TSEL: Temporal Semantic Evidence Learning for Language-Based Audio Moment Retrieval

Xiaokai Zhang Xiang Shang
Xi'an Jiaotong-Liverpool University
Emails: Xiaokai.Zhang24@student.xjtlu.edu.cn, Xiang.Shang24@student.xjtlu.edu.cn

***Abstract*—This technical report describes our submissions to DCASE 2026 Challenge Task 6, Audio Moment Retrieval from Long Audio. The task requires returning temporal windows in a long audio recording that match a natural-language query, with the official ranking emphasizing the top-ranked prediction at Recall1@0.7. Our system is based on Temporal Semantic Evidence Learning (TSEL): instead of directly regressing a single start/end pair, it first predicts query-conditioned temporal evidence and then decodes candidate windows. We package four systems: a conservative internal MS-CLAP evidence baseline** (task6-1)**, a TSEL/SBEC evidence system with a learned evidence decoder** (task6-2)**, a candidate-level evidence fusion system** (task6-3)**, and a risk-aware evidence scoring system** (task6-4)**. On the CASTELLA development-testing split, these systems obtain 16.11%, 21.97%, 23.42%, and 23.59% Recall1@0.7, respectively. TSEL-ECF is our practical main system; TSEL-RAES is retained as a risk-analysis variant. A clean five-seed validation rerun gives 29.83±0.45% Recall1@0.7 for ECF and 28.92±0.24% for RAES, with RAES producing fewer harmful interventions. The official evaluation set has no public ground truth, so no official hidden score is claimed here.**

***Index Terms*—audio moment retrieval, language-based audio retrieval, temporal grounding, CLAP, evidence learning, hard-negative mining**

I. Introduction

Audio Moment Retrieval (AMR) localizes the segment in a long audio recording that corresponds to a text query. This differs from clip-level audio-text retrieval because a system must identify not only whether a sound is present, but also when the relevant event occurs. DCASE 2026 Task 6 evaluates this ability with ranking metrics over predicted temporal windows, and the primary ranking metric is Recall1@0.7: the top-ranked predicted window is considered correct only if its intersection-over-union (IoU) with a ground-truth moment is at least 0.7.

Our main design choice is to treat AMR as evidence-driven temporal grounding. Given audio features $\mathbf{A}$, a text query $\mathbf{q}$, and audio duration T, the model predicts a dense evidence curve $S(t, \mathbf{q})$ and candidate windows $\hat{\mathcal{W}} = \{(\hat{s}_i, \hat{e}_i, \hat{c}_i)\}_{i=1}^{K}$. Candidate selection then chooses the most reliable window instead of relying on one direct start/end regression. This design was motivated by two common failure modes: semantic misses, where a model selects a plausible but wrong acoustic region, and boundary drift, where the correct event is found but the predicted window is too wide or shifted.

II. Method and Submitted Systems

We use TSEL as the umbrella name for an evidence-driven temporal grounding pipeline. MS-EB denotes our internal MS-CLAP evidence baseline, not the organizer-provided DETR baseline. SBEC denotes the Semantic-to-Boundary Evidence Curriculum, LD denotes the learned candidate decoder, ECF denotes Evidence Candidate Fusion, and RAES denotes Risk-Aware Evidence Scoring.

A. Evidence Target and MS-EB

For each query, let $G = \{(s_j, e_j)\}$ be the set of ground-truth moments and c_t be the center time of frame t. The soft evidence target is defined as a max over moment-wise targets:

$$g_t = \max_{(s,e)\in G} h_t(s, e), \tag{1}$$

$$h_t(s, e) = \begin{cases} 1, & s \le c_t \le e, \\ \alpha e^{-d_t^2/(2\sigma_e^2)}, & \text{otherwise}, \end{cases} \tag{2}$$

where $d_t = d(c_t, [s, e])$ is the distance to the nearest boundary, $\alpha = 0.25$, and $\sigma_e = 2.0$ frames. The value α is a fixed empirical smoothing factor that gives weak supervision to boundary-adjacent context while keeping inside-window frames dominant. Start and end targets are Gaussian boundary distributions centered at all annotated starts and ends; multiple moments are combined by max pooling and then normalized.

The evidence network predicts an evidence logit, a start logit, and an end logit for each frame:

$$z_t, y_t^s, y_t^e = f_\theta(\mathbf{A}_t, \mathbf{q}). \tag{3}$$

The evidence curve used by the decoders is $S_t = \sigma(z_t)$. Training uses masked BCE-with-logits (BCEWL):

$$\mathcal{L}_{\text{ev}} = \mathcal{L}_{\text{BCE}} + \lambda_b \mathcal{L}_{\text{bnd}}, \tag{4}$$

$$\mathcal{L}_{\text{BCE}} = \text{BCEWL}_{\text{mask}}(z, g), \tag{5}$$

$$\mathcal{L}_{\text{bnd}} = \frac{\text{CE}_{\text{mask}}(y^s, p^s) + \text{CE}_{\text{mask}}(y^e, p^e)}{2}, \tag{6}$$

with $\lambda_b = 0.5$. The BCE term uses a positive-frame weight computed from the valid-frame positive/negative ratio. task6-1 uses this MS-EB branch with heuristic evidence decoding.

B. SBEC and Learned Decoder

task6-2 adds SBEC and a learned decoder. Semantic false peaks (SFP) are mined high-evidence windows that have low overlap with any ground-truth window. Boundary-width

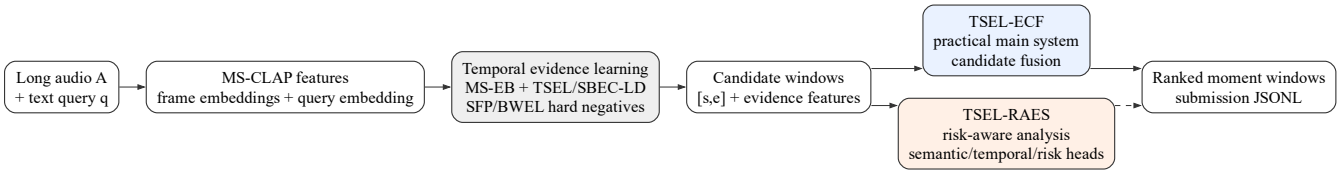

Fig. 1. Overview of the submitted Temporal Semantic Evidence Learning (TSEL) pipeline. The submitted systems share MS-CLAP features and temporal evidence learning. TSEL-ECF is the practical main candidate-fusion system, while TSEL-RAES is retained as a risk-aware analysis variant.

evidence learning (BWEL) uses hard negative windows that overlap an event but miss strict localization because they are shifted, too wide, or too narrow. The type-aware hard-negative term encourages the mean evidence inside ground-truth windows and the boundary logits at true starts/ends to outrank the corresponding hard negatives. Let $\bar{S}(w)$ be the mean evidence inside window w and let $B(w) = \frac{1}{2}(y^s_{\lfloor s_w \rfloor} + y^e_{\lfloor e_w \rfloor})$ be its boundary score. For a matched ground-truth window g and mined semantic and boundary negatives w^-_{sem} and w^-_{bdry}, we use

$$\mathcal{L}_{\text{SFP}} = \max(0, m_s - \bar{S}(g) + \bar{S}(w^-_{\text{sem}})), \tag{7}$$

$$\mathcal{L}_{\text{BWEL}} = \max(0, m_b - B(g) + B(w^-_{\text{bdry}})), \tag{8}$$

and add $\lambda_s \mathcal{L}_{\text{SFP}} + \lambda_w \mathcal{L}_{\text{BWEL}}$ to the evidence loss. The submitted setting uses $m_s = 0.1$, $m_b = 0.2$, top-3 semantic negatives, top-3 boundary negatives, $\lambda_s = 0.15$, and $\lambda_w = 0.1$.

Candidate windows are generated by deterministic evidence decoders: current start/end output, threshold-connected components at the mean/60th/70th percentiles, peak expansion, peak-drop expansion, fixed multiscale windows around evidence peaks with lengths 10/20/40/80/150 frames, and dense contrast-scored windows. For each candidate $w_i = (s_i, e_i)$, the quality target is

$$q_i = \max_{g \in G} \text{IoU}(w_i, g). \tag{9}$$

The learned decoder uses evidence-shape features, relative duration, peak position, candidate source, and source rank. It is trained with

$$\mathcal{L}_{\text{LD}} = \text{MSE}(\sigma(a_i), q_i) + \lambda_r \mathcal{L}_{\text{pair}}, \tag{10}$$

where $\mathcal{L}_{\text{pair}}$ is a within-query hinge ranking loss ($\lambda_r = 0.2$). Candidate lists are deduplicated with high-threshold NMS during candidate construction and finally filtered with temporal NMS at IoU 0.7, retaining the top 10 windows for each query.

C. ECF and RAES

task6-3 is TSEL-ECF. It keeps candidates from both the MS-EB branch and the SBEC/LD branch in a shared pool. The fusion scorer is an absolute candidate-quality predictor with anchor-relative features, not only a binary "better than anchor" classifier. Features include branch identity, branch confidence, branch rank, evidence-shape statistics, cross-branch agreement, and differences from the MS-EB anchor. All candidate features are standardized using the training candidate mean and standard deviation, then the same statistics are applied to validation, development-testing, and evaluation outputs.

task6-4 is TSEL-RAES. It uses the same candidate pool, but adds heads for anchor reliability, semantic gain, temporal gain, anchor risk, semantic risk, temporal risk, strict gain, and utility. Its score penalizes high-disagreement candidates when the anchor appears reliable. RAES is therefore reported as a risk-modeling variant, not as a statistically proven improvement over ECF; detailed RAES-head ablations are left for extended analysis rather than used as the main challenge claim.

III. Data, Features, and Rules

Our submitted systems are trained from organizer-provided MS-CLAP features. In this submission, the evidence models use the CASTELLA development-training split, containing 2182 annotated queries. Checkpoint and hyperparameter selection use the CASTELLA development-validation split, containing 352 annotated queries. Candidate decoders and candidate scorers are trained from candidate-label pairs generated on the development-training split and selected on development-validation. The CASTELLA development-testing split, containing 1347 annotated queries, is used only for the local results reported in Table I.

The DCASE evaluation set contains 177 queries over 100 audio recordings with provided MS-CLAP features and a submission template, but no public ground-truth windows. It is used only for output generation and format checking. All four submitted systems use only the organizer-provided MS-CLAP features and query embeddings for the evaluation set. We do not annotate the evaluation set, do not make subjective judgments about hidden labels, do not use visual information from original videos, and do not use any LLM API.

IV. Experimental Results

Table I reports local annotated-split results. The first row is the official organizer baseline trained with CASTELLA and Clotho-Moment; the remaining rows are our submitted systems trained under the protocol described above. Because the training-data configurations are not identical, the organizer-baseline row is included for context rather than as a controlled architecture comparison. MS-EB is our internal evidence baseline rather than the organizer-provided DETR baseline. mAP(avg) is computed over IoU thresholds 0.50, 0.55, ..., 0.95.

Under these different training-data configurations, MS-EB is slightly lower than the official organizer baseline at Recall1@0.5 but higher at the stricter Recall1@0.7 threshold. This suggests that the dense evidence formulation is promising for stricter boundary matching, while not constituting a controlled

TABLE I

CASTELLA DEVELOPMENT-TESTING RESULTS. ECF AND RAES USE FIVE-SEED MEAN AND STANDARD DEVIATION FOR THE LIGHTWEIGHT CANDIDATE SCORER ONLY; EVIDENCE MODELS AND CANDIDATE POOLS ARE FIXED.

System	Submission	R1@0.5	R1@0.7	mAP(avg)	Role
Official DETR baseline	–	25.61	13.59	12.06	organizer baseline
MS-EB	task6-1	23.83	16.11	12.77	backup
TSEL/SBEC-LD	task6-2	33.04	21.97	17.37	single-branch
TSEL-ECF	task6-3	34.65±0.45	23.42±0.60	18.66±0.31	candidate fusion
TSEL-RAES	task6-4	35.13±0.24	23.59±0.28	18.53±0.08	risk-aware scoring

TABLE II

CLEAN FIVE-SEED VALIDATION RERUN WITH FIXED EVIDENCE AND CANDIDATE POOL. VALUES ARE MEAN±SAMPLE STANDARD DEVIATION OVER SEEDS 2026–2030.

System	R1@0.7	Top1 IoU	Changed	Regressed
MS-EB+LD	27.56	34.73	–	–
SBEC+LD	27.56	34.91	–	–
TSEL-ECF	29.83±0.45	36.74±0.68	253.4±15.6	79.2±6.5
TSEL-RAES	28.92±0.24	36.38±0.60	217.4±38.4	61.6±16.0

architecture-only comparison. TSEL/SBEC-LD improves Recall1@0.7 by 5.86 absolute points over MS-EB on the same development-testing split. ECF adds another 1.45 absolute points by deciding between MS-EB and SBEC/LD candidates at the candidate level.

RAES obtains a numerically higher mean Recall1@0.7 than ECF, but the difference is only 0.17 points and is small relative to seed variation. We therefore do not claim that RAES is clearly better than ECF. For task6-2, the selected learned-decoder checkpoint reached 39.20% Recall1@0.5 and 30.11% Recall1@0.7 on validation, but the lower development-testing score indicates split sensitivity and possible decoder overfitting; the metadata reports development-testing results.

After completing the evaluation outputs, we reran the final ECF and RAES candidate scorers on a clean remote workspace using the same fixed MS-EB/SBEC evidence, decoder outputs, and candidate pool. This rerun is a validation stability check, not an official evaluation result. Table II shows that ECF is the stronger accuracy system, while RAES changes fewer examples and causes fewer regressions.

The validation rerun clarifies the roles of the final two submissions. ECF is the more accurate candidate-fusion system. RAES is better interpreted as a risk-aware analysis variant: it makes fewer replacements, reduces harmful regressions, and lowers good-anchor regressions from 13.0 to 8.6 cases on average. This is why we keep ECF as the practical main system and RAES as an explanatory risk-modeling submission.

V. OFFICIAL EVALUATION OUTPUTS

The official evaluation set is used only for output generation. All submitted files contain 177 JSONL rows and use the required fields for query ID, query text, audio duration, video ID, and predicted relevant windows. Each query has at least one predicted window; additional windows are sorted by confidence and are included for mAP computation. Predicted windows are clamped to the audio duration before writing the output file.

Because the evaluation set has no public labels, we only run format and boundary-sanity diagnostics on these outputs. The number of top-ranked windows ending exactly at the audio duration is 59/177 for task6-1, 122/177 for task6-2, 79/177 for task6-3, and 136/177 for task6-4. These counts are computed after duration clamping and therefore diagnose output geometry, not hidden accuracy. They suggest that ECF is the more stable fusion output, while RAES can still over-trust long or end-clamped candidates despite its risk heads.

The package contains one task-level report and four system folders:

- task6/ contains the technical report PDF.
- Four numbered system folders each contain one output JSONL file and one metadata YAML file.

VI. DISCUSSION

The main limitation is that all submitted systems still depend on candidate decoding from one-dimensional evidence curves. While the learned decoder and candidate fusion improve strict localization, some predictions remain biased toward long windows or clip boundaries. The packaged progression supports the combined SBEC-and-learned-decoding stage and candidate-level fusion, but it does not isolate every component or compare against all simple fusion baselines. RAES shows that explicit risk estimation is a useful research direction on annotated local splits, but the evaluation-output diagnostics indicate that its quality guard is not yet sufficient for all repeated or ambiguous scenes.

VII. CONCLUSION

We presented an evidence-based AMR system for DCASE 2026 Task 6. The single-branch TSEL/SBEC learned decoder improves development-testing Recall1@0.7 from 16.11% for the MS-CLAP evidence baseline to 21.97%, and candidate-level fusion further improves it to 23.42%. The method remains reproducible from provided MS-CLAP features and does not use prohibited evaluation-set annotation, visual information, or LLM APIs. A clean five-seed validation rerun further supports ECF as the stronger accuracy system and RAES as a more conservative risk-aware variant.

REFERENCES

[1] DCASE Community, "DCASE 2026 Challenge Task 6: Audio Moment Retrieval from Long Audio," 2026. [Online]. Available: https://dcase.community/challenge2026/task-audio-moment-retrieval-from-long-audio

[2] H. Munakata, T. Nishimura, S. Nakada, and T. Komatsu, "Language-based Audio Moment Retrieval," in *Proc. ICASSP*, 2025.

[3] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *Proc. ICASSP*, pp. 336–340, 2024.

[4] H. Munakata, T. Imamura, T. Nishimura, and T. Komatsu, "CASTELLA: Long Audio Dataset with Captions and Temporal Boundaries," arXiv:2511.15131, 2026.

[5] J. Lei, T. L. Berg, and M. Bansal, "Detecting Moments and Highlights in Videos via Natural Language Queries," in *Proc. NeurIPS*, 2021.

[6] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum Learning," in *Proc. ICML*, 2009.