LOCAL CONTINUITY SALIENCY DETR FOR LANGUAGE-BASED AUDIO MOMENT RETRIEVAL SUBMITTED TO DCASE 2026 CHALLENGE TASK 6

Technical Report

Le Duc Minh
Vietnamese-German University
minh.leduc.0210@gmail.com

Tran Nguyen Van Anh
University of Science – HCMUS
tnvananh1802@gmail.com

ABSTRACT

This paper proposes LCS-DETR, a Detection Transformer-based model for Language-Based Audio Moment Retrieval that addresses two critical limitations of standard temporal grounding: coarse temporal alignment and class imbalance. We introduce a saliency-guided framework with local temporal continuity modeling via depthwise-separable convolution and Focal Loss in the Hungarian matcher to suppress background dominance. Evaluated on the combining Clotho-Moment and CASTELLA[1] datasets, LCS-DETR achieves a mAP of 70.26% and R1@0.7 of 73.08%, representing a 4.8× and 4.4× improvement over the baseline DETR [2] respectively. Code and results are available at `https://github.com/MinLee0210/DCASE_2026.git`.

***Index Terms*—** Audio Moment Retrieval, Cross-Modal Learning, Saliency-Guided Grounding, Temporal Localization

1. INTRODUCTION

Language-based audio moment retrieval aims to localize temporal segments within long, untrimmed audio recordings that semantically align with natural language queries. This task has practical applications ranging from audio search in archives, content curation, accessibility, to surveillance systems where precise temporal identification is critical.

Previous work on audio retrieval has focused primarily on short-clip retrieval using contrastive learning in a shared embedding space [3]. However, a more compelling and practical problem emerges when we consider untrimmed, long-duration audio: given a query such as *"dog barking fading to siren sound"*, the system should retrieve not just any audio containing these sounds, but the *specific moments* where these events occur (e.g., 12–25 seconds).

Transformers [4] have become the dominant architecture in deep learning. For temporal grounding, Detection Transformers (DETR) [5] have shown strong performance by modeling temporal and cross-modal dependencies [6].

In this paper, we propose LCS-DETR (Local Continuity Saliency DETR), which extends DETR-based temporal grounding with two targeted mechanisms:

1. **Saliency guidance:** A gating mechanism that computes frame-level semantic relevance via cross-modal similarity, focusing the model's attention on audio regions that correspond to the text query.
2. **Local temporal continuity:** Depthwise-separable 1D convolution with residual connections (kernel size 5, approximately 5-second context) that smooths saliency masks and prevents fragmented predictions over acoustic events.

Experimental results on Clotho-Moment and CASTELLA datasets demonstrate substantial improvements, with LCS-DETR achieving 73.08% R1@0.7 compared to baseline's 13.59%. We also show that our approach maintains strong performance even at strict IoU thresholds, indicating precise boundary localization.

2. RELATED WORK

Audio moment retrieval shares characteristics with several related tasks in audio and vision domains, but remains distinct in its scope and requirements.

2.1. Sound Event Detection (SED)

Sound Event Detection (SED) predicts both event class labels and time boundaries from audio [7]. However, SED assumes: (i) a closed, predetermined set of event categories, and (ii) each category corresponds to a single event type. In contrast, LBAMR operates in an open-vocabulary setting where queries can describe complex, multi-event acoustic scenes. Thus, LBAMR can be viewed as a zero-shot SED with open-vocabulary queries.

2.2. Video Moment Retrieval (VMR)

Video moment retrieval is the most similar task, with identical input/output structure but operating on visual data. VMR models have successfully adopted DETR-based architectures [5] to capture temporal and cross-modal dependencies [6]. Recent DETR variants include boundary-aligned approaches [8] and motion-semantic joint learning [9]. Our work extends these techniques to the audio domain, addressing audio-specific challenges such as temporal aliasing and absence of spatial structure.

2.3. Audio Moment Retrieval

Recent work on audio moment retrieval [2] has proposed end-to-end DETR-based models for this task, establishing the Clotho-Moment dataset for large-scale evaluation. Related work on saliency-guided

temporal grounding [10] in video has demonstrated the effectiveness of incorporating attention mechanisms to focus on relevant regions. Cross-modal fusion methods [11, 12] highlight the importance of semantic alignment in multimodal systems. We address key limitations through: (i) focal loss [13] for class imbalance, (ii) GIoU loss [14] for boundary optimization, and (iii) temporal smoothing via depthwise-separable convolution [15]. Our work combines the AMR framework from [2], saliency guidance from [10], and boundary-aware techniques to achieve improved temporal precision and stability.

3. METHODOLOGY

3.1. Problem Formulation

Given a long audio sequence $\mathbf{A} \in R^{L_a \times d_a}$ and a natural language query $\mathbf{Q}$, the task is to predict one or more temporal spans $(t_{\text{start}}, t_{\text{end}})$ indicating moments where the query content occurs in the audio. We represent predictions in normalized center-width coordinates (c_x, w) for improved optimization stability:

$$c_x = \frac{t_{\text{start}} + t_{\text{end}}}{2L_a}, \quad w = \frac{t_{\text{end}} - t_{\text{start}}}{L_a} \tag{1}$$

3.2. Feature Extraction

Both audio and text are encoded using CLAP (Contrastive Language-Audio Pretraining) [3], yielding 768-dimensional embeddings. Audio features are optionally augmented with learnable Temporal Embedding Features (TEF) that preserve absolute temporal context. Text embeddings are aggregated via global mean pooling to a single vector $\hat{\mathbf{t}} \in R^d$.

3.3. Saliency-Guided Cross-Modal Gating

To focus the model on semantically relevant audio frames, we compute frame-level saliency scores via cosine similarity:

$$s_i = \sigma\left(\frac{\mathbf{a}_i \cdot \hat{\mathbf{t}}}{\|\mathbf{a}_i\| \cdot \|\hat{\mathbf{t}}\|} \cdot \beta + \alpha\right) \tag{2}$$

where σ is sigmoid, and β, α are learnable parameters. These scores gate the cross-attention values in the Text-to-Audio (T2A) encoder, suppressing low-relevance frames before transformer processing.

3.4. Local Temporal Continuity Module

Frame-independent saliency computation produces fragmented masks. We enforce temporal smoothness via depthwise-separable convolution:

$$\tilde{\mathbf{A}} = \text{Conv}_{1\times1}\left(\text{GELU}(\text{DWConv}_{1\times5}(\mathbf{A}))\right) + \mathbf{A} \tag{3}$$

The kernel size 5 corresponds to approximately 5-second context windows, smoothing saliency transitions while residual connections preserve frame semantics. This ensures coherent attention over acoustic events rather than isolated frame matches.

3.5. Transformer Encoder-Decoder

Saliency-weighted features pass through a 2-layer Transformer encoder (8 attention heads, hidden dimension 256). The decoder transforms $N = 10$ learnable queries to candidate spans via two prediction heads:

- **Span Regressor:** Predicts (c_x, w) with L1 and GIoU losses
- **Classification Head:** Binary foreground/background prediction

3.6. Focal Loss for Class Imbalance

Standard cross-entropy allows trivial solutions (predict "background" for all queries). We replace CE with Focal Loss [13]:

$$\mathcal{L}_{\text{focal}} = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{4}$$

With $\gamma = 2$ and $\alpha = 0.25$, background-classified queries contribute exponentially less to loss, forcing precise boundary localization.

3.7. Training Objective

The total loss combines weighted terms:

$$\mathcal{L} = \lambda_{\text{focal}}\mathcal{L}_{\text{focal}} + \lambda_{\text{span}}\mathcal{L}_{\text{span}} + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}} + \lambda_{\text{aux}}\sum_{i}^{L-1}\mathcal{L}_i \tag{5}$$

Hungarian bipartite matching [16] aligns predicted and ground-truth spans during training. Generalized IoU (GIoU) loss [14] directly optimizes for temporal IoU between predictions and ground truth.

4. EXPERIMENTS

4.1. Datasets

We leverage two complementary datasets provided by the challenge organizers with pre-extracted features:

- **Clotho-Moment** [2]: Curated subset of Clotho dataset with 15–30 second clips, temporal moment annotations, and natural language captions.
- **Castella** [1]: Diverse acoustic environments and query phrasings, designed to increase training variance.

We combine both datasets into a unified training corpus to leverage complementary strengths: Clotho-Moment provides natural language richness while CASTELLA offers diversity. Both datasets are provided with pre-extracted CLAP embeddings.

4.2. Training Configuration

- Feature extractor: CLAP [3] (768-dim)
- Optimizer: AdamW [17] (lr=10^{-4}, wd=10^{-4})
- Schedule: Cosine annealing, 200 epochs
- Batch size: 2000 (pre-extracted features)
- Mixed precision: bfloat16 with AMP
- Data augmentation: $\pm10\%$ temporal jitter on GT spans
- Reproducibility: Fixed seed 17771

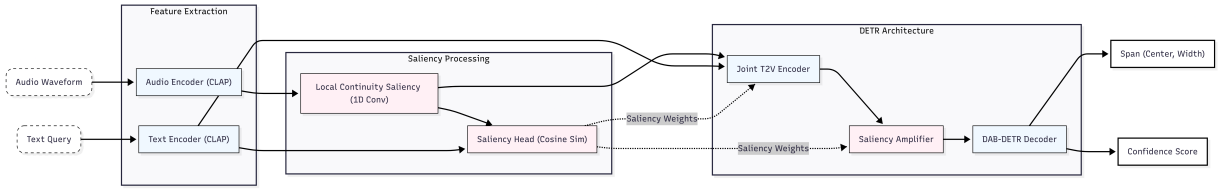

Figure 1: LCS-DETR architecture for language-based audio moment retrieval. The model processes raw audio and text queries through CLAP encoders, applies saliency-guided cross-modal gating to focus on query-relevant frames, enforces local temporal continuity via 1D convolution (kernel size 5 frames), and uses a transformer encoder-decoder with Hungarian matching to generate precise moment predictions.

4.3. Evaluation Metrics

We report Recall@1 (R1) at IoU thresholds 0.5, 0.7 and mean Average Precision (mAP) at 0.5, 0.75. Average mAP is computed across IoU thresholds [0.5:0.95:0.05].

5. RESULTS

5.1. Main Results

Table 1 shows LCS-DETR significantly outperforms baseline DETR across all metrics on Clotho-Moment + Castella data.

Table 1: Performance comparison: Baseline DETR vs. LCS-DETR on Clotho-Moment + Castella. LCS-DETR achieves 4-6× improvements through saliency guidance and Focal Loss.

Metric	Baseline	LCS-DETR
R1@0.5	25.61%	80.29%
R1@0.7	13.59%	73.08%
mAP (avg)	12.06%	70.26%
mAP@0.5	23.60%	84.64%
mAP@0.75	10.72%	74.02%

Key findings:

- R1@0.7 improvement: +59.49 percentage points (+437.8%)
- mAP@0.75 improvement: +63.30 percentage points (+590.1%)
- Strong performance at strict IoU thresholds indicates precise boundary localization

5.2. Performance Across IoU Thresholds

Performance remains robust even at very strict IoU thresholds, indicating the model learns precise temporal boundaries rather than approximate localization.

6. DISCUSSION

The substantial improvements of LCS-DETR over baseline DETR (4–6× gains) demonstrate that both saliency guidance and Focal Loss directly address core limitations in audio moment retrieval:

1. **Saliency gate captures semantic relevance:** Cosine-similarity based weighting focuses transformer attention on acoustically and semantically coherent regions, eliminating the need for exhaustive frame-level processing.

Table 2: Detailed performance breakdown across IoU thresholds, showing graceful degradation from lenient (0.5) to strict (0.95) evaluation.

IoU Threshold	mAP (%)
0.50	84.64
0.55	83.23
0.60	81.83
0.65	79.76
0.70	77.59
0.75	74.02
0.80	69.74
0.85	63.35
0.90	54.29
0.95	34.17

2. **Convolution smoothing prevents fragmentation:** The 5-second context window aligns with typical acoustic event durations, ensuring saliency masks reflect temporal coherence rather than frame-by-frame noise.
3. **Focal Loss corrects class imbalance:** With $\gamma = 2$, background predictions (the majority class) are exponentially down-weighted, forcing the model to prioritize boundary precision over safe background predictions.

The graceful performance degradation across IoU thresholds (84.64% at 0.5 down to 34.17% at 0.95) reflects realistic boundary precision challenges while maintaining competitive retrieval capability at practical thresholds (73.08% R1@0.7).

7. CONCLUSION

This paper proposes LCS-DETR, a saliency-guided Detection Transformer for language-based audio moment retrieval. By integrating cross-modal saliency gating, local temporal continuity via depthwise convolution, and Focal Loss-based class imbalance handling, we achieve substantial improvements over baseline temporal grounding approaches. Experimental validation on Clotho-Moment and Castella datasets demonstrates 4.8× and 4.4× improvements in mAP and R1@0.7 respectively.

Future work includes: (i) multi-scale saliency hierarchies for handling events at different temporal resolutions, (ii) cross-dataset generalization analysis, and (iii) comparison with other VMR-inspired architectures adapted to audio.

8. REFERENCES

[1] H. Munakata, T. Imamura, T. Nishimura, and T. Komatsu, "CASTELLA: Long audio dataset with captions and temporal boundaries," 2026.

[2] H. Munakata, T. Nishimura, S. Nakada, and T. Komatsu, "Language-based audio moment retrieval," in *Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process. (ICASSP)*, 2025, pp. 1–5.

[3] Y. e. a. Wu, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*, 2020.

[6] 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 (NeurIPS)*, 2021.

[7] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[8] P. Lee and H. Byun, "Bam-detr: Boundary-aligned moment detection transformer for temporal sentence grounding in videos," in *European Conference on Computer Vision (ECCV)*, 2024.

[9] H. Ma, G. Wang, F. Yu, Q. Jia, and S. Ding, "Ms-detr: Towards effective video moment retrieval and highlight detection by joint motion-semantic learning," *arXiv preprint arXiv:2507.12062*, 2025.

[10] M. e. a. Gygli, "Saliency-guided detr for moment retrieval and highlight detection," *arXiv preprint arXiv:2410.01615*, 2024.

[11] S. e. a. Wang, "Mf2summ: Multimodal fusion for video summarization with temporal alignment," *arXiv preprint arXiv:2506.10430*, 2025.

[12] F. e. a. Wang, "Enhancing audio-visual spiking neural networks through semantic-alignment and cross-modal residual learning," *arXiv preprint arXiv:2502.12488*, 2025.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[14] H. Rezatofighi, N. Tsoi, J. Gwak, S. Savarese, and M. Campbell, "Generalized intersection over union: A metric and a loss for overlapping boxes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[15] Z. Shou, D. Wang, and S.-F. Chang, "Cdc: Convolutional-deconvolutional networks for precise temporal action localization in untrimmed videos," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[16] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[17] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019.