STELA: STABLE TEMPORAL EVIDENCE LEARNING FOR AUDIO FORGERY DETECTION AND LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio forgery detection and localization aims to identify spoofed frames in manipulated speech and determine the forgery boundaries. Existing methods have primarily improved frame discrimination through stronger feature learning and boundary-aware modeling. However, local acoustic fluctuations within manipulated regions can still compromise prediction, leading to fragmented and discontinuous localization. We propose STELA, a **S**table **T**emporal **E**vidence **L**earning framework for **A**udio forgery detection and localization to address this challenge. Specifically, for stable evidence learning, a segment-oriented acoustic augmentation and Paired State Modeling (PSM) provide complementary constraints against local acoustic variability. The augmentation introduces segment-level perturbations to discourage the detector from conflating incidental acoustic variations with manipulation boundaries. PSM further regularizes the representations within each utterance, encouraging acoustically varied frames sharing the same authenticity state to remain consistent. By mitigating representation drift and enhancing temporal coherence, these complementary constraints provide a more reliable basis for subsequent manipulation localization. Building on this improved evidence, Structured Manipulation Localization (SML) extends threshold-based proposal generation with structured scoring of alternative interval hypotheses, using interior, boundary, and contextual evidence to recover coherent manipulation intervals from fragmented predictions. Extensive experiments on PartialEdit, PartialSpoof, HAD, and LAV-DF demonstrate that STELA achieves state-of-the-art performance in both forgery frame detection and temporal localization, with strong generalization to unseen conditions.

1 INTRODUCTION

Recent advances in text-to-speech, voice conversion, and neural speech editing have made speech manipulation increasingly accessible (Li et al., 2025). Speech synthesis and editing systems such as A^3T and VoiceCraft support localized content replacement (Bai et al., 2022; Peng et al., 2024), motivating dedicated evaluation of neural speech edits (Zhang et al., 2025). By replacing only a few words while leaving the remainder of an utterance unchanged, partial manipulation can alter semantic content while preserving most of the original signal. This poses a challenge beyond utterance-level spoof detection: a forensic system must not only determine whether manipulation is present, but also identify when it occurs. This need has motivated the development of partially spoofed speech datasets and fine-grained localization benchmarks, including HAD, PartialSpoof, and, more recently, PartialEdit for neural speech editing (Yi et al., 2021; Zhang et al., 2023a; 2025).

Existing methods typically perform frame-level authenticity prediction and then infer manipulation intervals from frame or boundary evidence (Cai & Li, 2024; Xie et al., 2024; Zhong et al., 2024; Wu et al., 2024). Prior analyses identify sensitivity to transition artifacts (Liu et al., 2024; Mao et al., 2026); in our qualitative comparisons, continuous manipulations can also yield fragmented predictions, as illustrated in Fig. 1. Our analysis suggests that such fragmentation is closely associated with acoustic fluctuations within segments, which may be mistaken for authenticity changes. From a data perspective, augmentation strategies inherited from utterance-level anti-spoofing provide limited exposure to localized acoustic variations; from a representation perspective, existing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

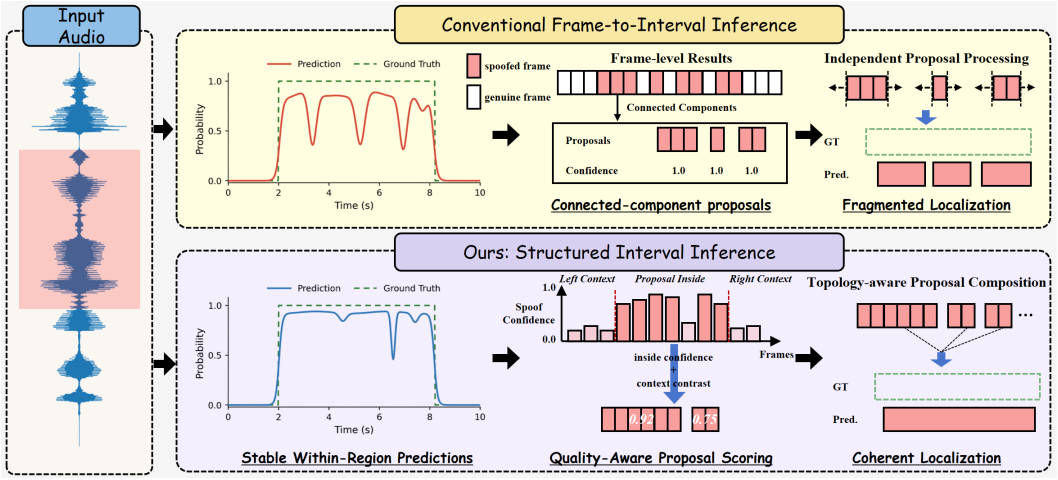

Figure 1: Comparison between conventional connected-component localization and STELA. Conventional inference preserves fragmented frame predictions, whereas STELA produces more coherent manipulation intervals through structured proposal modeling.

supervision focuses on authenticity discrimination, boundary awareness, or segment-level consistency, but lacks explicit constraints on embedding consistency under such acoustic variations. Beyond frame-level discrimination, the conversion of frame results into manipulation intervals remains underexplored. Conventional pipelines obtain candidate regions from posteriors or detected boundaries, largely fixing the interval structure at an early stage (Wu et al., 2024; Rimon et al., 2026). Subsequent refinement can rarely reunify them into a coherent interval, as illustrated in Fig. 1. Consequently, improvements in frame discrimination do not necessarily translate into equally reliable temporal localization. Robust partial-spoof localization therefore requires greater resilience to local acoustic variability while preserving coherent temporal evidence for interval inference.

Motivated by these observations, we propose a unified framework for stable frame-level detection and structured temporal localization. For forgery frame detection, our key principle is to prevent local acoustic variations from being confused with authenticity changes. At the data level, we introduce segment-oriented acoustic augmentation that applies diverse augmentation operations to randomly selected subsegments. We further construct boundary-free interior samples from extended manipulated regions, encouraging the model to focus on within-region evidence rather than manipulation boundaries. At the representation level, we introduce Paired State Modeling (PSM), which dynamically estimates a pair of utterance-specific state prototypes and encourages frames sharing the same authenticity state to form consistent representations within each utterance. These data- and representation-level designs jointly improve the stability of frame-level predictions. For temporal localization, we develop Structured Manipulation Localization (SML), which performs structured inference over alternative interval hypotheses rather than relying on a fixed proposal topology. By jointly considering candidate quality, boundary information, and temporal consistency, SML recovers coherent manipulation regions from fragmented predictions.

Our main contributions are summarized as follows:

1. We develop a **S**table **T**emporal **E**vidence **L**earning framework for **A**udio forgery detection and localization (STELA) that jointly addresses two coupled challenges: unstable frame-level predictions caused by acoustic variability and fragmented frame-to-interval conversion. The framework consistently improves both detection and localization across multiple benchmarks.

2. We introduce complementary data and representation strategies for robust frame-level evidence learning. Segment-oriented acoustic augmentation with randomized cross-fading and boundary-free interior sampling expose the detector to diverse within-region variations, while Paired State Modeling uses prototype-guided supervision to maintain consistent representations for frames sharing the same authenticity state.

3. We propose SML, which performs structured inference over alternative interval hypotheses instead of directly inheriting fragmented frame predictions. By jointly considering candidate quality, boundary evidence, and temporal consistency, SML recovers coherent manipulation intervals and substantially improves temporal localization accuracy.

2 Related Work

2.1 Frame-Level Partial-Spoof Detection

Utterance-level anti-spoofing includes spectro-temporal graph models such as AASIST (Jung et al., 2022) and self-supervised front ends with data augmentation (Tak et al., 2022b). More generally, wav2vec 2.0 and HuBERT provide pretrained speech representations (Baevski et al., 2020; Hsu et al., 2021). Partial spoofing requires within-utterance decisions, explored through early investigations and joint utterance–segment learning (Zhang et al., 2021b;a), with benchmarks including HAD, PartialSpoof, and PartialEdit (Yi et al., 2021; Zhang et al., 2023a; 2025). Cai & Li (2024) integrate frame-level boundary and deepfake detection, while TDL combines embedding similarity learning with temporal convolution (Xie et al., 2024). BAM regulates frame interactions through boundary-aware attention (Zhong et al., 2024), and BFC-Net combines inter-frame differences with boundary–frame graph modeling (Zhou et al., 2025). SAIL investigates multi-scale semantic inconsistencies and robustness to acoustic interference (Cao et al., 2025). Interpretability analysis and SAL examine transition-artifact dependence, with SAL introducing segment-aware supervision and augmentation (Liu et al., 2024; Mao et al., 2026).

Nevertheless, existing studies mainly emphasize authenticity discrimination, boundary awareness, or segment structural consistency, leaving the effects of natural within-segment acoustic fluctuations relatively underexplored. Brief pauses and local phonetic or prosodic variations are inherent in speech but may disturb learned representations and be mistaken for authenticity changes, causing unstable frame predictions and fragmented manipulation regions. Common augmentation strategies largely focus on utterance-level acoustic corruption, post-processing, or segment composition, with less emphasis on localized, label-preserving acoustic variations. Meanwhile, explicit constraints on intra-utterance embedding consistency under these variations remain less explored. These limitations motivate complementary data- and representation-level strategies for robustness to natural acoustic variability in partial-spoof detection.

2.2 Temporal Manipulation Localization

Partial-spoof localization converts frame evidence into manipulation intervals. CFPRF generates coarse proposals and refines their confidence and boundaries (Wu et al., 2024). Split and Conquer partitions utterances using detected boundaries and classifies the resulting segments (Rimon et al., 2026), while LOCO progressively refines localization under utterance-level weak supervision (Wu et al., 2025). In audio–visual localization, LAV-DF introduces boundary-aware temporal forgery detection (Cai et al., 2023), whereas UMMAFormer develops multimodal-adaptive temporal modeling and a cross-attention feature pyramid (Zhang et al., 2023b). GEM-TFL combines weakly supervised decomposition with temporal–semantic proposal refinement (Zhu et al., 2026).

Temporal action localization provides precedents for candidate generation and quality estimation: BSN combines boundary-sensitive proposal generation with proposal evaluation, while BMN jointly models boundary probabilities and proposal confidence (Lin et al., 2018; 2019). SML retains the original connected components while adding endpoint variants and complete spans across adjacent fragments. Its region-structured scoring compares these hypotheses using evidence from candidate interiors, boundaries, and external contexts.

3 Proposed Method

3.1 Overview

Given a partially spoofed utterance, our goal is to predict frame-level authenticity and further localize manipulated temporal intervals. As illustrated in Fig. 2, the proposed framework consists of

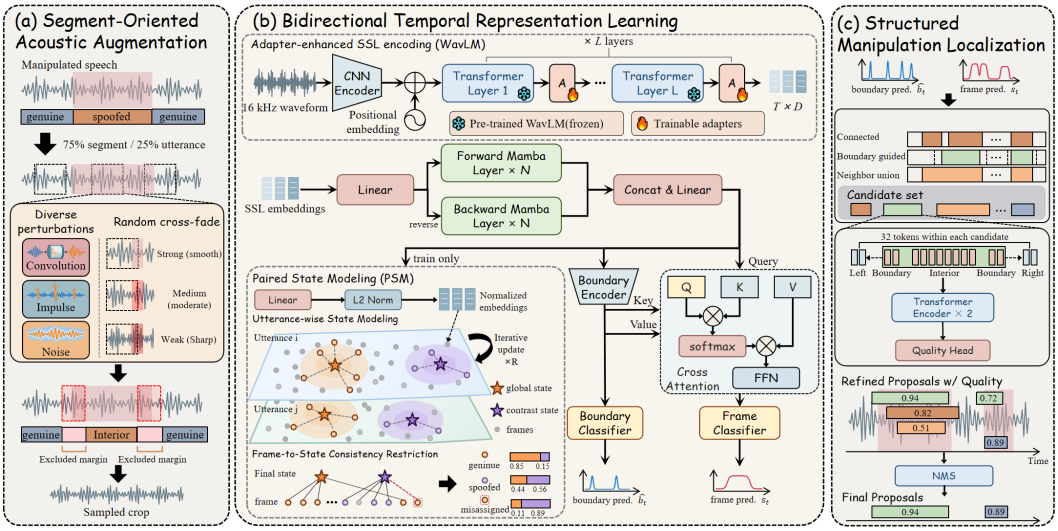

Figure 2: Overview of STELA. **(a)** Segment-oriented augmentation and interior sampling expose local acoustic variability. **(b)** Bidirectional temporal modeling captures contextual dependencies, while PSM regularizes same-state representations within each utterance. **(c)** SML jointly scores and assembles candidate proposals for coherent localization.

segment-oriented acoustic augmentation, bidirectional temporal representation learning, and Structured Manipulation Localization (SML). The first two components improve the stability of frame-level evidence through localized acoustic augmentation, boundary-free interior sampling, and Paired State Modeling (PSM), while SML converts the resulting frame and boundary evidence into coherent manipulation intervals through structured reasoning over candidate proposals.

3.2 Segment-Oriented Acoustic Augmentation

Given an input waveform $\mathbf{x}$ with frame-level authenticity labels $\mathbf{y}$, we introduce segment-oriented acoustic augmentation to enrich local acoustic variability and reduce reliance on manipulation-boundary cues. We randomly sample local waveform segments $\mathcal{S} = \{[s_k, e_k]\}_{k=1}^{K}$ and independently augment them using the acoustic perturbation algorithms from RawBoost (Tak et al., 2022a). For each selected segment,

$$\tilde{\mathbf{x}}_{s_k:e_k} = \mathcal{A}_k\left(\mathbf{x}_{s_k:e_k}\right), \tag{1}$$

where $\mathcal{A}_k(\cdot)$ denotes a sampled augmentation operation or their composition. Cross-fading with varying strengths is applied at augmentation boundaries to smooth pseudo-transitions and reduce spurious boundary cues.

Since local prediction interruptions occur more frequently within extended manipulated regions, we additionally construct boundary-free interior samples. For a manipulated interval $\mathcal{I} = [s, e]$, we exclude a margin m from both boundaries,

$$\mathcal{I}_{\text{int}} = [s + m, e - m], \tag{2}$$

and recombine the resulting interior segment with the original waveform to form a new training sample.

3.3 Bidirectional Temporal Representation Learning

We employ WavLM (Chen et al., 2022) as the SSL encoder and keep its pre-trained backbone frozen, with lightweight bottleneck adapters inserted for task adaptation. This adaptation strategy provides competitive in-domain performance and strong average cross-domain generalization without full-backbone fine-tuning; detailed comparisons with alternative WavLM adaptation schemes are provided in Appendix A.1. The SSL embeddings are then modeled by forward and backward Mamba layers (Gu & Dao, 2024), whose outputs are fused into the bidirectional temporal representation

$\mathbf{H} = \{\mathbf{h}_t\}_{t=1}^{T}$. A boundary encoder predicts manipulation boundaries, and its features are incorporated into $\mathbf{H}$ through cross-attention before frame-level authenticity classification.

Paired State Modeling. Beyond frame and boundary supervision, we introduce Paired State Modeling (PSM) to encourage intra-utterance state consistency. Each frame representation is projected and ℓ_2-normalized as

$$\mathbf{e}_t = \mathrm{Norm}(\mathbf{W}_p \mathbf{h}_t). \tag{3}$$

For each utterance, PSM dynamically initializes two unordered state prototypes. The first is obtained from the normalized coordinate-wise median of all frame embeddings, while the second is initialized by the frame farthest from the first:

$$\mathbf{s}_1^{(0)} = \mathrm{Norm}\left(\mathrm{Median}_t(\mathbf{e}_t)\right), \qquad \mathbf{s}_2^{(0)} = \mathbf{e}_{\arg\max_t (1-\mathbf{e}_t^\top \mathbf{s}_1^{(0)})}. \tag{4}$$

The two states are iteratively refined through soft frame-to-state assignment and weighted aggregation:

$$a_{t,k}^{(r)} = \mathrm{Softmax}_k\left(\frac{\mathbf{e}_t^\top \mathbf{s}_k^{(r)}}{\tau}\right), \qquad \mathbf{s}_k^{(r+1)} = \mathrm{Norm}\left(\sum_{t=1}^{T} a_{t,k}^{(r)} \mathbf{e}_t\right). \tag{5}$$

Since the two states have no predefined genuine or spoofed identity, we employ utterance-level permutation-invariant supervision:

$$\mathcal{L}_{\mathrm{PSM}} = \min\left[\mathrm{CE}(\mathbf{Z}, \mathbf{y}), \mathrm{CE}(\mathbf{Z}, 1-\mathbf{y})\right], \tag{6}$$

where $\mathbf{Z}$ contains the final state logits. Unlike global class-compactness objectives, PSM models the relative state structure within each utterance and encourages frames sharing the same authenticity state to form consistent representations. PSM is used only during training.

Finally, the overall frame-level training objective is

$$\mathcal{L} = \lambda_{\mathrm{bd}}\mathcal{L}_{\mathrm{bd}} + \lambda_{\mathrm{frame}}\mathcal{L}_{\mathrm{frame}} + \lambda_{\mathrm{PSM}}\mathcal{L}_{\mathrm{PSM}}, \tag{7}$$

where $\mathcal{L}_{\mathrm{bd}}$ and $\mathcal{L}_{\mathrm{frame}}$ denote the focal loss (Lin et al., 2017) for boundary prediction and the MSE loss for frame-level authenticity prediction, respectively.

3.4 Structured Manipulation Localization

Given the frame-level spoof posterior $\mathbf{p}^F$, SML converts dense frame predictions into manipulation intervals through candidate proposal generation and proposal-level quality estimation. We first threshold $\mathbf{p}^F$ and group consecutive positive frames into contiguous intervals, yielding an initial set of proposals. To handle potential fragmentation and misalignment caused by local posterior fluctuations, we augment the initial proposals with two complementary types of candidates. Boundary-adjusted proposals locally refine either endpoint of an initial proposal to account for boundary uncertainty, while merge proposals form a single interval from the start of one proposal to the end of a subsequent proposal, thereby covering the intermediate gaps and providing a complete-interval hypothesis for fragmented manipulations. The initial proposals are retained alongside these candidates for subsequent quality estimation.

To evaluate these competing proposals, we fuse the detection features $\mathbf{F}_{\mathrm{DA}}$, boundary-aware features $\mathbf{F}_{\mathrm{BA}}$, spoof posterior $\mathbf{p}^F$, and boundary confidence $\mathbf{b}$ into a frame-wise localization representation $\mathbf{Z}$. For each proposal c_i, a fixed-length token sequence is sampled from five regions covering the left context, start-boundary neighborhood, interval interior, end-boundary neighborhood, and right context. Positional and region embeddings are added to encode the relative location and semantic role of each token, after which a lightweight Transformer (Vaswani et al., 2017) models the resulting sequence within each candidate:

$$\mathbf{X}_i = \mathrm{Sample}(\mathbf{Z}, c_i) + \mathbf{E}_{\mathrm{pos}} + \mathbf{E}_{\mathrm{reg}}, \qquad \hat{q}_i = Q_\theta(\mathrm{Trans}(\mathbf{X}_i)), \tag{8}$$

where Q_θ denotes the proposal-level readout and quality predictor. This representation enables the predicted quality to jointly reflect interior manipulation support, boundary alignment, and contrast with the surrounding context, rather than relying on pooled frame confidence alone.

Candidate quality is supervised by its maximum temporal IoU with the ground-truth intervals:

$$q_i^\star = \max_{g \in \mathcal{G}} \mathrm{IoU}(c_i, g), \qquad \mathcal{L}_{\mathrm{SML}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{BCE}(\hat{q}_i, q_i^\star), \tag{9}$$

Table 1: In-domain frame-level detection performance (%). Each entry is reported as EER / Acc. / F1. Best and second-best results for each metric are shown in **bold** and underline, respectively.

Method	PS	HAD	PE-E1	LAV-DF
CFPRF	9.98 / 92.14 / 93.90	0.18 / 99.91 / 99.95	0.82 / 99.45 / 99.65	2.18 / 98.80 / 99.37
BAM	11.72 / 90.11 / 92.42	0.18 / 99.93 / 99.96	0.54 / 99.57 / 99.73	1.19 / 99.45 / 99.71
TDL†	10.43 / 89.51 / 89.82	0.90 / 99.31 / 99.57	6.30 / 94.51 / 96.37	1.47 / 99.41 / 99.69
SAL	3.62 / 96.81 / 97.43	0.47 / 99.59 / 99.75	0.36 / 99.75 / 99.84	0.77 / 99.56 / 99.77
Ours	**3.32 / 97.01 / 97.59**	**0.002 / 99.999 / 99.999**	**0.17 / 99.88 / 99.92**	**0.51 / 99.62 / 99.80**

†TDL operates on its native 160 ms temporal grid.

Table 2: Cross-domain frame-level detection performance (%). All models are trained on PE-E1 and directly evaluated on unseen target domains. Results are reported as Acc. / F1. Best and second-best results are shown in **bold** and underline, respectively.

Method	PE-E2	PS	HAD	ADD2023	Avg.
CFPRF	81.55 / 87.44	70.89 / 79.52	86.60 / 91.46	63.93 / 75.69	75.74 / 83.53
BAM	68.40 / 76.52	70.51 / 79.29	29.07 / 32.89	26.68 / 30.44	48.66 / 54.78
SAL	86.32 / 91.00	**74.17 / 82.04**	85.45 / 90.68	68.49 / 79.38	78.61 / 85.77
Ours	**87.99 / 92.26**	69.53 / 79.92	**89.67 / 93.81**	**77.72 / 86.71**	**81.23 / 88.18**

where $\mathcal{G}$ denotes the ground-truth interval set. During inference, proposals are ranked by $\hat{q}_i$, followed by standard NMS to remove redundant overlaps.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate STELA on PartialSpoof (PS) (Zhang et al., 2023a), Half-truth Audio Detection (HAD) (Yi et al., 2021), PartialEdit (PE) (Zhang et al., 2025), LAV-DF (Cai et al., 2023), and ADD2023 (Yi et al., 2023). In-domain experiments are conducted on PS, HAD, PE, and LAV-DF following their respective protocols. For cross-domain evaluation, models are trained on PE-E1 and directly tested on PE-E2, HAD, PS, and ADD2023 to assess generalization across unseen manipulation conditions and datasets.

Baselines. We compare STELA with TDL (Xie et al., 2024), BAM (Zhong et al., 2024), CFPRF (Wu et al., 2024), SAL (Mao et al., 2026), and UMMAFormer (Zhang et al., 2023b). These methods are retrained from their public implementations using the same data splits while retaining their original model-specific settings when applicable. We additionally report the published PartialSpoof results of Split and Conquer (Rimon et al., 2026), for which no usable public implementation is available.

Evaluation Metrics. For frame-level detection, we report EER, Accuracy (Acc), and F1-score, with Acc and F1 computed at a fixed threshold of 0.5. For temporal localization, we report AP at multiple temporal IoU thresholds, mAP over $0.50{:}0.05{:}0.95$, and AR under different proposal budgets.

4.2 Comparison with State-of-the-Art Methods

In-domain frame-level detection. Table 1 compares frame-level detection performance under in-domain conditions. Our method achieves the lowest EER and the highest Acc. and F1 across all four datasets, consistently outperforming the competing methods. The improvement is particularly evident on PS, while performance on HAD is already close to saturation for most methods. Overall, these results demonstrate consistent gains in frame-level discrimination across different manipulation conditions.

Cross-domain frame-level detection. Table 2 evaluates cross-domain generalization by training all methods on PE-E1 and directly testing them on four target datasets without adaptation. Our method achieves the highest Acc. and F1 on PE-E2, HAD, and ADD2023, while SAL performs best

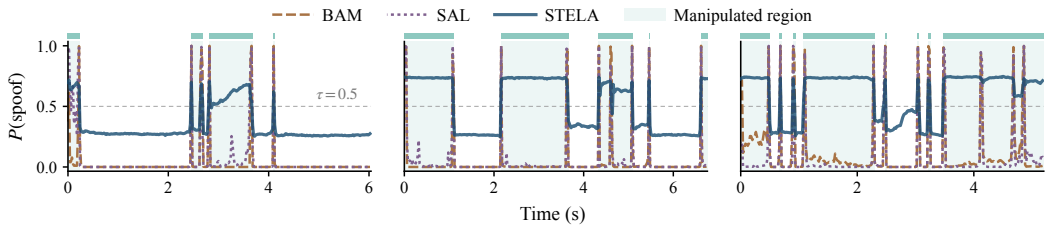

Figure 3: Qualitative comparison of frame-level spoof predictions on three representative utterances. Shaded areas denote ground-truth manipulated regions, and the horizontal dashed line indicates the decision threshold ($\tau = 0.5$).

on PS. The improvement is particularly evident on ADD2023, where our method improves Acc. from 68.49% to 77.72% and F1 from 79.38% to 86.71% over SAL. On HAD, it also increases Acc. and F1 to 89.67% and 93.81%, respectively. These consistent improvements in both in-domain and cross-domain evaluations demonstrate the effectiveness of the proposed learning strategies.

Qualitative frame-level analysis. To further examine the temporal behavior of frame-level predictions, we compare the spoof posteriors produced by BAM, SAL, and STELA on three representative utterances in Figure 3. For several long manipulated regions, BAM and SAL exhibit strong responses around manipulation boundaries, while their confidence substantially decreases in the region interior and is only occasionally restored by brief, isolated peaks. Such prediction patterns suggest sensitivity not only to boundary-localized cues but also to local acoustic variations within manipulated segments, making it difficult to sustain coherent spoof evidence over extended regions. In contrast, STELA maintains more stable and continuous spoof posteriors across the corresponding manipulated intervals. This behavior is consistent with the intended effect of the proposed segment-level acoustic augmentation and PSM, which jointly improve robustness to local acoustic variability and promote temporally consistent representations.

Temporal manipulation localization. Across the four benchmarks, our method achieves the highest mAP and remains particularly strong under stringent IoU criteria (Table 3). On PS, it improves mAP from 71.27% to 83.47% and AP@.95 from 56.72% to 77.11% over Split & Conquer. Although Split & Conquer is slightly better at smaller proposal budgets, our method achieves higher AR@20. On PE-E1, our method further improves mAP to 99.39% and AP@.95 to 98.03%, while HAD is already close to saturation. On LAV-DF, UMMAFormer performs better at looser IoU thresholds and larger recall budgets, whereas our method achieves higher AP@.9, AP@.95, and mAP. Overall, these results show that SML mainly improves precise temporal alignment under stringent overlap criteria rather than simply increasing proposal recall.

4.3 Ablation Study

To assess the contributions of the proposed designs, we conduct ablations on both frame-level learning and temporal localization. For frame-level learning, DA0–DA4 vary the combination of RawBoost-based acoustic augmentation and boundary-free interior cropping: DA0 disables both components; DA1 and DA2 use utterance-level and mixed segment-/utterance-level augmentation, respectively; and DA3 and DA4 further introduce interior cropping on top of DA1 and DA2. PSM is enabled throughout these data-level comparisons. With DA4 fixed, we further compare PSM with CRL and a baseline using neither representation regularizer. For temporal localization, we first establish direct proposal baselines by assigning different confidence scores to the same proposals obtained from contiguous spoofed frame predictions. We then evaluate the two main components aof SML: proposal generation and proposal quality estimation. Proposal generation is analyzed by ablating boundary adjustment and proposal merging, while proposal quality estimation is examined by replacing the learned quality estimator and Transformer-based proposal representation with simpler alternatives.

Frame-level detection ablation. Among the data-level variants in Table 4, DA4 achieves the strongest overall performance, supporting the complementarity of mixed utterance/segment augmentation and interior-focused training, although DA1 retains the highest cross-domain Macro-F1. With DA4 fixed, PSM improves all reported cross-domain averages over both CRL and the baseline

Table 3: Temporal manipulation localization performance (%). Best and second-best results are shown in **bold** and underline, respectively.

Dataset	Method	AP@.5	AP@.75	AP@.9	AP@.95	mAP	AR@1	AR@5	AR@20
PS	CFPRF-PRN	63.03	53.27	46.50	40.46	53.03	18.15	56.56	64.67
	BAM (CC)	51.69	46.21	43.26	41.42	46.74	20.73	61.18	72.60
	TDL (CC)*	8.68	2.47	0.37	0.07	3.60	6.78	15.99	16.66
	SAL (CC)	64.30	57.90	54.01	51.11	58.35	22.64	67.92	79.29
	UMMAFormer	52.99	31.89	17.69	9.04	33.09	17.37	39.57	55.53
	Split & Conquer †	78.51	72.06	63.93	56.72	71.27	**28.19**	**76.41**	87.42
	Ours	**86.45**	**83.99**	**80.54**	**77.11**	**83.47**	26.23	75.39	**90.43**
HAD	CFPRF-PRN	99.59	99.37	98.98	96.82	99.11	99.16	99.25	99.25
	BAM (CC)	98.64	98.38	97.92	96.34	98.21	99.24	99.44	99.44
	TDL (CC)*	97.09	61.79	10.25	2.07	59.63	72.37	72.44	72.44
	SAL (CC)	66.15	64.87	63.14	60.83	64.59	83.85	98.40	98.54
	UMMAFormer	99.98	99.86	98.01	88.17	98.49	98.68	98.84	98.86
	Ours	**100.00**	**99.99**	**99.99**	**99.91**	**99.99**	**99.99**	**99.99**	**99.99**
PE-E1	CFPRF-PRN	98.45	96.63	92.42	85.36	95.46	96.24	96.47	96.47
	BAM (CC)	95.18	94.04	91.15	86.43	93.20	97.48	98.25	98.33
	TDL (CC)*	77.88	24.60	1.18	0.10	35.05	52.41	52.92	52.93
	SAL (CC)	97.30	96.70	95.35	92.06	96.24	98.44	99.06	99.09
	UMMAFormer	98.57	91.78	57.79	19.44	81.90	84.43	88.13	89.11
	Ours	**99.75**	**99.54**	**99.01**	**98.03**	**99.39**	**98.96**	**99.39**	**99.41**
LAV-DF	CFPRF-PRN	77.14	76.47	76.10	74.17	76.35	76.40	76.40	76.40
	BAM (CC)	63.23	57.71	55.39	49.90	58.16	81.96	86.02	86.09
	TDL (CC)*	86.45	56.21	2.12	0.16	51.15	60.91	64.81	64.81
	SAL (CC)	89.05	86.70	83.20	74.78	85.72	87.37	93.43	93.45
	UMMAFormer	**97.29**	**95.67**	89.92	61.97	92.04	85.67	**94.89**	**96.14**
	Ours	95.49	92.66	**90.58**	**88.99**	**92.80**	**88.03**	93.57	93.57

CC denotes connected-component conversion of thresholded frame predictions with unit proposal confidence. *TDL operates on its native 160ms temporal grid. †Split & Conquer results are taken from the original paper and are available only on PS.

Table 4: Frame-level ablation results (%). Cross-domain results are averaged over PE-E2, PS, HAD, and ADD2023 using models trained on PE-E1. Best and second-best results within each comparison block are shown in **bold** and underline, respectively, with ties determined at the reported precision.

Setting	In-domain EER↓			Cross-domain Average			
	PS	HAD	PE-E1	EER↓	Acc.↑	F1↑	Macro-F1↑
Data-level design (PSM enabled)							
DA0	5.775	0.004	0.244	26.576	76.572	84.110	68.834
DA1	4.725	0.009	0.217	26.262	79.904	86.994	**70.560**
DA2	4.848	0.003	0.175	27.107	76.351	84.344	67.761
DA3	3.732	0.003	0.201	26.236	74.351	82.782	66.149
DA4	**3.323**	**0.002**	**0.171**	**25.009**	**81.225**	**88.177**	70.309
Representation-level design (DA4 fixed)							
w/o PSM & CRL	4.308	**0.002**	0.165	29.595	78.079	86.036	67.047
w/ CRL	4.471	0.003	**0.156**	25.279	77.039	84.877	68.401
Ours (w/ PSM)	**3.323**	**0.002**	0.171	**25.009**	**81.225**	**88.177**	**70.309**

using neither regularizer, while achieving the lowest in-domain EER on PS. The comparison with CRL supports the effectiveness of PSM relative to an alternative representation constraint, rather than only to the absence of regularization. A small EER increase on PE-E1 nevertheless indicates that its advantage is not uniform across datasets. The four configurations are visualized in Appendix A.2 (Fig. 4). Across the displayed utterances, the alternatives produce reliable responses in some regions but exhibit suppressed spoof responses, interior confidence decay, or prediction in-

Table 5: Temporal localization ablation on PS (%) using fixed frame-level predictions. Confidence scoring baselines assign proposal-level scores directly from frame-level predictions on connected regions without learned quality estimation. Higher is better for all metrics. Best and second-best results are shown in textbfbold and underline, respectively.

Setting	AP@.5	AP@.75	AP@.9	AP@.95	mAP	AR@1	AR@5	AR@20
Confidence Scoring Baselines (w/o SML)								
Uniform ($\hat{q} = 1$)	69.310	63.549	60.110	57.430	63.888	22.979	70.675	81.131
Interior Mean	80.477	76.077	72.736	69.649	76.069	25.796	73.101	81.147
Context Scoring	81.133	77.040	74.068	71.317	77.070	26.065	73.210	81.146
SML Ablations								
w/o Boundary Adjustment	82.351	78.673	75.569	72.695	78.577	26.210	73.716	82.911
w/o Proposal Merging	86.309	83.591	80.229	76.843	83.170	26.218	75.349	89.527
w/o Learned Quality	82.140	79.116	73.354	69.344	78.122	25.527	73.060	90.251
w/o Transformer	84.860	82.140	78.670	74.755	81.623	25.943	74.369	90.226
Full SML	**86.448**	**83.988**	**80.544**	**77.106**	**83.473**	**26.225**	**75.386**	**90.430**

terruptions elsewhere. The full configuration generally sustains above-threshold responses within manipulated intervals and low responses in intervening genuine regions, indicating more consistent within-region evidence across the shown samples rather than merely stronger responses on an individual example. Nevertheless, a short manipulated interval in the first example remains fragmented, showing that local prediction failures are not eliminated.

Temporal localization ablation. To separate the contributions of proposal generation and proposal-level quality estimation, we compare Full SML with direct proposal-scoring baselines and controlled component ablations (Table 5). The direct baselines derive intervals from contiguous spoof predictions and differ only in how confidence is computed from frame-level scores. They underperform Full SML, particularly under strict IoU thresholds and at larger proposal budgets, suggesting that rescoring fixed proposals alone is insufficient. Removing boundary adjustment substantially reduces mAP and AR@20, supporting explicit treatment of endpoint uncertainty. Replacing learned quality estimation with non-learned scoring causes the largest drops in mAP and AP@.95, while removing the Transformer also reduces localization accuracy, supporting within-proposal token interaction for quality estimation. Proposal merging has a modest effect on aggregate metrics but a pronounced role in long-fragmentation analysis: full-output recall at IoU 0.95 falls from 58.99% to 4.89% without merging (Appendix A.3). Complete cross-dataset SML ablations on PE-E1, HAD, and LAV-DF are provided in Appendix A.4.

5 Conclusion

In this paper, we introduced STELA, a unified framework for audio temporal forgery detection and localization that addresses unstable frame-level evidence and fragmented interval predictions. STELA improves frame-level robustness through segment-oriented acoustic augmentation, boundary-free interior sampling, and Paired State Modeling, which reduce sensitivity to local acoustic variations and encourage consistent representations for frames sharing the same authenticity state. Structured Manipulation Localization further reasons over alternative interval hypotheses by jointly considering candidate quality, boundary information, and temporal compatibility to recover coherent manipulation regions. Experiments on PS, HAD, PE, and LAV-DF demonstrate a consistent improvements in both frame-level detection and temporal localization, with particularly strong performance under stringent IoU criteria. Ablation and cross-domain analyses further confirm the complementary benefits of robust frame-level representation learning and structured interval inference. A remaining challenge concerns utterances that contain both extended and extremely brief manipulations, where short genuine manipulations can be confused with transient prediction fluctuations near longer manipulated regions. Future work may explore duration-aware temporal modeling to improve localization across diverse manipulation scales.

AI USE STATEMENT

In this work, we used generative AI tools during experimental preparation and manuscript development. At the initial stage of the project, these tools helped us identify relevant prior work on audio forgery detection and localization and locate publicly available implementations. We also used them to assist with the implementation and reproduction of baseline methods for our comparative experiments. During manuscript preparation, generative AI tools assisted in drafting selected passages based on the authors' technical descriptions and experimental findings, and in refining the language, clarity, and readability of the manuscript. The authors checked the retrieved references and code repositories against their original sources, reviewed and tested the AI-assisted baseline implementations, and reviewed and revised all AI-assisted text to ensure consistency with the methods and experimental results reported in the paper. We take full responsibility for the final content of this work, including the manuscript, code, experimental results, and scientific claims, as well as any content developed with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We provide implementation and training details for STELA in the appendix. **The anonymized source code is also included in the supplementary material.** Appendix A.5.1 describes audio preprocessing, frame-label resolution, and data augmentation. The frame encoder and Paired State Modeling are detailed in Appendices A.5.2 and A.5.3, respectively. Appendix A.5.4 specifies the frame-encoder training objectives, optimization settings, random seed, and development-set checkpoint-selection criteria. Appendix A.5.5 describes the SML candidate construction, tokenization, quality prediction, and NMS procedures, together with its training configuration, temporal-IoU evaluation thresholds, and matching rule. Learned localization results are averaged over two fixed seeds, with SML checkpoints selected on a held-out development calibration split.

REFERENCES

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

He Bai, Renjie Zheng, Junkun Chen, Mingbo Ma, Xintong Li, and Liang Huang. A^3T: Alignment-aware acoustic and text pretraining for speech synthesis and editing. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1399–1411, 2022. URL https://proceedings.mlr.press/v162/bai22d.html.

Zexin Cai and Ming Li. Integrating frame-level boundary detection and deepfake detection for locating manipulated regions in partially spoofed audio forgery attacks. *Computer Speech & Language*, 85:101597, 2024. doi: 10.1016/j.csl.2023.101597.

Zhixi Cai, Shreya Ghosh, Abhinav Dhall, Tom Gedeon, Kalin Stefanov, and Munawar Hayat. Glitch in the matrix: A large scale benchmark for content driven audio–visual forgery detection and localization. *Computer Vision and Image Understanding*, 236:103818, 2023. doi: 10.1016/j.cviu.2023.103818.

Jialu Cao, Hui Tian, Peng Tian, Haizhou Li, and Jianzong Wang. Robust detection of partially spoofed audio using semantic-aware inconsistency learning. *IEEE Transactions on Audio, Speech, and Language Processing*, 33:4590–4603, 2025. doi: 10.1109/TASLPRO.2025.3617241.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022. URL https://arxiv.org/abs/2110.01200.

Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. A survey on speech deepfake detection. *ACM Comput. Surv.*, 57(7), February 2025. ISSN 0360-0300. doi: 10.1145/3714458. URL https://doi.org/10.1145/3714458.

Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. BSN: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision*, 2018. URL https://arxiv.org/abs/1806.02964.

Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. URL https://arxiv.org/abs/1907.09702.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. URL https://arxiv.org/abs/1708.02002.

Hanxiao Liu, Zihang Dai, David So, and Quoc V. Le. Pay attention to MLPs. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.nips.cc/paper/2021/hash/4cc05b35c2f937c5bd9e7d41d3686fff-Abstract.html.

Tianchi Liu, Lin Zhang, Rohan Kumar Das, Yi Ma, Ruijie Tao, and Haizhou Li. How do neural spoofing countermeasures detect partially spoofed audio? In *Proceedings of Interspeech*, pp. 1105–1109, 2024. doi: 10.21437/Interspeech.2024-2009.

Yuchen Mao, Wen Huang, and Yanmin Qian. Localizing speech deepfakes beyond transitions via segment-aware learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2026. URL https://arxiv.org/abs/2601.21925.

Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. VoiceCraft: Zero-shot speech editing and text-to-speech in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12442–12462, 2024. doi: 10.18653/v1/2024.acl-long.673. URL https://aclanthology.org/2024.acl-long.673/.

Inbal Rimon, Oren Gal, and Haim Permuter. Split and conquer partial deepfake speech, 2026. URL https://arxiv.org/abs/2604.02913.

Hemlata Tak, Madhu R. Kamble, Jose Patino, Massimiliano Todisco, and Nicholas W. D. Evans. RawBoost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6382–6386, 2022a. doi: 10.1109/ICASSP43922.2022.9746213.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic speaker verification spoofing and deepfake detection using Wav2vec2.0 and data augmentation. In *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, pp. 112–119, 2022b. doi: 10.21437/Odyssey.2022-16. URL https://www.isca-archive.org/odyssey_2022/tak22_odyssey.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Junyan Wu, Wei Lu, Xiangyang Luo, Rui Yang, Qian Wang, and Xiaochun Cao. Coarse-to-fine proposal refinement framework for audio temporal forgery detection and localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7395–7403, 2024. doi: 10.1145/3664647.3680585.

Junyan Wu, Wenbo Xu, Wei Lu, Xiangyang Luo, Rui Yang, and Shize Guo. Weakly-supervised audio temporal forgery localization via progressive audio-language co-learning network. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 574–582, 2025. doi: 10.24963/ijcai.2025/65. URL https://www.ijcai.org/proceedings/2025/65.

Yuankun Xie, Haonan Cheng, Yutian Wang, and Long Ye. An efficient temporary deepfake location approach based embeddings for partially spoofed audio detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 966–970, 2024. doi: 10.1109/ICASSP48485.2024.10448196.

Jiangyan Yi, Ye Bai, Jianhua Tao, Haoxin Ma, Zhengkun Tian, Chenglong Wang, Tao Wang, and Ruibo Fu. Half-truth: A partially fake audio detection dataset. In *Proceedings of Interspeech*, pp. 1654–1658, 2021. doi: 10.21437/Interspeech.2021-930.

Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, Le Xu, Junzuo Zhou, Hao Gu, Zhengqi Wen, Shan Liang, Zheng Lian, Shuai Nie, and Haizhou Li. ADD 2023: The second audio deepfake detection challenge. In *Proceedings of the IJCAI Workshop on Deepfake Audio Detection and Analysis*, volume 3597 of *CEUR Workshop Proceedings*, pp. 125–130, 2023. URL https://ceur-ws.org/Vol-3597/paper21.pdf.

Lin Zhang, Xin Wang, Erica Cooper, and Junichi Yamagishi. Multi-task learning in utterance-level and segmental-level spoof detection. In *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pp. 9–15, 2021a. doi: 10.21437/ASVSPOOF.2021-2. URL https://www.isca-archive.org/asvspoof_2021/zhang21_asvspoof.html.

Lin Zhang, Xin Wang, Erica Cooper, Junichi Yamagishi, Jose Patino, and Nicholas Evans. An initial investigation for detecting partially spoofed audio. In *Proceedings of Interspeech*, pp. 4264–4268, 2021b. doi: 10.21437/Interspeech.2021-738. URL https://www.isca-archive.org/interspeech_2021/zhang21ca_interspeech.html.

Lin Zhang, Xin Wang, Erica Cooper, Nicholas Evans, and Junichi Yamagishi. The PartialSpoof database and countermeasures for the detection of short fake speech segments embedded in an utterance. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:813–825, 2023a. doi: 10.1109/TASLP.2022.3233236.

Rui Zhang, Hongxia Wang, Mingshan Du, Hanqing Liu, Yang Zhou, and Qiang Zeng. UMMAFormer: A universal multimodal-adaptive transformer framework for temporal forgery localization. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8749–8759, 2023b. doi: 10.1145/3581783.3613767.

Xiao Zhang, Rui Zhao, Junjie Yan, Mengya Gao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. P2SGrad: Refined gradients for optimizing deep face models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1905.02479.

You Zhang, Baotong Tian, Lin Zhang, and Zhiyao Duan. PartialEdit: Identifying partial deepfakes in the era of neural speech editing. In *Proceedings of Interspeech*, pp. 5353–5357, 2025. doi: 10.21437/Interspeech.2025-942.

Jiafeng Zhong, Bin Li, and Jiangyan Yi. Enhancing partially spoofed audio localization with boundary-aware attention mechanism. In *Proceedings of Interspeech*, pp. 4838–4842, 2024. doi: 10.21437/Interspeech.2024-587.

Yi Zhou, Zhaodong Xue, Lotfi Senhadji, Huazhong Shu, and Jiasong Wu. BFC-Net: Boundary-frame cross graph attention network for partially spoofed audio localization. *Neurocomputing*, 650:130867, 2025. doi: 10.1016/j.neucom.2025.130867.

Xiaodong Zhu, Yuanming Zheng, Suting Wang, Junqi Yang, Yuhong Yang, Weiping Tu, and Zhongyuan Wang. GEM-TFL: Bridging weak and full supervision for forgery localization through EM-guided decomposition and temporal refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42932–42941, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhu_GEM-TFL_Bridging_Weak_and_Full_Supervision_for_Forgery_Localization_through_CVPR_2026_paper.html.

A APPENDIX

A.1 EFFECT OF WAVLM ADAPTATION STRATEGY

To examine the choice of backbone adaptation used in STELA, we compare several WavLM adaptation strategies under both in-domain and cross-domain settings. This comparison serves as an implementation analysis rather than a primary contribution of the proposed framework.

Table 6 compares different WavLM adaptation strategies under in-domain evaluation. Freezing WavLM without task adaptation leads to a clear degradation, whereas full fine-tuning and parameter-efficient adaptation substantially improve detection. On PE-E1, Ours-I0 performs comparably to full WavLM fine-tuning, while Houlsby Adapter also achieves competitive performance. The budget-matched LoRA-QV variant is less effective under the in-domain setting.

Table 7 further compares all adaptation strategies under the unified fixed-0.5 cross-domain protocol. Different strategies exhibit domain-dependent behavior: LoRA-QV performs best on PE-E2, whereas Houlsby Adapter performs best on PS. Ours-I0 achieves the strongest performance on HAD and ADD2023 and, more importantly, the highest average Acc. and F1 across the four unseen domains. Full fine-tuning remains the second-best strategy on average. These results suggest that the adaptation strategy affects cross-domain transfer substantially, while the lightweight adaptation adopted in STELA provides strong average generalization without requiring full-backbone fine-tuning.

Table 6: In-domain frame-level detection under different WavLM adaptation strategies (%). Each entry is reported as EER / Acc. / F1. LoRA-QV and Houlsby Adapter were trained only on PE-E1. Best and second-best available results for each metric are shown in **bold** and underline, respectively.

Adaptation	PS	HAD	PE-E1
Frozen WavLM	9.339 / 90.878 / 92.823	0.136 / 99.894 / 99.935	2.548 / 97.855 / 98.640
Full WavLM FT	5.000 / 96.013 / 96.822	0.003 / 99.992 / 99.995	**0.168** / 99.859 / 99.911
LoRA-QV†	–	–	2.561 / 97.439 / 98.358
Houlsby†	–	–	0.191 / 99.809 / 99.879
Ours-Plain	**3.323 / 97.010 / 97.587**	**0.002 / 99.999 / 99.999**	0.171 / **99.880 / 99.923**

†LoRA-QV and Houlsby Adapter have independently trained in-domain models only for PE-E1.

Table 7: Cross-domain frame-level detection under different WavLM adaptation strategies (%). All models are trained on PE-E1 and directly evaluated on unseen target domains using a fixed decision threshold of 0.5. Results are reported as Acc. / F1. Best and second-best results are shown in **bold** and underline, respectively.

Adaptation	PE-E2	PS	HAD	ADD2023	Avg.
Frozen WavLM	88.11 / 92.84	61.71 / 75.20	76.94 / 85.67	73.46 / 83.90	75.05 / 84.40
Full WavLM FT	87.66 / 92.09	70.72 / 80.33	85.82 / 91.14	74.74 / 84.33	79.73 / 86.97
LoRA-QV	**88.30 / 92.95**	61.70 / 75.18	76.53 / 85.39	72.41 / 83.13	74.73 / 84.16
Houlsby	79.37 / 85.54	**72.42 / 81.28**	76.54 / 84.51	60.32 / 72.88	72.16 / 81.05
Ours-Plain	87.99 / 92.26	69.53 / 79.92	**89.67 / 93.81**	**77.72 / 86.71**	**81.23 / 88.18**

A.2 QUALITATIVE FRAME-LEVEL ABLATION

To examine how the data-level and representation-level designs affect frame-level predictions, we compare four configurations on the same four utterances in Figure 4. DA1 retains PSM but uses utterance-level augmentation without interior cropping. The other three configurations use DA4, with CRL, neither CRL nor PSM, or PSM (Ours), respectively. Thus, comparing DA1 with Ours examines the combined data-level design with PSM retained, whereas comparing the three DA4 configurations examines the choice of representation supervision.

Across the displayed examples, the alternative configurations produce reliable responses in some regions but exhibit failures in others, including suppressed spoof responses, interior confidence decay,

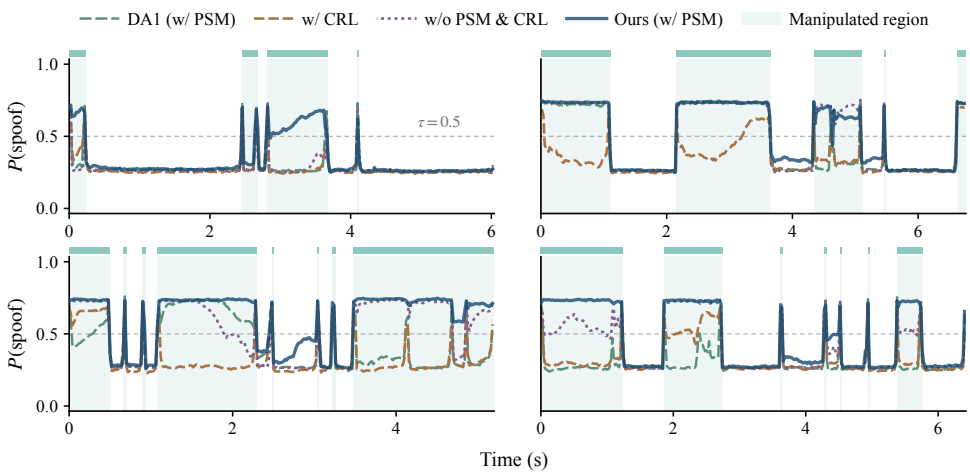

Figure 4: Qualitative frame-level ablation on four utterances. Each panel compares DA1 (w/ PSM), DA4 w/ CRL, DA4 w/o PSM & CRL, and DA4 w/ PSM (Ours) on the same utterance. Shaded areas denote ground-truth manipulated regions, and the horizontal reference line indicates the decision threshold $\tau = 0.5$.

and prediction interruptions. In contrast, the full configuration generally sustains above-threshold responses within manipulated regions while preserving low responses in intervening genuine regions. The key distinction is therefore more consistent prediction behavior across the displayed utterances and regions, rather than merely stronger responses on an individual example. Nevertheless, a short manipulated interval in the upper-left panel still exhibits an interior prediction interruption under the full configuration. These observations complement the quantitative ablation by illustrating more stable within-region evidence alongside a remaining local failure.

A.3 Diagnostic Analysis of Proposal Merging

The aggregate ablation in Table 5 shows only a modest mAP decrease when merged proposals are removed. To better characterize their role, we perform an event-level diagnostic analysis using the saved frame predictions for all 228,333 ground-truth manipulation events from 71,237 utterances in the PS test set. We focus on fragmentation patterns for which directly thresholded frame predictions are unlikely to provide a complete interval. All diagnostic recalls use the strict criterion $\mathrm{IoU} > \tau$.

We define three subsets in Table 8. Long-split contains sufficiently long manipulations whose frame predictions form at least two separated positive components. Endpoint-only describes long manipulations with responses near both ends but little effective support in the interior, whereas Interior-pulse captures long manipulations containing only short interior response pulses. Endpoint-only is fully contained in Long-split; Interior-pulse is independently defined, with 637 of its 651 events also belonging to Long-split. These cases constitute only a small fraction of the full test set, which is consistent with the relatively small change in aggregate mAP when proposal merging is removed.

Table 8: Diagnostic subsets used to analyze fragmented manipulation events in PS. Ratios are computed over all 228,333 ground-truth manipulation events.

Subset	Definition	# Events	Ratio
Long-split	Length ≥ 25 frames with at least two separated positive-response components inside the ground-truth interval	3,916	1.715%
Endpoint-only	Long manipulation with responses near both endpoints but no effective interior response	824	0.361%
Interior-pulse	Long manipulation containing an interior positive-response pulse of at most three frames	651	0.285%

Table 9 evaluates proposal merging from two complementary perspectives. $R_{\text{all}}@.95$ measures whether any output proposal achieves IoU > 0.95, reflecting the geometric coverage provided by proposal generation, whereas $R_{\text{Top5}}@.95$ restricts the evaluation to the five highest-ranked outputs. Proposal merging substantially increases the availability of accurately localized intervals. On Long-split events, $R_{\text{all}}@.95$ increases from 4.89% to 58.99%; for Endpoint-only and Interior-pulse events, the model without merging produces no proposal above this threshold, whereas Full SML reaches 81.19% and 36.56%, respectively.

The smaller gains after Top-5 ranking reveal a second aspect of the problem: proposal merging frequently generates geometrically accurate intervals, but these proposals are not always assigned sufficiently high quality scores to survive a tight output budget. The effect nevertheless remains strongly asymmetric on Long-split events. At IoU 0.95, enabling proposal merging recovers 14.77% of events that otherwise fail, while only 0.03% exhibit the opposite behavior, corresponding to a net recovery of 14.75 percentage points with a 95% confidence interval of $[13.74, 15.79]$. Thus, the small aggregate mAP difference masks a substantially larger effect on the fragmentation pattern that proposal merging is designed to address.

Figure 5 provides controlled examples of these behaviors. The two learned variants use the same frame predictions, proposal scorer, NMS, and output budget, differing only in whether merged proposals are available. Panels (a)–(c) illustrate recovery from ordinary fragmentation, endpoint-dominated responses, and a short interior pulse, respectively. In each case, merged proposals provide a complete interval that is unavailable without merging. Panel (d) shows a shared failure in which only 2.15% of the ground-truth interval retains positive frame-level responses. This example highlights an important limitation: proposal merging can reconnect fragmented evidence, but cannot recover a manipulation when the underlying frame-level evidence is largely absent.

The dependence on the underlying frame evidence is further quantified in Table 10. Proposal merging is most effective when a manipulation is divided into a small number of fragments and a substantial portion of the ground-truth interval still retains positive frame-level responses. For Long-split events consisting of two components, $R@.95$ improves by 17.04 percentage points, whereas the gain decreases as fragmentation becomes more severe. More importantly, events with at least 50% frame-level coverage obtain a 23.44-point improvement, compared with only 1.87 points when coverage falls below 50%. These results characterize proposal merging as a mechanism for reorganizing partially preserved evidence rather than compensating for missing frame-level detection.

Table 9: Effect of proposal merging on difficult fragmentation patterns (%). $R_{\text{all}}@.95$ considers all output proposals, while $R_{\text{Top5}}@.95$ considers only the five highest-ranked outputs.

Subset	$R_{\text{all}}@.95$			$R_{\text{Top5}}@.95$		
	w/o Merge	**Full**	**Δ**	**w/o Merge**	**Full**	**Δ**
Long-split	4.89	**58.99**	**+54.10**	4.66	**19.41**	**+14.75**
Endpoint-only	0.00	**81.19**	**+81.19**	0.00	**0.85**	**+0.85**
Interior-pulse	0.00	**36.56**	**+36.56**	0.00	**4.07**	**+4.07**

Table 10: Proposal-merging performance under different Long-split conditions (%). $R@.95$ is evaluated using the Top-5 outputs.

Condition	**# Events**	**w/o Merge**	**Full**	**ΔR@.95**
2 components	3,096	5.78	22.82	**+17.04**
3 components	608	0.49	8.47	**+7.98**
≥ 4 components	212	0.24	0.94	**+0.71**
Frame coverage $< 50\%$	1,578	0.00	1.87	**+1.87**
Frame coverage $\geq 50\%$	2,338	7.81	31.24	**+23.44**

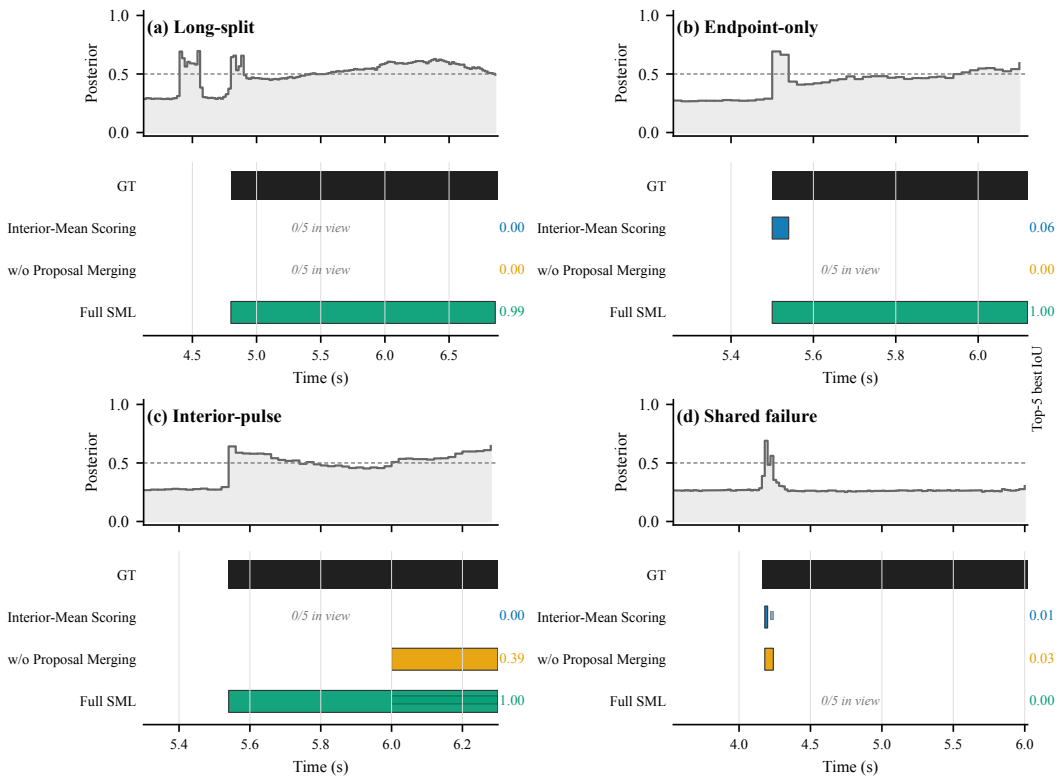

Figure 5: Controlled examples illustrating proposal merging. (a) Long-split fragmentation (CON_E_0051987); (b) Endpoint-only response (CON_E_0012985); (c) Interior-pulse response (CON_E_0047660); and (d) a shared failure under extremely sparse frame-level evidence (CON_E_0006717). Bars denote ground-truth and predicted intervals, and the numbers on the right indicate the best IoU among the Top-5 outputs.

A.4 Cross-Dataset SML Ablation

The main paper reports the SML ablation on PS, where differences among localization strategies are most pronounced. Here, we provide the complete results on PS, PE-E1, HAD, and LAV-DF. Frame-level predictions are fixed within each dataset. Full SML and its learned ablations are averaged over two seeds, whereas the confidence-scoring baselines are deterministic.

These results provide a more complete view of the SML ablation. PS exhibits the clearest component effects: removing boundary adjustment or learned quality estimation produces substantial degradation, while Transformer-based proposal encoding also provides a consistent improvement. Proposal merging has only a small effect on aggregate PS metrics, but the dedicated fragmentation analysis in Appendix A.3 shows that its contribution is concentrated on difficult fragmented events.

On PE-E1 and HAD, localization performance is already close to saturation. Full SML achieves the highest mAP on PE-E1 and ties for the highest mAP on HAD, while several ablated variants slightly outperform it on individual metrics at the reported precision. These results indicate that the additional proposal modeling does not degrade localization on comparatively simple cases, but the saturated operating regime limits the discriminative value of component-wise comparisons.

LAV-DF provides a complementary setting with greater variation. Full SML achieves the highest mAP, AP@.5, AP@.75, and the strongest recall at larger proposal budgets. Replacing transformer encoders into mean-pool encoding slightly exceeds Full SML at AP@.9 and AP@.95, while removing proposal merging yields a marginally higher AR@1.

Table 11: Temporal localization ablation on PS and PE-E1 (%). Frame-level predictions are fixed for all settings. Higher is better for all metrics. Best and second-best results within each dataset are shown in **bold** and underline, respectively.

Setting	AP@.5	AP@.75	AP@.9	AP@.95	mAP	AR@1	AR@2	AR@5	AR@20
PS									
Uniform ($\hat{q} = 1$)	69.310	63.549	60.110	57.430	63.888	22.979	43.048	70.675	81.131
Interior Mean	80.477	76.077	72.736	69.649	76.069	25.796	47.232	73.101	81.147
Context Scoring	81.133	77.040	74.068	71.317	77.070	26.065	47.546	73.210	81.146
w/o Boundary Adjustment	82.351	78.673	75.569	72.695	78.577	26.210	47.652	73.716	82.911
w/o Proposal Merging	86.309	83.591	80.229	76.843	83.170	26.218	**47.866**	75.349	89.527
w/o Learned Quality	82.140	79.116	73.354	69.344	78.122	25.527	46.591	73.060	90.251
w/o Transformer	84.860	82.140	78.670	74.755	81.623	25.943	47.226	74.369	90.226
Full SML	**86.448**	**83.988**	**80.544**	**77.106**	**83.473**	**26.225**	47.799	**75.386**	**90.430**
PE-E1									
Uniform ($\hat{q} = 1$)	98.905	98.552	97.525	95.801	98.236	98.865	99.248	99.323	99.331
Interior Mean	99.670	99.453	98.752	97.437	99.221	98.991	99.325	99.331	99.331
Context Scoring	99.670	99.476	98.937	97.907	99.304	98.991	**99.328**	99.330	99.331
w/o Boundary Adjustment	99.736	**99.540**	**99.007**	**98.037**	99.382	**99.000**	99.327	**99.392**	99.396
w/o Proposal Merging	99.688	99.481	98.959	98.005	99.333	98.959	99.314	99.346	99.347
w/o Learned Quality	99.737	99.526	98.980	97.966	99.366	98.994	99.307	99.365	99.399
w/o Transformer	99.738	99.530	98.995	97.915	99.367	**99.000**	99.312	99.368	99.401
Full SML	**99.745**	99.539	**99.007**	98.033	**99.389**	98.961	99.317	99.387	**99.405**

Table 12: Temporal localization ablation on HAD and LAV-DF (%). Frame-level predictions are fixed for all settings. Higher is better for all metrics. Best and second-best results are shown in **bold** and underline, respectively, for LAV-DF; highlighting is omitted for HAD due to performance saturation and extensive ties at the reported precision.

Setting	AP@.5	AP@.75	AP@.9	AP@.95	mAP	AR@1	AR@2	AR@5	AR@20
HAD									
Uniform ($\hat{q} = 1$)	100.000	99.987	99.987	99.895	99.984	99.987	99.987	99.987	99.987
Interior Mean	100.000	99.989	99.989	99.847	99.980	99.987	99.987	99.987	99.987
Context Scoring	100.000	99.989	99.989	99.849	99.980	99.987	99.987	99.987	99.987
w/o Boundary Adjustment	100.000	99.989	99.989	99.911	99.987	99.987	99.987	99.987	99.987
w/o Proposal Merging	100.000	99.989	99.989	99.911	99.987	99.987	99.987	99.987	99.987
w/o Learned Quality	100.000	99.989	99.989	99.838	99.979	99.986	99.986	99.986	99.986
w/o Transformer	100.000	99.989	99.989	99.908	99.986	99.987	99.987	99.987	99.987
Full SML	100.000	99.989	99.989	99.911	99.987	99.987	99.987	99.987	99.987
LAV-DF									
Uniform ($\hat{q} = 1$)	79.135	74.519	72.089	70.453	74.963	87.976	91.184	91.637	91.649
Interior Mean	94.159	91.135	89.489	88.293	91.402	88.023	91.605	91.649	91.649
Context Scoring	94.185	91.290	89.698	88.541	91.544	88.023	91.606	91.649	91.649
w/o Boundary Adjustment	95.444	92.623	90.429	88.811	92.723	88.029	92.481	92.846	92.858
w/o Proposal Merging	94.342	91.452	89.928	88.761	91.699	**88.039**	91.743	91.771	91.772
w/o Learned Quality	95.057	92.244	90.354	89.047	92.444	88.031	91.363	92.495	92.697
w/o Transformer	95.345	92.552	**90.602**	**89.070**	92.735	87.993	92.059	92.869	92.915
Full SML	**95.490**	**92.660**	90.579	88.991	**92.797**	88.034	**92.532**	**93.573**	**93.573**

A.5 Implementation Details

A.5.1 Input representation and data augmentation

All audio is resampled to 16 kHz with frame labels at 20-ms resolution. Training utterances are padded or truncated to 12 s. We apply RawBoost algorithm 5, which sequentially introduces linear–nonlinear convolutive noise and impulsive signal-dependent noise, either locally with probability 0.75 or globally with probability 0.25. Local augmentation operates on two to four non-overlapping spans, each covering 5–25% of the utterance and constrained to 0.2–2.0 s, with 20-ms cross-fades at both ends. This operation preserves waveform length and frame-label alignment. Development and test audio are not augmented.

With probability 0.25, the augmented utterance is replaced by a 0.4–1.2-s crop sampled from a constant-label region after excluding eight frames (160 ms) at each boundary. If no valid interior exists, the utterance is retained. Waveforms and frame targets are cropped consistently, and boundary targets are regenerated. The resulting crop contains no internal authenticity transition, although fake-only crops retain positive boundary targets at their utterance edges.

A.5.2 Adapter-based frame encoder

We use a frozen WavLM-Large encoder with 24 Transformer blocks and hidden dimension 1024. A trainable Plain Adapter is inserted after the feed-forward sublayer of each block:

$$\mathrm{Adapter}(\mathbf{x}) = \mathbf{x} + \mathbf{W}_{\mathrm{up}}\,\mathrm{GELU}(\mathbf{W}_{\mathrm{down}}\mathbf{x} + \mathbf{b}_{\mathrm{down}}) + \mathbf{b}_{\mathrm{up}}, \tag{10}$$

with bottleneck dimension 641 and no adapter dropout. The up-projection weights and biases are initialized to zero, and adapter updates at padded positions are masked. Only the adapters and task-specific modules are optimized.

The encoder output is projected to 144 dimensions with batch normalization and SELU, processed by a 16-layer state-space `MixerModel`, and followed by layer normalization, SELU, and sinusoidal positional encoding to obtain $\mathbf{F}^D$. A boundary aMLP, based on the attention-augmented gated-MLP design (Liu et al., 2021), produces $\mathbf{F}^b$ for two-class boundary prediction. The aMLP uses a single residual gated block with a 72-dimensional projection, temporal spatial gating, and a 64-dimensional auxiliary attention path.

Single-head cross-attention uses $\mathbf{F}^D$ as queries and $\mathbf{F}^b$ as keys; the value projection is applied to the key-projected boundary features. A 144-dimensional feed-forward block, residual connection, layer normalization, and a second aMLP produce the boundary-aware representation $\mathbf{F}^B$ for frame classification. Both classification heads use dropout 0.25.

A.5.3 Paired State Modeling

Paired State Modeling (PSM) operates on $\mathbf{F}^B$ during training only. Frame features undergo non-affine layer normalization, a bias-free $144 \to 64$ projection, and ℓ_2 normalization. For each utterance, the first state is initialized by the coordinate-wise median feature and the second by the frame farthest from it in cosine distance. We perform three iterations of cosine-similarity-based soft assignment and normalized weighted-mean state updates. The learnable assignment temperature is initialized to 0.2 and clamped to $[0.03, 2.0]$.

The two states have no fixed authenticity identities and are supervised through permutation-invariant cross-entropy:

$$\mathcal{L}_{\mathrm{PSM}} = \frac{1}{B}\sum_{u=1}^{B} \min\left\{\mathrm{CE}(\mathbf{o}_u, \mathbf{y}_u), \mathrm{CE}(\mathbf{o}_u, 1-\mathbf{y}_u)\right\}. \tag{11}$$

A.5.4 Training objectives and optimization

Frame classification uses the P2SGrad-style squared error (Zhang et al., 2019) $\mathcal{L}_{\mathrm{frame}} = \frac{1}{2N}\sum_{t=1}^{N} \|\mathbf{z}_t - \mathbf{e}_{y_t}\|_2^2$, where $\mathbf{z}_t$ contains the two-class scores and $\mathbf{e}_{y_t}$ is the one-hot target. Boundary targets mark the endpoints of constant-label runs, excluding utterance edges adjacent to genuine

regions. For boundary probability $\hat{b}_t$, the boundary loss is

$$\mathcal{L}_{\mathrm{bd}} = -\frac{1}{N_+} \sum_t \left[b_t (1-\hat{b}_t)^2 \log \hat{b}_t + (1-b_t)\hat{b}_t^2 \log(1-\hat{b}_t) \right], \tag{12}$$

where N_+ is the number of positive boundary frames. If a mini-batch contains no positive boundary, we use the unnormalized negative sum. The overall objective is

$$\mathcal{L}_{\mathrm{front}} = \mathcal{L}_{\mathrm{frame}} + 0.5\,\mathcal{L}_{\mathrm{bd}} + 0.05\,\mathcal{L}_{\mathrm{PSM}}. \tag{13}$$

We train for at most 30 epochs using AdamW, batch size 6, learning rate 10^{-5}, and seed 1234. Weight decay is 10^{-6} for task-specific modules and zero for adapters. Evaluation on the clean development set begins after the first five epochs, with early stopping after seven evaluated epochs without improvement. We retain the checkpoint with the lowest development frame EER, breaking ties by higher development mAP.

A.5.5 Structured manipulation localization

After freezing the frame encoder, we cache the 144-dimensional detector and boundary-aware features, $\mathbf{h}_t^D$ and $\mathbf{h}_t^B$, together with spoof posterior p_t and boundary confidence b_t. Their concatenation is fused pointwise:

$$\mathbf{z}_t = \mathrm{LN}\big(\mathbf{h}_t^B + \mathbf{W}_r\,\mathrm{GELU}\big(\mathbf{W}_p\,\mathrm{LN}[\mathbf{h}_t^D; \mathbf{h}_t^B; p_t; b_t]\big)\big), \tag{14}$$

where $\mathbf{W}_p : 290 \to 144$ and $\mathbf{W}_r : 144 \to 144$. The weights and biases of $\mathbf{W}_r$ are initialized to zero.

At inference, thresholding p_t at 0.5 produces connected components. The candidate pool includes each original component, its one-sided endpoint variants with offsets of ± 1 or ± 2 frames, and intervals spanning two or three consecutive components. Candidates are clipped and deduplicated; only multi-component candidates are subject to the 600-frame length limit.

Each candidate is represented by 32 tokens sampled from the left context, start boundary, interior, end boundary, and right context, with respective counts $(4, 4, 16, 4, 4)$ and candidate-normalized ranges $[-0.25, 0]$, $[-0.10, 0.10]$, $[0, 1]$, $[0.90, 1.10]$, and $[1, 1.25]$. Projected tokens receive learned region and position embeddings and are processed by a two-layer pre-norm Transformer with width 128, four heads, feed-forward width 512, and dropout 0.1. Three learned queries aggregate the tokens, followed by a $384 \to 128$ projection, GELU, dropout 0.1, and layer normalization. We append six proposal statistics: $\log(1+\ell_i)/7$, $\log(1+n_i^{\mathrm{cc}})/4$, mean posterior, threshold occupancy, maximum posterior, and minimum posterior. A $134 \to 128 \to 1$ MLP predicts candidate confidence $\hat{q}_i$. The token projection, embeddings, Transformer, and query-aggregation modules are initialized from the existing D2 scorer and fine-tuned, with the original output heads replaced.

The scoring target is the maximum temporal IoU with any ground-truth interval, $q_i^\star = \max_j \mathrm{tIoU}(c_i, g_j)$, optimized using soft-target binary cross-entropy. We uniformly sample up to 128 candidates per utterance with importance weighting to preserve the full-candidate objective. Training uses the broader endpoint-offset grid $\{-4, -2, -1, 0, 1, 2, 4\}^2$ and permits candidates to end at the same component or any of the next 32 components, retaining the 600-frame limit for multi-component candidates.

The scorer is trained for eight epochs using AdamW, learning rate 10^{-4}, weight decay 10^{-4}, batch size 8, and gradient clipping at 5. Checkpoints from epochs 2, 4, and 8 are evaluated on a held-out development calibration split and selected by mAP, with AP@.95 as the tie-breaker. Learned localization results are averaged over seeds 925 and 926.

Inference uses greedy temporal-IoU NMS in descending confidence order, suppressing candidates with IoU strictly greater than 0.7 with an already selected interval. No confidence threshold or endpoint regression is applied. The final scorer contains 617,125 trainable parameters. AP and AR are evaluated over temporal-IoU thresholds 0.50:0.05:0.95, using strict $\mathrm{tIoU} > \tau$ matching.