Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085338

ARTICLE

Boundary Measure Alignment via Optimal Transport for Temporal Action Detection

Tiyao Zhang and Xue Yuan*

School of Automation and Intelligence, Beijing Jiaotong University, Beijing, China

*Corresponding Author: Xue Yuan. Email: xyuan@bjtu.edu.cn

Received: 09 May 2026; Accepted: 16 July 2026; Published: 15 September 2026

ABSTRACT: Temporal action detection aims to localize action instances in untrimmed videos and recognize their categories. Although recent detectors have achieved strong performance, accurate boundary localization remains challenging due to gradual action transitions, temporal ambiguity, and annotation uncertainty. Existing boundary supervision usually relies on point-wise classification or local regression losses, which compare predictions and targets at corresponding temporal positions but do not explicitly model temporal displacement between misaligned boundary responses. To address this issue, this paper proposes Boundary Measure Alignment (BMA), a training-stage auxiliary objective for temporal action detection. BMA represents predicted and annotated action starts and ends as probability measures over the feature-level temporal domain. Predicted boundary measures are constructed from the rising and falling edges of the foreground probability trajectory, while target measures are generated by Gaussian smoothing around annotated boundaries. Entropic optimal transport is then used to align predicted and target measures, allowing the loss to explicitly encode temporal boundary displacement. During inference, the BMA branch is removed, so the original detector pipeline remains unchanged. We integrate BMA into AFSD, FCOS, and an ActionFormer-style detector. Experiments on THUMOS14 and ActivityNet-v1.3 show consistent average mAP improvements. On THUMOS14, BMA improves the average mAP of AFSD, FCOS, and the ActionFormer-style detector from 52.0%, 45.3%, and 66.8% to 53.3%, 47.1%, and 69.1%, respectively. On ActivityNet-v1.3, the corresponding average mAP improves from 34.40%, 32.30%, and 35.60% to 34.83%, 32.82%, and 36.04%.

KEYWORDS: Temporal action detection; boundary localization; optimal transport; distribution alignment; video understanding

1 Introduction

Temporal Action Detection (TAD) aims to localize the start and end times of action instances in untrimmed videos and recognize their categories [1,2]. As a fundamental task in video understanding, TAD has broad applications in surveillance, human-computer interaction, sports analysis, and other real-world scenarios [3]. Unlike video-level action classification, TAD requires both semantic discrimination and accurate temporal boundary localization [4,5]. Although recent detectors have achieved strong performance through improved temporal modeling and Transformer-based architectures [6–9], accurate boundary localization remains challenging due to gradual action transitions, ambiguous visual changes, local noise in long videos, and annotation uncertainty [10–12]. Under strict temporal Intersection-over-Union (tIoU) thresholds, even small boundary errors can noticeably affect detection performance [12].

Existing boundary supervision commonly relies on point-wise classification, start/end confidence prediction, or local regression losses, such as cross-entropy, L1 distance, and IoU-based losses [13,14]. These objectives mainly compare predictions and targets at corresponding temporal positions. When predicted boundary responses are shifted from ground-truth boundaries, conventional local supervision does not explicitly encode how far the predicted response should move along the temporal axis [15,16].

To address this limitation, we propose Boundary Measure Alignment (BMA), a training-stage auxiliary objective for TAD. BMA represents predicted and annotated starts and ends as probability measures over the feature-level temporal domain. Predicted measures are constructed from the rising and falling edges of the foreground probability trajectory, while target measures are generated by Gaussian smoothing around annotated boundaries [1,2]. Entropic Optimal Transport (OT) is then used to align predicted and target measures, allowing the loss to explicitly model temporal boundary displacement [17,18]. Unlike OT-based assignment methods that use OT for candidate-instance matching, BMA applies OT to temporal boundary measures, where the transport cost directly reflects boundary displacement [15,19,20]. Fig. 1 illustrates the difference between conventional point-wise boundary supervision and the proposed measure-level boundary alignment.

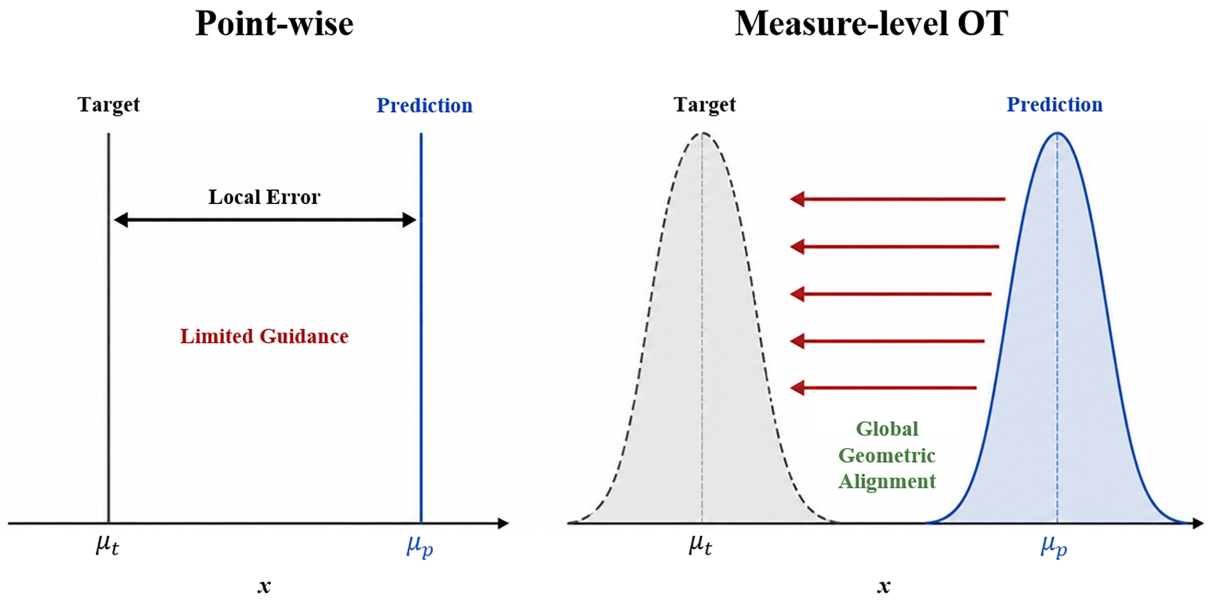

Figure 1: Comparison between point-wise boundary supervision and measure-level alignment. Conventional methods rely on point-wise matching at local temporal positions, whereas the proposed method represents boundaries as distributions and performs global alignment through optimal transport, thereby providing more effective geometry-aware optimization signals under temporal misalignment.

BMA is used only during training and is removed during inference. Therefore, it keeps the detector architecture, model parameters, and post-processing pipeline unchanged at test time. We integrate BMA into AFSD, FCOS, and an ActionFormer-style detector. Experiments on THUMOS14 and ActivityNet-v1.3 show consistent average mAP improvements across these detectors, indicating that BMA provides useful auxiliary boundary supervision for detectors with temporal foreground or boundary response outputs.

The main contributions of this paper are summarized as follows:

(1) We introduce a boundary measure representation for TAD, where action starts and ends are separately represented as probability measures on the one-dimensional temporal domain.

(2) We propose an entropic OT-based boundary measure alignment loss that explicitly encodes temporal displacement between predicted and target boundary measures.

(3) We integrate BMA into AFSD, FCOS, and an ActionFormer-style detector. Experiments on THUMOS14 and ActivityNet-v1.3 show consistent improvements while keeping the original inference pipelines unchanged.

2 Related Work

2.1 Temporal Action Detection

Temporal Action Detection (TAD) aims to localize and classify action instances in untrimmed videos. Existing methods can be broadly grouped into proposal-based methods, anchor-free dense prediction methods, and Transformer-based detectors. Proposal-based methods first generate temporal candidates and then refine or classify them [5,21,22]. Anchor-free methods remove predefined temporal anchors and predict action categories, boundary scores, or localization offsets at dense temporal locations [23,24]. More recent Transformer-based detectors improve long-range temporal modeling through self-attention, multi-scale feature aggregation, or query-based decoding [5–8]. Although these detectors have improved temporal representation and detection accuracy, their boundary learning objectives are still commonly formulated as local classification, regression, or IoU-related losses. These local objectives provide useful supervision but do not explicitly encode the temporal displacement between shifted boundary responses and annotated boundaries.

2.2 Boundary Modeling in TAD

Accurate boundary modeling is critical for high-quality temporal localization, especially under strict tIoU thresholds. Existing methods enhance boundary localization by predicting frame-level foreground probabilities, start/end confidence curves, boundary offsets, salient boundary features, or IoU-based localization scores [3,14,25,26]. These designs make detectors more sensitive to action transitions and improve proposal refinement. However, most of them still supervise boundary responses at corresponding temporal positions. When a predicted boundary peak is shifted from the annotated boundary, point-wise losses can penalize the mismatch but do not directly describe how far the response should move along the temporal axis. This motivates us to represent action starts and ends as temporal probability measures and align them with a displacement-aware OT objective.

2.3 Distribution and Set-Based Learning

Distributional and set-based learning have been widely used to reduce the limitations of point-wise supervision. DETR-style methods formulate detection as set prediction and perform global matching between predictions and targets [16,27,28]. Other methods model localization uncertainty using discrete probability distributions over spatial coordinates [29]. These approaches show that distributional representations and global matching can provide richer supervision than independent point-wise losses. However, they mainly focus on spatial object detection, bounding-box uncertainty, or candidate-level assignment. They do not directly address the alignment of one-dimensional temporal boundary responses in TAD.

2.4 Optimal Transport in Deep Learning

Optimal Transport (OT) provides a principled framework for measuring distribution discrepancy with an explicit transport cost. It has been used for domain adaptation and feature alignment [30,31], generative modeling [32], correspondence estimation in dynamic visual scenes [33], and dynamic label assignment in object detection [15]. Entropic regularization and Sinkhorn iterations make OT differentiable and efficient

for deep learning [34,35]. In detection tasks, OT is often used to match candidates with ground-truth instances. Different from these applications, BMA applies OT to predicted and target start/end boundary measures defined on the same one-dimensional temporal domain. Therefore, the transport cost has a direct temporal interpretation: it reflects how much boundary probability mass should be moved along the temporal axis. This distinction makes BMA different from general feature distribution alignment and OT-based candidate assignment.

3 Methodology

3.1 Overview

This section presents Boundary Measure Alignment (BMA), a training-stage auxiliary objective for temporal action detection. As shown in Fig. 2, BMA constructs predicted start/end measures from detector outputs and aligns them with target measures generated from annotated boundaries. The OT computation is used only during training and is removed during inference, so the original detector pipeline remains unchanged.

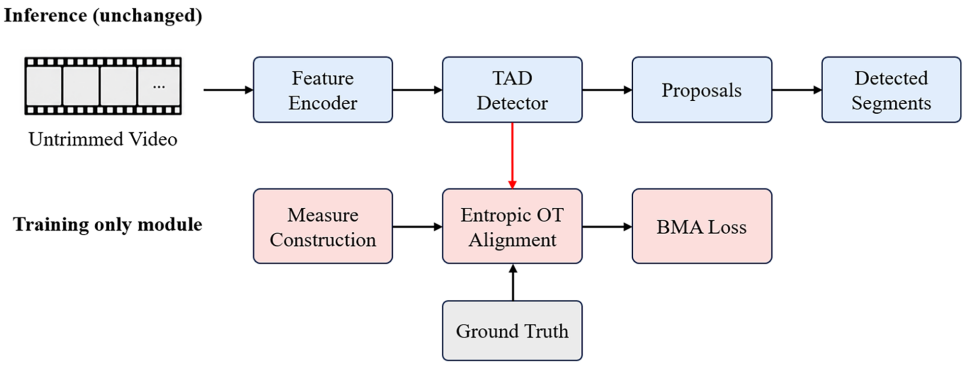

Figure 2: Overall pipeline of the proposed BMA module. Given an untrimmed video, temporal features are first extracted by the backbone network, and the temporal action detector generates action proposals and foreground probability trajectories. The BMA module constructs start and end boundary measures from the predicted trajectories and aligns them with target measures generated from ground-truth boundaries through entropy-regularized optimal transport. This module is used only as an auxiliary supervision branch during training and does not introduce additional optimal transport computation during inference.

BMA is not used for proposal assignment or detection-result matching. Instead, it measures the temporal displacement between predicted and target boundary measures defined on the same feature-level temporal domain. This makes the transport plan temporally interpretable and distinguishes BMA from general distribution matching or OT-based assignment. The overall training and inference pipeline of BMA is shown in Fig. 2.

3.2 Boundary Measure Representation

Given an untrimmed video, the detector extracts a temporal feature sequence of length T'. We define the feature-level temporal domain as

$$\Omega = \{1, 2, \ldots, T'\} \tag{1}$$

For the k-th annotated action instance with original boundaries (s_k, e_k), the boundaries are mapped to the feature-level domain according to the feature stride, yielding $(\hat{s}_k, \hat{e}_k)$. In implementation, the mapped positions are clipped to $[1, T']$ and rounded to the nearest feature index.

Let $l_t \in \mathbb{R}^C$ denote the classification logits at temporal location t. For a multi-label sigmoid head, the class-agnostic foreground probability is defined as

$$p_t = \max_{c \in \{1,\ldots,C\}} \sigma(l_{t,c}) \tag{2}$$

For detectors with an explicit background category, it can also be defined as

$$p_t = 1 - P_{\text{bg}}(t) \tag{3}$$

The maximum operation is performed independently at each temporal location over the category dimension. The same foreground trajectory is used for start and end measure construction: positive temporal variations indicate potential starts, while negative variations indicate potential ends.

An action start usually corresponds to a rising edge of the foreground trajectory, while an action end corresponds to a falling edge. We compute the first-order temporal difference as

$$\Delta p_t = p_t - p_{t-1} \tag{4}$$

with $\Delta p_1 = 0$. The predicted start and end responses are

$$r_s(t) = \text{ReLU}\,(\Delta p_t) \tag{5}$$
$$r_e(t) = \text{ReLU}\,(-\Delta p_t) \tag{6}$$

The responses are normalized into probability measures:

$$\mu_s(t) = \frac{r_s(t) + \varepsilon}{\sum_{\tau=1}^{T'} (r_s(\tau) + \varepsilon)} \tag{7}$$
$$\mu_e(t) = \frac{r_e(t) + \varepsilon}{\sum_{\tau=1}^{T'} (r_e(\tau) + \varepsilon)} \tag{8}$$

Thus, μ_s and μ_e satisfy

$$\sum_{t=1}^{T'} \mu_s(t) = 1 \tag{9}$$
$$\sum_{t=1}^{T'} \mu_e(t) = 1 \tag{10}$$

For target measures, discrete annotated boundaries are converted into Gaussian-smoothed distributions. Given K action instances, the target start and end measures are

$$v_s(t) = \frac{\sum_{k=1}^{K} \exp\left(-\frac{(t - \hat{s}_k)^2}{2\sigma_b^2}\right)}{\sum_{\tau=1}^{T'} \sum_{k=1}^{K} \exp\left(-\frac{(\tau - \hat{s}_k)^2}{2\sigma_b^2}\right)} \tag{11}$$

$$v_e(t) = \frac{\sum_{k=1}^{K} \exp\left(-\frac{(t-\hat{e}_k)^2}{2\sigma_b^2}\right)}{\sum_{\tau=1}^{T'} \sum_{k=1}^{K} \exp\left(-\frac{(\tau-\hat{e}_k)^2}{2\sigma_b^2}\right)} \tag{12}$$

here, σ_b controls the smoothness of the target boundary measures. The negative sign in the exponential term ensures that each target measure peaks around the annotated boundary and decays with temporal distance.

The target measures aggregate all annotated boundaries in the video. Therefore, BMA provides class-agnostic global boundary regularization rather than instance-specific boundary matching. When multiple instances overlap or boundaries are densely located, a single foreground trajectory may produce entangled responses; this limitation is discussed in the qualitative analysis.

Fig. 3 illustrates the construction of predicted and target boundary measures.

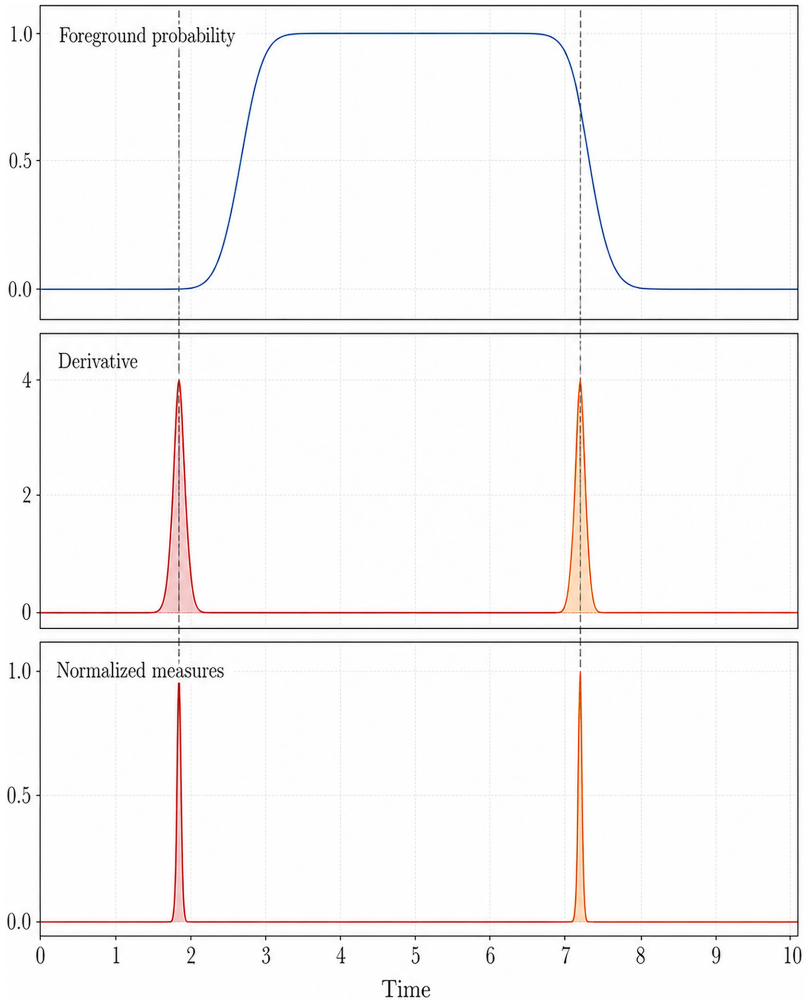

Figure 3: Illustration of boundary measure construction. The foreground probability trajectory shows a rising trend at the action start and a falling trend at the action end. By extracting the positive and negative temporal derivatives, the start and end variation intensities are respectively characterized. After non-negative mapping and normalization, they are converted into probability measures defined over the temporal domain.

3.3 Geometry-Aware Alignment via Optimal Transport

When the predicted boundary response is shifted from the annotated boundary, point-wise losses mainly penalize values at corresponding temporal locations. If a predicted boundary peak lies around t_p while the target peak lies around t_g, such losses do not explicitly encode the displacement $\mid t_p - t_g \mid$. OT provides a displacement-aware objective by transporting probability mass from the predicted boundary measure to the target measure.

For a predicted measure m and target measure m^*, we define the temporal cost matrix as $C_{ij} = \mid t_i - t_j \mid^p$. The entropic OT alignment loss is

$$\mathcal{L}_{OT}(m, m^\star) = \min_{\pi \in \Pi(m, m^\star)} \sum_{i,j} \pi_{ij} C_{ij} + \varepsilon_{OT} \sum_{i,j} \pi_{ij} (\log \pi_{ij} - 1) \tag{13}$$

where π is the transport plan, and ε_{OT} controls the smoothness of the plan. When two measures are close to Dirac-like peaks at t_p and t_g, the transport cost is directly related to $\mid t_p - t_g \mid^p$. Thus, OT directly penalizes temporal displacement between boundary measures.

BMA does not directly optimize tIoU, which is computed from final detected segments after classification, regression, and post-processing. Instead, it reduces boundary-response displacement. Since strict tIoU thresholds are sensitive to start/end errors, this auxiliary supervision can indirectly improve boundary-sensitive detection. Fig. 4 visualizes the measure-level alignment process under temporal boundary offset.

3.4 Numerical Solution with Sinkhorn Iterations

We use Sinkhorn iterations to obtain a differentiable approximation of the entropic OT objective. Given the Gibbs kernel $K = \exp(-C/\varepsilon_{OT})$, the dual vectors are updated as

$$u^{(l+1)} = a \oslash (K v^{(l)}) \tag{14}$$

$$v^{(l+1)} = b \oslash (K^\top u^{(l+1)}) \tag{15}$$

where a and b are the predicted and target measure vectors. After M iterations, the transport plan is estimated and $\langle \pi^*, C \rangle$ is used as the alignment cost. This approximation provides stable gradients while keeping training computationally feasible.

3.5 Joint Training Objective

BMA is added to the original detection objective as an auxiliary loss. The baseline detector loss is written as

$$\mathcal{L}_{det} = \mathcal{L}_{cls} + \lambda_{reg} \mathcal{L}_{reg} + \lambda_{aux} \mathcal{L}_{aux} \tag{16}$$

where $\mathcal{L}_{cls}$, $\mathcal{L}_{reg}$, and $\mathcal{L}_{aux}$ denote classification, regression, and detector-specific auxiliary losses. The BMA loss is

$$\mathcal{L}_{BMA} = \mathcal{L}_{ot}(m_s, m_s^*) + \mathcal{L}_{ot}(m_e, m_e^*) \tag{17}$$

The final joint training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{det} + \lambda_{BMA} \mathcal{L}_{BMA} \tag{18}$$

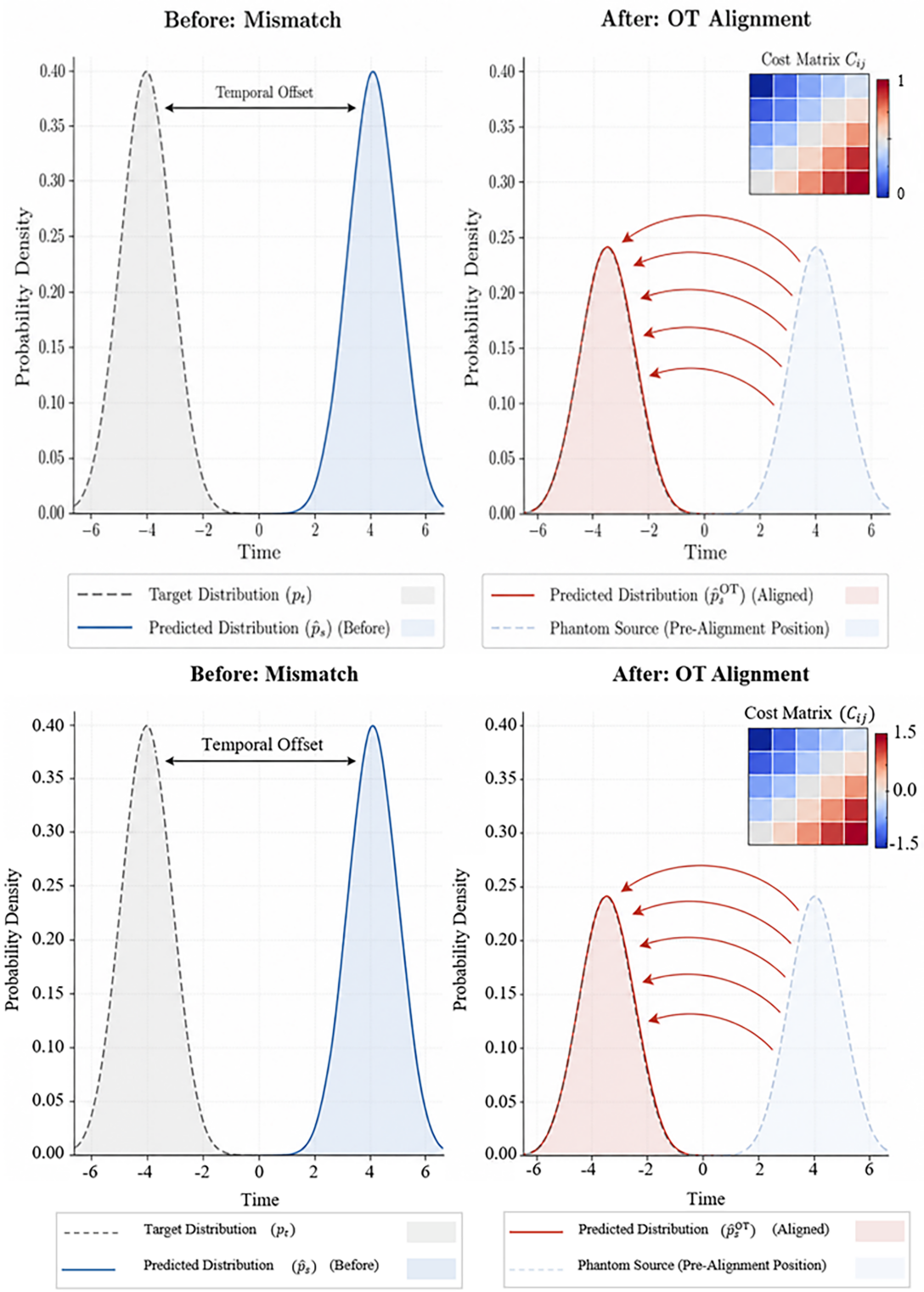

Figure 4: Illustration of boundary measure alignment under support mismatch. When the predicted measure deviates from the target measure along the temporal axis, conventional local losses may struggle to provide effective optimization signals. In contrast, optimal transport explicitly models the temporal transport cost and moves probability mass from the predicted distribution toward the target distribution, thereby achieving geometry-aware boundary alignment.

BMA does not replace the original classification, regression, or assignment objectives. It only regularizes global start/end boundary responses during training, while the original detector remains responsible for category prediction and final segment localization.

4 Experiments

We evaluate BMA on THUMOS14 and ActivityNet-v1.3. To examine detector compatibility, BMA is integrated into AFSD, FCOS, and an ActionFormer-style detector. For each detector, BMA is used only during training, while the original inference and post-processing pipeline remains unchanged. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

4.1 Experimental Setup

4.1.1 Dataset

THUMOS14. THUMOS14 contains untrimmed videos from 20 action categories. Following the standard protocol, we train on the validation set and evaluate on the test set. Its dense action instances make it suitable for evaluating fine-grained temporal boundary localization.

ActivityNet-v1.3. ActivityNet-v1.3 is a large-scale benchmark with 200 action categories. We train on the training set and evaluate on the validation set. Its longer videos and diverse action categories are used to evaluate the generalization ability of the proposed BMA module.

4.1.2 Evaluation Metrics

We use mean Average Precision (mAP) as the main evaluation metric. For THUMOS14, we report mAP at tIoU thresholds of 0.3, 0.4, 0.5, 0.6, and 0.7, together with their average. For ActivityNet-v1.3, we follow the standard protocol and report mAP at tIoU thresholds from 0.5 to 0.95 with a step size of 0.05, as well as the average mAP.

4.1.3 Implementation Details

We use pre-extracted I3D features as input whenever the corresponding baseline adopts I3D features. The original losses of each detector are kept unchanged, and BMA is added as an auxiliary boundary measure alignment loss during training. During inference, the BMA branch and OT computation are disabled.

For detectors with multi-label sigmoid classification heads, such as FCOS and the ActionFormer-style detector, we construct the foreground probability trajectory by taking the maximum action probability over all classes at each temporal location. For detectors with explicit boundary confidence outputs, such as AFSD, the start and end confidence sequences are normalized as predicted boundary measures. Target measures are generated by Gaussian smoothing around annotated boundaries in the feature-level temporal domain. Unless otherwise specified, we set $\varepsilon = 10^{-6}$, $\sigma_b = 1.0$, $\lambda_{BMA} = 0.05$, $\varepsilon_{OT} = 0.05$, use $C_{ij} = \mid i - j \mid$, and perform 20 Sinkhorn iterations.

4.2 Main Results

The main comparison results are reported in Table 1.

The main comparison results are reported in Table 1. On THUMOS14, BMA consistently improves all three detectors. The average mAP of AFSD, FCOS, and the ActionFormer-style detector increases from 52.0%, 45.3%, and 66.8% to 53.3%, 47.1%, and 69.1%, respectively. The gains are more evident at stricter thresholds; for example, mAP@0.7 improves by 1.6%, 2.2%, and 2.9% for the three detectors.

Table 1: Comparison with baseline and existing methods on THUMOS14 and ActivityNet-v1.3.

Method	Feat.	THUMOS14						ActivityNet-v1.3			
		0.3	0.4	0.5	0.6	0.7	Avg.	0.5	0.75	0.95	Avg.
BSN [36]	TSN	53.5	45.0	36.9	28.4	20.0	36.8	46.46	29.96	8.02	29.17
BMN [2]	TSN	56.0	47.4	38.8	29.7	20.5	38.5	50.07	34.78	8.29	33.85
PBRNet [37]	I3D	58.5	54.6	51.3	41.8	29.5	47.1	53.96	34.97	8.98	35.01
VSGN [38]	*	66.7	60.4	52.4	41.0	30.4	50.2	52.38	36.01	8.37	35.07
RCL [39]	*	70.1	62.3	52.9	42.7	30.7	51.0	54.19	36.19	9.17	35.98
TAGS [40]	I3D	68.6	63.8	57.0	46.3	31.8	52.8	56.30	36.80	**9.60**	36.50
TALLFormer [41]	Swin	76.0	–	63.2	–	34.5	59.2	54.10	36.20	7.90	35.60
AFSD [42]	I3D	67.3	62.4	55.5	43.7	31.1	52.0	52.40	35.30	6.50	34.40
+BMA (Ours)	I3D	**68.4**	**63.5**	**56.8**	**45.1**	**32.7**	**53.3**	**53.05**	**35.76**	**6.58**	**34.83**
FCOS [43]	I3D	62.6	56.3	47.9	36.5	23.2	45.3	50.22	33.51	5.57	32.30
+BMA (Ours)	I3D	**64.0**	**57.8**	**49.7**	**38.4**	**25.4**	**47.1**	**50.91**	**34.2**	**5.68**	**32.8**
ActionFormer [4]	I3D	82.1	77.8	71.0	59.4	43.9	66.8	53.50	36.20	**8.20**	35.60
+BMA (Ours)	I3D	**84.4**	**79.5**	**73.0**	**61.5**	**46.8**	**69.1**	**54.61**	**36.90**	7.34	**36.04**

Note: Bold numbers indicate the best performance among all compared methods. The symbol * denotes that the feature extraction architecture used by the corresponding method is not specified in its original paper.

On ActivityNet-v1.3, BMA also improves the average mAP of AFSD, FCOS, and the ActionFormer-style detector from 34.40%, 32.30%, and 35.60% to 34.83%, 32.82%, and 36.04%, respectively. The gains are smaller than on THUMOS14, and the ActionFormer-style detector does not improve at mAP@0.95. This suggests that BMA is affected by long video duration, coarser annotations, action-duration variation, and feature-level temporal resolution.

4.3 Cross-Detector and Cross-Dataset Analysis

Table 1 shows that BMA is not limited to the ActionFormer-style detector. It brings average mAP gains on AFSD, FCOS, and the ActionFormer-style detector under the same experimental setting. This indicates that BMA can provide auxiliary boundary supervision for detectors with temporal foreground or boundary response outputs.

The gains are larger on THUMOS14 than on ActivityNet-v1.3. THUMOS14 contains denser action instances, where reducing boundary displacement more directly improves tIoU. ActivityNet-v1.3 contains longer videos and coarser boundary annotations; therefore, extremely strict metrics such as mAP@0.95 are more sensitive to annotation ambiguity and feature-level discretization. Thus, BMA improves average detection performance across detectors, but its effect under extremely strict boundary matching on ActivityNet-v1.3 remains limited.

4.4 Boundary Localization Analysis

Since the objective of BMA is to improve temporal boundary alignment quality, we further evaluate its effect using boundary localization errors. For each ground-truth action instance, we select the prediction segment with the same category and the highest tIoU after non-maximum suppression as its matched prediction.

All boundary errors are computed in the feature-level temporal domain rather than on the raw video-frame index. In our THUMOS14 setting, I3D features are extracted using 16-frame clips at approximately 30 fps with a temporal stride of 4 frames. Therefore, one feature-level temporal step corresponds to 4 original

video frames, i.e., approximately 0.133 s. Accordingly, an error value of e in Table 2 corresponds to 4e original frames, or 0.133e s.

Table 2: Feature-level boundary localization error analysis on THUMOS14.

Method	Start Error	End Error	Avg. Boundary Error	mAP@0.7	Avg. mAP
Baseline	1.3/0.173 s	1.1/0.147 s	1.2/0.160 s	43.9	66.8
Baseline + BMA	0.9/0.120 s	0.7/0.093 s	0.8/0.107 s	46.8	69.1

Given the i-th matched prediction segment $(\hat{s}_i, \hat{e}_i)$ and its corresponding ground-truth action segment (s_i, e_i), the start boundary error and end boundary error are respectively defined as:

$$E_s = \frac{1}{N}\sum_{i=1}^{N} \mid \hat{s}_i - s_i \mid \tag{19}$$

$$E_e = \frac{1}{N}\sum_{i=1}^{N} \mid \hat{e}_i - e_i \mid \tag{20}$$

The average boundary error is defined as:

$$E_b = \frac{1}{2}(E_s + E_e) \tag{21}$$

As shown in Table 2, compared with the baseline model, introducing BMA reduces the start error from 1.3 to 0.9 feature steps and the end error from 1.1 to 0.7 feature steps. Since one feature step corresponds to 4 original frames, i.e., approximately 0.133 s, the average boundary error is reduced from 1.2 feature steps, or 0.160 s, to 0.8 feature steps, or 0.107 s. This corresponds to an average reduction of 0.4 feature steps, namely 1.6 original frames or approximately 0.053 s. These results indicate that the performance improvement comes not only from better action category discrimination, but also from more accurate temporal boundary localization. The reduction in boundary errors is consistent with the performance improvement at higher tIoU thresholds on THUMOS14.

4.5 Ablation Experiment

4.5.1 Start/End Measure Ablation

As shown in Table 3, aligning either start or end measures improves the baseline, and jointly aligning both achieves the best performance. This confirms that start and end boundaries provide complementary localization cues. Therefore, we use both start and end BMA in the final model.

Table 3: Ablation study on start and end boundary measures on THUMOS14.

Model	Start Measure	End Measure	mAP@0.7	Avg. mAP
Baseline	–	–	43.9	66.8
+Start BMA	✓	–	45.2	67.3
+End BMA	–	✓	45.6	67.8
+Start & End BMA	✓	✓	46.8	69.1

Joint start-end alignment provides the most balanced boundary improvement, although a single-side alignment may occasionally yield comparable mAP due to metric variance.

4.5.2 Loss Function Comparison

Table 4 compares different boundary alignment losses.

Table 4: Comparison of boundary alignment losses on THUMOS14.

Boundary Alignment Loss	mAP@0.7	Avg. mAP
None	43.9	66.8
BCE/CE	46.2	68.5
L2 Distribution Loss	46.4	68.5
KL Divergence	46.2	68.6
Entropic OT, Ours	46.8	69.1

Table 4 compares different boundary alignment losses. BCE/CE, L2, and KL losses improve the baseline, showing that boundary distribution supervision is useful. Entropic OT achieves the best mAP@0.7 and average mAP because it additionally models temporal transport cost between different boundary locations.

4.5.3 Robustness to Noisy Foreground Trajectories

Since BMA uses temporal differences of the foreground probability trajectory, we evaluate its robustness to noisy foreground responses. Gaussian perturbations are added to the trajectory used for BMA loss construction, where α controls the noise strength.

As shown in Table 5, BMA remains effective under small and moderate perturbations. When $\alpha = 0.03$, it still achieves 68.6% average mAP and 46.1% mAP@0.7, higher than the baseline results of 66.8% and 43.9%. With larger noise, performance decreases but remains above the baseline. This indicates that BMA tolerates moderate foreground fluctuations, while still relying on reasonably reliable foreground estimates.

Table 5: Robustness analysis under noisy foreground probability trajectories on THUMOS14.

Method	Noise Level α	mAP@0.5	mAP@0.6	mAP@0.7	Avg. mAP
Baseline	–	71.0	59.4	43.9	66.8
+BMA	0	73.0	61.5	46.8	69.1
+BMA	0.01	72.8	61.3	46.5	68.9
+BMA	0.03	72.5	60.9	46.1	68.6
+BMA	0.05	72.0	60.2	45.4	68.0

4.5.4 Analysis on Dense and Overlapping Boundaries

We further evaluate the behavior of BMA under different boundary-density conditions. As shown in Table 6, BMA improves the average mAP from 68.2% to 70.8% on sparse-boundary videos and from 61.4% to 62.2% on dense-boundary videos, corresponding to absolute gains of 2.6 and 0.8 percentage points, respectively. At mAP@0.7, BMA improves sparse-boundary videos from 45.1% to 48.4%, with an absolute gain of 3.3 percentage points, while improving dense-boundary videos from 36.7% to 37.5%, with an absolute gain of 0.8 percentage points. These results indicate that BMA remains effective in dense-boundary videos, but adjacent or overlapping action boundaries may entangle foreground variations and weaken the effect of global boundary-measure alignment.

Table 6: Analysis on sparse-boundary and dense-boundary videos on THUMOS14.

Subset	Method	mAP@0.5	mAP@0.6	mAP@0.7	Avg. mAP
Sparse-boundary videos	Baseline	72.4	61.0	45.1	68.2
Sparse-boundary videos	+BMA	74.8	63.6	48.4	70.8
Dense-boundary videos	Baseline	66.5	52.8	36.7	61.4
Dense-boundary videos	+BMA	67.2	53.6	37.5	62.2

4.5.5 Fine-Grained Analysis under Different Action Durations

To further analyze the behavior of BMA under different temporal conditions, we divide action instances in THUMOS14 into short, medium, and long groups according to their temporal duration. Short actions usually have more sensitive boundary changes, while long actions are more likely to contain internal foreground fluctuations. Table 7 reports the results of the ActionFormer-style detector on different duration groups.

Table 7: Fine-grained analysis under different action durations on THUMOS14.

Action Duration	Method	mAP@0.5	mAP@0.6	mAP@0.7	Avg. mAP
Short	Baseline	67.8	55.6	39.2	63.5
Short	+BMA	70.5	58.8	42.5	66.4
Medium	Baseline	72.1	60.4	44.1	67.2
Medium	+BMA	74.2	62.8	47.0	69.5
Long	Baseline	73.4	61.8	45.0	68.6
Long	+BMA	75.0	63.2	46.3	70.2

As shown in Table 7, BMA improves the baseline in all duration groups. For short, medium, and long actions, the average mAP increases from 63.5% to 66.4%, from 67.2% to 69.5%, and from 68.6% to 70.2%, corresponding to absolute gains of 2.9, 2.3, and 1.6 percentage points, respectively. At the stricter mAP@0.7 threshold, the gains are 3.3, 2.9, and 1.3 percentage points for short, medium, and long actions, respectively. These numerical results show that BMA is more beneficial for short and medium actions, where a small boundary shift can cause a larger tIoU change. For long actions, BMA still improves detection performance, but the gain is relatively smaller because internal foreground fluctuations may make boundary response construction less stable.

4.6 Hyperparameter Analysis

4.6.1 Effect of λ_{BMA}

The experimental results with different λ_{BMA} are shown in Table 8. A moderate BMA weight provides the best trade-off between boundary alignment and the original detection objective. Too small weights fail to provide sufficient geometric supervision, whereas overly large weights may disturb the optimization of the detector.

4.6.2 Effect of Gaussian Bandwidth σ

The experimental results with different σ are shown in Table 9. An appropriate Gaussian bandwidth is important for constructing target boundary measures. A small bandwidth makes the target distribution too sharp, while an overly large bandwidth weakens boundary localization by over-smoothing the supervision.

Table 8: Effect of BMA loss weight on THUMOS14.

λ_{BMA}	Avg mAP
0.01	67.6
0.02	68.4
0.05	69.1
0.10	68.6

Table 9: Effect of Gaussian bandwidth on THUMOS14.

σ	mAP@0.7	Avg mAP
0.5	46.7	69.0
1.0	46.8	69.1
2.0	46.8	69.1
4.0	46.4	68.8

4.6.3 OT-Related Hyperparameter Analysis

We further analyze the influence of OT-related hyperparameters, including the number of Sinkhorn iterations, the entropic regularization coefficient, and the temporal cost function.

As shown in Table 10, 20 Sinkhorn iterations provide a good trade-off between accuracy and efficiency. The best performance is obtained with $\varepsilon_{OT} = 0.05$ and the linear cost $C_{ij} = | i - j |$, which are used as default settings.

Table 10: Ablation study on OT-related hyperparameters on THUMOS14.

Setting	Value	mAP@0.7	Avg. mAP
Sinkhorn iterations	5	46.0	68.4
	10	46.4	68.8
	20	46.8	69.1
	50	46.7	69.0
Entropy coefficient $\varepsilon_{\mathrm{OT}}$	0.01	46.2	68.6
	0.05	46.8	69.1
	0.10	46.3	68.7
Temporal cost C_{ij}	$\lvert i-j\rvert$	46.8	69.1
	$(i-j)^2$	46.2	68.6
	$\lvert i-j\rvert/\mathrm{T}'$	46.5	68.9

4.7 Complexity and Efficiency Analysis

Since BMA is used only during training, it does not change inference-time parameters, FLOPs, or post-processing. As shown in Table 11, the BMA-enhanced detector has the same parameter count and FLOPs as the baseline, and the FPS remains nearly unchanged.

Although BMA introduces Sinkhorn iterations during training, the overhead is moderate. As shown in Table 12, the training time per epoch increases from 18.6 to 21.4 min, and GPU memory increases from 13.2 to 14.1 GB. This corresponds to a 15.1% training-time overhead, while the average mAP improves from

66.8% to 69.1%. Therefore, BMA improves boundary localization without increasing inference cost, with only moderate additional training cost.

Table 11: Inference efficiency comparison on THUMOS14.

Method	Params (M)	FLOPs (G)	FPS	Avg mAP
Baseline	28.27	85.53	28.9	66.8
Baseline + BMA	28.27	85.53	28.7	69.1

Table 12: Training efficiency comparison on THUMOS14.

Method	Training Time/Epoch	GPU Memory	Sinkhorn Iterations	Training Overhead	Avg. mAP
Baseline	18.6 min	13.2 GB	–	–	66.8
Baseline + BMA	21.4 min	14.1 GB	20	+15.1%	69.1

4.8 Qualitative Analysis

Fig. 5 visualizes the effect of BMA on temporal boundary localization and training-stage measure alignment. Compared with the baseline, BMA produces segments closer to the annotated boundaries. The predicted boundary measures, target measures, and OT plans further show that BMA encourages predicted boundary mass to align with target boundary regions during training.

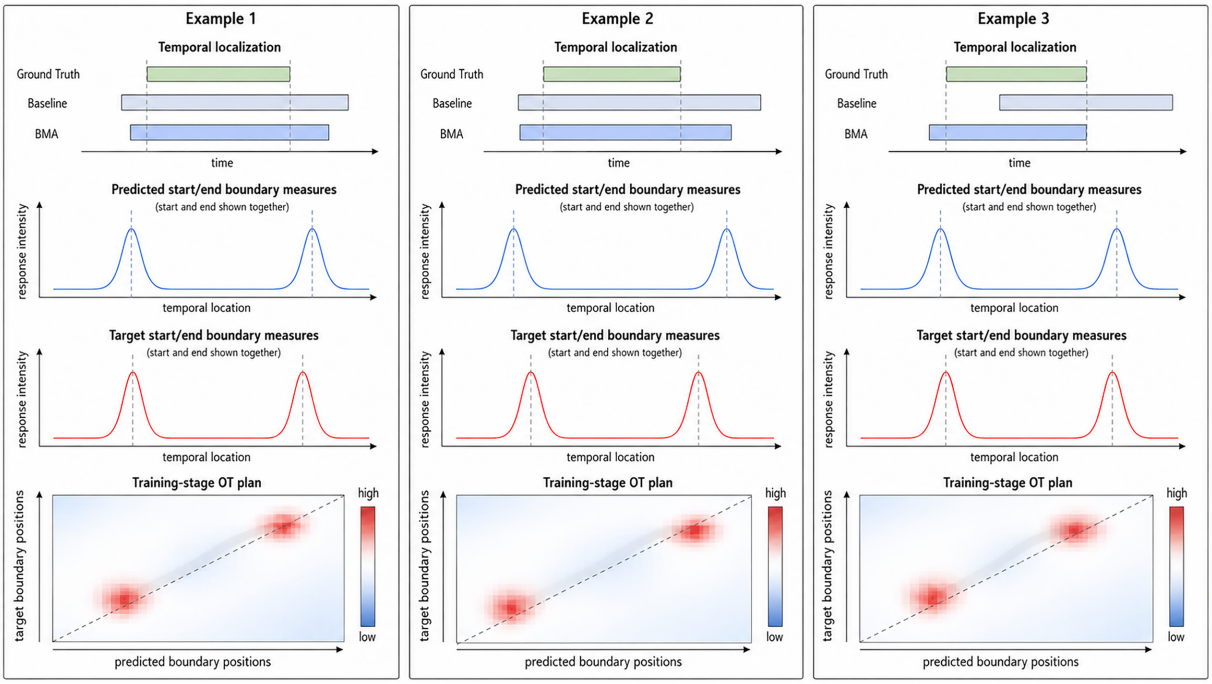

Figure 5: Qualitative illustration of boundary localization and training-stage measure alignment. Each example shows the ground-truth segment, baseline prediction, BMA-enhanced prediction, predicted start/end boundary measures, Gaussian-smoothed target measures, and the training-stage OT plan. Start and end measures are shown together for compact visualization. The gray curve schematically indicates the dominant transport path.

BMA may still fail when foreground responses contain strong internal fluctuations, when multiple actions are densely adjacent or overlapping, or when long-video annotations are coarse. These cases suggest that BMA is more suitable as a global boundary regularization term than as an instance-level boundary matching mechanism.

5 Conclusion

This paper proposes Boundary Measure Alignment (BMA), a training-stage auxiliary objective for temporal action detection. BMA represents predicted and annotated action starts and ends as temporal probability measures and aligns them with entropic optimal transport, thereby explicitly modeling boundary displacement. Since BMA is removed during inference, it does not change the original detector pipeline or introduce additional inference cost.

Experiments on THUMOS14 and ActivityNet-v1.3 show that BMA improves AFSD, FCOS, and an ActionFormer-style detector. The gains are more evident on THUMOS14 and under stricter tIoU thresholds, while improvements on ActivityNet-v1.3 are more moderate due to longer videos, coarser annotations, and larger action-duration variations. Robustness, ablation, efficiency, and qualitative analyses further verify the effectiveness and limitations of BMA. Future work will explore more robust foreground trajectory estimation, class-wise or instance-aware boundary measures, and broader adaptation to other TAD architectures.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Tiyao Zhang; methodology, Tiyao Zhang; software, Tiyao Zhang; validation, Tiyao Zhang; formal analysis, Tiyao Zhang; investigation, Tiyao Zhang; resources, Xue Yuan; data curation, Tiyao Zhang; writing—original draft preparation, Tiyao Zhang; writing—review and editing, Xue Yuan; visualization, Tiyao Zhang; supervision, Xue Yuan; project administration, Tiyao Zhang; funding acquisition, Xue Yuan. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data available on request from the authors.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhao Y, Xiong Y, Wang L, Wu Z, Tang X, Lin D. Temporal action detection with structured segment networks. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. New York, NY, USA: IEEE; 2017. p. 2933–42. doi:10.1109/ICCV.2017.317.
2. Lin T, Liu X, Li X, Ding E, Wen S. BMN: boundary-matching network for temporal action proposal generation. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Repblic of Korea. New York, NY, USA: IEEE; 2019. p. 3888–97. doi:10.1109/iccv.2019.00399.
3. Xu M, Zhao C, Rojas DS, Thabet A, Ghanem B. G-TAD: sub-graph localization for temporal action detection. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. New York, NY, USA: IEEE; 2020. p. 10153–62. doi:10.1109/cvpr42600.2020.01017.
4. Zhang CL, Wu J, Li Y. ActionFormer: localizing moments of actions with transformers. In: Computer Vision—ECCV 2022. ECCV 2022. Lecture Notes in Computer Science. Vol. 13664. Cham, Switzerland: Springer; 2022. p. 492–510. doi:10.1007/978-3-031-19772-7_29.
5. Liu X, Wang Q, Hu Y, Tang X, Zhang S, Bai S, et al. End-to-end temporal action detection with transformer. IEEE Trans Image Process. 2022;31:5427–41. doi:10.1109/TIP.2022.3195321.

6. Shi D, Zhong Y, Cao Q, Zhang J, Ma L, Li J, et al. ReAct: temporal action detection with relational queries. In: Computer Vision—ECCV 2022. Cham, Switzerland: Springer; 2022. p. 105–21. doi:10.1007/978-3-031-20080-9_7.
7. Liberatori B, Conti A, Rota P, Wang Y, Ricci E. Test-time zero-shot temporal action localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 17–21; Seattle, WA, USA. New York, NY, USA: IEEE; 2024. p. 18720–29.
8. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, et al. Video swin transformer. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. New York, NY, USA: IEEE; 2022. p. 3192–201. doi:10.1109/CVPR52688.2022.00320.
9. Gritsenko AA, Xiong X, Djolonga J, Dehghani M, Sun C, Lucic M, et al. End-to-end spatio-temporal action localisation with video transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 17–21; Seattle, WA, USA. New York, NY, USA: IEEE; 2024. p. 18373–83.
10. Qing Z, Su H, Gan W, Wang D, Wu W, Wang X, et al. Temporal context aggregation network for temporal action proposal refinement. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. New York, NY, USA: IEEE; 2021. p. 485–94. doi:10.1109/CVPR46437.2021.00055.
11. Lu Z, Elhamifar E. FACT: frame-action cross-attention temporal modeling for efficient action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 17–21; Seattle, WA, USA. New York, NY, USA: IEEE; 2024. p. 18175–85.
12. Alwassel H, Caba Heilbron F, Escorcia V, Ghanem B. Diagnosing error in temporal action detectors. In: Computer Vision—ECCV 2018. Cham, Switzerland: Springer; 2018. p. 264–80. doi:10.1007/978-3-030-01219-9_16.
13. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. New York, NY, USA: IEEE; 2020. p. 658–66. doi:10.1109/CVPR.2019.00075.
14. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-IoU loss: faster and better learning for bounding box regression. Proc AAAI Conf Artif Intell. 2020;34(7):12993–3000. doi:10.1609/aaai.v34i07.6999.
15. Ge Z, Liu S, Li Z, Yoshie O, Sun J. OTA: optimal transport assignment for object detection. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. New York, NY, USA: IEEE; 2021. p. 303–12. doi:10.1109/cvpr46437.2021.00037.
16. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Proceedings of the Computer Vision—ECCV 2020. Cham, Switzerland: Springer International Publishing; 2020. p. 213–29. doi:10.1007/978-3-030-58452-8_13.
17. Cuturi M. Sinkhorn distances: lightspeed computation of optimal transportation distances. arXiv:1306.0895. 2013.
18. Peyré G, Cuturi M. Computational optimal transport with applications to data sciences. Found Trends® Mach Learn. 2019;11(5–6):355–607. doi:10.1561/2200000073.
19. Frogner C, Zhang C, Mobahi H, Araya M, Poggio TA. Learning with a Wasserstein loss. Adv Neural Inf Process Syst. 2018.
20. Torres LC, Pereira LM, Amini MH. A survey on optimal transport for machine learning: theory and applications. arXiv:2106.01963. 2021.
21. Tan J, Tang J, Wang L, Wu G. Relaxed transformer decoders for direct action proposal generation. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. New York, NY, USA: IEEE; 2022. p. 13506–15. doi:10.1109/ICCV48922.2021.01327.
22. Song Y, Kim D, Cho M, Kwak S. Online temporal action localization with memory-augmented transformer. arXiv:2408.02957. 2024.
23. Lin C, Li J, Wang Y, Tai Y, Luo D, Cui Z, et al. Fast learning of temporal action proposal via dense boundary generator. Proc AAAI Conf Artif Intell. 2020;34(7):11499–506. doi:10.1609/aaai.v34i07.6815.
24. Shi D, Zhong Y, Cao Q, Ma L, Lit J, Tao D. TriDet: temporal action detection with relative boundary modeling. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. New York, NY, USA: IEEE; 2023. p. 18857–66. doi:10.1109/CVPR52729.2023.01808.

25. Zhao P, Xie L, Ju C, Zhang Y, Wang Y, Tian Q. Bottom-up temporal action localization with mutual regularization. In: Computer Vision—ECCV 2020. Cham, Switzerland: Springer; 2020. p. 539–55. doi:10.1007/978-3-030-58598-3_32.
26. Yang L, Peng H, Zhang D, Fu J, Han J. Revisiting anchor mechanisms for temporal action localization. IEEE Trans Image Process. 2020;29:8535–48. doi:10.1109/TIP.2020.3016486.
27. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: deformable transformers for end-to-end object detection. arXiv:2010.04159. 2020.
28. Sun P, Zhang R, Jiang Y, Kong T, Xu C, Zhan W, et al. Sparse R-CNN: end-to-end object detection with learnable proposals. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. New York, NY, USA: IEEE; 2021. p. 14449–58. doi:10.1109/cvpr46437.2021.01422.
29. Li X, Wang W, Wu L, Chen S, Hu X, Li J, et al. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection. arXiv:2006.04388. 2020.
30. Courty N, Flamary R, Tuia D, Rakotomamonjy A. Optimal transport for domain adaptation. IEEE Trans Pattern Anal Mach Intell. 2017;39(9):1853–65. doi:10.1109/TPAMI.2016.2615921.
31. Damodaran BB, Kellenberger B, Flamary R, Tuia D, Courty N. DeepJDOT: deep joint distribution optimal transport for unsupervised domain adaptation. In: Proceedings of the Computer Vision—ECCV 2018. Cham, Switzerland: Springer; 2018. p. 467–83. doi:10.1007/978-3-030-01225-0_28.
32. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, Australia. p. 214–23.
33. Puy G, Boulch A, Marlet R. FLOT: scene flow on point clouds guided by optimal transport. arXiv:2007.11142. 2020.
34. Kolouri S, Park SR, Thorpe M, Slepcev D, Rohde GK. Optimal mass transport: signal processing and machine-learning applications. IEEE Signal Process Mag. 2017;34(4):43–59. doi:10.1109/msp.2017.2695801.
35. Villani C. Optimal transport: old and new. Berlin/Heidelberg, Germany: Springer; 2009. 976 p.
36. Lin T, Zhao X, Su H, Wang C, Yang M. BSN: boundary sensitive network for temporal action proposal generation. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany. Berlin/Heidelberg, Germany: Springer; 2018. p. 3–19.
37. Liu Q, Wang Z. Progressive boundary refinement network for temporal action detection. Proc AAAI Conf Artif Intell. 2020;34(7):11612–9. doi:10.1609/aaai.v34i07.6829.
38. Zhao C, Thabet AK, Ghanem B. Video self-stitching graph network for temporal action localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. New York, NY, USA: IEEE; 2021. p. 13658–67.
39. Wang Q, Zhang Y, Zheng Y, Pan P. RCL: recurrent continuous localization for temporal action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. New York, NY, USA: IEEE; 2022. p. 13566–75.
40. Nag S, Zhu X, Song YZ, Xiang T. Proposal-free temporal action detection via global segmentation mask learning. In: Proceedings of the European Conference on Computer Vision (ECCV); 2022 Jun 18–24; New Orleans, LA, USA. New York, NY, USA: IEEE; 2022. p. 645–62.
41. Cheng F, Bertasius G. TallFormer: temporal action localization with a long-memory transformer. In: Proceedings of the European Conference on Computer Vision (ECCV); 2022 Jun 18–24; New Orleans, LA, USA. New York, NY, USA: IEEE; 2022. p. 503–21.
42. Lin C, Xu C, Luo D, Wang Y, Tai Y, Wang C, et al. Learning salient boundary feature for anchor-free temporal action localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Oct 10–17; Montreal, QC, Canada. New York, NY, USA: IEEE; 2021. p. 3320–9.
43. Tian Z, Shen C, Chen H, He T. FCOS: fully convolutional one-stage object detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 October 27–Nov 2; Seoul, Republic of Korea. p. 9627–36.