Temporal Evidence-Guided Multiset-Preserving Reallocation for Cross Domain Weakly Supervised Video Anomaly Localization

Marwan Tayseer Al Zawahra[1], Mohammad Subhi Al-Batah[2]*, Mowafaq Salem Alzboon[3]
Department of Computer Science-Faculty of Information Technology, Jadara University, Irbid 21110, Jordan[1, 2]
Department of Medical Laboratory Science-Faculty of Applied Medical Sciences, Amman Arab University, Amman, Jordan[3]

***Abstract*—Cross-domain weakly supervised video anomaly localization may fail even when a detector produces useful anomaly evidence, because high anomaly scores can be attached to the wrong temporal segments. This failure mode is studied through TEMRAL, a deterministic post-inference procedure for frozen weakly supervised detectors. TEMRAL measures local changes between neighboring segment representations, combines this evidence with the original detector score to rank temporal positions, and then reassigns the original scores according to that ranking. The score multiset for each video is preserved exactly at the detector's native temporal resolution; the method therefore cannot create larger scores, recalibrate the output distribution, or introduce additional anomaly evidence. Configuration selection uses only UCF Crime validation data and is fixed before transfer. In experiments from UCF Crime to ShanghaiTech, TEMRAL improves frame level ROC AUC and average precision for mean, top k, and attention multiple instance learning detectors using both I3D and R3D representations. All thirty model seed evaluations improve on both metrics, with AUC gains between 0.0302 and 0.0569 and average precision gains between 0.0068 and 0.0131. The unchanged configuration also improves precision recall AUC and ROC AUC in all fifteen evaluations on XD Violence. Clustered bootstrap analysis and controls based on temporal smoothing and random score permutations support the interpretation that temporal score misallocation contributes to cross domain localization error and can sometimes be corrected without retraining or target domain adaptation. Additional diagnostic analyses reveal partial source-selection stability, substantial score displacement, and sensitivity to temporal boundary handling; these delimit the interpretation of the observed gains.**

***Keywords*—Weakly supervised video anomaly detection; cross domain generalization; temporal anomaly localization; multiple instance learning; post inference reallocation; multiset preservation; temporal representation change**

I. Introduction

Video anomaly detection (VAD) seeks to locate unusual events in long, untrimmed video streams. Frame-level supervision can make this task expensive to scale, which has driven widespread use of weakly supervised video anomaly detection (WS-VAD). In the common multiple instance learning formulation, a video is treated as a bag of temporal snippets and only a video level normal or anomalous label is available during training [1]. The model must therefore learn both which snippets carry abnormal evidence and how that evidence should be placed in time.

Much of the recent WS-VAD literature improves the quality of the evidence itself. Self-training derives stronger instance supervision from coarse labels [2], robust temporal feature magnitude learning increases abnormal normal separability [3], and normality guided learning adds prototype or similarity cues [4]. Prompt enhanced and vision language approaches further inject semantic context into anomaly scoring [5]–[9]. These methods are valuable, but they primarily alter the representation, the training objective, or the score generation mechanism.

Cross domain deployment introduces a separate source of error. A detector trained on one surveillance distribution may encounter different viewpoints, scene layouts, motion patterns, activity frequencies, and anomaly types at test time. Existing work has studied transfer with limited target supervision [10] and zero-shot transfer without target adaptation in an unsupervised setting [11]. In practical use, however, a pretrained weakly supervised detector may also be frozen: no target labels, no retraining, and no target specific parameter search are available.

This setting motivates a question that is distinct from asking whether the detector produced enough anomaly evidence. Some failures may arise because useful high scores exist but are attached to suboptimal temporal positions. Temporal representation changes are already known to carry useful dynamics information [12]; the objective is not to present local feature change as a new anomaly cue. Instead, it is used only to decide where an already fixed inventory of detector scores should be placed.

Temporal Evidence Guided Multiset Preserving Reallocation for Anomaly Localization (TEMRAL) is a deterministic post-inference procedure for frozen WS-VAD detectors. TEMRAL combines each original segment score with a rank derived from local cosine changes in the frozen representation, uses the combined quantity only to order temporal locations, and then assigns the detector's original scores to those locations. The resulting sequence is an exact permutation of the original per-video score sequence. Consequently, its score distribution is unchanged at the detector's native temporal resolution; only temporal assignment is modified.

* Corresponding author.

The authors are grateful to the Deanship of Scientific Research at Jadara University for providing financial support for this publication.

All design choices are selected on a dedicated UCF-Crime validation split and frozen before target evaluation. The same configuration is tested under UCF-Crime to ShanghaiTech transfer with mean, top-k, and attention MIL detectors using I3D and R3D features. Across six detector backbone settings and five seeds, TEMRAL improves frame level ROC-AUC and average precision in all 30 evaluations. The configuration is then transferred unchanged to XD-Violence, where both reported metrics improve in all 15 evaluations.

To test whether the effect is merely a consequence of smoothing or arbitrary score shuffling, TEMRAL is compared with moving average, bidirectional exponential moving average, Gaussian smoothing, and random multiset preserving permutations. A two-way clustered bootstrap also resamples model seeds and eligible videos. These controls make the proposed mechanism testable: multiset preservation alone is insufficient, and improvements must come from the representation guided ordering.

The main contributions are:

- Temporal evidence allocation is formulated as a distinct post inference factor in cross domain WS-VAD and provide a training free procedure that acts only on frozen outputs.
- TEMRAL enforces exact per video score multiset preservation, separating temporal reassignment from score inflation, recalibration, smoothing, or the creation of new anomaly values.
- One source selected configuration is evaluated without target tuning across three MIL pooling strategies, two feature backbones on ShanghaiTech, five random seeds, and a second unseen target, XD-Violence.
- The evaluation includes source safety checks, fixed smoother controls, random permutation falsification, and seed-by-video clustered bootstrap analysis.

The study therefore does not claim that temporal reallocation replaces representation learning or domain adaptation. It tests the narrower proposition that some cross domain localization error can persist even when useful evidence already exists, and that part of this error can be corrected by changing where that evidence is assigned.

The claim tested here concerns controlled reassignment of a frozen score inventory, not a guarantee that the inventory is discriminative, or that every reassigned score remains within the same semantic event.

II. Related Work

A. Weakly Supervised Video Anomaly Detection

WS-VAD reduces annotation cost by learning snippet level behavior from video level labels. The ranking-based MIL formulation of Sultani et al. [1] established a widely used baseline, after which work increasingly focused on more reliable instance evidence. MIST derives stronger instance supervision through self training [2], RTFM exploits feature magnitude structure [3], and NG-MIL combines a normality based similarity signal with a MIL classifier [4].

Recent methods also use semantic and multimodal context. PE-MIL introduces anomaly aware prompts [5]; prompt enhanced context features [6], VadCLIP [8], and normality guided text prompting [9] use language or vision language priors to reduce ambiguity under coarse labels. BN-WVAD adds an abnormality signal derived from BatchNorm statistics [13].

Work published in 2026 further explores heterogeneous anomaly structure. LAS-VAD models connected temporal components and intention cues [14], MoME uses a mixture of memory experts to represent general and diverse patterns [7], and REBA couples temporal experts with bidirectional video text alignment [15]. Although these methods differ considerably, they all intervene upstream by learning better representations, supervision, or anomaly predictions. TEMRAL begins after those components are frozen.

B. Temporal Dynamics, Context, and Score Refinement

Temporal structure has long been used in anomaly localization. Pu and Wu [12] explicitly relate adjacent feature changes to anomaly score dynamics during training. OE-CTST models interactions across temporal scales to capture both abrupt and gradually evolving anomalies [16], while segment exploration methods use temporal and semantic similarity to select informative training instances [17]. These studies motivate representation change as a useful signal, but they do not impose a post-inference conservation constraint.

Other approaches modify the prediction sequence itself. BN-WVAD [13] and related refinement strategies amend anomaly scores with auxiliary evidence, while DyAnNet combines scene dynamicity and pseudo anomaly scores with iterative regressor training and segment remapping [18]. These mechanisms can change score magnitudes as well as temporal placement.

TEMRAL deliberately separates these effects. Local representation change is converted to a within video ordering signal; it is never used as an additional anomaly probability. The final output is a permutation of the original detector scores, which makes it possible to attribute changes at native resolution to temporal reassignment rather than to a new score distribution.

C. Cross-Domain Video Anomaly Detection

Cross-domain VAD differs in how much target information is available. Jain et al. [10] use limited target supervision, whereas Aich et al. [11] study zero-shot cross-domain VAD without target domain adaptation in an unsupervised setting. These studies show that restricted or absent target access is not by itself novel.

The evaluated setting combines a frozen weakly supervised detector with source only configuration selection and a post inference intervention. No target video or target annotation is used to train the detector, choose a checkpoint, adapt features, or select TEMRAL parameters. The method therefore does not align domains or learn invariant representations; it asks whether downstream temporal allocation can recover some localization quality after domain shift.

D. Post Hoc Temporal Processing and Positioning of TEMRAL

Model-agnostic post-processing is also established in neighboring temporal tasks. GAP improves temporal action

detection by refining predicted boundaries at inference time [19], and weakly supervised temporal action localization has used Gaussian-weighted fusion and self-correcting pseudo-labels to improve temporal boundaries [20].

TEMRAL shares only the general idea that frozen temporal predictions can still be revised. Its operation is different: it neither averages neighboring scores nor changes their magnitudes. Representation evidence ranks locations, and the detector's existing score values are reassigned under exact per video multiset preservation. This narrow combination of post inference ordering and strict score conservation defines the method's contribution.

III. Proposed Method

Fig. 1 summarizes the complete TEMRAL post-inference pipeline and highlights the exact score-multiset-preservation constraint.

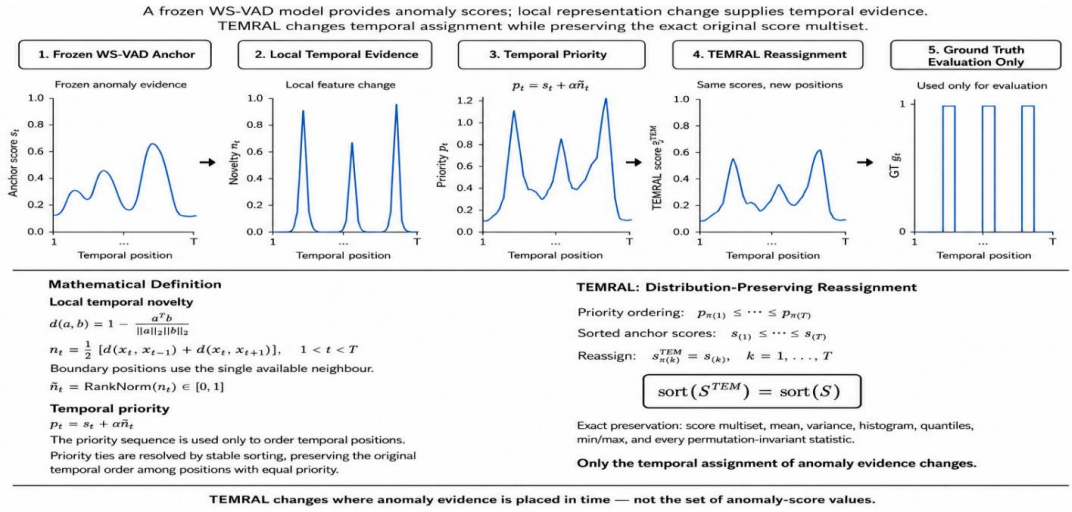

Fig. 1. Overview of TEMRAL. Local representation-change evidence defines a temporal priority ordering used to reassign the frozen detector's original anomaly scores while exactly preserving the per-video score multiset at the native temporal resolution.

A. Problem Formulation and Frozen Detector

Let an untrimmed video be represented at the native temporal resolution of a pretrained weakly supervised video anomaly detector by T temporal segments. For each segment t, the frozen detector provides a feature representation xt and an anchor anomaly score st. The corresponding representation and score sequences are denoted as

$$X = (x_1, x_2, \dots, x_T), \quad S = (s_1, s_2, \dots, s_T). \qquad (1)$$

Each temporal representation is a D-dimensional vector,

$$x_t \in \mathbb{R}^D. \qquad (2)$$

TEMRAL operates entirely after the base detector has completed inference. The detector parameters are not updated, and no additional training or target-domain optimization is performed. The frozen detector therefore serves as an anchor that supplies two quantities: the temporal representation sequence X and the original anomaly-score sequence S.

The objective of TEMRAL is deliberately different from learning new anomaly evidence. Instead, it asks whether the temporal allocation of the evidence already produced by a frozen detector can be improved. The collection of original anomaly-score values is therefore retained, while their temporal assignment is allowed to change.

This formulation separates two components of anomaly localization. Evidence generation determines which anomaly-score values are produced by the detector, whereas evidence allocation determines where those values occur in time. TEMRAL modifies only the latter.

B. Local Representation Change Evidence

TEMRAL derives temporal evidence from changes between neighboring frozen representations. For two representations xi and xj, their cosine distance is defined as

$$d(x_i, x_j) = 1 - \frac{x_i^T x_j}{|x_i|_2 |x_j|_2}. \qquad (3)$$

For numerical stability, the implementation lower-bounds the cosine-similarity denominator by 10^{-12} and clips the resulting similarity to the interval [−1,1].

The frozen TEMRAL configuration uses symmetric local cosine change. For an interior temporal position t, the local representation-change evidence is computed as the average distance to the immediately preceding and following representations:

$$n_t = \frac{d(x_t, x_{t-1}) + d(x_t, x_{t+1})}{2}. \qquad (4)$$

For the first and last positions, only one distinct adjacent transition is available. The implementation duplicates that available transition so that the same symmetric computation remains defined at the temporal boundaries.

The resulting local-change sequence is then converted to a deterministic within-video rank representation. Let ρ_t denote the ordinal rank of nt, ranging from 0 for the lowest-ranked temporal position to T−1 for the highest-ranked position. Ties are

resolved deterministically using stable sorting in temporal order; therefore, equal local-change values receive distinct ordinal ranks rather than averaged ranks. The normalized representation-change evidence is

$$r_t = \frac{\rho_t}{T-1}, \quad T > 1. \tag{5}$$

Thus,

$$0 \le r_t \le 1. \tag{6}$$

For a single-segment sequence, the implementation assigns rt=0.

Importantly, rt is not an anomaly probability or an additional anomaly score. It measures only the relative strength of local representation change within the video. TEMRAL subsequently uses this quantity as evidence for ranking temporal positions.

C. Evidence Guided Temporal Priority

TEMRAL combines the frozen detector's original anomaly score with the local representation-change evidence to construct a temporal priority value:

$$p_t = s_t + \alpha r_t \tag{7}$$

Here, α determines the contribution of the representation-change evidence. Source-only validation selected the frozen value

$$\alpha = 1.25 \tag{8}$$

Consequently, the final TEMRAL priority becomes

$$p_t = s_t + 1.25 r_t. \tag{9}$$

The priority sequence $P=(p_1,\ldots,p_T)$ is used only to order temporal positions. Priority ties are resolved by stable sorting, preserving the original temporal order among positions with equal priority. It is not treated as the final anomaly-score sequence and need not be interpreted as a calibrated probability.

The final frozen configuration is therefore characterized by three source-selected choices: symmetric local cosine representation change (local_cosine), additive score-based priority (score_add), and α=1.25. These choices are fixed before evaluation on the target domains. The complete source-only selection protocol is described in Section IV.

D. Exact Multiset Preserving Temporal Reallocation

The central operation of TEMRAL is to reassign the original anomaly-score values according to the temporal priority ordering.

Let

$\pi P(1),\pi P(2),\ldots,\pi P(T)$

denote the temporal indices sorted from the lowest to the highest value of pt. Separately, let the original detector scores sorted from smallest to largest be

$$s_1 \le s_2 \le \cdots \le s_T. \tag{10}$$

TEMRAL assigns the k-th smallest original anomaly score to the k-th lowest-priority temporal position:

$$s^*_{\pi_{P(k)}} = s_k, \quad k = 1, \ldots, T. \tag{11}$$

Accordingly, positions with larger priority values receive larger scores from the detector's existing score inventory, whereas low-priority positions receive lower original scores.

The transformed sequence can therefore differ temporally from the anchor sequence:

$$S^* \neq S \quad \text{in general} \tag{12}$$

However, TEMRAL preserves the complete collection of score values exactly. The central conservation property is

$$\text{sort}(S^*) = \text{sort}(S). \tag{13}$$

Eq. (13) means that TEMRAL can change the temporal location of a score, but it cannot create a new score, delete an existing one, or change its magnitude.

The implementation performs this operation directly. First, temporal positions are sorted according to their priority values. Second, the original anchor scores are sorted independently. Finally, the sorted scores are assigned to the corresponding priority-ordered temporal locations.

A hard numerical audit additionally verifies

$$\max_k |\text{sort}(S^*)_k - \text{sort}(S)_k| \le 10^{-12}. \tag{14}$$

If this condition is violated, the implementation raises an error rather than accepting the transformed sequence.

E. Conservation Properties and Inference Procedure

Because S∗ is an exact permutation of S at the detector's native temporal resolution, any statistic that depends only on the collection of score values and not on their temporal positions is preserved.

In particular,

$$\min(S^*) = \min(S), \quad \max(S^*) = \max(S) \tag{15}$$

$$\text{mean}(S^*) = \text{mean}(S), \quad \text{var}(S^*) = \text{var}(S). \tag{16}$$

Moreover, for any fixed threshold τ, the number of temporal positions whose anomaly score meets or exceeds that threshold is preserved:

$$\sum_{t-1}^{T} I(s_t^* \ge \tau) = \sum_{t-1}^{T} I(s_t \ge \tau) \tag{17}$$

where, I(·) denotes the indicator function, which equals 1 when the enclosed condition is satisfied and 0 otherwise.

These conservation properties prevent TEMRAL from obtaining localization gains by globally increasing anomaly confidence, rescaling the detector output, or manufacturing additional high-valued predictions. Any change produced at the detector's native temporal resolution is attributable to where the existing anomaly evidence is placed.

For each video, TEMRAL inference consists of the following steps:

1) Obtain the frozen temporal representations X and anchor anomaly scores S.
2) Compute cosine distances between neighboring representations.

3) Average the preceding and following distances to obtain local representation-change evidence nt.
4) Rank the local-change values within the video and normalize their ordinal ranks to obtain rt.
5) Compute the temporal priority pt=st+1.25rt.
6) Sort temporal locations according to pt.
7) Sort the original detector scores and reassign them according to the priority ordering.
8) Verify the exact score-preservation condition in Eq. (13).
9) Return the reallocated anomaly-score sequence S∗.

No gradient-based learning, detector fine-tuning, or target-domain parameter selection is performed during this inference procedure. TEMRAL is therefore a deterministic, training-free post-inference transformation of a frozen detector.

IV. Experimental Setup

A. Datasets and Cross-Domain Protocol

TEMRAL is evaluated in a source-to-target cross-domain setting using UCF-Crime as the labeled source domain and ShanghaiTech and XD-Violence as unseen target domains. All detector training, checkpoint selection, and TEMRAL configuration selection are performed using source-domain data only. Target-domain videos and annotations are reserved strictly for final evaluation.

UCF-Crime. UCF-Crime is a large-scale surveillance video anomaly-detection benchmark containing approximately 1,900 long untrimmed videos and 13 categories of real-world anomalous events [1]. The official partition is followed of 1,610 training videos and 290 test videos. Two source-domain splitting protocols are used for distinct purposes. For the detector evaluation ensemble, each run forms a stratified 80:20 split of the official 1,610-video training partition using the run seed; therefore, the five evaluation runs do not share a single train-validation partition. For TEMRAL configuration selection, a separate Mean-I3D development ensemble is trained on one fixed stratified 80:20 split with split seed 42, yielding 1,288 training videos and 322 validation videos. Only model/training randomness varies across the five development runs.

The official 290-video test partition is held out from detector training and TEMRAL configuration selection and is used later only for source-safety evaluation. The fixed 322-video development validation subset is used for source-only TEMRAL configuration selection and checkpoint selection within the separate development ensemble. The evaluation checkpoints used for ShanghaiTech, XD-Violence, and source-safety reporting retain their original seed-specific source splits.

ShanghaiTech. ShanghaiTech is a surveillance anomaly-detection benchmark containing diverse scenes and abnormal activities [21]. Under the weakly supervised evaluation split used in this study, the target test set contains 199 videos. The corresponding frame-level evaluation set contains 142,060 frames, including 8,317 anomalous and 133,743 normal frames. Of the 199 test videos, 44 contain both normal and anomalous temporal regions, whereas 155 are entirely normal.

ShanghaiTech is used exclusively as an unseen target domain. Neither its videos nor its frame-level annotations are used for detector training, checkpoint selection, TEMRAL hyperparameter selection, feature adaptation, or post-hoc parameter tuning.

XD-Violence. XD-Violence is a large-scale multimodal violence-detection benchmark containing unconstrained real-world videos from diverse sources [22]. The official 800-video test split is used as a second unseen target domain, consisting of 300 normal and 500 abnormal videos. The visual representations used in this evaluation are five-crop I3D RGB features with shape [T,5,2048]. The five spatial crops are averaged before frozen detector inference, resulting in a T×2048 temporal representation for each video.

As with ShanghaiTech, XD-Violence does not participate in model training or TEMRAL configuration selection. The complete configuration selected using UCF-Crime validation data is transferred to XD-Violence without modification.

The resulting protocol is therefore source-selected and target-unseen: all learnable detector parameters and all TEMRAL design choices are determined before evaluating either target dataset.

B. Frozen WS VAD Detectors and Feature Representations

TEMRAL is applied to pretrained WS-VAD detectors that are frozen before post-inference temporal reallocation. To examine whether the proposed mechanism depends on a particular MIL aggregation strategy, three detector families are considered: mean pooling, top-k pooling, and attention pooling.

For UCF-Crime and ShanghaiTech, each pooling strategy is evaluated with two feature representations, I3D and R3D, producing six detector-backbone configurations. The I3D features are 2,048-dimensional, whereas the R3D features are 512-dimensional. UCF-Crime I3D videos are represented by 32 temporal segments, while UCF-Crime R3D videos use 10 temporal segments. Target-video temporal lengths may vary according to the corresponding pre-extracted representation.

The I3D representation belongs to the inflated 3D convolutional family introduced by Carreira and Zisserman [23]. The term R3D is retained for the second 512-dimensional feature representation used in the frozen experimental pipeline; no architectural equivalence between this representation and R(2+1)D is assumed in the present manuscript.

All detector families use a 256-dimensional hidden representation with ReLU activation and dropout of 0.30. For top-k pooling, the number of selected temporal instances is defined as 20% of the valid sequence length, with at least one instance retained. The attention model uses a 256→128→1 attention network with a tanh activation and masked softmax.

The detectors are trained at the bag level using binary cross-entropy with logits and video-level labels only; no frame-level anomaly annotations or temporal boundaries are used during training. Optimization uses Adam with a learning rate of 1×10−4, zero weight decay, batch size 16, and a fixed 20-epoch schedule. Gradient norms are clipped at 5.0. After each epoch, video-level ROC-AUC is computed on the validation subset associated with that training run, and the checkpoint with the

highest source-validation AUC is retained. The detector architecture and optimization settings are identical across the evaluation and development ensembles; the difference lies in how the source train-validation split is controlled.

Five model seeds are used throughout the study:

42, 52, 62, 72, and 82.

For the evaluation checkpoints, the run seed parameterizes both the source train-validation split and training stochasticity, so run-to-run variability includes both split and optimization variation. In the separate development ensemble used for TEMRAL configuration selection, the source split is fixed at split seed 42, and only model/training randomness varies across model seeds 42, 52, 62, 72, and 82.

Once selected, the detector checkpoints are frozen. TEMRAL therefore receives only the frozen temporal representations and segment-level anomaly scores and does not alter detector weights.

For XD-Violence, the same mean, top-k, and attention families are evaluated using I3D representations. The five spatial crops are averaged from [T,5,2048] to [T,2048] before being supplied to the corresponding frozen UCF-trained detector.

C. Source Only TEMRAL Configuration Selection

A central requirement of the experimental design is that no target-domain information is used to select TEMRAL hyperparameters.

Configuration selection is performed using the 322-video UCF-Crime validation subset of the separate fixed-split Mean-I3D development ensemble. All five development models use the same stratified source split (split seed 42), while model/training randomness varies across seeds 42, 52, 62, 72, and 82. The final source-only search evaluates four temporal-evidence variants—previous cosine, local cosine, global cosine, and their ensemble—together with three priority formulations—rank-add, score-add, and centered-rank-add—and 12 positive values of the evidence-weight parameter α. This produces 144 candidate configurations per development seed. Candidate configurations are ranked primarily by mean UCF-Crime validation AUC across the five development runs, with mean validation AP and positive-seed consistency used as successive tie-breakers. No target-domain data are used during this search.

A subtle issue arises because the multiset-preserving TEMRAL output is a permutation of the original per-video score sequence. A permutation-invariant video-level source objective cannot directly distinguish alternative temporal assignments that contain the same score inventory. The unconstrained priority-based variant is therefore used, referred to as Raw TEMRAL, as the source-only selection surrogate. The selected ranking configuration is subsequently transferred unchanged to the multiset-preserving TEMRAL formulation.

The selected source-only configuration is:

Selected setting: temporal evidence = local_cosine; priority rule = score_add; evidence weight $\alpha = 1.25$.

Across the five fixed-split development runs, the selected configuration achieves a mean UCF-Crime validation AUC of 0.913742±0.000868 and AP of 0.913806±0.001723, corresponding to mean improvements of +0.014583 AUC and +0.013661 AP over the corresponding source-validation anchors. Both metrics improve in all five development seeds.

After this source-only selection, the configuration is frozen before any ShanghaiTech or XD-Violence evaluation. The separate evaluation checkpoints are not retrained or reselected as part of TEMRAL configuration selection.

Importantly, no configuration is reselected when the MIL pooling strategy, backbone, or target dataset changes. In particular, the configuration selected using Mean-I3D source validation is transferred directly to R3D without R3D-specific tuning. The same frozen setting is also reused for mean, top-k, and attention detectors and for the XD-Violence target.

This protocol ensures that any target-domain performance changes cannot be attributed to target-label-driven parameter selection.

D. Evaluation Metrics and Temporal Alignment

For ShanghaiTech, the primary evaluation measures are frame-level ROC-AUC and average precision (AP). ROC-AUC evaluates the ranking of anomalous versus normal frames across decision thresholds, whereas AP summarizes precision-recall performance and is particularly informative under the strong class imbalance characteristic of anomaly-detection benchmarks.

For ShanghaiTech, the frozen Anchor and TEMRAL segment scores are expanded to each video's exact annotated length N by endpoint-aligned linear interpolation. The 199 test videos contain 142,060 annotated frames in total. Frame-level ROC-AUC and step-based average precision (AP) are then computed from concatenated predictions and binary annotations. The mapping is applied identically to Anchor and TEMRAL, and the resulting global metrics are not averages of per-video AUC or AP.

For XD-Violence, area under the precision-recall curve (PR-AUC), calculated by trapezoidal integration over the precision-recall curve, is the primary metric; frame-level ROC-AUC is also reported. AP on ShanghaiTech and trapezoidal PR-AUC on XD-Violence are different numerical definitions and are not treated as interchangeable.

For XD-Violence, the frozen detector outputs T segment scores, which TEMRAL reassigns at native resolution. Each sequence is linearly resampled to a reference grid R=N/16, and the R reference scores are repeated 16 times to match the N-frame annotation grid. In the audited 800-video set, native T differs from R for every video (median T=32.5, median R=128.5). Therefore, direct 16-frame repetition of the native T scores, as used by some benchmark-specific 16-frame-rate models, cannot be substituted for the required T-to-R mapping without changing temporal alignment.

Exact score-multiset preservation is guaranteed only at the detector's native segment resolution. The subsequent interpolation used for frame-level evaluation is an evaluation

mapping, and no exact multiset-preservation claim is made after this temporal expansion.

For all target evaluations, the primary comparison is between the original Anchor score sequence and the multiset-preserving TEMRAL sequence. Raw TEMRAL is retained as an ablation that isolates the effect of the preservation constraint rather than as the principal proposed output.

Protocol comparison and limitations. The publicly released RTFM [3] evaluator repeats each native clip score over 16 frames, whereas ShanghaiTech scores in the present study are expanded directly from variable-length segment outputs to each video's N annotations. For XD-Violence, the final 16-frame reference grid and trapezoidal PR-AUC definition are consistent with the reference-style evaluation, but the preceding T-to-R resampling is specific to the supplied representation. Neither the 45/45 internal reproduction checks nor the alignment-sensitivity analysis demonstrates byte-for-byte equivalence to published external model code; test-list ordering and ground-truth provenance require separate verification.

E. Source Safety Analysis and Falsification Controls

Cross-domain gains alone are insufficient to establish that the proposed mechanism behaves sensibly. Accordingly, the evaluation includes both source-domain safety analysis and targeted falsification controls.

Source safety is evaluated on the official 290-video UCF-Crime test partition using frame-level temporal annotations. This analysis is conducted only after TEMRAL configuration selection and does not feed back into hyperparameter tuning. The primary source-safety protocol uses the frozen historical UCF temporal ground-truth representation. As a sensitivity check, the evaluation is repeated using an independently encoded GlanceVAD ground-truth variant. The two annotation encodings preserve the same per-video lengths and differ in only 149 binary frame labels; the qualitative TEMRAL conclusions remain unchanged. This comparison is therefore treated as a ground-truth sensitivity analysis rather than an additional model-selection path.

To test whether TEMRAL simply behaves as a generic temporal smoother, it is compared with three fixed post-hoc smoothing controls: a moving average with window size 3, a bidirectional exponential moving average with $\beta=0.6$, and Gaussian smoothing with $\sigma=1.0$. These control parameters are fixed a priori and are not tuned using ShanghaiTech annotations.

A Random-DP null control is also constructed. For each frozen detector checkpoint, Random-DP performs random temporal permutations of the original anomaly scores while preserving the same per-video score multiset. Fifty independent random permutations are evaluated per checkpoint. This control isolates a critical question: whether exact score preservation alone is sufficient to improve localization, or whether the representation-guided temporal ordering used by TEMRAL is necessary.

The smoothing and permutation controls are evaluated across all six ShanghaiTech detector-backbone configurations—mean, top-k, and attention combined with I3D and R3D—using all five seeds. No model training, target-domain hyperparameter search, or target-based control selection is performed.

Finally, the frozen Anchor predictions are independently reproduced inside the post-hoc evaluation pipeline. This audit checks that the TEMRAL evaluation procedure itself does not inadvertently alter the baseline predictions before reallocation.

F. Statistical Analysis

Global frame-level metrics summarize overall localization performance but do not directly quantify how consistently the effect occurs across independently trained models and individual videos. The global analysis is therefore complemented with paired per-video statistical inference.

Because the experimental observations are crossed over both model seeds and videos, a two-way clustered bootstrap is used that independently resamples these two dimensions. This approach follows established bootstrap procedures for crossed and multi-factor data structures [24], [25].

For ShanghaiTech, per-video AUC and AP comparisons are restricted to the 44 videos containing both normal and anomalous frames, because a single-class video does not support a meaningful within-video ROC-AUC calculation. The bootstrap therefore operates over five seed clusters and 44 eligible video clusters.

For XD-Violence, the corresponding per-video analysis uses the five model seeds and 475 eligible mixed-class videos. Only videos for which both normal and anomalous frame labels are present enter the per-video clustered inference.

Ten thousand bootstrap replicates are used to estimate the mean paired per-video difference and its 95% clustered confidence interval. The canonical implementation independently resamples the seed and video dimensions using the clean-v3 batched procedure with batch size 500 and bootstrap random seed 20260817. The bootstrap probability that the paired effect is positive is also reported. This quantity is explicitly interpreted as a bootstrap positivity probability, rather than as a classical hypothesis-testing p-value.

The reported clustered confidence intervals quantify the mean per-video localization effect under seed-and-video resampling. They are not confidence intervals for the concatenated global frame-level AUC, AP, or PR-AUC values. This distinction is maintained throughout all statistical reporting.

G. Reviewer-Directed Diagnostics and Sensitivity Protocol

To directly probe the allocation hypothesis, a ground-truth-informed (GT-guided) diagnostic comparator ranks temporal segments by the observed fraction of anomalous frames and then reassigns the original score multiset. Target labels are used only for this diagnostic, never for selecting or deploying TEMRAL. Because frame interpolation and global aggregation intervene, this comparator is not claimed to be a proven mathematical upper bound on final frame-level AUC or AP.

Representation-guidance controls replace the local-cosine priority evidence with reversed or reproducibly shuffled ranks. Two spatially restricted baselines preserve the original score inventory: local assignment within nonoverlapping blocks of

approximately 10% of a video, and an assignment constrained to at most 10% source-to-destination displacement. The latter uses a Hungarian solution to a stated surrogate objective for $T \leq 192$; longer sequences use a feasible local-block fallback that is not optimal. These approaches are controls rather than alternative target-selected TEMRAL configurations.

For every video, deterministic stable matching of original score tokens and priority-ranked destinations yields absolute displacement in segments, displacement divided by max(1, T-1), and proportions moved beyond 10% of video duration. Edge evidence is examined using alternative zero-valued or neighbor-copied boundary conditions. Ordinal ties are evaluated with stable, reversed-stable, and averaged ranks. Segment-to-frame sensitivity is assessed using linear, nearest-neighbor, and left-step expansion; this does not by itself establish equivalence with any external official evaluation implementation.

Source-selection stability is assessed using three additional independently drawn stratified 80:20 UCF-Crime splits (seeds 101, 202 and 303), distinct from the original split seed 42. Five newly trained Mean-I3D models (seeds 42, 52, 62, 72 and 82; 20 epochs each) are evaluated per split. For each split, the original 144-configuration source-only search is repeated using the unconstrained Raw score-priority surrogate, with winner frequency, AUC uncertainty across five training seeds, and the frozen original configuration's within-split rank reported. No target examples or labels participate in selection.

Computational profiling separates a CPU/NumPy-only post-inference microbenchmark from a T4 CUDA frozen-detector benchmark with synthetic pre-extracted features. Post-processing, detector inference, feature extraction, and dataset I/O are explicitly distinguished. Python-traced peak allocations are not equated with GPU VRAM or total process memory. Temporal operations require O(TD+T log T) time, with O(TD) possible NumPy temporary storage in addition to the T-dimensional score/rank arrays.

V. Results

A. Cross Domain Localization on ShanghaiTech

The first evaluation tests whether the source-selected TEMRAL configuration improves localization on the unseen ShanghaiTech target domain. Table I summarizes the results across three MIL aggregation strategies and two feature backbones.

TABLE I. Cross-Domain Localization Performance on ShanghaiTech Averaged Over Five Seeds

Detector	Anchor AUC	TEMRAL AUC	ΔAUC	Anchor AP	TEMRAL AP	ΔAP
Mean-I3D	0.613855	0.654430	+0.040575	0.102553	0.110605	+0.008052
Topk-I3D	0.594259	0.645709	+0.051450	0.086582	0.099698	+0.013116
Attention-I3D	0.583985	0.640851	+0.056866	0.105799	0.113283	+0.007484
Mean-R3D	0.557941	0.589616	+0.031675	0.072236	0.079086	+0.006849
Topk-R3D	0.544219	0.579757	+0.035538	0.069461	0.076792	+0.007331
Attention-R3D	0.561741	0.591978	+0.030237	0.072801	0.079641	+0.006840

TEMRAL improves both frame-level AUC and AP for all six detector-backbone configurations. The AUC gains range from +0.030237 for Attention-R3D to +0.056866 for Attention-I3D, while the AP gains range from +0.006840 to +0.013116.

The direction of improvement is also stable across random seeds. For every detector-backbone configuration, TEMRAL improves AUC in all five seeds and AP in all five seeds. Across the six configurations, this corresponds to 30/30 positive model-seed changes for AUC and 30/30 for AP.

This consistency is important because the same configuration—local cosine evidence, score-add priority, and α=1.25—is used throughout without any ShanghaiTech-specific adjustment. The observed improvements therefore arise after the complete source-only selection procedure has already been frozen.

The effect is not restricted to one representation family. I3D configurations exhibit larger absolute AUC gains overall, but all three R3D configurations also improve on both metrics. This indicates that the reallocation mechanism is not tied exclusively to the feature space on which the configuration was originally selected.

B. Effect of the Multiset Preservation Constraint

To examine the role of exact score preservation, the frozen Anchor is compared with the unconstrained Raw variant and TEMRAL under Mean-MIL for both I3D and R3D. Table II reports the corresponding source- and target-domain results.

TABLE II. Mean-MIL Source-Target Comparison of Anchor, Raw, and TEMRAL

Backbone / Domain	Anchor AUC	Raw AUC	TEMRAL AUC	Anchor AP	Raw AP	TEMRAL AP
I3D / UCF-Crime	0.774523	0.729924	0.761710	0.188312	0.212691	0.197136
I3D / ShanghaiTech	0.613855	0.629153	0.654430	0.102553	0.091715	0.110605
R3D / UCF-Crime	0.776929	0.693752	0.756347	0.208046	0.137033	0.174009
R3D / ShanghaiTech	0.557941	0.618944	0.589616	0.072236	0.092817	0.079086

The unconstrained Raw variant demonstrates that temporal representation evidence can strongly alter detector behavior. On ShanghaiTech, Raw increases AUC for both I3D and R3D. However, its behavior is less controlled: for Mean-I3D, Raw improves target AUC but reduces target AP relative to the Anchor, whereas TEMRAL improves both metrics.

The preservation constraint also reduces the magnitude of source-domain perturbation. For I3D, Raw decreases source AUC by 0.044599, whereas TEMRAL decreases it by 0.012813. For R3D, the corresponding AUC changes are −0.083177 for Raw and −0.020582 for TEMRAL. The same moderation is visible in the magnitude of the source AP change: for I3D, Raw changes AP by +0.024379 versus +0.008824 for TEMRAL; for R3D, the changes are −0.071013 and −0.034036, respectively. This is consistent with the exact score-inventory constraint limiting the intervention, although exact multiset preservation does not imply preservation of frame-level AUC or AP.

These results should not be interpreted as universal source improvement. TEMRAL can still change frame-level source localization because temporal permutation changes where scores occur, even when their values are preserved exactly. Rather, the result supports the narrower conclusion that the conservation constraint substantially limits source-domain disruption compared with unconstrained score modification.

C. Clustered Statistical Evidence on ShanghaiTech

The global results in Table I are complemented by two-way clustered bootstrap inference over model seeds and eligible ShanghaiTech videos.

The canonical clustered analysis yields positive 95% confidence intervals for all 12 TEMRAL-versus-Anchor architecture-metric comparisons across the six ShanghaiTech detector-backbone configurations. Table III reports the corresponding Mean-MIL I3D and R3D results as representative backbone-level cases. Fig. 2 summarizes the canonical paired clustered effects of TEMRAL relative to the frozen Anchor across all six ShanghaiTech detector–backbone configurations and both evaluation metrics.

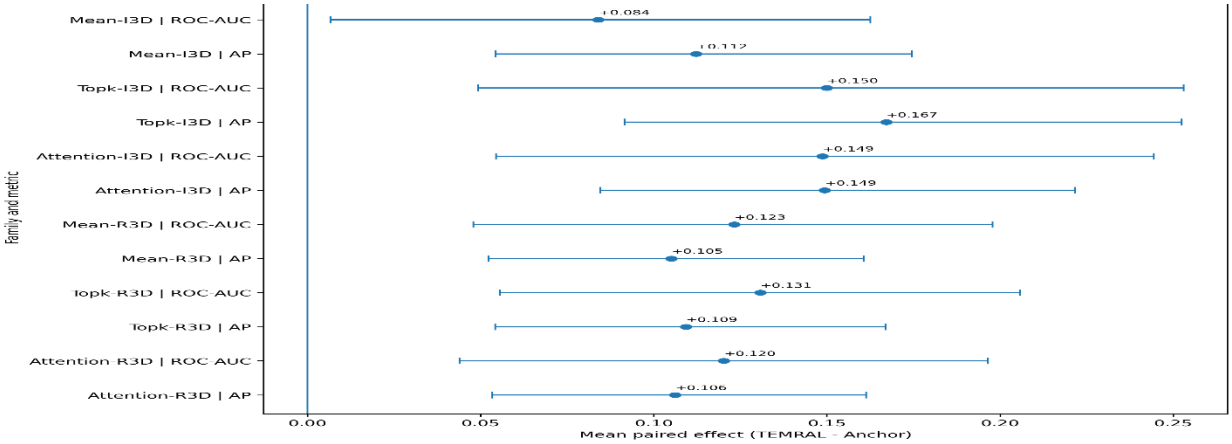

Fig. 2. Canonical two-way clustered bootstrap effects of TEMRAL versus the frozen Anchor on ShanghaiTech. Points indicate the mean paired per-video effect, while horizontal bars denote the corresponding 95% confidence intervals obtained using crossed seed-by-video resampling. Results are shown for all six detector–backbone configurations and both ROC-AUC and AP. All 12 confidence intervals lie entirely above zero.

TABLE III. REPRESENTATIVE MEAN-MIL TWO-WAY CLUSTERED BOOTSTRAP ANALYSIS OF TEMRAL VERSUS ANCHOR ON SHANGHAITECH

Rep.	M.	Mean per-video effect	95% CI	P(Δ > 0)
I3D	AUC	+0.084021	[+0.006725, +0.162515]	0.9833
I3D	AP	+0.112214	[+0.054414, +0.174519]	1.0000
R3D	AUC	+0.123259	[+0.047921, +0.197806]	0.9992
R3D	AP	+0.105126	[+0.052329, +0.160607]	1.0000

All four Mean-MIL confidence intervals reported in Table III lie entirely above zero. For Mean-I3D, the lower confidence bound is +0.006725 for AUC and +0.054414 for AP. For Mean-R3D, the corresponding lower bounds are +0.047921 and +0.052329. Across the complete six-family analysis, all 12 architecture-metric TEMRAL-versus-Anchor intervals are positive.

These quantities represent mean per-video effects under crossed seed-by-video resampling. They are not confidence intervals for the concatenated global AUC or AP values reported in Table I.

Taken together with the 30/30 seed-level directional consistency, these results indicate that the observed effect is not driven solely by one favorable checkpoint or by a small subset of global concatenated predictions.

D. Falsification Tests Against Temporal Smoothing and Random Reallocation

A central alternative explanation is that the improvements might result from generic smoothing rather than representation-guided temporal reallocation. TEMRAL is therefore compared against moving-average, bidirectional EMA, and Gaussian smoothing controls across all six ShanghaiTech detector-backbone configurations and both metrics.

Across these comparisons, all 36 TEMRAL-versus-fixed-smoother clustered 95% confidence intervals are positive. The smallest lower confidence bound across the 36 comparisons is +0.001449, and the minimum bootstrap probability of a positive effect is 0.9772.

This result makes a purely smoothing-based explanation insufficient. Although smoothing can alter temporal continuity and suppress local fluctuations, it does not reproduce the localization behavior of TEMRAL consistently across the tested detector families, backbones, and metrics.

The Random-DP control provides a second falsification test. Because Random-DP preserves the same per-video anomaly-score multiset while randomly permuting its temporal positions, it isolates whether multiset preservation alone is enough to produce improvement.

In the architecture-level Random-DP null summaries, none of the 50 random-permutation null replicates reaches or exceeds the observed TEMRAL value for any of the six architecture families in either AUC or AP. With the prespecified +1 correction, the empirical exceedance rate for each architecture-metric comparison is therefore $1/51 \approx 0.0196$. This quantity is treated as an empirical null exceedance rate rather than a classical p-value.

The combination of the smoothing and Random-DP controls therefore supports a more specific interpretation: preservation of the original score inventory is not sufficient by itself; the representation-guided temporal ordering is the operative component.

Finally, the post-hoc Anchor reproduction audit yields a maximum AUC discrepancy of exactly 0 and a maximum AP discrepancy of approximately 8.33×10^{-17} across the 30 model-seed rows. This confirms that the evaluation pipeline reproduces the original frozen detector outputs before TEMRAL is applied.

E. Generalization to XD Violence as a Second Unseen Target

The second evaluation tests whether the effect observed on ShanghaiTech extends to a second unseen target dataset. The same UCF-selected TEMRAL configuration is transferred to XD-Violence without retraining or target-domain tuning.

Table IV summarizes the five-seed global results for Mean-, Topk-, and Attention-I3D detectors.

TABLE IV. FROZEN UCF-CRIME → XD-VIOLENCE CROSS-DOMAIN RESULTS

Detector	Anchor PR-AUC	TEMRAL PR-AUC	ΔPR	Anchor ROC-AUC	TEMRAL ROC-AUC	ΔROC
Mean-I3D	0.189203	0.196183	+0.006980	0.428424	0.444801	+0.016377
Topk-I3D	0.192910	0.199288	+0.006378	0.439092	0.453838	+0.014746
Attention-I3D	0.201537	0.207960	+0.006423	0.460562	0.472888	+0.012325

TEMRAL improves both PR-AUC and ROC-AUC for all three pooling strategies. The PR-AUC gains range from +0.006378 to +0.006980, while the ROC-AUC gains range from +0.012325 to +0.016377.

The improvement is again directionally stable across seeds: for each of the three detector families, all five seeds improve in PR-AUC and all five improve in ROC-AUC. This corresponds to 15/15 positive model-seed changes for each metric.

The canonical XD-Violence clustered analysis provides additional support. TEMRAL versus Anchor yields positive mean per-video effects for all three detector families and both metrics, with all 6/6 confidence intervals entirely above zero. Mean-I3D produces a PR-AUC per-video effect of +0.021112 with a 95% CI of [+0.004385, +0.037817], Topk-I3D yields +0.027144 with [+0.011764, +0.042748], and Attention-I3D yields +0.022039 with [+0.005697, +0.038622]. The corresponding ROC-AUC confidence intervals are also entirely positive.

However, the absolute XD-Violence scores remain comparatively low. The second-target experiment should therefore be interpreted as evidence that the direction of the reallocation effect transfers to another unseen domain, not as evidence that TEMRAL eliminates severe representation-level domain shift. Fig. 3 summarizes the cross-domain effect of the frozen source-selected TEMRAL configuration across the evaluated detector families and unseen target datasets.

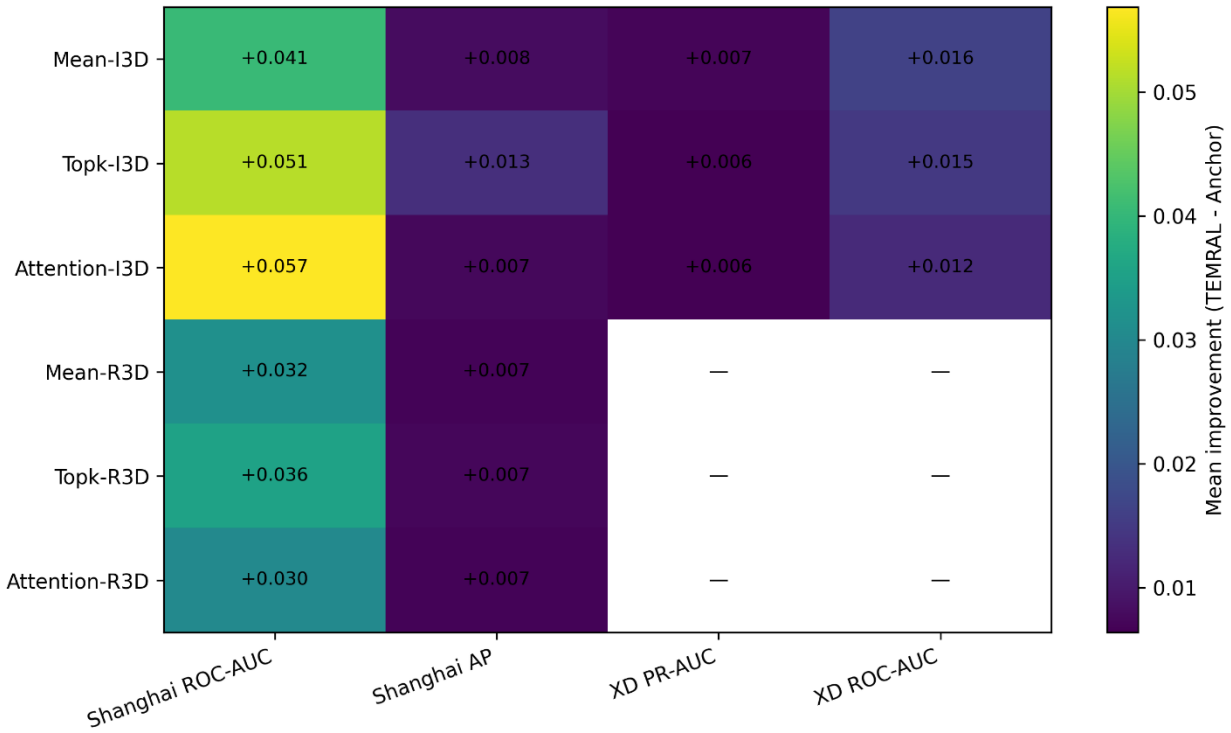

Fig. 3. Cross-domain summary of TEMRAL performance relative to the corresponding frozen Anchor detectors across ShanghaiTech and XD-Violence. The visualization summarizes the direction and magnitude of the reported target-domain improvements across detector families; formal paired seed-by-video clustered inference is reported separately in the corresponding results tables and text. Blank cells indicate detector–backbone configurations that were not evaluated on XD-Violence.

F. Summary of Empirical Findings

Across the complete frozen evaluation, five findings emerge.

First, TEMRAL improves both AUC and AP on ShanghaiTech for all six tested detector-backbone combinations and across every individual model seed. Second, these improvements extend beyond the I3D representation on which configuration selection was performed and remain positive for R3D without backbone-specific retuning. Third, exact multiset preservation substantially moderates the source-domain perturbation associated with unconstrained temporal score modification. Fourth, neither fixed temporal smoothing nor random multiset-preserving permutation reproduces the observed effect. Finally, the same frozen source-selected configuration yields consistent improvements on XD-Violence, providing a second unseen-target test of the proposed evidence-allocation hypothesis.

Taken together, the results are consistent with the hypothesis that temporal misallocation contributes to cross-domain WS-VAD localization error and that representation-guided reassignment can recover part of the lost localization performance without retraining, target-domain adaptation, or increasing anomaly-score magnitude.

G. Oracle, Ablation, and Restricted Reallocation Results

The five-seed global ROC-AUC comparisons in Table V show that the source-selected TEMRAL configuration improves on each anchor, but trails the GT-informed diagnostic comparator. On ShanghaiTech, reversing novelty lowers AUC to 0.4567–0.5186 across the six settings, whereas shuffled novelty yields 0.5246–0.5947; both are below the corresponding TEMRAL results. On XD-Violence, both controls can exceed the anchor yet remain below TEMRAL, so they do not establish anomaly-specific novelty evidence. The GT-informed comparator demonstrates residual rearrangement potential under ground-truth access, not an actionable or formally optimal method.

TABLE V. FIVE-SEED GLOBAL ROC-AUC: ANCHOR (A), TEMRAL (T), GT-GUIDED (GT), AND 10%-CONSTRAINED (C10); LISTED IN THAT ORDER

Target / family	A / T / GT / C10 (ROC-AUC)
SH Mean-I3D	0.6139 / 0.6544 / 0.7413 / 0.6269
SH Topk-I3D	0.5943 / 0.6457 / 0.7146 / 0.6136
SH Attn-I3D	0.5840 / 0.6409 / 0.7205 / 0.6024
SH Mean-R3D	0.5579 / 0.5896 / 0.6397 / 0.5661
SH Topk-R3D	0.5442 / 0.5798 / 0.6321 / 0.5525
SH Attn-R3D	0.5617 / 0.5920 / 0.6402 / 0.5698
XD Mean-I3D	0.4284 / 0.4448 / 0.5589 / 0.4349
XD Topk-I3D	0.4391 / 0.4538 / 0.5764 / 0.4457
XD Attn-I3D	0.4606 / 0.4729 / 0.5874 / 0.4655

Both 10%-limited methods underperform unrestricted TEMRAL on global ROC-AUC and the corresponding second metric in all nine model/target settings. This is not proof of superiority to all possible local alternatives: the constrained assignment optimizes a surrogate and uses a nonoptimal fallback

for long sequences. For Mean-I3D on ShanghaiTech, an exploratory crossed seed-by-video bootstrap estimates a per-video AUC effect of +0.0840 for TEMRAL versus Anchor (95% interval [+0.0067,+0.1625]), +0.0805 versus local 10% (95% interval [+0.0117,+0.1517]), and +0.0648 versus constrained 10% (95% interval [+0.0148,+0.1154]). These intervals concern the 44 eligible mixed-label videos, not pooled global AUC; five seed clusters and unadjusted exploratory multiple contrasts constrain inference.

Weight sensitivity provides an additional falsification check. For ShanghaiTech Mean-I3D, global ROC-AUC is 0.613855 at α=0, 0.654430 at the locked α=1.25, and 0.6572 at α=3.0; corresponding AP values are 0.102553, 0.110605, and approximately 0.1106. For XD-Violence Mean-I3D, ROC-AUC is 0.428424, 0.444801, and 0.4482 at the same weights; PR-AUC is 0.189203, 0.196183, and approximately 0.1973. These are exploratory target-domain sensitivity measurements, not source-only model-selection results, and neither α=3.0 nor another test-favored value replaces the frozen α=1.25 configuration.

H. Temporal Displacement, Boundaries, and Alignment

Measured score-token reassignment is substantial: averaged over video-by-model-seed records, mean absolute displacement is 21.29 segments (0.3259 of normalized video duration) on ShanghaiTech and 14.91 segments (0.3343) on XD-Violence. The mean fractions of scores displaced more than 10% of video duration are 0.7763 and 0.7870, respectively; 0.9641 and 0.9380 change position at all. The averages weight video/seed records equally, not individual frame or score tokens. Hence improvements in aggregate localization metrics do not imply plausible assignment within a single semantic event.

Boundary tests reveal non-negligible design sensitivity: setting edge novelty to zero raises both global AUC and AP/PR-AUC relative to original duplicated-transition boundaries in all nine model/target settings. This alternative was evaluated after the original configuration was locked and is not adopted or used for target-driven tuning. Priority ties were essentially absent (mean tied-position fraction 0 in both datasets); novelty ties averaged 0 on ShanghaiTech and 0.00125 on XD-Violence. Average-rank and reverse-stable tie controls produced nearly identical global metrics. Nearest-neighbor and left-step frame alignment retained positive TEMRAL-minus-Anchor changes in both global metrics across all nine settings, although external official-protocol code equivalence remains to be independently checked.

Temporal-alignment sensitivity was evaluated with three fixed mappings, without changing detector weights or the frozen TEMRAL configuration. For ShanghaiTech (30 model-seed evaluations), mean global TEMRAL-minus-Anchor ROC-AUC changes were +0.041057 (linear), +0.040876 (nearest-neighbor), and +0.041187 (left-step); corresponding AP changes were +0.008279, +0.008029, and +0.008806. For XD-Violence (15 evaluations), the ROC-AUC changes were +0.014483, +0.013748, and +0.013979, and the PR-AUC changes were +0.006593, +0.006266, and +0.006381, in the same order. Both evaluated metrics improved in all 30/30 ShanghaiTech and 15/15 XD-Violence runs under each mapping. These are descriptive cross-seed averages of changes in pooled frame-level metrics, not confidence intervals or evidence of exact correspondence with reference implementations.

I. Independent-Split Selection Stability

Across 15 additional Mean-I3D training runs (three independent source splits, five training seeds, 20 epochs), the score_add rule was selected on all three new splits. The complete original local_cosine/score_add/α=1.25 setting was selected on split 101 only; split 202 selected ensemble/score_add/α=0.75, and split 303 previous_cosine/score_add/α=2.00. The original setting ranked 1st, 19th and 9th of 144, with within-split validation-AUC shortfalls of 0, 0.005941 and 0.001752, respectively (Table VI). Thus the priority rule is relatively stable while novelty family and α remain sample-sensitive. Selection operates on the Raw priority surrogate, not directly on post-interpolation target localization.

TABLE VI. SOURCE-ONLY SELECTION STABILITY ACROSS THREE ADDITIONAL INDEPENDENT TRAIN/VALIDATION SPLITS

Split	Winner; AUC mean ± SD; original rank/gap
101	local / 1.25; 0.914267 ± .001164; rank 1; gap 0
202	ensemble / .75; 0.917847 ± .000720; rank 19; gap .005941
303	previous / 2.00; 0.917508 ± .001812; rank 9; gap .001752

J. Runtime and Memory Characterization

For CPU/NumPy core processing at D=2048, mean latency grows from 1.145 ms (T=10; 873 videos/s) to 2.042 ms (T=32; 490 videos/s), 5.557 ms (T=128; 180 videos/s), and 22.056 ms (T=512; 45 videos/s). Python-traced peak allocations rise from approximately 0.56 MB to 25.25 MB over the T=10–512 range. In a mixed-device synthetic-feature comparison at T=32, the T4 CUDA Mean-I3D frozen detector took 0.506 ms and CPU TEMRAL post-processing took 2.693 ms, or 5.32 times detector-only latency. Accordingly, TEMRAL overhead cannot be described as negligible in this setting. These measurements exclude feature extraction, disk I/O, and end-to-end raw-video processing, and they do not measure total process RSS or GPU VRAM.

VI. DISCUSSION AND LIMITATIONS

A. Temporal Misallocation as a Distinct Source of Cross-Domain Error

Across ShanghaiTech, TEMRAL improves both AUC and AP for all six detector backbone settings without retraining, feature adaptation, or any increase in available score values. The result is consistent with temporal misallocation being one contributor to cross-domain error. It does not show that misallocation is the only cause, nor that a frozen detector always contains enough useful evidence.

This distinction separates evidence quality from evidence placement. WS-VAD methods such as MIST, RTFM, NG-MIL, and BN-WVAD change how anomaly evidence is learned or produced [2]–[4], [13]. TEMRAL assumes that this stage is complete and asks whether a different placement of the same score inventory improves localization. The positive XD-Violence results, despite low absolute performance, suggest that

this downstream correction can remain useful even when representation shift is still substantial.

The GT-guided diagnostic comparator confirms that a frozen score inventory admits substantially better placements when target ground truth is supplied. Since this oracle-like analysis is label-informed and not an optimized bound for frame interpolation, it should be understood as a diagnostic gap rather than evidence of strong native discrimination or deployable improvement.

B. Role of Representation Change

Local representation change is not a new anomaly cue. Dynamics-aware learning has already linked adjacent feature variation with anomaly score behavior [12], and temporal context methods use related information at several stages of training. The role of local change in TEMRAL is therefore deliberately limited: it ranks temporal positions after inference and never becomes a new score.

This restriction also differentiates TEMRAL from explicit score refinement. Discriminative score suppression, for example, combines classifier output with prototype based dissimilarity to alter numerical anomaly predictions [26]. TEMRAL instead requires the complete per-video score multiset to remain unchanged. Its empirical question is whether a familiar temporal cue can improve placement while numerical evidence is held fixed.

C. Multiset Preservation as a Mechanistic Constraint

Exact multiset preservation is the main mechanistic control. At native segment resolution, the minimum, maximum, mean, variance, empirical distribution, and the count above any fixed threshold are identical before and after TEMRAL. Improvements therefore cannot be attributed to globally larger confidence values, rescaling, or a newly created set of peaks.

The constraint does not guarantee source performance. Reassigning a fixed set of scores can still worsen localization when the original temporal arrangement was already good. The source domain analysis shows this effect, and the Raw variant illustrates the opposite trade-off: unconstrained refinement may produce larger changes but makes it harder to isolate the cause. The clustered TEMRAL versus Raw intervals also include zero, so the data do not establish target domain statistical superiority of the constrained variant; the value of TEMRAL is the controlled interpretation it permits.

A further qualification is that preservation of the score distribution does not imply local temporal continuity: the measured mean normalized displacement exceeds 0.32 on both targets. Constrained and block-local assignments reduce temporal reach but yield smaller gains under the tested 10% configurations. Interpretation should therefore separate aggregate ranking improvements from per-event localization plausibility.

D. Beyond Generic Post Processing and Temporal Smoothing

Post processing alone cannot explain the observed gains. GAP [19] revises action boundaries, whereas the fixed smoothing controls in this study average or diffuse neighboring scores. TEMRAL does neither. Across ShanghaiTech, all 36 architecture metric clustered comparisons against the three fixed smoothers favor TEMRAL.

The Random-DP control is complementary because it satisfies the same multiset preservation rule while discarding representation guidance. None of the 50 random permutations for any architecture metric comparison matches or exceeds the observed TEMRAL value; with the prespecified plus one correction, the empirical exceedance rate is 1/51, or approximately 0.0196. This quantity is used only as a falsification summary, not as a conventional p value.

E. Cross Domain Interpretation

Prior cross-domain work either adapts with limited target information [10] or studies transfer without target adaptation under other supervision assumptions [11]. TEMRAL should not be described as a domain adaptation method. It uses no target-specific fitting and does not align distributions or learn a new decision boundary. The low absolute XD-Violence scores make this limitation clear: reallocation improves the frozen anchor, but it does not solve the representation gap.

Accordingly, TEMRAL is complementary to domain generalization. Representation-level methods seek better evidence under distribution shift; TEMRAL asks whether the evidence already produced by a frozen detector can be placed more effectively in time.

F. Generalization Across Pooling Strategies, Backbones, and Targets

Robustness across the tested pooling and representation choices is notable because the configuration was selected only with Mean-I3D source validation. The same setting improves mean, top-k, and attention detectors with I3D and R3D on ShanghaiTech, with positive changes in both metrics for all 30 seeds. No R3D specific tuning was performed.

XD-Violence provides a second transfer test with different content and evaluation mapping. The direction of improvement remains positive for all 15 model seed evaluations in both reported metrics. These findings support robustness within the tested scope, but they do not establish architecture independent or dataset independent performance.

The selected source-only configuration does not consistently win on additional source validation splits (one of three new splits), despite retaining a competitive within-split rank and a stable score_add rule. Boundary sensitivity is also material, as a zero-edge alternative improves the target metrics across all model/target settings; because this was observed post hoc, no retuning is claimed. Some unconstrained Raw results surpass multiset-preserving TEMRAL on target AUC (including Mean-R3D on ShanghaiTech); thus score conservation is a controlled attribution restriction, not a guarantee of higher predictive performance.

G. Limitations

TEMRAL has several clear limitations. First, it can only rearrange evidence that already exists; nearly uninformative score sequences cannot be repaired by permutation. Second, local cosine change is not anomaly-specific and may respond to benign camera motion, illumination shifts, or scene transitions. Third, exact preservation is guaranteed only at the detector's

native segment resolution, not after interpolation to frame level benchmark grids. Temporal-alignment sensitivity indicates directional robustness for the three tested mappings, but neither exact external evaluator parity nor author-released test-list/annotation hash identity has been established.

Fourth, score preservation does not guarantee preservation of source domain localization, because temporal assignment can still deteriorate. Fifth, the evaluation covers three MIL pooling strategies, two backbones on ShanghaiTech, and two target datasets, but remains narrower than the current WS-VAD landscape; XD-Violence is evaluated with I3D only. Finally, the method uses immediate neighbor cosine change, so it cannot model longer temporal dependencies or recover information already lost through coarse feature extraction.

The lower-than-chance XD-Violence ROC-AUC values (0.428–0.473 across the reported Anchor and TEMRAL means) are a material limitation of absolute cross-domain discrimination. Small relative improvements should not be read as adequate transfer performance. The per-video clustered intervals exclude normal-only ShanghaiTech videos and use only five training-seed clusters; they do not justify universal population or pooled-metric significance claims. The current temporal alignment has been sensitivity-tested, but independently verified equivalence to official benchmark evaluation implementations is still outstanding.

H. Implications and Future Directions

Future work can retain the conservation principle while using stronger source generalized representations, multi-scale temporal evidence, or window-limited reallocation. Restricting how far a score may move would create a useful middle ground between no reallocation and video-wide permutation. The same controlled idea may also help other sequence tasks distinguish errors in confidence magnitude from errors in temporal placement.

VII. Conclusion

This study examined temporal evidence placement as a possible source of error in cross domain weakly supervised video anomaly localization. TEMRAL uses local changes in frozen representations to rank temporal positions and then reallocates only the detector's original segment scores. Because the final sequence is an exact per video permutation at native resolution, the method changes where anomaly evidence appears without changing which score values exist.

All configuration choices were made on UCF-Crime validation data before target evaluation. Under transfer to ShanghaiTech, TEMRAL improved ROC-AUC and average precision for mean, top-k, and attention MIL detectors with I3D and R3D representations in every model seed evaluation. The same fixed configuration improved PR-AUC and ROC-AUC across all tested I3D detectors on XD-Violence. Clustered bootstrap analysis, fixed smoother comparisons, and random multiset preserving controls support the interpretation that the observed effect is tied to evidence guided temporal ordering rather than score inflation or arbitrary permutation.

TEMRAL does not replace representation learning, domain generalization, or domain adaptation, and it cannot create discrimination that a frozen detector failed to produce. Its contribution is a controlled way to separate score generation from score placement. The results indicate that temporal misallocation can be a distinct and partially correctable component of cross domain localization error, motivating future work on multi scale and locally constrained evidence reallocation.

Additional diagnostics with source-independent splits, GT-guided score placement, restricted temporal movement, boundary/tie sensitivity, alternative frame alignments, and CPU/GPU timing clarify the scope of this finding. Source selection is only partially stable, and unrestricted placement often moves scores substantial distances; improved global metrics should therefore not be taken as proof of semantically local assignments or strong absolute transfer discrimination.

Acknowledgment

The authors are grateful to the Deanship of Scientific Research at Jadara University for providing financial support for this publication.

References

[1] W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 6479–6488, doi: 10.1109/CVPR.2018.00678.

[2] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, “MIST: Multiple instance self-training framework for video anomaly detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 14009–14018.

[3] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, “Weakly-supervised video anomaly detection with robust temporal feature magnitude learning,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 4975–4986.

[4] S. Park, H. Kim, M. Kim, D. Kim, and K. Sohn, “Normality-guided multiple instance learning for weakly supervised video anomaly detection,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2023, pp. 2665–2674.

[5] J. Chen, L. Li, L. Su, Z.-J. Zha, and Q. Huang, “Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 18319–18329, doi: 10.1109/CVPR52733.2024.01734.

[6] Y. Pu, X. Wu, L. Yang, and S. Wang, “Learning prompt-enhanced context features for weakly-supervised video anomaly detection,” IEEE Trans. Image Process., vol. 33, pp. 4923–4936, 2024, doi: 10.1109/TIP.2024.3451935.

[7] B. Sun, J. Chen, Z. Wu, F. Gao, F. Yang, L. Su, and Y. Wang, “Joint learning of general and diverse patterns with mixture of memory experts for weakly-supervised video anomaly detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2026, pp. 35638–35647.

[8] P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, P. Wang, and Y. Zhang, “VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection,” Proc. AAAI Conf. Artif. Intell., vol. 38, no. 6, pp. 6074–6082, 2024, doi: 10.1609/aaai.v38i6.28423.

[9] Z. Yang, J. Liu, and P. Wu, “Text prompt with normality guidance for weakly supervised video anomaly detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 18899–18908, doi: 10.1109/CVPR52733.2024.01788.

[10] Y. Jain, A. Dabouei, and M. Xu, “Cross-domain learning for video anomaly detection with limited supervision,” in Computer Vision – ECCV 2024. Springer, 2024, pp. 468–484, doi: 10.1007/978-3-031-73404-5_27.

[11] A. Aich, K.-C. Peng, and A. K. Roy-Chowdhury, “Cross-domain video anomaly detection without target domain adaptation,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2023, doi: 10.1109/WACV56688.2023.00261.

[12] Y. Pu and X. Wu, "Locality-aware attention network with discriminative dynamics learning for weakly supervised anomaly detection," in Proc. IEEE Int. Conf. Multimedia Expo (ICME), 2022, pp. 1–6, doi: 10.1109/ICME52920.2022.9859718.

[13] Y. Zhou, Y. Qu, X. Xu, F. Shen, J. Song, and H. T. Shen, "BatchNorm-based weakly supervised video anomaly detection," IEEE Trans. Circuits Syst. Video Technol., vol. 34, no. 12, pp. 13642–13654, 2024, doi: 10.1109/TCSVT.2024.3450734.

[14] Y. Wang and S. Zhao, "Weakly supervised video anomaly detection with anomaly-connected components and intention reasoning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2026, pp. 28545–28556.

[15] C. Chu, N. Japar, and C. K. Lim, "REBA: Residual mixture-of-experts and bidirectional video-text alignment for better fine-grained weakly supervised video anomaly detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Findings, 2026, pp. 8280–8290.

[16] S. Majhi, R. Dai, Q. Kong, L. Garattoni, G. Francesca, and F. Brémond, "OE-CTST: Outlier-embedded cross temporal scale transformer for weakly-supervised video anomaly detection," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2024, pp. 8574–8583.

[17] A. Acharya, H. Sapkota, Q. Yu, and X. Liu, "The road less seen: Segment exploration for weakly supervised video anomaly detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2026, pp. 14147–14156.

[18] K. V. Thakare, Y. Raghuwanshi, D. P. Dogra, H. Choi, and I.-J. Kim, "DyAnNet: A scene dynamicity guided self-trained video anomaly detection network," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2023, pp. 5541–5550.

[19] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang, "Post-processing temporal action detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 18837–18845.

[20] J. Zhou, L. Huang, L. Wang, S. Liu, and H. Li, "Improving weakly supervised temporal action localization by bridging train-test gap in pseudo labels," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 23003–23012.

[21] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding-based anomaly detection in stacked RNN framework," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 341–349, doi: 10.1109/ICCV.2017.45.

[22] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in Computer Vision – ECCV 2020. Springer, 2020, pp. 322–339, doi: 10.1007/978-3-030-58577-8_20.

[23] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the Kinetics dataset," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 6299–6308.

[24] A. B. Owen, "The pigeonhole bootstrap," Ann. Appl. Stat., vol. 1, no. 2, pp. 386–411, 2007, doi: 10.1214/07-AOAS122.

[25] A. B. Owen and D. Eckles, "Bootstrapping data arrays of arbitrary order," Ann. Appl. Stat., vol. 6, no. 3, pp. 895–927, 2012, doi: 10.1214/12-AOAS547.

[26] C. Xu, C. Li, and H. Xing, "Discriminative score suppression for weakly supervised video anomaly detection," in Proc. Winter Conf. Appl. Comput. Vis. (WACV), 2025, pp. 9569–9578.