcal{L}_{\text{main}} + \gamma\mathcal{L}_{\text{aux}} + \alpha\mathcal{L}_{\text{KD}} + \alpha_{\text{self}}\mathcal{L}_{\text{self}}, \tag{3}$$

with $\gamma = 0.05$, $\alpha = 0.10$, and $\alpha_{\text{self}} = 0.05$. For the 4c model, 30% of spatial tokens are masked during training to reduce dependence on any single channel.

2.3 Qwen2-Audio semantic distillation

The Qwen-distilled systems use a frozen Qwen2-Audio encoder as a semantic teacher. During training, the teacher receives the sum of the clean dry foreground sources, while the M2D student receives the reverberant noisy mixture. This asymmetric teacher–student design provides a clean semantic target without adding Qwen2-Audio to the inference pipeline.

Student tokens are projected to the Qwen representation dimension and both sequences are pooled to 32 temporal bins. In the 4c v2 system, the tagging head uses all four channels, while semantic matching is computed from the reference channel to avoid spatial duplication. We optimize the task loss together with linear centered kernel alignment (CKA) [9]. CKA-based objectives have also been used recently for audio-language model distillation under representation mismatch [10]. We combine CKA with cosine distance:

$$\mathcal{L}_{\mathrm{Qwen}} = \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{CKA}}(1 - \mathrm{CKA}(\mathbf{S}, \mathbf{T})) + \lambda_{\cos}(1 - \cos(\bar{\mathbf{S}}, \bar{\mathbf{T}})). \tag{4}$$

For 1c, $(\lambda_{\mathrm{CKA}}, \lambda_{\cos}) = (0.05, 0.02)$. For 4c, we use the more conservative weights $(0.02, 0.01)$. Distillation weights and the student projector learning rate are warmed up for five epochs. The frozen teacher is removed from the inference checkpoint.

3 Experiments

3.1 Dataset and evaluation

We use the DCASE 2026 Task 4 development data [1]. Spatial mixtures are generated from 18 classes of anechoic event recordings, FOA room impulse responses, FOA background noise, and interfering sounds. Audio is sampled at 32 kHz. Each generated 10-s scene contains zero to three target events for tagging and one to three target events for separator training. Event SNR is sampled from 5 to 20 dB, with up to two interference events at 0 to 15 dB.

We report class-aware permutation-invariant SDR improvement (CAPI-SDRi) [2], exact mixture-level label accuracy, and source-level label accuracy. All tagger comparisons use the same downloaded ResUNetK checkpoint.

3.2 Training setup

All M2D variants use AdamW with a backbone learning rate of 10^{-5} and gradient clipping at 0.5. The semantic projector uses a learning rate of 10^{-4}. The 1c SD, 1c Qwen, 4c SD, and 4c Qwen runs use batch sizes 64, 16, 24, and 4, respectively; gradient accumulation is used where required. Models are selected by validation classification loss or the final training checkpoint. The selected epochs are 22 (1c Masked SD), 33 (1c Qwen), 20 (4c Masked SD), and 21 (4c Qwen).

3.3 Results

We evaluate the 1c and 4c tracks separately because they use different tagging inputs and have distinct DCASE baseline systems. All rows within each track use the same released ResUNetK separator. Table 1 reports the 1c results. Our submitted 1c system uses the last checkpoint at epoch 33 and improves over the 1c DCASE baseline by 0.386 dB CAPI-SDRi, 2.910 percentage points in mixture-level accuracy, and 2.872 percentage points in source-level accuracy.

Table 1: Development-set results for the 1c track with the released ResUNetK. "Mix." and "Src." are label accuracies in percent.

System	CAPI (dB)	Mix.	Src.
DCASE baseline	8.171	57.143	67.147
Masked SD	8.354	58.069	68.311
Semantic KD (epoch 33)	**8.557**	**60.053**	**70.019**

Table 2: Development-set results for the 4c track with the released ResUNetK.

System	CAPI (dB)	Mix.	Src.
DCASE baseline	8.489	60.714	70.394
Masked SD	8.546	60.053	70.293
Semantic KD (epoch 21)	**8.807**	**61.442**	**72.535**

Table 2 reports the 4c results. Our submitted 4c system uses the best checkpoint at epoch 21. It improves over the 4c DCASE baseline by 0.318 dB CAPI-SDRi, 0.728 percentage points in mixture-level accuracy, and 2.141 percentage points in source-level accuracy.

4 Conclusion

We presented a two-stage DCASE 2026 Task 4 system with distillation-enhanced M2D tagging and label-queried ResUNetK separation. The experiments show that four-channel tagging is important for the joint task and that semantic distillation from Qwen2-Audio yields the strongest tagger in both tracks. The submitted 1c system uses epoch 33 and obtains 8.557 dB CAPI-SDRi, 60.053% mixture-level accuracy, and 70.019% source-level accuracy. The submitted 4c system uses epoch 21 and obtains 8.807 dB CAPI-SDRi, 61.442% mixture-level accuracy, and 72.535% source-level accuracy.

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," 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2026.

[3] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked modeling duo: Towards a universal audio pre-training framework," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[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 *Proceedings of the IEEE International Conference*

on Acoustics, Speech and Signal Processing, 2017, pp. 776–780.

[5] T. Morocutti, J. Greif, P. Primus, F. Schmid, and G. Widmer, "On temporal guidance and iterative refinement in audio source separation," *arXiv preprint arXiv:2507.17297*, 2025.

[6] F. Wu and Z.-Q. Wang, "TS-TFGridNet: Extending TFGridNet for label-queried target sound extraction via embedding concatenation," DCASE Challenge," DCASE 2025 Challenge Technical Report, 2025. [Online]. Available: https://dcase.community/documents/challenge2025/technical_reports/DCASE2025_Wu_45_t4.pdf

[7] H. Yin, J. Bai, Y. Xiao, H. Wang, S. Zheng, Y. Chen, R. K. Das, C. Deng, and J. Chen, "Exploring text-queried sound event detection with audio source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025.

[8] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[9] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 3519–3529.

[10] Q. Yang, B. Zhao, Z. Kang, X. Li, Y. He, C. Liu, X. Zhang, X. Qu, J. Peng, and J. Wang, "Attention-weighted centered kernel alignment for knowledge distillation in large audio-language models applied to speech emotion recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2026.