arXiv:2610.35516v1 [cs.CV] 26 Oct 2026

Learning High-Quality Pseudo-Labels for Point-Supervised Temporal Action Localization

Zhenying Fang
Independent Researcher
zhenying.fang@foxmail.com

Abstract

Point-supervised temporal action localization (PTAL) aims to recognize and localize action instances in untrimmed videos using only point-level annotations. Existing pseudo-label-based methods have made substantial progress in PTAL. However, confidence-based initialization can overlook well-localized proposals. Meanwhile, existing methods either keep pseudo-labels fixed during detector training or update them without fully exploiting alternative boundary evidence from the original proposal pool. These limitations constrain supervision quality, thereby hindering detection performance. To learn high-quality pseudo-labels, we propose a framework that integrates prior-guided initialization with consensus-guided refinement during detector training. Our framework comprises two modules: Prior-Guided Initialization (PGI) and Consensus-Guided Refinement (CGR). PGI combines proposal confidence with point-derived priors to establish instance-specific reliability for initial pseudo-label selection and candidate retention. CGR continually uses the retained original candidates to construct boundary targets, combining their consensus with evolving detector predictions to refine pseudo-labels throughout training. Experiments on four benchmarks demonstrate that our method achieves competitive detection performance compared with existing methods. In particular, our method achieves 68.5% average mAP over tIoU thresholds 0.1–0.7 on THUMOS14, surpassing PseR by 3.7 percentage points. Code is available at https://github.com/zhenyingfang/LHQP.

1. Introduction

Temporal action localization (TAL) aims to recognize action instances and determine their temporal boundaries in untrimmed videos [8, 47], supporting applications such as visual question answering [17], video summarization [16], and highlight detection [39]. Fully supervised methods [6, 35, 37] have achieved impressive performance, but require precise temporal boundaries for individual action instances, which are costly and labor-intensive to annotate. This annotation burden has motivated weakly supervised temporal action localization (WTAL) [32, 33], which learns from video-level category labels. Without instance-level temporal cues, however, recovering complete action intervals remains difficult, leaving a substantial performance gap to fully supervised methods.

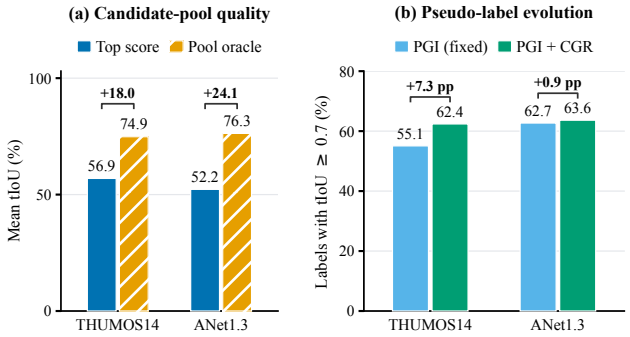


Figure 1. **Proposal quality and pseudo-label evolution.** (a) Mean tIoU of top-score and oracle (maximum ground-truth tIoU) proposals among same-class candidates containing each annotated point. (b) Fraction of pseudo-labels with tIoU ≥ 0.7 for fixed PGI and PGI+CGR at the final update, evaluated on the same instances. Brackets indicate differences in percentage points.

To balance detection performance and annotation cost, point-supervised temporal action localization (PTAL) has been proposed as an alternative supervision setting. PTAL requires only one annotated frame and its category per action instance. Early methods [15, 27, 44] follow a *localization-by-classification* paradigm, inferring action intervals from Class Activation Sequences (CAS) learned under point supervision. However, CAS are learned using classification objectives, and the mismatch between classification and localization limits the detection performance of these methods [10]. To provide explicit localization supervision, *pseudo-label-based* methods construct temporal pseudo-labels to train localization models [5, 7, 48]. POTLoc [36] generates pseudo-labels using point-compatible proposals and proposal statistics.

PseR [5] propagates seed proposals through temporal perturbations and refines pseudo-labels using proposal selection and ranking.

Despite this progress, pseudo-label quality remains constrained at both initialization and subsequent detector training. First, proposal confidence does not necessarily reflect localization quality for the annotated instance. Fig. 1(a) reveals a substantial gap between highest-confidence and oracle selection from the same point-compatible candidate pool, showing that well-localized proposals can be overlooked during initialization. Second, pseudo-labels either remain fixed during detector training or evolve without fully exploiting alternative boundaries from the original pool. Fixed labels cannot incorporate new localization evidence, while reducing the pool to pseudo-label intervals limits direct access to individual original alternatives during later updates. Together, these limitations restrict supervision quality and the detector's ability to learn accurate action boundaries.

We address these limitations by learning high-quality pseudo-labels through reliable initialization and iterative boundary refinement. Our framework comprises two modules: Prior-Guided Initialization (PGI) and Consensus-Guided Refinement (CGR). PGI assesses the reliability of candidate proposals associated with each annotated point by combining confidence, geometric, and conflict priors. Using these reliability scores, PGI selects initial pseudo-labels while retaining alternative candidates with their original boundaries and reliability scores to support subsequent pseudo-label evolution.

CGR reuses the candidates retained by PGI to progressively refine pseudo-label boundaries during detector training. At each update round, it combines a reliability-weighted consensus of these candidates with current detector predictions to construct boundary targets, which guide consistency-controlled pseudo-label updates. This allows alternative boundaries beyond the initial pseudo-label to provide supporting evidence throughout training. As shown in Fig. 1(b), PGI+CGR increases the fraction of pseudo-labels with tIoU ≥ 0.7 over fixed PGI supervision by 7.3 and 0.9 percentage points on THUMOS14 and ActivityNet1.3, respectively.

Experiments on THUMOS14, ActivityNet1.3, GTEA, and BEOID demonstrate competitive detection performance compared with existing methods. Our contributions are summarized as follows:

- We introduce PGI to assess candidate reliability using point annotations, enabling instance-aware pseudo-label initialization while retaining alternative boundary hypotheses for subsequent evolution.
- We introduce CGR to combine consensus from retained original candidates with evolving detector predictions for boundary target construction, allowing alternative boundary evidence to continually support pseudo-label refinement.
- Extensive experiments on four benchmarks demonstrate improved pseudo-label quality and competitive detection performance, while ablation studies validate the complementary roles of PGI and CGR.

2. Related Work

2.1. Temporal Action Localization

Fully supervised temporal action localization (TAL) relies on instance-level category labels and precise temporal boundary annotations. Representative detectors follow anchor-based, anchor-free, hybrid, or query-based designs. Anchor-based methods associate predictions with predefined temporal windows. Within anchor-based designs, one-stage detectors [22] jointly predict action categories and boundary offsets, while two-stage detectors [40] first generate temporal proposals and subsequently refine and classify them. Temporal anchors provide useful duration priors, but their predefined scales can restrict flexibility when action durations vary substantially [41]. To alleviate these restrictions, anchor-free methods [1, 11, 35, 47] eliminate manually designed temporal anchors. They predict action categories and distances to temporal boundaries directly at temporal locations. Hybrid approaches [6, 41] further integrate anchor-based and anchor-free branches to exploit their complementary strengths. Across these dense prediction designs, post-processing such as non-maximum suppression is typically used to remove duplicate detections. Query-based detectors [14, 25] simplify this pipeline by formulating TAL as set prediction with learnable action queries and one-to-one matching. Despite these architectural advances, training still relies on costly temporal boundary annotations, motivating localization under weaker supervision.

2.2. Weakly-Supervised Temporal Action Localization

Weakly supervised temporal action localization (WTAL) reduces annotation costs by learning from video-level category labels without temporal boundary annotations [18, 32, 43]. A common approach follows the *localization-by-classification* paradigm [34, 42, 43], which trains a classifier to estimate snippet-level class scores and derives action intervals by thresholding the resulting activation sequences. However, classification objectives tend to emphasize the most discriminative snippets, while less distinctive parts of an action may receive weak responses. This discrepancy makes it difficult to recover complete action instances from classification scores alone.

Pseudo-label-based methods [20, 26, 33, 46] address this discrepancy by generating temporal pseudo-labels to supervise an additional localization model, enabling direct optimization of temporal boundaries. Despite their progress, video-level labels provide no explicit indication of instance

positions, making reliable pseudo-label construction challenging and leaving a performance gap relative to fully supervised methods. Point annotations offer sparse temporal cues for individual instances. Our method uses these cues to establish instance-specific proposal reliability for initial pseudo-label construction and to preserve relevant original boundary hypotheses for subsequent label evolution.

2.3. Point-Supervised Temporal Action Localization

Point-supervised temporal action localization (PTAL) requires only a single annotated frame and its category for each action instance [27]. One line of research expands sparse point supervision and strengthens action representations. Early *localization-by-classification* methods include SF-Net [27], which mines labels for neighboring snippets, and LACP [15], which learns action completeness through pseudo-label sequence search. CRRC-Net [10] improves representation compactness and snippet-label reliability, while proposal-level learning [44] generates dense supervision through clustering. Related efforts model temporal structure through auxiliary understanding tasks [29] or enrich visual representations with refined textual descriptions [30].

A complementary line of research constructs temporal pseudo-labels for explicit localization learning [5, 7, 23, 36, 48, 49]. These methods infer interval supervision from sparse annotations, enabling direct optimization of temporal boundaries. Among closely related approaches, POTLoc [36] refines point-supervised proposals into pseudo-labels for self-training. NoCo [49] introduces online label revision through a teacher–student framework and is also evaluated under point supervision.

Building on advances in pseudo-label generation and online correction, we focus on the continued use of original proposal evidence after initialization. Our framework uses point-derived reliability to connect initial pseudo-label selection with candidate retention, then dynamically reweights the retained boundary hypotheses using evolving detector predictions. The original boundaries remain fixed, while their contributions to supervision evolve, allowing candidates beyond the initially selected interval to support subsequent updates.

3. Method

3.1. Problem Formulation and Overview

Given an untrimmed video v of duration T_v, point supervision provides annotations $\mathcal{A}_v = \{(p_i, c_i)\}_{i=1}^{N_v}$, where N_v is the number of annotated action instances, p_i is a timestamp within the i-th instance, and c_i is its category. Instance boundaries are unavailable during training. Our goal is to construct temporal pseudo-labels from these annotations and train a detector to recognize and localize action instances in unseen videos.

Following the setup of existing pseudo-label-based methods [5, 7, 32, 49], we use a point-supervised proposal generator to obtain an original pool of candidate proposals $\mathcal{P}_v = \{(b_j, z_j, \hat{c}_j)\}_{j=1}^{M_v}$, where M_v is the number of proposals for video v. The j-th proposal contains an interval $b_j = (s_j, e_j)$, a confidence score z_j, and a predicted category $\hat{c}_j$. Here s_j and e_j are its start and end timestamps, with $s_j < e_j$. Unless explicitly shown, the video index is omitted for brevity, and all point and proposal indices refer to video v.

Given the original proposal pool $\mathcal{P}_v$ and point annotations $\mathcal{A}_v$, our framework improves pseudo-label quality through two complementary modules: Prior-Guided Initialization (PGI) and Consensus-Guided Refinement (CGR) (Fig. 2(a)). For the i-th point annotation in video v, PGI uses point-derived proposal reliability to select an initial pseudo-label $\mathbf{y}_i^0$ from $\mathcal{P}_v$ and retain a candidate subset indexed by $\mathcal{P}_{v,i}^{\text{ret}}$. CGR then combines a reliability-weighted consensus of these retained candidates with evolving detector predictions to construct boundary targets for consistency-controlled label updates. The original proposal pool thus supports both initial supervision and subsequent boundary refinement during detector training.

3.2. Prior-Guided Initialization (PGI)

PGI aims to improve initial pseudo-label quality while retaining alternative candidates for subsequent refinement (Fig. 2(b)). For each annotated point, it assesses candidate reliability using three complementary priors: a model-derived confidence prior and point-derived geometric and conflict priors. The resulting scores guide both initial pseudo-label selection and candidate retention, linking initialization with later refinement.

For the i-th annotation (p_i, c_i), we collect from $\mathcal{P}_v$ the same-class proposals containing p_i and define their index set as:

$$\mathcal{P}_{v,i} = \{j \in \{1, \dots, M_v\} : \hat{c}_j = c_i,\ s_j \le p_i \le e_j\}. \quad (1)$$

Confidence prior. For candidate $j \in \mathcal{P}_{v,i}$, we define its model-derived confidence prior π_{ij} by applying a softmax to the source scores within this set:

$$\pi_{ij} = \frac{\exp(z_j)}{\sum_{q \in \mathcal{P}_{v,i}} \exp(z_q)}. \quad (2)$$

However, π_{ij} is obtained by normalizing CAS-derived scores, which primarily reflect snippet-level action classification confidence and do not fully capture instance-level localization quality. We therefore complement this confidence prior with geometric and conflict priors to assess each candidate's compatibility with the annotated instance.

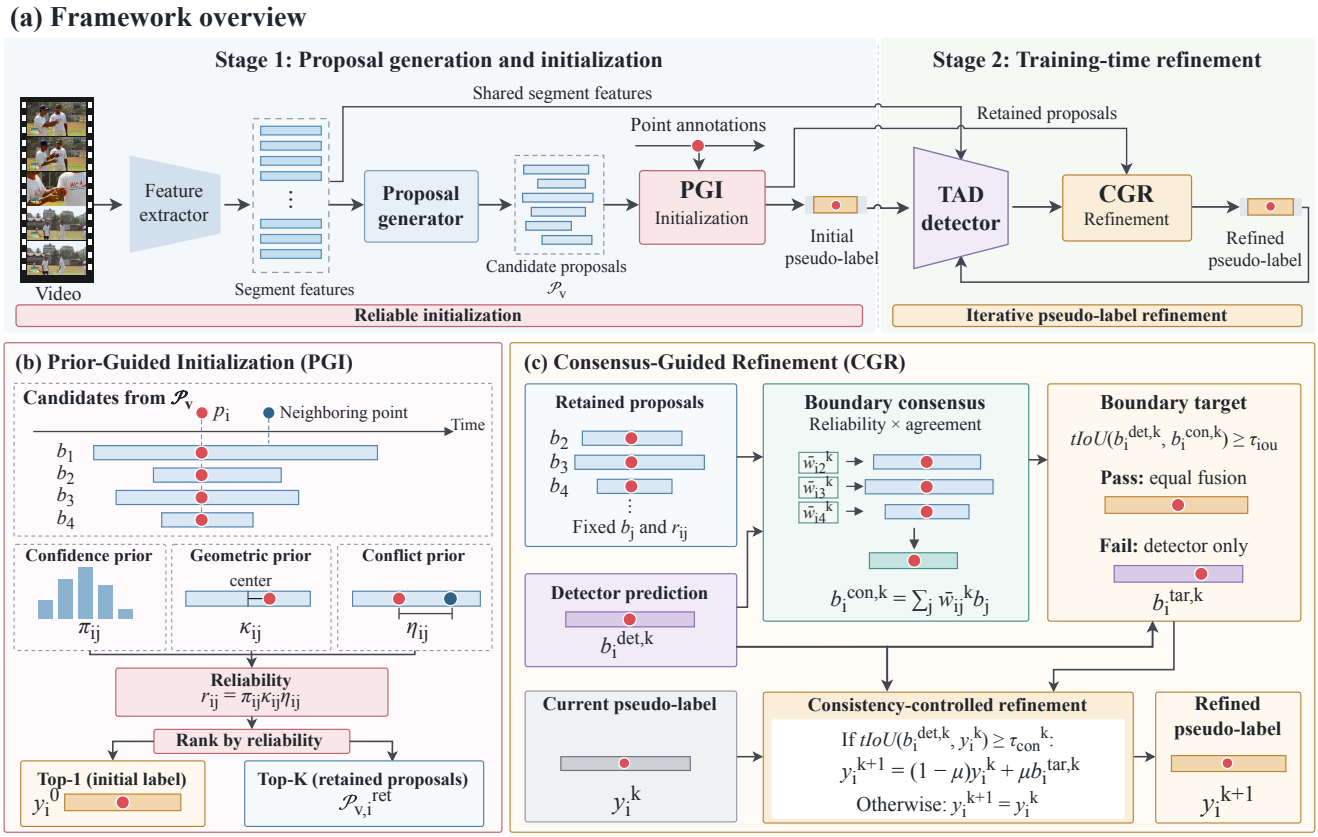


Figure 2. **Overview of proposal-guided pseudo-label learning.** (a) Two-stage framework: candidate proposal generation and pseudo-label initialization, followed by detector training with iterative refinement. (b) PGI combines confidence, geometric, and conflict priors to select initial pseudo-labels and retain top-K candidates. (c) CGR constructs gated boundary targets from retained-proposal consensus and detector predictions, then refines pseudo-labels through consistency-controlled momentum updates.

Geometric prior. TSPNet [38] has demonstrated the effectiveness of point-derived center cues for proposal confidence estimation. Unlike its learning-based center-score estimation, PGI formulates a geometric prior by directly measuring point centrality from the annotated timestamp and proposal boundaries, without additional training. For the j-th proposal $b_j = (s_j, e_j)$, we define this prior as:

$$\kappa_{ij} = \mathrm{clip}_{[0,1]}\left(1 - \frac{|p_i - (s_j + e_j)/2|}{(e_j - s_j)/2}\right), \tag{3}$$

where $\mathrm{clip}_{[0,1]}(x) = \min\{1, \max\{0, x\}\}$ clips its input to the interval $[0, 1]$. This geometric prior favors intervals whose center lies near the annotated point and attenuates those placing it close to a boundary.

Conflict prior. A proposal covering additional annotated points can span neighboring instances rather than isolate the target. We collect the other points covered by candidate j as:

$$\mathcal{C}_{ij} = \{\ell \in \{1, \ldots, N_v\} : \ell \neq i,\ s_j \leq p_\ell \leq e_j\}, \tag{4}$$

and measure their accumulated distance from p_i:

$$d_{ij} = \sum_{\ell \in \mathcal{C}_{ij}} |p_\ell - p_i|. \tag{5}$$

With timestamps expressed in seconds, we define the conflict prior as:

$$\eta_{ij} = \begin{cases} 1, & d_{ij} = 0, \\ \tanh(1/d_{ij}), & d_{ij} > 0. \end{cases} \tag{6}$$

This prior decreases with positive accumulated distance, attenuating intervals that cover additional or more distant annotated instances. It equals one when the candidate covers no other point.

Shared reliability and initial pseudo-labels. We combine the three priors into an instance-specific reliability score and

select the initial pseudo-label:

$$\begin{aligned} r_{ij} &= \pi_{ij}\kappa_{ij}\eta_{ij}, \\ j_i^* &= \arg\max_{j\in\mathcal{P}_{v,i}} r_{ij}, \\ \mathbf{y}_i^0 &= (s_{j_i^*}, e_{j_i^*}). \end{aligned} \tag{7}$$

The initial pseudo-label inherits the annotated category c_i. Let $\mathcal{I}$ contain the point indices for which $\mathcal{P}_{v,i} \neq \varnothing$; these points have initial temporal supervision. Points with empty candidate sets are omitted from detector supervision and subsequent label evolution.

Retention of original hypotheses. For each $i \in \mathcal{I}$, we form the retained candidate index set $\mathcal{P}_{v,i}^{\mathrm{ret}} \subseteq \mathcal{P}_{v,i}$ by selecting the K highest-reliability candidates, retaining all candidates when $|\mathcal{P}_{v,i}| < K$. Their original boundaries $\{b_j\}_{j\in\mathcal{P}_{v,i}^{\mathrm{ret}}}$ and reliability scores $\{r_{ij}\}_{j\in\mathcal{P}_{v,i}^{\mathrm{ret}}}$ remain fixed throughout training. Thus, PGI provides initial supervision together with alternative boundary hypotheses that CGR reuses in subsequent updates.

3.3. Consensus-Guided Refinement (CGR)

CGR aims to progressively refine pseudo-label boundaries by combining retained original proposal evidence with evolving detector predictions (Fig. 2(c)). At each update round, it aggregates the candidates retained by PGI into a reliability-weighted boundary consensus and combines this consensus with the current detector prediction to construct an instance-specific boundary target. Pseudo-labels are then refined through momentum updates under a progressively relaxed consistency criterion, allowing alternative boundaries from the original proposals to continually guide supervision during detector training.

Current detector predictions. At update round k, we obtain detector predictions covering p_i and apply the PGI reliability rule to select the highest-reliability interval $b_i^{\mathrm{det},k}$. This prediction provides the current localization estimate for boundary target construction. The annotated category c_i remains the class supervision.

Consensus from retained original proposals. We aggregate the boundaries of candidates indexed by $\mathcal{P}_{v,i}^{\mathrm{ret}}$ to obtain an instance-specific proposal consensus. Each candidate is weighted by its fixed reliability and temporal agreement with $b_i^{\mathrm{det},k}$, followed by normalization over the retained set:

$$\begin{aligned} w_{ij}^k &= r_{ij}\exp\Big(\mathrm{tIoU}(b_j, b_i^{\mathrm{det},k})\Big), \\ \bar{w}_{ij}^k &= \frac{w_{ij}^k}{\sum_{q\in\mathcal{P}_{v,i}^{\mathrm{ret}}} w_{iq}^k}, \\ b_i^{\mathrm{con},k} &= \sum_{j\in\mathcal{P}_{v,i}^{\mathrm{ret}}} \bar{w}_{ij}^k b_j, \end{aligned} \tag{8}$$

where $b_i^{\mathrm{con},k}$ denotes the consensus interval for the i-th point annotation at update round k, computed as a weighted average of the retained original proposal intervals. The normalized weights $\bar{w}_{ij}^k$ favor reliable candidates that agree temporally with the current detector prediction.

The original intervals b_j and reliability scores r_{ij} remain fixed throughout training, while their consensus is recomputed at each update using the current detector prediction. This enables retained alternatives to the initial pseudo-label $\mathbf{y}_i^0$ to continually support boundary target construction.

Boundary target construction. We fuse the current detector prediction with the proposal consensus when the candidate evidence satisfies a compatibility criterion. Let $g_i^k \in \{0,1\}$ denote this gate. The boundary target is:

$$b_i^{\mathrm{tar},k} = b_i^{\mathrm{det},k} + \tfrac{1}{2} g_i^k \left(b_i^{\mathrm{con},k} - b_i^{\mathrm{det},k}\right), \tag{9}$$

where $b_i^{\mathrm{tar},k}$ denotes the target interval for refining the pseudo-label associated with the i-th point annotation at update round k. When $g_i^k = 1$, the target equally weights the detector prediction and proposal consensus for both temporal boundaries. When $g_i^k = 0$, the target is the detector prediction.

We determine g_i^k from the temporal agreement between the detector prediction and proposal consensus:

$$g_i^k = \mathbf{1}\Big[\mathrm{tIoU}(b_i^{\mathrm{det},k}, b_i^{\mathrm{con},k}) \geq \tau_{\mathrm{iou}}\Big], \tag{10}$$

where $\mathbf{1}[\cdot]$ is the indicator function, and $\tau_{\mathrm{iou}} \in [0,1]$ specifies the minimum required temporal overlap between the detector prediction and proposal consensus. Consensus fusion is activated when this overlap meets the threshold; otherwise, the target uses the detector prediction alone.

Consistency-controlled pseudo-label refinement. Given the boundary target $b_i^{\mathrm{tar},k}$, we refine the current pseudo-label $\mathbf{y}_i^k$ through a consistency-controlled update.

At round k, a point annotation is eligible for refinement if its detector prediction $b_i^{\mathrm{det},k}$ is available and $\mathrm{tIoU}(b_i^{\mathrm{det},k}, \mathbf{y}_i^k) \geq \tau_{\mathrm{con}}^k$, where $\tau_{\mathrm{con}}^k \in [0,1]$ is the consistency threshold specifying the minimum required temporal overlap between the detector prediction and the current

pseudo-label at this round. Let $\mathcal{I}_k$ denote the eligible points selected for this round's update. We refine the pseudo-labels of selected points through a momentum update:

$$\mathbf{y}_i^{k+1} = (1-\mu)\mathbf{y}_i^k + \mu b_i^{\mathrm{tar},k}, \qquad i \in \mathcal{I}_k, \tag{11}$$

where $\mu \in (0, 1]$ controls the contribution of the boundary target to the refined pseudo-label. Pseudo-labels associated with points outside $\mathcal{I}_k$ remain unchanged, and all annotated categories are preserved. The refined pseudo-labels supervise the next training period. At subsequent updates, CGR constructs boundary targets from retained original proposals and current detector predictions to progressively improve pseudo-label quality.

3.4. Training and Inference

Training. We train the detector with PGI-initialized pseudo-labels and annotated categories, replacing the temporal supervision with refined pseudo-labels after each CGR update. The retained original proposals and their reliability scores remain fixed throughout training and support boundary target construction at subsequent updates.

Inference. During inference, the trained detector directly predicts action categories, temporal boundaries, and confidence scores for an unseen video. PGI, CGR, point annotations, and the original candidate pool are used only during training.

4. Experiment

4.1. Experimental Setup

Datasets. We evaluate our method on four benchmark datasets. For fair comparison, we use the point annotations provided by HR-Pro [48].

- **THUMOS14** [12] contains 413 untrimmed videos spanning 20 action categories. We use 200 videos for training and the remaining 213 for testing.
- **ActivityNet1.3** [2] comprises 10,024 training, 4,926 validation, and 5,044 test videos. Following the standard protocol [5, 48], we train on the training split and evaluate on the validation split.
- **GTEA** [9] consists of 28 videos depicting seven fine-grained daily activities in a kitchen, with approximately 1,800 frames per video.
- **BEOID** [4] contains 58 videos covering 30 action categories, with an average of 12.5 action instances per video. We follow the training and testing split used in [31].

Evaluation metric. We follow the standard evaluation protocols and report mean average precision (mAP) at multiple temporal intersection-over-union (tIoU) thresholds. We additionally evaluate pseudo-label quality using mean tIoU with ground truth and the fraction of pseudo-labels with tIoU ≥ 0.7.

Implementation details. For a fair comparison, we follow existing methods [5, 48] to divide each video into 16-frame snippets and extract video features using the I3D network pre-trained on Kinetics-400 [3]. Given the extracted features, we use HR-Pro [48] as the proposal generator to obtain the candidate proposal pool $\mathcal{P}_v$ for each video v.

Across all datasets, PGI retains up to $K = 10$ candidates per point, and CGR uses $\tau_{\mathrm{iou}} = 0.5$ for consensus gating and $\mu = 0.9$ for momentum updates. CGR refines pseudo-labels between 20% and 80% of training, with the consistency threshold decreasing linearly from 1 to 0:

$$\tau_{\mathrm{con}}^k = 1 - \frac{e_k - 0.2E}{0.6E}, \tag{12}$$

where E is the number of training epochs and e_k is the epoch of update k.

We adopt ActionFormer [47] as the detector and train it with AdamW using a weight decay of 0.05. The initial learning rate is 10^{-3} on ActivityNet1.3 and 10^{-4} on the other datasets. For THUMOS14, ActivityNet1.3, GTEA, and BEOID, the batch sizes are 2, 16, 1, and 2, and the training durations are 35, 10, 60, and 50 epochs, respectively. At inference, we apply Soft-NMS to suppress redundant predictions.

4.2. Comparison with State-of-The-Art Methods

Results on THUMOS14. In Table 1, we compare our framework with state-of-the-art methods under different levels of supervision on THUMOS14. Among the compared point-supervised methods, our framework achieves the highest average mAP over all three threshold ranges, reaching 68.5% over tIoU thresholds 0.1–0.7 and surpassing PseR by 3.7 percentage points. It also achieves 78.2% and 61.8% average mAP over 0.1–0.5 and 0.3–0.7, respectively. At the stricter tIoU thresholds of 0.5 and 0.7, our method improves over PseR by 5.4 and 5.9 percentage points, indicating stronger temporal localization performance.

Results on ActivityNet1.3. As shown in Table 2, our method achieves 30.9% average mAP over tIoU thresholds 0.5–0.95, the highest among the compared methods and 0.3 percentage points above PseR. At tIoU 0.75, it obtains 30.7% mAP, improving over PseR by 1.1 percentage points. These results show that proposal-guided pseudo-label learning also improves overall detection performance on ActivityNet1.3.

Results on GTEA. As shown in Table 3, our method achieves 62.2% average mAP over tIoU thresholds 0.1–0.7, surpassing TRA and PseR by 8.1 and 9.9 percentage points,

Table 1. Comparison with state-of-the-art methods on THUMOS14. Results are reported as mAP (%) at different tIoU thresholds.

Supervision	Method	mAP@IoU(%)							AVG (0.1:0.5)	AVG (0.3:0.7)	AVG (0.1:0.7)
		0.1	0.2	0.3	0.4	0.5	0.6	0.7			
Frame-level	ActionFormer ECCV'22 [47]	-	-	82.1	77.8	71.0	59.4	43.9	-	66.8	-
	TriDet CVPR'23 [35]	-	-	83.6	80.1	72.9	62.4	47.4	-	69.3	-
	MDVLM-TAL NeurIPS'26 [37]	-	-	84.6	80.8	74.4	63.5	60.1	-	72.7	-
Video-level	P-MIL CVPR'23 [32]	71.8	67.5	58.9	49.0	40.0	27.1	15.1	57.4	38.0	47.0
	PivoTAL CVPR'23 [33]	74.1	69.6	61.7	52.1	42.8	30.6	16.7	60.1	40.8	49.6
	ISSF AAAI'24 [46]	72.4	66.9	58.4	49.7	41.8	25.5	12.8	57.8	37.6	46.8
	PVLR MM'24 [21]	74.9	69.9	61.4	53.1	45.1	30.5	17.1	60.9	41.4	50.3
	PseudoFormer CVPR'25 [26]	76.8	72.6	65.4	55.6	44.1	31.3	17.1	62.9	42.7	51.9
	CRC TPAMI'25 [42]	72.9	68.0	59.8	51.3	42.7	27.8	15.1	58.9	39.3	48.2
	ACL-Net TIP'26 [20]	81.0	76.9	70.0	60.4	51.3	36.8	21.5	67.9	48.0	56.8
Point-level	SF-Net ECCV'20 [27]	68.3	62.3	52.8	42.2	30.5	20.6	12.0	51.2	31.6	41.2
	Ju *et al.* ICCV'21 [13]	72.3	64.7	58.2	47.1	35.9	23.0	12.8	55.6	35.4	44.9
	LACP ICCV'21 [15]	75.7	71.4	64.6	56.5	45.3	34.5	21.8	62.7	44.5	52.8
	CRRC-Net TIP'22 [10]	77.8	73.5	67.1	57.9	46.6	33.7	19.8	64.6	45.1	53.8
	LPR MMSJ'23 [7]	85.3	82.5	76.6	67.8	55.3	38.5	21.7	73.5	52.0	61.1
	Yin *et al.* BMVC'23 [44]	77.1	72.6	65.9	54.4	44.9	33.1	20.2	63.3	43.9	52.6
	HR-Pro AAAI'24 [48]	85.6	81.6	74.3	64.3	52.2	39.8	24.8	71.6	51.1	60.3
	TSPNet CVPR'24 [38]	82.3	77.6	70.1	60.0	49.4	37.6	24.5	67.9	48.3	57.4
	SMBD ECCV'24 [24]	-	-	66.0	57.9	47.0	36.0	22.0	64.2	45.7	-
	SNPR TIP'24 [19]	77.9	73.9	66.6	59.4	48.6	36.7	22.7	65.3	46.8	55.1
	AAPL AAAI'25 [45]	64.3	-	54.6	-	35.2	-	14.0	-	-	42.8
	NoCo AAAI'25 [49]	86.0	82.1	75.6	65.7	53.6	40.6	25.3	72.6	52.2	61.3
	QROT CVPR'25 [23]	-	-	73.1	64.4	54.3	41.3	27.4	72.3	52.1	-
	TRA ESWA'26 [28]	86.6	83.6	76.3	66.7	55.2	41.4	26.5	73.7	53.2	62.3
	PseR TMM'26 [5]	85.5	83.2	78.7	70.9	60.1	45.4	29.6	75.7	56.9	64.8
	Ours	86.6	83.8	80.4	74.8	65.5	52.6	35.5	78.2	61.8	68.5

Table 2. Comparison with state-of-the-art methods on ActivityNet1.3. Results are reported as mAP (%) at different tIoU thresholds.

Supervision	Method	mAP@tIoU (%)			
		0.5	0.75	0.95	AVG[0.5:0.95]
Video-level	PivoTAL CVPR'23	45.1	28.2	5.0	28.1
	ISSF AAAI'24	39.4	25.8	6.4	25.8
	PVLR MM'24	43.6	27.4	6.5	27.4
	PseudoFormer CVPR'25	46.9	28.6	6.5	29.0
	CRC TPAMI'25	42.0	26.2	5.4	26.2
	ACL-Net TIP'26	43.4	27.6	6.6	27.6
Point-level	LACP ICCV'21	40.4	24.6	5.7	25.1
	CRRC-Net TIP'22	39.8	24.1	5.9	24.0
	Yin *et al.* BMVC'23	48.3	27.8	7.0	29.1
	HR-Pro AAAI'24	42.8	27.2	8.0	27.1
	SNPR TIP'24	41.3	30.9	4.8	26.5
	AAPL AAAI'25	39.6	24.3	5.6	24.7
	TRA ESWA'26	45.3	27.4	6.8	28.0
	PseR TMM'26	49.5	29.6	7.8	30.6
	Ours	49.0	30.7	7.2	30.9

Table 3. Comparison with state-of-the-art methods on GTEA. Results are reported as mAP (%) at different tIoU thresholds.

Method	mAP@tIoU (%)				
	0.1	0.3	0.5	0.7	AVG[0.1:0.7]
SF-Net ECCV'20	58.0	37.9	19.3	11.9	31.0
Ju *et al.* ICCV'21	59.7	38.3	21.9	18.1	33.7
LACP ICCV'21	63.9	55.7	33.9	20.8	43.5
Yin *et al.* BMVC'23	63.1	52.1	37.3	22.2	45.1
HR-Pro AAAI'24	72.6	61.1	37.3	17.5	47.3
TSPNet CVPR'24	74.6	60.9	39.5	16.6	49.0
SMBD ECCV'24	75.0	61.3	41.1	14.2	47.4
SNPR TIP'24	74.3	62.8	35.7	13.7	46.6
AAPL AAAI'25	70.3	54.4	37.7	23.4	46.3
TRA ESWA'26	77.9	67.8	47.2	22.4	54.1
PseR TMM'26	72.7	67.7	42.9	19.9	52.3
Ours	84.2	77.7	57.9	25.5	62.2

respectively. It also achieves the highest mAP at every reported threshold. At tIoU 0.5, our method obtains 57.9% mAP, compared with TRA's 47.2%, demonstrating stronger localization performance on this fine-grained action benchmark.

Table 4. Comparison with state-of-the-art methods on BEOID. Results are reported as mAP (%) at different tIoU thresholds.

Method	mAP@tIoU (%)				
	0.1	0.3	0.5	0.7	AVG[0.1:0.7]
SF-Net ECCV'20	62.9	40.6	16.7	3.5	30.9
Ju *et al.* ICCV'21	63.2	46.8	20.9	5.8	34.9
LACP ICCV'21	76.9	61.4	42.7	25.1	51.8
Yin *et al.* BMVC'23	78.2	65.3	45.1	26.6	54.2
HR-Pro AAAI'24	78.5	72.1	55.3	26.1	59.4
TSPNet CVPR'24	83.8	73.0	51.1	23.8	59.6
SMBD ECCV'24	78.2	71.0	52.5	25.2	57.4
SNPR TIP'24	77.2	64.3	44.0	24.5	53.1
AAPL AAAI'25	75.5	67.6	48.5	26.3	55.2
TRA ESWA'26	83.3	74.5	58.9	31.4	63.7
PseR TMM'26	88.9	83.9	66.1	28.9	68.9
Ours	89.8	85.7	73.3	42.3	74.2

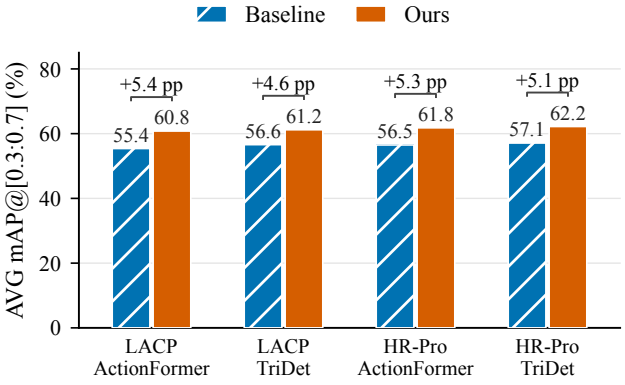


Figure 3. **Generalization across proposal generators and detectors on THUMOS14.** Baselines train matched detectors with each generator's original pseudo-labels; Ours adds PGI and CGR. AVG mAP is evaluated over tIoU thresholds 0.3–0.7. Brackets show gains in percentage points.

Results on BEOID. Table 4 shows that our method achieves 74.2% average mAP over tIoU thresholds 0.1–0.7, improving over PseR by 5.3 percentage points. Our method outperforms all compared approaches at every reported threshold. In particular, mAP at tIoU 0.7 increases from PseR's 28.9% to 42.3%, a gain of 13.4 percentage points, demonstrating improved detection performance under a strict temporal overlap criterion.

4.3. Generalization across Proposal Generators and Detectors

We evaluate the applicability of our framework to different proposal generators and detector architectures on THUMOS14. Specifically, we use LACP and HR-Pro to generate candidate proposals and adopt ActionFormer and TriDet as detectors. Figure 3 compares our framework with baselines trained using each generator's original pseudo-labels under the four matched generator–detector combinations.

With LACP proposals, our framework improves average mAP over tIoU thresholds 0.3–0.7 by 5.4 and 4.6 percentage points when using ActionFormer and TriDet, respectively. With HR-Pro proposals, the corresponding gains are 5.3 and 5.1 percentage points. The consistent improvements across all four combinations demonstrate that proposal-guided pseudo-label learning benefits different candidate sources and detector architectures.

Table 5. Component ablation of PGI and CGR. Results are average mAP (%) over tIoU thresholds 0.1–0.7 for THUMOS14, GTEA, and BEOID, and 0.5–0.95 for ActivityNet1.3. A checkmark indicates an enabled module.

PGI	CGR	THUMOS14	ActivityNet1.3	GTEA	BEOID
–	–	64.1	29.1	56.5	69.5
✓	–	67.0	30.0	57.0	72.2
✓	✓	68.5	30.9	62.2	74.2

4.4. Ablation Studies

Effectiveness of PGI and CGR. We evaluate the contributions of PGI and CGR through progressive component ablation on all four benchmarks. All configurations use HR-Pro as the proposal generator and ActionFormer as the detector. We compare a baseline trained directly with HR-Pro proposals, a variant using PGI initialization with fixed pseudo-labels, and the full framework incorporating CGR.

As shown in Table 5, adding PGI improves average mAP by 2.9, 0.9, 0.5, and 2.7 percentage points on THUMOS14, ActivityNet1.3, GTEA, and BEOID, respectively. Incorporating CGR further improves these results by 1.5, 0.9, 5.2, and 2.0 percentage points, respectively. The consistent gains from both stages demonstrate the effectiveness of prior-guided initialization and consensus-guided refinement across the four benchmarks.

Contribution of Point-Derived Priors. We examine the effects of the geometric and conflict priors on initial pseudo-label quality on THUMOS14. We retain the confidence prior in all variants and selectively enable the two point-derived priors, keeping the candidate pool fixed and evaluating labels before CGR refinement. Ground-truth boundaries are used only for evaluation.

Table 6 shows that both point-derived priors improve pseudo-label quality over confidence-only selection. The geometric prior increases mean tIoU from 0.571 to 0.601 and the fraction of labels with tIoU $\geq$ 0.7 from 39.9% to 50.9%. The conflict prior yields a larger improvement in mean tIoU, reaching 0.651, while the geometric prior produces a higher fraction of well-localized labels. Combining both priors achieves the best results on both metrics, with

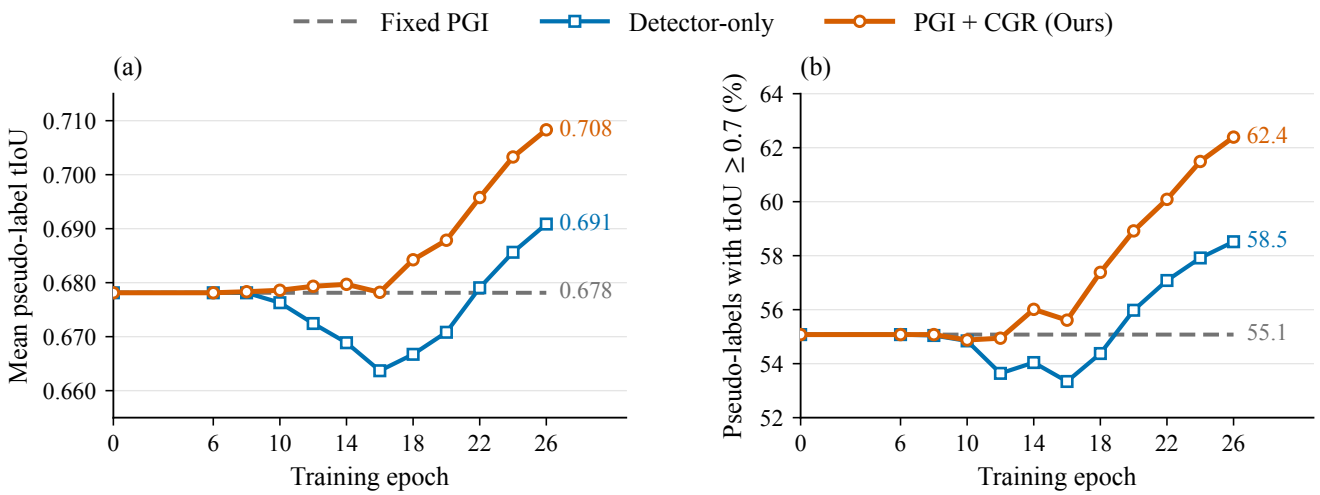


Figure 4. **Pseudo-label refinement dynamics on THUMOS14.** (a) Mean pseudo-label tIoU. (b) Fraction of pseudo-labels with tIoU ≥ 0.7. Fixed PGI, detector-only refinement, and full CGR share the same initialization and are evaluated against official ground truth on the same instances. Epoch 0 denotes initialization; markers indicate refinement checkpoints. Final values are annotated.

Table 6. Ablation of PGI priors on THUMOS14. Initial pseudo-label quality is measured by mean tIoU and the fraction of labels with tIoU ≥ 0.7.

Confidence	Geometric	Conflict	Mean tIoU	tIoU ≥ 0.7 (%)
✓	–	–	0.571	39.9
✓	✓	–	0.601	50.9
✓	–	✓	0.651	47.0
✓	✓	✓	0.678	55.1

Table 7. Ablation of consistency control in CGR on THUMOS14. AVG mAP (%) is computed over tIoU thresholds 0.1–0.7. Pseudo-label quality is evaluated against official ground truth at the final update.

Consistency control	τ_{con}^{k}	AVG mAP (%)	Mean tIoU	tIoU ≥ 0.7 (%)
None	0	64.3	0.674	56.4
Fixed	0.5	67.4	0.684	58.4
Progressive	$1 \rightarrow 0$	68.5	0.708	62.4

a mean tIoU of 0.678 and a high-quality label fraction of 55.1%. These results support their complementary roles in identifying reliable initial pseudo-labels.

Effectiveness of CGR. On THUMOS14, we assess CGR by comparing fixed PGI labels, detector-only updates using detector predictions alone, and full CGR, all with the same PGI initialization and ActionFormer detector. As shown in Figure 4, CGR better preserves initialization quality during the middle of training and improves both quality metrics in later updates. At the final update, CGR reaches a mean pseudo-label tIoU of 0.708, with 62.4% of labels achieving tIoU ≥ 0.7. Compared with fixed PGI labels, CGR improves mean tIoU by 0.030 and the high-quality label fraction by 7.3 percentage points. Compared with detector-only updates, it improves mean tIoU by 0.017 and the high-quality label fraction by 3.9 percentage points. These results demonstrate that combining retained proposal evidence with detector predictions improves pseudo-label quality during training.

Consistency-Controlled Refinement. We further examine the role of consistency control in CGR on THUMOS14. With the same PGI initialization, proposal consensus, fusion parameters, and update epochs, we compare three settings: no consistency constraint ($\tau_{\text{con}}^{k} = 0$), a fixed threshold ($\tau_{\text{con}}^{k} = 0.5$), and the progressive schedule that decreases the threshold from 1 to 0. Detection performance is measured at the final training epoch, and pseudo-label quality is evaluated against official ground truth after the final update.

Table 7 shows that progressive consistency control achieves 68.5% average mAP, outperforming unconstrained updates and the fixed threshold by 4.2 and 1.1 percentage points, respectively. It also yields the highest mean pseudo-label tIoU of 0.708 and a high-quality label fraction of 62.4%, exceeding the fixed-threshold variant by 4.0 percentage points on the latter metric. These results support strict agreement early in refinement followed by gradual relaxation, which improves both temporal supervision and detection performance.

Table 8. Contribution of proposal consensus on THUMOS14. AVG mAP (%) is computed over tIoU thresholds 0.1–0.7 at the final training epoch. Mean tIoU and the high-quality label fraction are evaluated against official GT at the final update. The highlighted row denotes full CGR.

Boundary evidence	AVG mAP (%)	Mean tIoU	tIoU ≥ 0.7 (%)
Detector-only	67.4	0.691	58.5
Initial boundary	68.0	0.699	61.5
Proposal consensus	68.5	0.708	62.4

Contribution of Proposal Consensus. On THUMOS14, we compare detector-only refinement, refinement guided by the initial PGI boundary ($K = 1$), and full CGR with proposal consensus ($K = 10$), using the same PGI initialization. Detector-only uses detector predictions alone; the other two variants introduce original boundary evidence from a single initial pseudo-label or multiple retained candidates, respectively. Table 8 shows that incorporating original boundary evidence improves both detection performance and pseudo-label quality over detector-only refinement. Full CGR achieves 68.5% average mAP, compared with 67.4% for detector-only refinement, while increasing mean pseudo-label tIoU from 0.691 to 0.708 and the fraction of labels with tIoU ≥ 0.7 from 58.5% to 62.4%. These results demonstrate the effectiveness of retained proposal evidence in guiding pseudo-label refinement.

Number of Retained Proposals. We evaluate the sensitivity to the number of retained proposals K on THUMOS14, keeping PGI initialization and all other refinement settings unchanged. We vary $K \in \{1, 5, 10, 20\}$, where $K = 1$ corresponds to the initial-boundary variant in Table 8. As shown in Figure 5, AVG mAP increases from 68.0% at $K = 1$ to 68.2% at $K = 5$ and 68.5% at $K = 10$, and remains at 68.5% when K increases to 20. This performance plateau supports $K = 10$ as a compact default.

Consensus Fusion Threshold. We vary the consensus fusion threshold $\tau_{\text{iou}} \in \{0.1, 0.3, 0.5, 0.7\}$ on THUMOS14 and ActivityNet1.3, keeping the PGI initialization and all other refinement settings fixed within each dataset. This threshold controls whether proposal consensus is fused with the detector prediction. Table 9 shows that thresholds of 0.3 and 0.5 yield comparable AVG mAP on both datasets, reaching 68.5% on THUMOS14 and 30.9% on ActivityNet1.3. The default threshold of 0.5 also achieves 35.5% mAP@0.7 on THUMOS14 and the highest mAP@0.95 of 7.2% on ActivityNet1.3 among the tested settings. Increasing the threshold to 0.7 reduces both average and strict-tIoU detection performance on the two datasets. These results support 0.5 as a shared default for consensus fusion.

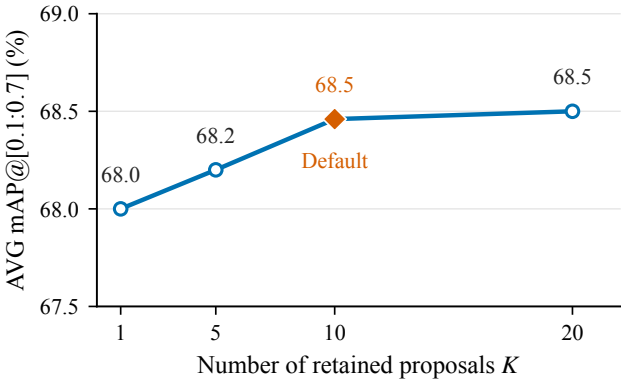


Figure 5. Sensitivity to the number of retained proposals K on THUMOS14, measured by AVG mAP (%) over tIoU thresholds 0.1–0.7. The $K = 1$ result is the initial-boundary variant in Table 8. The orange diamond marks the default $K = 10$.

Table 9. Sensitivity to the consensus fusion threshold τ_{iou}. All results are final-epoch mAP (%). AVG denotes the average over tIoU thresholds 0.1–0.7 on THUMOS14 and 0.5–0.95 on ActivityNet1.3. The highlighted row uses the default threshold of 0.5.

τ_{iou}	THUMOS14		ActivityNet1.3	
	mAP@0.7	AVG	mAP@0.95	AVG
0.1	35.5	68.1	7.0	30.8
0.3	35.1	68.5	7.1	30.9
0.5	35.5	68.5	7.2	30.9
0.7	33.7	67.8	7.0	30.7

4.5. Qualitative Analysis

We visualize pseudo-label refinement on training clips and final detection results on test videos from THUMOS14.

Pseudo-label refinement. Figure 6 compares top-score pseudo-labels, PGI initialization, and final CGR labels for two training clips. In the HammerThrow example, confidence-based selection extends intervals into neighboring instances, whereas PGI selects boundaries closer to the target actions. CGR further refines their temporal extent, increasing mean tIoU across the two highlighted instances from 0.424 to 0.854 and then to 0.911. The LongJump example exhibits both truncated and overextended top-score intervals; PGI and CGR progressively improve their alignment with GT, with mean tIoU increasing from 0.525 to 0.839 and 0.903. These examples illustrate the complementary roles of prior-guided initialization and consensus-guided

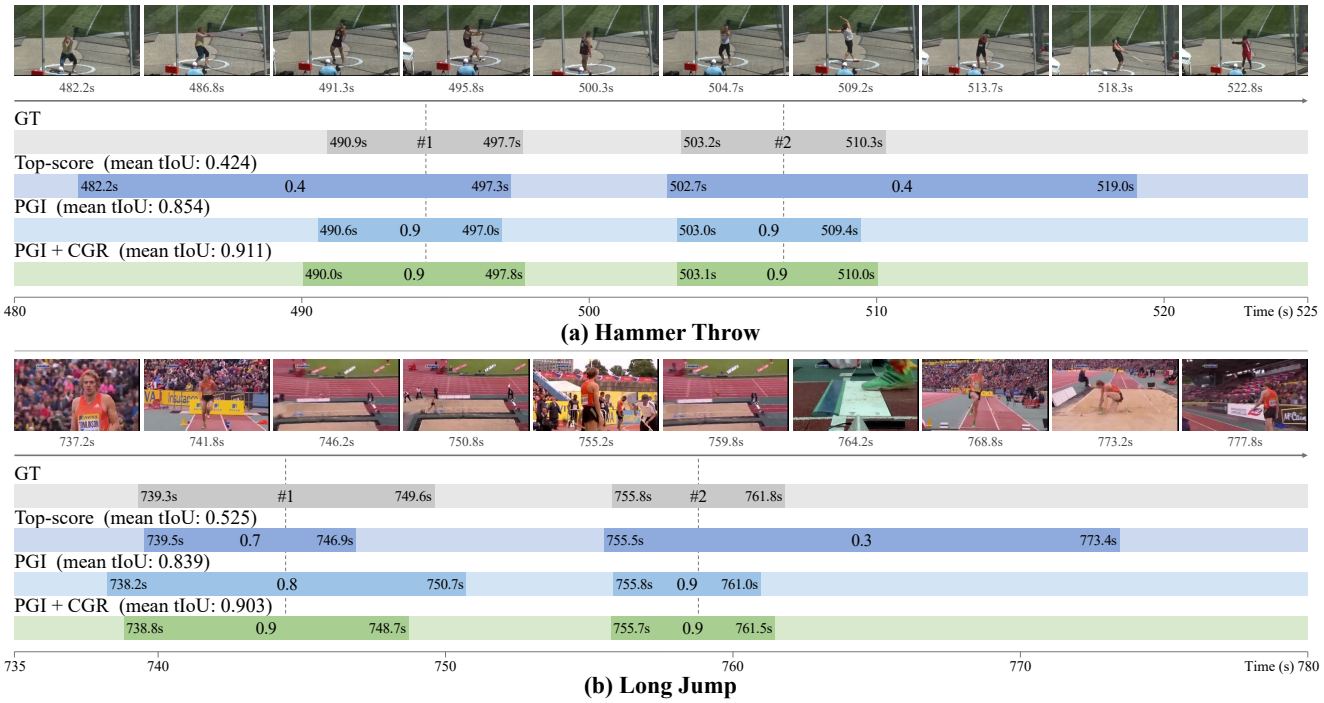


Figure 6. **Qualitative comparison of pseudo-labels on THUMOS14 training clips.** Each panel tracks two highlighted instances across top-score selection, PGI initialization, and the final CGR update. Top-score selection uses the highest-confidence same-class proposal containing each annotated point. Dashed lines mark annotated points, and numbers inside pseudo-label intervals denote tIoU with official GT. Instance-level tIoUs are rounded to one decimal place; row means are computed before rounding.

refinement in improving temporal supervision.

Temporal localization results. Figure 7 compares the final detector predictions with GT on two unseen test videos, using the same confidence threshold of 0.5 for visualization. The CleanAndJerk predictions closely follow the annotated intervals, achieving tIoU values of 0.9 and 1.0. The BasketballDunk predictions recover both action occurrences with tIoU values of 0.7 and 0.8. In both examples, the detector distinguishes separate instances and localizes their temporal extent.

5. Conclusion

We presented a framework for learning high-quality pseudo-labels in point-supervised temporal action localization through prior-guided initialization and consensus-guided refinement. Our central idea is to reuse original proposals as boundary evidence for both supervision initialization and subsequent refinement. PGI combines proposal confidence with point-derived geometric and conflict priors to select reliable initial pseudo-labels and retain alternative candidates. CGR integrates consensus from these candidates with evolving detector predictions to refine supervision through consistency-controlled updates. Experiments on four benchmarks demonstrate improved detection performance, while ablations and pseudo-label analyses validate the effectiveness of both modules. The framework is also effective with different proposal generators and detector architectures. These findings highlight the value of retaining and reusing original proposal evidence throughout pseudo-label learning.

References

[1] Tanay Agrawal, Abid Ali, Antitza Dantcheva, and Francois Bremond. Scaling action detection: Adatad++ with transformer-enhanced temporal-spatial adaptation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12222–12231. IEEE, 2025. 2

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 6

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[4] Dima Damen, Teesid Leelasawassuk, and Walterio Mayol-Cuevas. You-do, i-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based

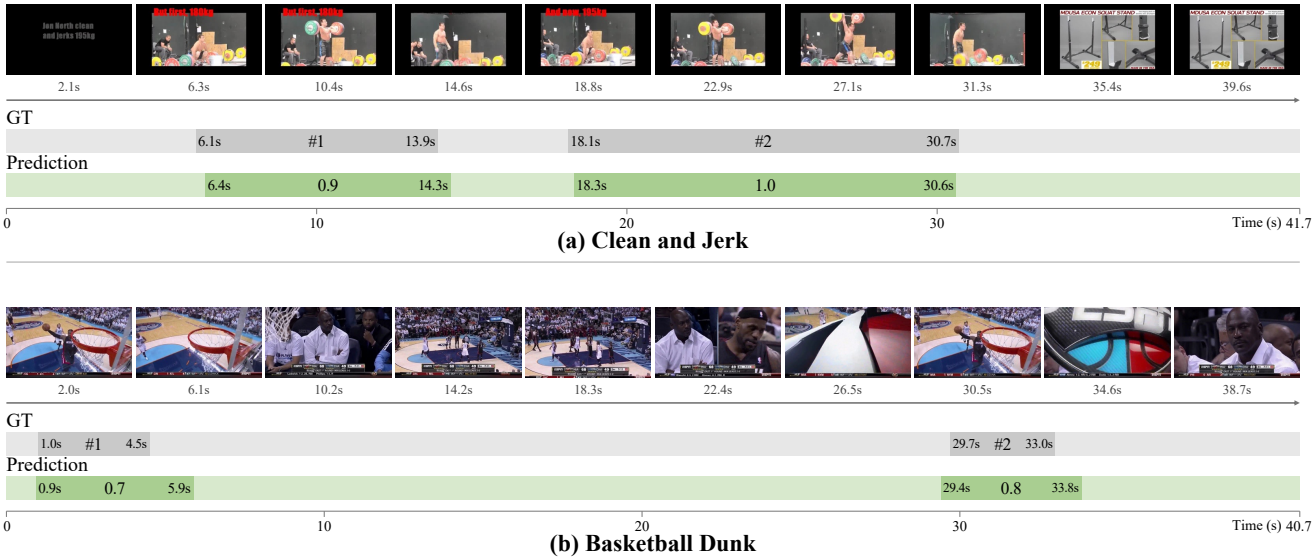


Figure 7. **Final detection results on THUMOS14 test videos.** GT and predicted intervals are shown for CleanAndJerk and BasketballDunk. All saved predictions with confidence ≥ 0.5 are displayed. Numbers inside prediction intervals denote tIoU with the corresponding official GT, rounded to one decimal place.

guidance. *Computer Vision and Image Understanding*, 149: 98–112, 2016. 6

[5] Zhenying Fang and Richang Hong. Pser: Pseudo-label refinement for point-supervised temporal action detection. *IEEE Transactions on Multimedia*, 28:128–139, 2026. 1, 2, 3, 6, 7

[6] Zhenying Fang, Suguo Zhu, Jun Yu, and Qi Tian. Pcpcad: proposal complementary action detector. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 424–429. IEEE, 2019. 1, 2

[7] Zhenying Fang, Jianping Fan, and Jun Yu. Lpr: learning point-level temporal action localization through re-training. *Multimedia Systems*, 29(5):2545–2562, 2023. 1, 3, 7

[8] Zhenying Fang, Jun Yu, and Richang Hong. Boundary discretization and reliable classification network for temporal action detection. *IEEE Transactions on Multimedia*, 27:5198–5211, 2025. 1

[9] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011. 6

[10] Jie Fu, Junyu Gao, and Changsheng Xu. Compact representation and reliable classification learning for point-level weakly-supervised action localization. *IEEE Transactions on Image Processing*, 31:7363–7377, 2022. 1, 3, 7

[11] Zan Gao, Weilin Yang, Yibo Zhao, Chunjie Ma, Chen Li, and Riwei Wang. Lsgnet: A local-pattern separation and global-aware network for temporal action detection. *IEEE Transactions on Image Processing*, 35:4643–4658, 2026. 2

[12] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 6

[13] Chen Ju, Peisen Zhao, Siheng Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Divide and conquer for single-frame temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13455–13464, 2021. 7

[14] Ho-Joong Kim, Yearang Lee, Jung-Ho Hong, and Seong-Whan Lee. Digit: Multi-dilated gated encoder and central-adjacent region integrated decoder for temporal action detection transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24286–24296. IEEE, 2025. 2

[15] Pilhyeon Lee and Hyeran Byun. Learning action completeness from points for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13648–13657, 2021. 1, 3, 7

[16] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE, 2012. 1

[17] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1369–1379, 2018. 1

[18] Guozhang Li, De Cheng, Xinpeng Ding, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. Boosting weakly-supervised temporal action localization with text information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10648–10657, 2023. 2

[19] Guozhang Li, De Cheng, Nannan Wang, Jie Li, and Xinbo Gao. Neighbor-guided pseudo-label generation and refinement for single-frame supervised temporal action localiza-

tion. *IEEE Transactions on Image Processing*, 33:2419–2430, 2024. 7

[20] Maodong Li, Zhihao Wang, Tian Wang, Jingxiong Wang, Jian Wang, and Bing Li. Weakly supervised temporal action localization with proposal-level action consistency learning. *IEEE Transactions on Image Processing*, 2026. 2, 7

[21] GeunTaek Lim, Hyunwoo Kim, Joonsoo Kim, and Yukyung Choi. Probabilistic vision-language representation for weakly supervised temporal action localization. In *ACM Multimedia 2024*, 2024. 7

[22] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 988–996, 2017. 2

[23] Mengnan Liu, Le Wang, Sanping Zhou, Kun Xia, Xiaolong Sun, and Gang Hua. Boosting point-supervised temporal action localization through integrating query reformation and optimal transport. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13865–13875. IEEE, 2025. 3, 7

[24] Mengnan Liu, Le Wang, Sanping Zhou, Kun Xia, Qi Wu, Qilin Zhang, and Gang Hua. Stepwise multi-grained boundary detector for point-supervised temporal action localization. In *European Conference on Computer Vision*, pages 333–349. Springer, 2025. 7

[25] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31:5427–5441, 2022. 2

[26] Ziyi Liu and Yangcen Liu. Bridge the gap: From weak to full supervision for temporal action localization with pseudoformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8711–8720. IEEE, 2025. 2, 7

[27] Fan Ma, Linchao Zhu, Yi Yang, Shengxin Zha, Gourab Kundu, Matt Feiszli, and Zheng Shou. Sf-net: Single-frame supervision for temporal action localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 420–437. Springer, 2020. 1, 3, 7

[28] Yunchuan Ma, Laiyun Qing, Guorong Li, Yuqing Liu, Yuankai Qi, and Qingming Huang. Boosting point-supervised temporal action localization via text refinement and alignment. *Expert Systems with Applications*, 334:134333, 2026. 7

[29] Yunchuan Ma, Laiyun Qing, Guorong Li, Yuqing Liu, Yuankai Qi, and Qingming Huang. Exploring the temporal consistency for point-level weakly-supervised temporal action localization. *arXiv preprint arXiv:2602.05718*, 2026. 3

[30] Yunchuan Ma, Laiyun Qing, Guorong Li, Yuqing Liu, Yuankai Qi, and Qingming Huang. Boosting point-supervised temporal action localization via text refinement and alignment. *Expert Systems with Applications*, 334:134333, 2027. 3

[31] Davide Moltisanti, Sanja Fidler, and Dima Damen. Action recognition from single timestamp supervision in untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9915–9924, 2019. 6

[32] Huan Ren, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. Proposal-based multiple instance learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2394–2404, 2023. 1, 2, 3, 7

[33] Mamshad Nayeem Rizve, Gaurav Mittal, Ye Yu, Matthew Hall, Sandra Sajeev, Mubarak Shah, and Mei Chen. Pivotal: Prior-driven supervision for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22992–23002, 2023. 1, 2, 7

[34] Yuxiang Shao, Feifei Zhang, and Changsheng Xu. Modeling semantic and localization uncertainty for weakly supervised temporal action localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026. 2

[35] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18857–18866, 2023. 1, 2, 7

[36] Elahe Vahdani and Yingli Tian. Potloc: Pseudo-label oriented transformer for point-supervised temporal action localization. *Computer Vision and Image Understanding*, 246: 104044, 2024. 1, 3

[37] Fengshun Wang, Zhengbo Zhang, and Zhigang Tu. Masked diffusion vision-language models for temporal action localization. *arXiv preprint arXiv:2605.29858*, 2026. 1, 7

[38] Ziying Xia, Jian Cheng, Siyu Liu, Yongxiang Hu, Shiguang Wang, Yijie Zhang, and Liwan Dang. Realigning confidence with temporal saliency information for point-level weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18440–18450, 2024. 4, 7

[39] Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. Less is more: Learning highlight detection from video duration. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1258–1267. IEEE, 2019. 1

[40] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE international conference on computer vision*, pages 5783–5792, 2017. 2

[41] Le Yang, Houwen Peng, Dingwen Zhang, Jianlong Fu, and Junwei Han. Revisiting anchor mechanisms for temporal action localization. *IEEE Transactions on Image Processing*, 29:8535–8548, 2020. 2

[42] Wenfei Yang, Huan Ren, Tianzhu Zhang, Zhe Zhang, Yongdong Zhang, and Feng Wu. Cross-task relation-aware consistency for weakly supervised temporal action detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2, 7

[43] Zichen Yang, Jie Qin, and Di Huang. Acgnet: Action complement graph network for weakly-supervised temporal action localization. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3090–3098, 2022. 2

[44] Yuan Yin, Yifei Huang, Ryosuke Furuta, and Yoichi Sato. Proposal-based temporal action localization with point-level supervision. In *BMVC*, pages 820–821, 2023. 1, 3, 7

[45] Shuhei M Yoshida, Takashi Shibata, Makoto Terao, Takayuki Okatani, and Masashi Sugiyama. Action-agnostic point-level supervision for temporal action detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9571–9579, 2025. 7

[46] Wulian Yun, Mengshi Qi, Chuanming Wang, and Huadong Ma. Weakly-supervised temporal action localization by inferring salient snippet-feature. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6908–6916, 2024. 2, 7

[47] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022. 1, 2, 6, 7

[48] Huaxin Zhang, Xiang Wang, Xiaohao Xu, Zhiwu Qing, Changxin Gao, and Nong Sang. Hr-pro: Point-supervised temporal action localization via hierarchical reliability propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7115–7123, 2024. 1, 3, 6, 7

[49] Quan Zhang, Yuxin Qi, Xi Tang, Rui Yuan, Xi Lin, Ke Zhang, and Chun Yuan. Rethinking pseudo-label guided learning for weakly supervised temporal action localization from the perspective of noise correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10085–10093, 2025. 3, 7