TRAINING-FREE AUDIO MOMENT RETRIEVAL VIA BACKGROUND-CONTRASTIVE GAUSSIAN MIXTURE LOCALIZATION

Technical Report

Meghan Kret

The Cooper Union, New York, USA meghan.kret@gmail.com

ABSTRACT

This report describes two submitted systems for DCASE 2026 Challenge Task 6 (Audio Moment Retrieval from Long Audio [1]). Both systems use no supervised temporal training and no labeled data. **System 1** combines background contrast normalization with per-query two-component Gaussian mixture model (GMM) [2] inference over frozen MS-CLAP 2023 [3], [4] similarity traces. **System 2** is an ablation without the contrast step. On the CASTELLA [5] development-test set, System 1 achieves 10.03% mAP and 13.51% R1@0.7 – surpassing the single-dataset supervised DETR [6], [7] baseline (9.11%, 10.32%) [8] using identical frozen features. On Clotho-Moment [8], System 1 achieves 44.28% mAP against the supervised baseline's 6.32%, a 37.96 pp cross-domain gap explained by the supervised decoder's domain-specific prior mismatch.

***Index Terms*—Audio moment retrieval, GMM, background contrast, no supervised training, DCASE 2026.**

I. SYSTEM OVERVIEW

Audio moment retrieval (AMR) asks for the temporal segment of a long recording that matches a free-form text query [8]. The DCASE 2026 Task 6 baseline [1] uses a QD-DETR [7] decoder trained on CASTELLA [5] temporal boundary annotations, with frozen MS-CLAP 2023 [3] features as input. Our systems replace this supervised decoder with per-query probabilistic inference, using no labeled data.

Both systems share the same core pipeline: CLAP feature projection, (optional) background contrast normalization, GMM fitting by EM [2], and moment extraction from the posterior. Fig. 1 shows the two systems side by side.

II. METHOD

II-A. CLAP Similarity Trace

MS-CLAP 2023 [3], [4] is a contrastive audio-language model trained on matched audio-caption pairs, using an HTS-AT [9] audio encoder. For a text query q and recording of duration T seconds, the pre-extracted 768-dim audio window features (1 s windows, 1 s hop) [1] are projected through the CLAP [3] projection head to 1024-dim L2-normalized embeddings $\{\hat{a}_t\}$. The query text embedding $\hat{q}$ is similarly projected. Raw per-window cosine similarity:

$$s_t = \hat{q} \cdot \hat{a}_t, \quad t = 1, \ldots, T. \tag{1}$$

This produces a time-varying similarity trace: frames where the audio matches the query score high; background frames score low.

II-B. Background Contrast Normalization (System 1)

Inspired by contrastive decoding in language generation [10], we subtract a neutral background similarity from the raw trace. A fixed neutral embedding $\hat{n}$ is the L2-normalized mean of "background sound", "ambient noise", and "background audio" – three phrases that are closely clustered in CLAP space (pairwise cosine similarity > 0.88). The contrast score:

$$c_t = s_t - (\hat{n} \cdot \hat{a}_t) = (\hat{q} - \hat{n}) \cdot \hat{a}_t \tag{2}$$

removes the component of each window's similarity explained by generic background-ness, increasing the separation between relevant and background score populations. On CASTELLA [5] and Clotho-Moment [8], this increases the GMM mean separation $\Delta = |\mu_{\text{rel}} - \mu_{\text{bg}}|$ on 66% of queries (mean Δ: $0.049 \rightarrow 0.055$). System 2 skips this step and uses s_t directly.

II-C. Per-Query GMM Inference

The trace (contrast for System 1, raw for System 2) is smoothed with a Gaussian kernel ($\sigma = 1.5$ frames, SciPy [11]) and modeled as a two-component Gaussian mixture model (GMM) [2]:

$$p(\tilde{c}_t \mid \theta) = \pi_{\text{rel}}\, \mathcal{N}(\tilde{c}_t \mid \mu_{\text{rel}}, \sigma^2_{\text{rel}}) + \pi_{\text{bg}}\, \mathcal{N}(\tilde{c}_t \mid \mu_{\text{bg}}, \sigma^2_{\text{bg}}). \tag{3}$$

The parameters θ are estimated by EM [2] (scikit-learn [12]) with 5 random initializations; the maximum-likelihood solution is retained to reduce sensitivity to local optima. The component with higher mean μ is labeled *relevant*. The per-frame posterior:

$$\gamma_t = \frac{\pi_{\text{rel}}\, \mathcal{N}(\tilde{c}_t \mid \mu_{\text{rel}}, \sigma^2_{\text{rel}})}{p(\tilde{c}_t \mid \theta)} \tag{4}$$

gives a soft relevance probability for each second, calibrated to the current trace's score distribution. No parameters are shared across queries or datasets.

II-D. Moment Extraction

Contiguous frames where $\gamma_t \geq 0.45$ are collected as candidate segments, each scored by $\overline{\gamma}_{a:b} + 0.3 \cdot \max_{a \leq t \leq b} \tilde{c}_t$. A peak-centered fallback handles flat traces. The top prediction is submitted.

II-E. Why No Supervised Training Is Needed

CLAP [3], [4] is trained with clip-level contrastive objectives on audio-caption pairs. Despite no explicit temporal supervision, the sliding-window similarity trace encodes useful temporal structure [13]: frames that match the query consistently score higher than background frames. The GMM exploits this structure by fitting the score distribution of each individual test trace. The DETR baseline [8], [7], [6] additionally trains a temporal decoder on CASTELLA [5] boundary annotations to learn event duration

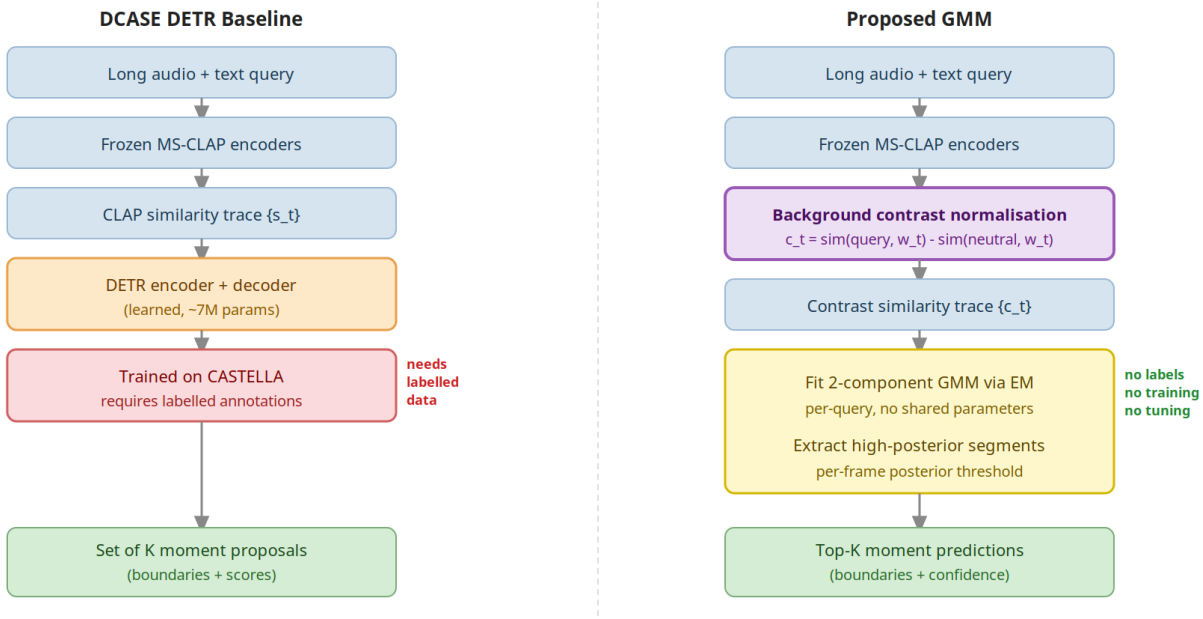

Fig. 1. Supervised DETR baseline (left) vs. proposed GMM systems (right). Both use identical frozen MS-CLAP 2023 [3] features. System 1 adds background contrast normalization (purple) before GMM fitting. System 2 uses the raw similarity trace.

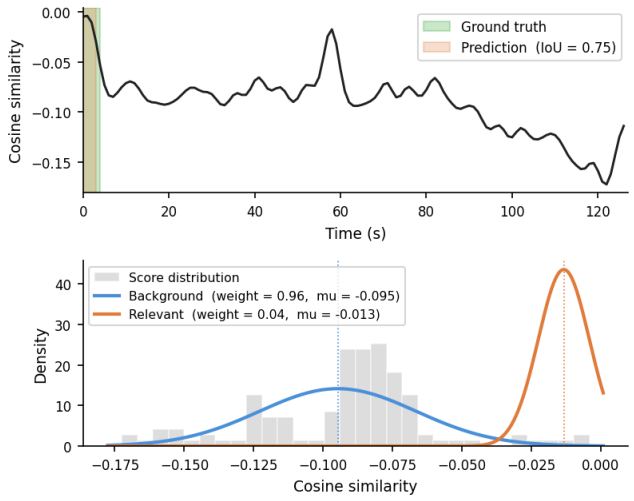

Fig. 2. Example from CASTELLA [5] (query: "Men laugh"). **Top:** contrast trace with ground truth (green) and prediction (orange, IoU= 0.88). **Bottom:** fitted 2-component GMM [2]. The relevant component has higher mean and lower weight (0.19), consistent with a brief event in a long recording.

priors and boundary placement conventions. This helps in-domain but hurts cross-domain when those priors are miscalibrated for a new dataset.

III. SUBMITTED SYSTEMS

TABLE I
SUBMITTED SYSTEMS.

Label	System	Description
`task6_1`	GMM + contrast	With background subtraction
`task6_2`	GMM simple	Without background subtraction

IV. DEVELOPMENT SET RESULTS

TABLE II
RESULTS ON CASTELLA [5] AND CLOTHO-MOMENT [8] DEVELOPMENT-TEST SETS (%). PRIMARY METRIC IS R1@0.7 (BOLD). METRICS FOLLOW THE MOMENT RETRIEVAL CONVENTION [14]. DETR BASELINE RESULTS FROM [1].

System	Test	R1@.5	**R1@.7**	mAP	@.5	@.75
DETR (CAS) [1]	CAS	**23.16**	10.32	9.11	**20.34**	6.96
DETR (CAS+CM) [1]	CAS	**25.61**	13.59	**12.06**	**23.60**	**10.72**
Sys. 2 GMM simple	CAS	17.67	11.88	9.20	15.99	8.74
Sys. 1 GMM+contrast	CAS	18.49	**13.51**	10.03	16.70	9.80
DETR (CAS) [1]	CM	11.54	3.78	6.32	15.81	4.60
Sys. 2 GMM simple	CM	55.44	44.19	40.88	59.80	43.32
Sys. 1 GMM+contrast	CM	**58.52**	**47.84**	**44.28**	**62.80**	**46.90**

On CASTELLA [5], System 1 surpasses the DETR baseline trained on CASTELLA only [1] at R1@0.7 (13.51% vs 10.32%) and mAP (10.03% vs 9.11%), while trailing at R1@0.5 where the baseline's learned coarse-location priors help. The multi-dataset DETR variant leads overall.

On Clotho-Moment [8], System 1 reaches 44.28% mAP against the baseline's 6.32% (+37.96 pp). Both systems use no labels; the gap is attributable entirely to the DETR decoder's CASTELLA-calibrated priors being miscalibrated for Clotho-Moment's programmatic boundaries and Walking Tours backgrounds.

Table III provides mechanistic evidence. If the cross-domain gap

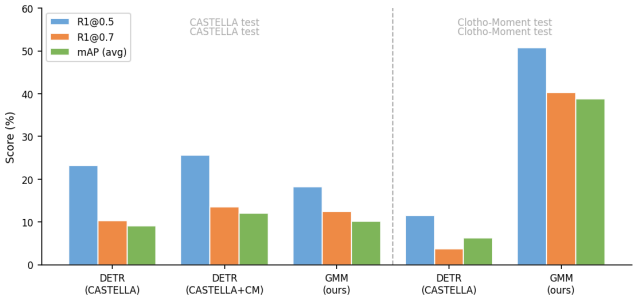

Fig. 3. Development-test results on CASTELLA [5] (left) and Clotho-Moment [8] (right). System 1 (dark outline) is competitive with single-dataset DETR [1] on CASTELLA and dominant cross-domain.

TABLE III
MEAN IOU BY SIGNAL QUALITY (Δ) ON CLOTHO-MOMENT [8]. ADVANTAGE GROWS WITH SIGNAL QUALITY – CONFIRMING PRIOR MISMATCH AS DETR'S CROSS-DOMAIN FAILURE CAUSE.

Δ range	n	Sys. 1	DETR [1]	Adv.
Low ($\Delta < 0.030$)	1556	0.183	0.146	+0.037
Mid (0.030–0.070)	3338	0.491	0.183	+0.308
High ($\Delta > 0.070$)	1755	0.768	0.256	+0.512

were due to weak signal in Clotho-Moment, the GMM advantage should shrink as signal quality improves. Instead it grows: at high Δ (clean CLAP trace), DETR achieves 0.256 mean IoU against System 1's 0.768. This confirms prior mismatch [5], not signal weakness, as the cause.

V. HYPERPARAMETERS

TABLE IV
FIXED HYPERPARAMETERS. NO VALUES WERE TUNED ON LABELED DATA.

Parameter	Value
MS-CLAP version [3]	2023
Audio window / hop	1 s / 1 s
Smoothing σ	1.5 frames
EM initializations [2], [12]	5
Posterior threshold	0.45
Neutral texts (sys. 1)	see Section II-B
Random seed	42

VI. EXTERNAL RESOURCES

Both systems use only the organizer-provided MS-CLAP 2023 [3], [4] audio-text features. No additional external datasets, pre-trained models, or labeled data were used.

VII. CONCLUSION

Two no-supervised-training systems are submitted for DCASE 2026 Task 6 [1]. System 1 (GMM + background contrast) is competitive with the single-dataset supervised baseline [8] at R1@0.7 on CASTELLA [5] and dominates cross-domain on Clotho-Moment [8] (+37.96 pp mAP). System 2 (GMM simple) is an ablation confirming the contribution of contrast normalization. Both systems confirm that frozen MS-CLAP 2023 [3], [4] traces already encode substantial temporal signal [13], and that per-query GMM inference [2] can exploit it without any labeled training data from CASTELLA [5] or Clotho-Moment [8].

VIII. REFERENCES

[1] DCASE Challenge, "DCASE 2026 challenge task 6: Audio moment retrieval from long audio," Available: https://dcase.community/challenge2026/index, 2026.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[3] Y. Wu *et al.*, "Large-scale contrastive language-audio pre-training with feature fusion and keyword-to-caption augmentation," *arXiv preprint arXiv:2211.06687*, 2022.

[4] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," *arXiv preprint arXiv:2206.04769*, 2022.

[5] H. Munakata *et al.*, "CASTELLA: Long audio dataset with captions and temporal boundaries," *arXiv preprint arXiv:2511.15131*, 2025.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. European Conference on Computer Vision*, 2020, pp. 213–229.

[7] W. Moon, S. Hyun, S. Park, D. Park, and J.-P. Heo, "Query-dependent video representation for moment retrieval and highlight detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 023–23 033.

[8] H. Munakata *et al.*, "Language-based audio moment retrieval," *arXiv preprint arXiv:2409.15672*, 2024.

[9] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2022, pp. 646–650.

[10] X. L. Li, A. Holtzman, D. Fried, P. Liang, J. Eisner, T. Hashimoto, L. Zettlemoyer, and M. Lewis, "Contrastive decoding: Open-ended text generation as optimization," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

[11] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . Contributors, "SciPy 1.0: Fundamental algorithms for scientific computing in python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,

M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[13] S. Ghosh *et al.*, "CompA: Addressing the gap in compositional reasoning in audio-language models," in *Proc. ICLR*, 2024.

[14] J. Lei, T. L. Berg, and M. Bansal, "QVHighlights: Detecting moments and highlights in videos via natural language queries," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 11 846–11 858.