LLM-GUIDED DISTRIBUTION PROPOSAL REFINEMENT FOR WEAKLY SUPERVISED TEMPORAL ACTION LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Weakly supervised temporal action localization aims to localize action instances in untrimmed videos using only the video-level label. However, the semantic mismatch between weak supervision and background frames may yield inaccurate action scores, which in turn causes inaccurate action boundaries. To address this, we introduce an LLM-guided distribution proposal refinement model, which explicitly models subaction semantics via four mechanisms. *First*, to reduce the semantic mismatch in weak supervision, we design an LLM-guided proposal learning module. It uses LLM prompts to extract the textual semantics of start, middle, and end subactions. A proposal structure loss ensures that subactions are properly ordered. *Second*, to alleviate ambiguous prediction among subactions, we design a contrastive proposal learning module. The discriminativeness of subaction representations is optimized via intra-action and inter-action contrastive losses. *Third*, to alleviate ambiguous boundary localization caused by smooth transitions, we design a proposal transition learning module. The score changes at the boundary are encouraged to be separable via a boundary transition loss. *Fourth*, to alleviate ambiguous boundary localization among gradual transition segments, we propose a proposal distribution refinement module. The boundary is refined to the location by considering the difference between left- and right-window distributions (measured via the first-order Wasserstein distance). Experiments show that our method achieves state-of-the-art performance on THUMOS'14 and ActivityNet 1.3 datasets.

1 INTRODUCTION

Temporal action localization (TAL) aims to localize action instances in untrimmed videos. Unlike fully supervised TAL, which requires frame-level annotations, weakly supervised TAL (WTAL) requires only video-level action labels, making it much easier to build large-scale datasets. Action localization can be enhanced through consistency learning (Li et al. (2026); Ren et al. (2023); Li et al. (2025; 2024b)), contrastive learning (Zhang et al. (2021); Gao et al. (2022); Li et al. (2022)), temporal difference learning (Zhou et al. (2025); Yun et al. (2024)), and pseudo-label learning (Zhang et al. (2025b); Luo et al. (2025); Zou et al. (2025)). Large language models (LLMs) provide rich textual semantic information to enhance action localization (Shao et al. (2025; 2026); Zhang et al. (2025a)). *However*, due to the lack of frame-wise annotations, the above methods struggle to capture the semantic variations across different frames. Such semantically diverse frames, jointly with background frames, are hard to match to weak video-level labels, leading to semantic mismatch and inaccurate action localization. Therefore, modeling such diverse semantics remains a significant challenge for WTAL.

In **Figure 1 (a)**, the weak video-level label supervision is applied to both action and background frames. However, this supervision mainly captures salient frames within the core action segment (e.g., the middle segment), while providing ambiguous guidance for transitional frames between action and background. To ensure complete localization, existing method (Li et al. (2025)) aims to enforce consistency among raw proposals by penalizing proposals whose centers deviate from the action center. Nevertheless, it neglects to directly examine frame-level transition features via their semantic alignment. Then, the detected start boundary exhibits a large offset from the ground truth.

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

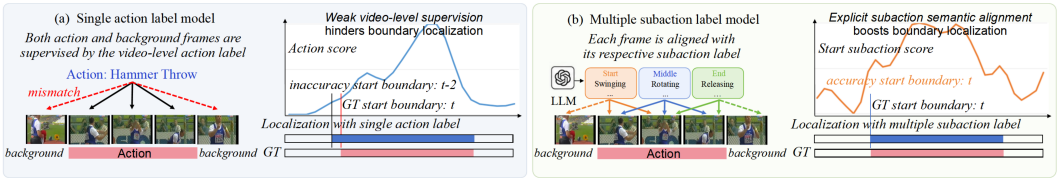

Figure 1: Illustration of inaccurate action boundaries caused by semantic mismatch. (a) In weak supervision, when matching background features to a single video-level label, this semantic mismatch may cause inaccurate boundary localization. (b) Our model introduces LLM-guided subaction labels for explicit semantic alignment, enabling accurate boundary localization.

To alleviate the boundary offset problem, relying solely on visual analysis of the core action is insufficient. It is crucial to identify segments that match the transitional semantics. However, existing methods (Shao et al. (2025; 2026); Zhang et al. (2025a)) learn text semantics from a single action label, lacking specific textual guidance for transition segments. To this end, we address this by explicitly defining transition subactions that are distinct from the core action, decomposing each action proposal into start, middle, and end subactions. In **Figure 1 (b)**, using LLM-based subaction prompts, we generate these fine-grained descriptions from the video-level action name (e.g., "Throw Hammer" has "swinging hammer" as the start subaction). Accordingly, we observe that significant changes in visual-text similarity correlate strongly with action–background boundary points.

To alleviate boundary errors caused by semantic mismatch, we propose an LLM-guided Distribution Proposal Refinement (LDPR) model, as shown in **Figure 2**. LDPR learns subaction semantics to refine boundary of raw proposals by introducing four mechanisms (i.e., subaction structure learning, discriminative subaction learning, separable transition learning, and distribution boundary refinement). *First*, to detect semantic subaction segments, we design an LLM-guided proposal contrastive learner. This module employs an LLM-based subaction prompt to decompose action names into texts of start/middle/end subactions. The proposal structure loss ensures that subactions are properly ordered. *Second*, to ensure discriminative subaction representations, we design a contrastive proposal learner. The intra-action contrastive loss separates the subactions within the same action, while the inter-action contrastive loss separates the middle subaction across different actions. *Third*, to alleviate ambiguous boundary localization caused by smooth transitions, we design a proposal transition learner. The proposal transition loss enforces a separable transition between the left and right windows around each predicted boundary. *Fourth*, to alleviate ambiguous boundary localization among gradual transition segments, we design a proposal distribution refiner. Since Wasserstein distance effectively models the distributional difference and is more stable than point-level differences, it serves as a robust indicator of temporal transitions. Therefore, we refine the boundary by computing a weighted average of candidate timestamps, where the weights are determined by the first-order Wasserstein distance, thus enhancing boundary localization.

Our contributions are summarized as follows:

(1) We propose the LLM-guided distribution proposal refinement (LDPR) model, which aims to alleviate the semantic mismatch with subaction semantic guidance in weakly supervised temporal action localization. To ensure subaction semantic learning, we design an LLM-guided proposal learning module to extract start, middle, and end segments. The proposal structure loss ensures that subactions are properly ordered.

(2) To ensure discriminative subaction representations, we design a contrastive proposal learning module. The discriminative subaction representations are ensured using intra-action contrastive learning and inter-action contrastive learning.

(3) To alleviate ambiguous boundary localization caused by smooth transitions, we design a proposal transition learning module. The proposal transition loss estimates the separable transitions at both start and end boundaries, ensuring proper boundary localization.

(4) To alleviate ambiguous boundary localization among gradual transition segments, we design a proposal distribution refinement module. It refines the boundary point by considering first-order Wasserstein distance between the left and right window distributions.

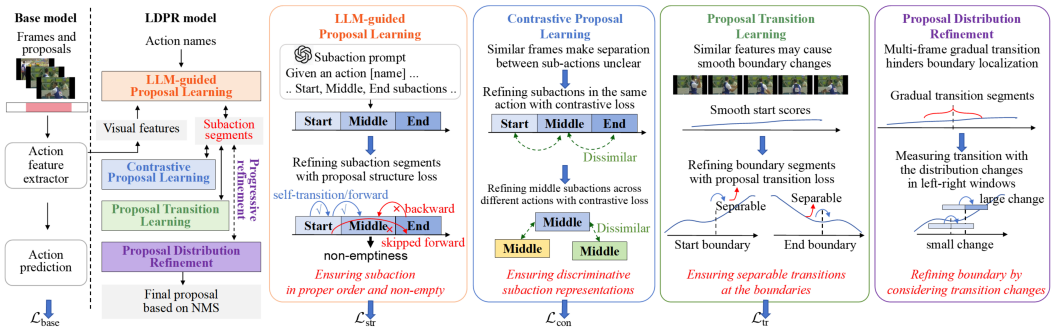

Figure 2: The LLM-guided Distribution Proposal Refinement (LDPR) Model. To alleviate inaccurate action boundaries caused by semantic mismatch, LDPR decomposes action semantics into subaction semantics and refines subaction segments by introducing four mechanisms, i.e., subaction structure learning, discriminative subaction learning, separable transition learning, and distribution boundary refinement. These mechanisms are implemented via four corresponding modules, i.e., LLM-guided proposal learner, contrastive proposal learner, proposal transition learner, and proposal distribution refiner.

2 Related Work

Weakly supervised temporal action localization aims to identify action instances in untrimmed videos using only video-level labels (Wang et al. (2017)). Existing methods fall into three categories. The first category concentrates on feature enhancement, leveraging cross-modal consensus (Hong et al. (2021)), attention mechanisms (He et al. (2022)), uncertainty guidance (Yang et al. (2021)), discriminability modeling (Tang et al. (2023)), causal modeling (Li et al. (2024a)), or evidential learning (Chen et al. (2022)). The second category tackles foreground-background separation, employing contrastive learning strategies (Zhang et al. (2021); Gao et al. (2022); Li et al. (2022)), semantic dictionaries (Wang et al. (2023b)), or prototype-based memory networks (Luo et al. (2021); Liu et al. (2023); Huang et al. (2022)). The third category enforces temporal or semantic consistency, using constraints at the foreground-action (Huang et al. (2021)), proposal-level (Li et al. (2026)), instance-level (Ren et al. (2023)), process-level (Li et al. (2025; 2024b)), or change-point detection (Zhou et al. (2025); Yun et al. (2024); Wang et al. (2023a)). The above methods rely solely on visual features, neglecting fine-grained text-semantic matching. This neglect often results in ambiguous action boundaries. Thus, resolving boundary ambiguity via LLM-guided text-visual alignment remains highly important.

Vision-language models have established a powerful paradigm for aligning visual and textual semantics, enabling downstream tasks to benefit from richly structured cross-modal representations (Radford et al. (2021)), unlike static word embedding models (Li et al. (2023)). Existing vision-language methods inject global text knowledge to enhance feature discriminability (Ju et al. (2023); Lim et al. (2024)), while others further enrich textual priors via class descriptions and external knowledge (Shao et al. (2025; 2026); Zhang et al. (2025a)). However, these methods ignore temporal action transitions, causing semantic mismatches near boundaries. To this end, we introduce LLM-guided prompts to decompose actions into start/middle/end texts, reducing boundary false predictions.

Pseudo-label Learning aims to address weakly supervised tasks by iteratively training the model with its own predicted labels. For weakly supervised temporal action localization, some methods focus on pseudo-label extraction, including soft snippet assignment (Rizve et al. (2023)), context-aware augmentation (Zhang et al. (2025b)), optimal transport matching (Luo et al. (2025)), and binary mask diffusion (Zou et al. (2025)). Others analyze pseudo-label optimization, with complementary learning constraints (Du et al. (2025)), pairwise IoU consistency (Feng et al. (2025)), and discrepancy-based penalties (Zhou et al. (2023)). In a unified framework consisting of a weak branch and a full branch, PseudoFormer (Liu & Liu (2025)) uses pseudo labels to learn both snippet-level and proposal-level information. However, the above methods formulate snippet-level pseudo-labels, which can be affected by noisy features, particularly near action boundaries where visual ambiguities are most severe. In contrast, our work aims to alleviate ambiguous predictions by refining pseudo-labels around action boundaries.

3 METHOD

3.1 BASE MODEL

The base model adopts a similar modality enhancement and action consistency learning (Li et al. (2025)). It calculates the similarity between the snippet representation and the average representation of a specific class. This similarity indicates the high/middle confidence of snippet-level pseudo labels. We use the model to optimize raw proposals from (Yun et al. (2024)). Each proposal is denoted as $Q = (\hat{t}^{st}, \hat{t}^{ed}, \hat{c})$, where $\hat{t}^{st}$ and $\hat{t}^{ed}$ indicate the start and end snippet indices, and $\hat{c}$ represents the predicted category. Consequently, the foreground snippet set is $Fg^Q = \{t|\hat{t}^{st} \leq t \leq \hat{t}^{ed}\}$.

Base loss contains four terms: The classification loss $\mathcal{L}_{cls}$ reduces the classification error of video-level prediction. The similar modality enhancement loss $\mathcal{L}_{sme}$ reduces the prediction error with snippet-level pseudo labels. The process consistency loss $\mathcal{L}_{pc}$ encourages proposals to cover action instances with a small center offset. The semantic consistency loss $\mathcal{L}_{cs}$ promotes similarity among proposals of the same class and suppresses scores on background snippets. Then, the base loss is $\mathcal{L}_{base} = \mathcal{L}_{cls} + \lambda_{sme}\mathcal{L}_{sme} + \lambda_{pc}\mathcal{L}_{pc} + \lambda_{cs}\mathcal{L}_{cs}$, where $\lambda_{sme}, \lambda_{pc}, \lambda_{cs}$ are their hyperparameters.

Discussion. The base model relies on pseudo labels derived from averaged class representations, which may neglect diverse subaction semantics, especially in transition segments. To address this, we introduce LLM-generated prompts to extract distinct subaction texts, and refine subaction representations to be properly ordered and discriminative.

3.2 LLM-GUIDED PROPOSAL CONTRASTIVE LEARNING

Existing methods (Shao et al. (2025; 2026); Zhang et al. (2025a)) rely on a single action label for text supervision, which may inadequately capture the semantics of transitional segments. To address this, we explicitly treat transition subactions as distinct from the core action using LLM.

The subaction prompt template generates subaction texts offline before training. Given the large language model GPT-5 (OpenAI (2026)), we design the same template across all classes to ensure consistent text learning. To ensure boundary modeling, one action is decomposed into three subactions, i.e., start, middle, and end. Each subaction is described with multi-dimensional semantics, including subject, object, and manner, thus ensuring visual-text matching. The subaction prompt is,

> **[Task]** Given an action [Name], generate English semantic descriptions for the Start, Middle, and End subactions.
> **[Subaction Definitions]** - Start: First observable initiating movement. - Middle: Most typical, class-discriminative visual pattern during stable execution. - End: Final movement.
> **[Semantic Dimensions]** (Apply to all subactions) - Subject: Observable identity, posture, and body movement. - Object: State of key objects/targets and their spatial relationship with the subject. - Manner: Motion pattern, body coordination, and interaction dynamics.

The text encoder extracts the subaction tokens from generated subaction texts (see supplementary materials for the full set of subaction texts across non-background classes). Given action class index c and subaction index h ($h = 1, 2, 3$ denote the start/middle/end subaction), the generated subaction texts are formulated as a set $Prompt^c = \{prompt_h^c\}$. Each subaction description is encoded by a frozen GPT-2 Large $Enc(\cdot)$ (Radford et al. (2019)) and average-pooled into one sentence-level token. The subaction text token is $\hat{e}_h^c = averagepool(Enc(prompt_h^c))$.

Semantic prediction decomposes the foreground segment into subaction segments via visual-text cross attention. In the visual branch, an FC layer projects I3D features (Carreira & Zisserman (2017)) to learn the visual features $X = \{x_t\}$, where $t = 1, ..., T$ is the snippet index. In the text branch, for a non-background action c, an FC layer projects the text token to learn the text features $E^c = \{e_h^c\}$, $h = 1, ..., H$. A single-layer Transformer decoder performs cross-attention, where the visual Query (from X) is explicitly guided by the textual Key/Value (from E^c). This process injects the fine-grained semantics of start/middle/end subactions into the visual features, outputting semantic-enhanced features X^c. For each snippet feature $x_t^c \in X^c$, a Multi-Layer Perceptron predicts the subaction logits $s_{c,t}^{sub} = \{s_{c,t,h}^{sub}\}$. To naturally handle unreliable snippets, we

adopt a multi-label scoring mechanism that does not constrain the cross-subaction sum to 1. For all frames, each subaction channel independently performs softmax across non-background action classes to obtain prediction scores $p^{sub}_{c,t,h}$. For each base proposal $Q = (\hat{t}^{st}, \hat{t}^{ed}, \hat{c})$, $\hat{c}$ is the proposal action label. The subaction label is determined by $y^{sub}_{\hat{c},t} = argmax_h(\{p^{sub}_{\hat{c},t,h}\})$. In each proposal, the corresponding snippet set is $Tset_{\hat{c},h} = \{t | t \in Fg^Q, y^{sub}_{\hat{c},t} = h\}$.

Proposal structure loss is designed to ensure the order and non-emptiness in the subaction structure. To measure temporal order in each proposal, we define a transition matrix $\mathbb{A}$ that assigns 1 to all backward transitions and skipped forward transitions ($\mathbb{A}^{t-1}_t = 1, if \quad y^{sub}_{\hat{c},t-1} > y^{sub}_{\hat{c},t} \quad or \quad y^{sub}_{\hat{c},t-1}+2 = y^{sub}_{\hat{c},t},$), while assigning 0 to valid forward transitions (start to middle to end) and self-loops ($\mathbb{A}^{t-1}_t = 0, otherwise$). Invalid transitions are weighted by the transition confidence, which is estimated as $w^{t-1}_t = p^{sub}_{\hat{c},t-1,y^{sub}_{\hat{c},t-1}} \cdot p^{sub}_{\hat{c},t,y^{sub}_{\hat{c},t}}$. For a proposal with fewer than 3 snippets, it is difficult to ensure that all subactions are non-empty. We ensure the non-emptiness of the middle subaction by penalizing with its snippet count $|Tset_{\hat{c},2}|$. To address the non-differentiability of the hard assignments, we optimize both terms using the straight-through estimator (see supplementary materials). The proposal structure loss consists of the above two terms as,

$$\mathcal{L}_{str} = \begin{cases} max(0, 1 - |Tset_{\hat{c},2}|), & if \quad \hat{t}^{ed} = \hat{t}^{st} \\ \frac{1}{\hat{t}^{ed} - \hat{t}^{st} + 1} \sum_{t=\hat{t}^{st}+1}^{\hat{t}^{ed}} w^{t-1}_t \mathbb{A}^{t-1}_t + max(0, 1 - |Tset_{\hat{c},2}|), & if \quad \hat{t}^{ed} > \hat{t}^{st} \end{cases} \tag{1}$$

3.3 Contrastive Subaction Learning

Contrastive subaction learner aims to separate subaction representations by considering two terms. *First*, intra-action loss ensures that subactions within the same action are distinguishable from one another. For query feature $x_q \in X^{\hat{c}}$, other snippet with its predicted action and subaction $(\hat{c}, y^{sub}_{\hat{c},q})$ is sampled as a positive term $intra^+$, while other subaction snippets in the same ground truth action are sampled as negative terms, indexed by $intra^- \in Intra^-$. *Second*, inter-action loss enforces that the middle subaction (which captures core action semantics) is separable across different actions. For query feature $x_{q'} \in X^{\hat{c}}$ with middle subaction, other snippet with its predicted action and subaction $(\hat{c}, y^{sub}_{\hat{c},q} = 2)$ is sampled as a positive term $inter^+$, while middle subaction snippets from other ground truth actions are sampled as negative terms, indexed by $inter^- \in Inter^-$. The above snippets are sampled in the same batch. To amplify the contrast in feature space, we introduce a temperature parameter τ_{con}. Given the cosine similarity $sim(x, x')$, the contrastive proposal loss is $\mathcal{L}_{con} = \mathcal{L}_{intra} + \mathcal{L}_{inter}$, where the two terms are,

$$\begin{cases} \mathcal{L}_{intra} = -log(\frac{\exp(sim(x^T_q, x_{intra^+})/\tau_{con})}{\exp(sim(x^T_q, x_{intra^+})/\tau_{con}) + \sum_{intra^- \in Intra^-} \exp(sim(x^T_q, x_{intra^-})/\tau_{con})}) \\ \mathcal{L}_{inter} = -log(\frac{\exp(sim(x^T_{q'}, x_{inter^+})/\tau_{con})}{\exp(sim(x^T_{q'}, x_{inter^+})/\tau_{con}) + \sum_{inter^- \in Inter^-} \exp(sim(x^T_{q'}, x_{inter^-})/\tau_{con})}) \end{cases} \tag{2}$$

3.4 Proposal Transition Learner

The proposal transition learner addresses the gradual changes in transition segments. Such changes may make boundary localization difficult. To resolve this, we introduce a transition loss only for non-empty subactions. Specifically, let $\mathbb{I}^{st}$ and $\mathbb{I}^{ed}$ be indicator functions that equal 1 if the predicted start and end subaction segments are non-empty, respectively, and 0 otherwise. This loss explicitly encourages two windows around the boundary to have separated scores. Given the start timestamp $\hat{t}^{st}$ and window size ws, we introduce left/right windows, which can be truncated to the video range. The average prediction scores of the start subaction in these windows are $Ap^L_{\hat{c},\hat{t}^{st}}$ and $Ap^R_{\hat{c},\hat{t}^{st}}$. The start boundary transition can be measured by their difference $Dp^{st}_{\hat{c}} = Ap^R_{\hat{c},\hat{t}^{st}} - Ap^L_{\hat{c},\hat{t}^{st}}$. Unlike start timestamp, end timestamp usually has a reversed transition direction. Given the score of end subaction, the end boundary transition is $Dp^{ed}_{\hat{c}} = Ap^L_{\hat{c},\hat{t}^{ed}} - Ap^R_{\hat{c},\hat{t}^{ed}}$. As similar features across the boundary may produce a small transition, we introduce hinge loss with margin parameter τ_{tr} to encourage the transition to exceed the margin. The proposal transition loss contains the two terms for start and end transitions as,

$$\mathcal{L}_{tr} = \mathbb{I}^{st} \cdot max(0, \tau_{tr} - Dp^{st}_{\hat{c}}) + \mathbb{I}^{ed} \cdot max(0, \tau_{tr} - Dp^{ed}_{\hat{c}}) \tag{3}$$

3.5 PROPOSAL DISTRIBUTION REFINEMENT

Due to the mismatch between background snippets and the video-level label, the boundary points may deviate from their ground truth points. To refine these boundaries, we focus on a localized candidate segment around each initial boundary. The refinement is only performed on non-empty start/end subactions. For the start boundary point $\hat{t}^{st}$, its candidate segment is $Seg_{st} = \{t|\hat{t}^{st} - ws + 1 \leq t \leq \hat{t}^{st} + ws - 1\}$, where ws is the window size. For the snippet $t \in Seg_{st}$, we evaluate its boundary changes using a left window $Left_t = \{t'|t - ws + 1 \leq t' \leq t\}$ and right window $Right_t = \{t'|t \leq t' \leq t + ws - 1\}$. Candidate segments and left/right windows are truncated to the video range, and are updated in multiple steps.

The boundary transition term identifies the point where a large distribution change occurs. As the distribution difference is more stable than the point-level difference, we measure the boundary change between left and right windows using first-order Wasserstein distance. Given start subaction score, the Wasserstein distance is estimated using the Empirical Cumulative Distribution Function of left/right distribution $F^{st,L}_{\hat{c},t}(x), F^{st,R}_{\hat{c},t}(x)$ (x is the random variable sampled from [0,1]). This distance is $W1^{st}_{\hat{c},t} = \int_0^1 |F^{st,L}_{\hat{c},t}(x) - F^{st,R}_{\hat{c},t}(x)|dx$. For the start point, we notice that the right window belongs to the action proposal, while the left window belongs to the background snippets. To encourage the average score of right window $Ap^{st,R}_{\hat{c},t}$ to be larger than that of left window $Ap^{st,L}_{\hat{c},t}$, we quantify the directional difference $Diff^{st} = Ap^{st,R}_{\hat{c},t} - Ap^{st,L}_{\hat{c},t}$. For the start point, the boundary transition term is $Bt^{st}_{\hat{c},t} = max(0, Diff^{st} \cdot W1^{st}_{\hat{c},t})$.

Temporal distance term is used to limit the boundary search in a small temporal distance. For the start point, the temporal distance term considers a decay parameter τ_{dist} to emphasize snippets closer to the boundary as $Temp^{st}_{\hat{c},t} = exp(-|t - \hat{t}^{st}|/\tau_{dist})$.

The boundary refiner aims to shift the boundary towards the region with proper transitions by performing a weighted centroid estimation, using the product of the two above terms as weights. Consequently, points that exhibit a large boundary transition within a small temporal window are assigned higher weights. If all $Bt^{st}_{\hat{c},t}$ are zero, the original start point is retained. Otherwise, start boundary point is refined using the floor function ($\lfloor \cdot \rfloor$) as,

$$\hat{t}^{*,st} = \lfloor \frac{\sum_{t \in Seg^{st}} Bt^{st}_{\hat{c},t} \cdot Temp^{st}_{\hat{c},t} \cdot t}{\sum_{t \in Seg^{st}} Bt^{st}_{\hat{c},t} \cdot Temp^{st}_{\hat{c},t},} + 0.5 \rfloor, \quad if \sum_{t \in Seg^{st}} Bt^{st}_{\hat{c},t} \cdot Temp^{st}_{\hat{c},t} > 0 \tag{4}$$

Similarly, for the end point, we compute both the boundary transition term and the temporal distance term, where directional difference $Diff^{ed} = Ap^{ed,L}_{\hat{c},t} - Ap^{ed,R}_{\hat{c},t}$ is reversed due to the transition direction. The end boundary point is refined as $\hat{t}^{*,ed}$. If $\hat{t}^{*,st} < \hat{t}^{*,ed}$, the proposal is updated as $Q^* = (\hat{t}^{*,st}, \hat{t}^{*,ed}, \hat{c})$. Otherwise, the original proposal is kept for the next step.

3.6 MODEL OPTIMIZATION AND INFERENCE

Algorithm 1 shows our training process, which is a progressive refinement framework with multiple steps. The first step warms up the base model and initializes the boundary points. The remaining steps optimize the subaction segment representation and refine the boundary points progressively. In subaction representation optimization, beyond the base loss, the total loss optimizes the proposal contrastive loss, the proposal structure loss, and the proposal transition loss with their respective hyperparameters λ_{str}, λ_{con}, and λ_{tr}.

$$\mathcal{L}_{total} = \mathcal{L}_{base} + \lambda_{str}\mathcal{L}_{str} + \lambda_{con}\mathcal{L}_{con} + \lambda_{tr}\mathcal{L}_{tr} \tag{5}$$

In inference stage, the raw proposals provide initial action boundary points. Then, the trained LDPR model computes subaction scores across timestamps, which are used to refine the boundary points once using Eq. (4) to form refined proposals. For each refined proposal, we first obtain the maximum predicted subaction score per snippet, and then average these scores within the proposal as the proposal-level score. Non-Maximum Suppression (NMS) is used to select the final proposals from multiple overlapping proposals.

Algorithm 1 Training algorithm for LDPR

Input: Visual features, raw proposals, action names, subaction prompts
Output: Trained LDPR model

```
Initialize subaction tokens from prompt template as text guidance for subaction semantics
Initialize boundary points using raw proposals
for each training epoch in the first step do
    Warm up base model by minimizing the base loss L_base
end for
for the remaining steps do
    for each training epoch do
        Train LDPR model by minimizing the total loss L_total
    end for
    if not final step then
        Use proposal distribution refiner to update boundary points for the next step
    end if
end for
Output the trained LDPR model
```

4 EXPERIMENTS

4.1 DATASETS AND IMPLEMENTATION DETAILS

Datasets. We evaluate our method on two popular action localization datasets: THUMOS'14 (Idrees et al. (2017)) and ActivityNet v1.3 (Heilbron et al. (2015)). Only video-level category labels are used for network training. THUMOS'14 contains untrimmed videos with 20 categories. Following (He et al. (2022)), we use 200 videos for training and 212 videos for testing. ActivityNet v1.3 includes untrimmed videos with 200 categories. Following (He et al. (2022)), we use 10024 videos for training, and 4926 videos for validation.

Feature extraction. We extract RGB features using the I3D (Carreira & Zisserman (2017)) network pre-trained on Kinetics-400 (Carreira & Zisserman (2017)), and extract optical flow features from RGB frames using the TVL1 (Duval et al. (2009)) algorithm. Each video is divided into 16-frame non-overlapping snippets. Each snippet takes a 1024-dimensional RGB feature and a 1024-dimensional optical flow feature (Li et al. (2025)). After a temporal convolution layer, the output of the feature extraction module is a 2048-dimensional vector.

Training details. Following (Li et al. (2025)), each untrimmed video is divided into non-overlapping 16-frame snippets for THUMOS'14 and ActivityNet v1.3. In action proposal extraction, we set the video category threshold to θ_v = 0.2 for THUMOS'14 and θ_v = 0.1 for ActivityNet v1.3. Following (Li et al. (2025)), action proposals are localized using multiple snippet-level thresholds, θ_s = [0: 0.1: 0.9] for THUMOS'14 and θ_s = [0: 0.1: 0.9] for ActivityNet v1.3. The non-maximum suppression is performed with the intersection over union threshold of 0.7. To ensure weak supervision, the raw proposals used in our method are generated by (Yun et al. (2024)), which are obtained offline before training. In contrastive proposal learning, the temperature parameter is $\tau_{con} = 0.03$. The negative terms are sampled from videos in the same batch. For intra-action negatives, one negative sample is drawn from each of the other available subactions, and for inter-action negatives, one negative sample is drawn from each of the other available non-background actions. In proposal transition learning, hinge loss parameter is $\tau_{tr} = 0.15$. In proposal distribution refinement, window size is 4, and distance parameter is $\tau_{dist} = 0.7$. In each optimization step, the base loss decays λ_{sme} from 1 to 0 over 160 epochs, and decays λ_{cs} from 1 to 0 over 70 epochs (Li et al. (2025)). λ_{pc} is set to 3.4. Our losses use $\lambda_{str} = 1$, $\lambda_{con} = 1$, $\lambda_{tr} = 1$. In the training stage, our model is optimized in four steps with Adam at a learning rate of 0.0001. Each step uses 1000 epochs with a batch size of 16 for THUMOS'14, and uses 1000 epochs with a batch size of 64 for ActivityNet v1.3.

Testing metrics. In the inference stage, we use the same video category threshold and multiple snippet-level thresholds as in the training stage. For evaluation, we report the mean Average Precision (AVG) under different Intersection over Union (IoU) thresholds (He et al. (2022)). The IoU thresholds are set as [0.1: 0.1: 0.7] for THUMOS'14 and [0.5: 0.05: 0.95] for ActivityNet v1.3.

mAP@IoU (%)	THUMOS'14								ActivityNet v1.3
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG [0.1:0.1:0.7]	AVG [0.5:0.05:0.95]
ASM-Loc He et al. (2022)	71.2	65.5	57.1	46.8	36.6	25.2	13.4	45.1	25.1
CASE Liu et al. (2023)	72.3	67.1	59.2	49.4	37.7	24.2	13.7	46.2	26.8
Bi-SCC Li et al. (2024b)	71.7	66.9	57.2	48.0	40.4	27.5	14.4	46.6	26.4
ISSF Yun et al. (2024)	72.4	66.9	58.4	49.7	41.8	25.5	12.8	46.8	25.8
APL Luo et al. (2025)	71.7	66.7	58.4	48.5	41.1	27.7	14.4	46.9	26.1
P-MIL Ren et al. (2023)	71.8	67.5	58.9	49.0	40.0	27.1	15.1	47.0	25.5
DDGNet Tang et al. (2023)	72.5	67.7	58.2	49.0	41.4	27.6	14.8	47.3	–
SDANet Zhou et al. (2025)	72.9	67.3	58.6	49.8	40.6	28.0	15.8	47.6	26.6
ActionDiff Zou et al. (2025)	76.7	70.8	61.7	51.2	40.9	28.2	15.9	49.3	26.4
PivoTAL Rizve et al. (2023)	74.1	69.6	61.7	52.1	42.8	30.6	16.7	49.6	28.1
MLLM Zhang et al. (2025a)	74.3	69.8	61.8	52.3	43.0	30.8	16.6	49.8	–
PVLR Lim et al. (2024)	74.9	69.9	61.4	53.1	45.1	30.5	17.1	50.3	27.4
PACL Li et al. (2026)	77.0	72.5	63.1	52.6	43.0	28.9	15.8	50.4	27.6
TVKP Shao et al. (2025)	75.6	70.6	62.5	54.0	44.8	31.1	17.3	50.8	27.2
ASM-Loc+NoCo Zhang et al. (2025b)	75.2	70.7	63.0	54.1	44.0	31.7	17.7	50.9	–
MSLU Shao et al. (2026)	76.1	71.3	62.9	54.9	45.8	31.8	17.7	51.5	27.5
SEAL Li et al. (2025)	78.4	73.6	66.1	55.8	44.5	31.2	19.9	52.8	25.6
Ours	**79.3**	**75.4**	**66.5**	**57.3**	**47.9**	**34.1**	**20.9**	**54.5**	**28.5**

Table 1: Comparison with the state-of-the-art methods on THUMOS'14 and ActivityNet v1.3.

τ_{con}	0.01	0.03	0.05	0.10	τ_{tr}	0.10	0.15	0.20	0.30	τ_{dist}	0.2	0.5	0.7	1
AVG %	54.3	**54.5**	54.3	54.2	AVG %	54.4	**54.5**	54.4	54.3	AVG %	54.3	54.4	**54.5**	54.3

Table 2: Parameters for contrastive learning, temporal distance term, and transition learning.

ws	2	4	6	8	Steps	1	2	3	4	5	Text	CLIP	GPT-2S	GPT-2M	GPT-2L
AVG %	54.3	**54.5**	54.4	54.1	AVG %	52.0	54.1	54.4	**54.5**	**54.5**	AVG %	54.1	54.3	54.4	**54.5**

Table 3: Effect of distribution window size, refinement steps, and text encoder.

Training	Inference	Offset ↓	AVG (%) ↑
✗	✗	1.20	53.7
PDR	✗	1.19	53.8
✗	PDR	1.17	54.0
PDR	PDR	**1.14**	**54.5**

Table 4: PDR effects for training and inference stages of Step 4.

Refine	without	BT	Temp.	BT+Temp.
Offset ↓	1.20	1.22	1.20	**1.14**
AVG % ↑	53.7	53.6	53.7	**54.5**
Transition	Snip Diff	Avg Diff	(Avg Diff)^2	(Avg Diff)$\cdot W1$
Offset ↓	1.17	1.16	1.16	**1.14**
AVG % ↑	53.9	54.1	54.2	**54.5**

Table 5: Boundary refinement via Boundary Transition (BT) and Temporal distance (Temp.) term.

Steps	Base	LPL	CPL	PTL	PDR	Training (final step) Params / FLOPs / Hours	Test (final step) Params / FLOPs / Latency	Boundary Offset ↓	Action AVG (%) ↑
Step 1	✓					18.90 M / 17.21 G / 0.22 h	18.90 M / 4.58 G / 12.68 ms	1.45	52.0
Step 2	✓	✓				20.53 M / 17.74 G / 0.25 h	20.53 M / 5.16 G / 17.54 ms	1.42	52.4
Step 2	✓	✓	✓			20.53 M / 17.74 G / 0.26 h	20.53 M / 5.16 G / 17.54 ms	1.36	52.8
Step 2	✓	✓	✓	✓		20.53 M / 17.74 G / 0.27 h	20.53 M / 5.16 G / 17.54 ms	1.26	53.4
Step 2	✓	✓	✓	✓	✓	20.53 M / 17.74 G / 0.33 h	20.53 M / 5.16 G / 576.386 ms	1.16	54.1
Step 4	✓	✓	✓	✓	✓	20.53 M / 17.74 G / 0.33 h	20.53 M / 5.16 G / 576.386 ms	**1.14**	**54.5**

Table 6: Modules of LLM-guided Proposal Learning (LPL), Contrastive Proposal Learning (CPL), Proposal Transition Learning (PTL), and Proposal Distribution Refinement (PDR).

4.2 Comparison with State-of-the Art Methods

Table 1 shows comparisons on THUMOS'14 and ActivityNet v1.3. Existing methods fall into three categories: pseudo-label enhancement (Zhang et al. (2025b)), proposal-level consistency constraints (Li et al. (2026); Ren et al. (2023); Li et al. (2025)), and multi-modal representation learning (Li et al. (2023); Lim et al. (2024); Shao et al. (2025; 2026); Zhang et al. (2025a)). However, these methods overlook the semantic mismatch between background snippets and video-level labels, leading to improper action scores on backgrounds and blurred boundaries. To address this, our model introduces fine-grained sub-action semantics and refines transitional subaction segments via four modules: an LLM-guided proposal learner, contrastive proposal learner, a proposal transition learner, and a proposal distribution refiner. These modules improve action localization.

PDR	$\mathcal{L}_{base}$	$\mathcal{L}_{\text{str}}$	$\mathcal{L}_{\text{con}}$	$\mathcal{L}_{\text{tr}}$	AVG % ↑
✗	✓				52.0
✓	✓				52.7
✓	✓	✓			52.9
✓	✓		✓		53.1
✓	✓			✓	53.0
✓	✓	✓	✓		53.8
✓	✓	✓		✓	53.5
✓	✓		✓	✓	53.9
✓	✓	✓	✓	✓	**54.5**

Table 7: Loss effects of Step 4.

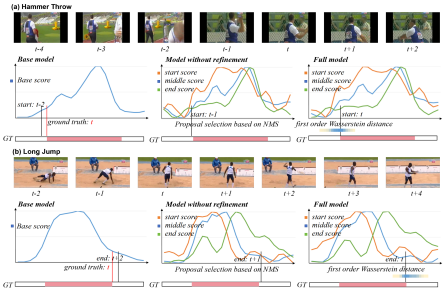

Figure 3: Subaction predictions of our LDPR model.

4.3 Ablation Studies

Ablation studies use two-stream features (RGB/optical flow) on THUMOS'14. (1) Empirical Parameters include the contrastive, distance, and transition parameters (**Table 2**), as well as the window size and steps (**Table 3**). (2) Effect of modules. Among CLIP, GPT-2 Small (GPT-2S), and GPT-2 Medium (GPT-2M), the model with a GPT-2 Large (GPT-2L) text encoder gets the best performance. For our full model, **Table 4** shows the effect of PDR during training and inference. In **Table 5**, given the predicted boundary points $\hat{t}$ and ground truth points t^{gt}, the model measures the *offset* as the average of $|\hat{t} - t^{gt}|$ (measured in snippets) over all boundary points, including start/end boundaries for each class. Only proposals with $IoU \geq 0.1$ are considered (see supplementary materials for *offset* with more IoU thresholds). The offset can be alleviated by boundary refinement. When comparing boundary transition strategies, snippet difference (Snip Diff) may cause large offsets from noisy features, while average snippet difference (Avg Diff) alleviates this. As first-order Wasserstein distance ($W1$) captures shape dissimilarity beyond Avg Diff ($W1 \geq |\text{Avg Diff}|$), the model with (Avg Diff)·$W1$ improves performance. In **Table 6**, starting from the base model (step 1), the following models add LPL, CPL, and PTL in step 2. The full model (the last row) adds PDR in training steps 2 and 3. For final step model in each row, Table 6 also reports training/test FLOPs and parameters (excluding offline text encoder and raw proposals), training hours for final step on the training set, and test latency averaged per video in the testing set. The large test latency of the full model is caused by PDR module, which uses CPU and does not increase FLOPs. (3) Effects of losses ($\mathcal{L}_{base}$, $\mathcal{L}_{str}$, $\mathcal{L}_{con}$, and $\mathcal{L}_{tr}$) are shown in **Table 7**. For the model with $\mathcal{L}_{base}$, the base scores can be used to learn the boundary transition term and to refine the boundary in the PDR module.

4.4 Visualization

Figure 3 visualizes the action prediction of the base model and our subaction prediction. (a) In "Hammer Throw", the base model (*left column*) mispredicts background as foreground (due to weak supervision), causing an early start boundary. *In the middle column*, the model introduces LLM-guided subaction texts and optimizes subaction representations with our total loss (without boundary point refinement). The model enhances proposal scores, and applying NMS to the proposals alleviates semantic mismatch. *In the right column*, our full model measures the boundary transition using the first-order Wasserstein distance between left/right windows at multiple snippets (darker color indicates larger distance). The distance sequence refines the start boundary, such that the action proposal covers the ground-truth action. (b) In "Long Jump", semantic mismatch of background causes a late end boundary in the base model. Our full model estimates the end subaction score and refines the boundary using the first-order Wasserstein distance. After refining the end boundary by considering Wasserstein distance, the end boundary aligns with the ground truth.

5 Conclusion

This work proposes an LLM-guided distribution proposal refinement model, which aims to alleviate inaccurate proposal localization caused by semantic mismatch in weak supervision. Specifically, we design an LLM-guided proposal learner to ensure a proper subaction structure, a contrastive proposal learner to ensure subaction discriminability, a proposal transition learner to encourage separable transitions, and a proposal distribution refiner to adjust the boundary for proper localization.

REFERENCES

João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 4724–4733. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.502. URL https://doi.org/10.1109/CVPR.2017.502.

Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, volume 13664 of *Lecture Notes in Computer Science*, pp. 192–208. Springer, 2022. doi: 10.1007/978-3-031-19772-7_12. URL https://doi.org/10.1007/978-3-031-19772-7_12.

Jia-Run Du, Jia-Chang Feng, Kun-Yu Lin, Fa-Ting Hong, Zhongang Qi, Ying Shan, Jian-Fang Hu, and Wei-Shi Zheng. Weakly-supervised temporal action localization by progressive complementary learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1):938–952, 2025. doi: 10.1109/TCSVT.2024.3456795.

Vincent Duval, Jean-François Aujol, and Yann Gousseau. The TVL1 model: A geometric point of view. *Multiscale Model. Simul.*, 8(1):154–189, 2009. doi: 10.1137/090757083. URL https://doi.org/10.1137/090757083.

Qianhan Feng, Wenshuo Li, Tong Lin, and Xinghao Chen. Full-stage pseudo label quality enhancement for weakly-supervised temporal action localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(11):11144–11157, 2025. doi: 10.1109/TCSVT.2025.3576615.

Junyu Gao, Mengyuan Chen, and Changsheng Xu. Fine-grained temporal contrastive learning for weakly-supervised temporal action localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 19967–19977, 2022. doi: 10.1109/CVPR52688.2022.01937. URL https://doi.org/10.1109/CVPR52688.2022.01937.

Bo He, Xitong Yang, Le Kang, Zhiyu Cheng, Xin Zhou, and Abhinav Shrivastava. Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 13915–13925. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01355. URL https://doi.org/10.1109/CVPR52688.2022.01355.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 961–970. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298698. URL https://doi.org/10.1109/CVPR.2015.7298698.

Fa-Ting Hong, Jia-Chang Feng, Dan Xu, Ying Shan, and Wei-Shi Zheng. Cross-modal consensus network for weakly supervised temporal action localization. In Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (eds.), *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pp. 1591–1599. ACM, 2021. doi: 10.1145/3474085.3475298. URL https://doi.org/10.1145/3474085.3475298.

Linjiang Huang, Liang Wang, and Hongsheng Li. Foreground-action consistency network for weakly supervised temporal action localization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 7982–7991. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00790. URL https://doi.org/10.1109/ICCV48922.2021.00790.

Linjiang Huang, Yan Huang, Wanli Ouyang, and Liang Wang. Two-branch relational prototypical network for weakly supervised temporal action localization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):5729–5746, 2022. doi: 10.1109/TPAMI.2021.3076172. URL https://doi.org/10.1109/TPAMI.2021.3076172.

Haroon Idrees, Amir Roshan Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The THUMOS challenge on action recognition for videos "in the wild". *Comput. Vis. Image Underst.*, 155:1–23, 2017. doi: 10.1016/j.cviu.2016.10.018. URL https://doi.org/10.1016/j.cviu.2016.10.018.

Chen Ju, Kunhao Zheng, Jinxiang Liu, Peisen Zhao, Ya Zhang, Jianlong Chang, Qi Tian, and Yanfeng Wang. Distilling vision-language pre-training to collaborate with weakly-supervised temporal action localization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14751–14762, 2023. doi: 10.1109/CVPR52729.2023.01417.

Ao Li, Huijun Liu, Jinrong Sheng, Zhongming Chen, and Yongxin Ge. Efficient dual-confounding eliminating for weakly-supervised temporal action localization. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (eds.), *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pp. 8179–8188. ACM, 2024a. doi: 10.1145/3664647.3681571. URL https://doi.org/10.1145/3664647.3681571.

Guozhang Li, De Cheng, Xinpeng Ding, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. Boosting weakly-supervised temporal action localization with text information. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10648–10657, 2023. doi: 10.1109/CVPR52729.2023.01026.

Guozhang Li, De Cheng, Xinpeng Ding, Nannan Wang, Jie Li, and Xinbo Gao. Weakly supervised temporal action localization with bidirectional semantic consistency constraint. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):13032–13045, 2024b. doi: 10.1109/TNNLS.2023.3266062.

Jingjing Li, Tianyu Yang, Wei Ji, Jue Wang, and Li Cheng. Exploring denoised cross-video contrast for weakly-supervised temporal action localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 19882–19892, 2022. doi: 10.1109/CVPR52688.2022.01929. URL https://doi.org/10.1109/CVPR52688.2022.01929.

Maodong Li, Chao Zheng, Jian Wang, and Bing Li. Similar modality enhancement and action consistency learning for weakly supervised temporal action localization. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 4842–4850. AAAI Press, 2025. doi: 10.1609/AAAI.V39I5.32512. URL https://doi.org/10.1609/aaai.v39i5.32512.

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*, 35:5894–5909, 2026. doi: 10.1109/TIP.2026.3697621.

Geuntaek Lim, Hyunwoo Kim, Joonsoo Kim, and Yukyung Choi. Probabilistic vision-language representation for weakly supervised temporal action localization. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (eds.), *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pp. 5507–5516. ACM, 2024. doi: 10.1145/3664647.3681537. URL https://doi.org/10.1145/3664647.3681537.

Qinying Liu, Zilei Wang, Shenghai Rong, Junjie Li, and Yixin Zhang. Revisiting foreground and background separation in weakly-supervised temporal action localization: A clustering-based approach. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10399–10409, 2023. doi: 10.1109/ICCV51070.2023.00957.

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)*, pp. 8711–8720, 2025. doi: 10.1109/CVPR52734.2025.00814.

Wang Luo, Tianzhu Zhang, Wenfei Yang, Jingen Liu, Tao Mei, Feng Wu, and Yongdong Zhang. Action unit memory network for weakly supervised temporal action localization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 9969–9979. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00984. URL https://openaccess.thecvf.com/content/CVPR2021/html/Luo_Action_Unit_Memory_Network_for_Weakly_Supervised_Temporal_Action_Localization_CVPR_2021_paper.html.

Wang Luo, Huan Ren, Tianzhu Zhang, Wenfei Yang, and Yongdong Zhang. Adaptive prototype learning for weakly-supervised temporal action localization. *IEEE Transactions on Image Processing*, 34:3154–3168, 2025. doi: 10.1109/TIP.2024.3431915.

OpenAI. Openai GPT-5 system card. *CoRR*, abs/2601.03267, 2026. doi: 10.48550/ARXIV.2601.03267. URL https://doi.org/10.48550/arXiv.2601.03267.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI Blog, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Technical Report.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Huan Ren, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. Proposal-based multiple instance learning for weakly-supervised temporal action localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 2394–2404. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00237. URL https://doi.org/10.1109/CVPR52729.2023.00237.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 22992–23002. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02202. URL https://doi.org/10.1109/CVPR52729.2023.02202.

Yuxiang Shao, Feifei Zhang, and Changsheng Xu. Text-video knowledge guided prompting for weakly supervised temporal action localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(5):4184–4197, 2025. doi: 10.1109/TCSVT.2024.3521125.

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*, 36(6):8156–8168, 2026. doi: 10.1109/TCSVT.2026.3660369.

Xiaojun Tang, Junsong Fan, Chuanchen Luo, Zhaoxiang Zhang, Man Zhang, and Zongyuan Yang. Ddg-net: Discriminability-driven graph network for weakly-supervised temporal action localization. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6599–6609, 2023. doi: 10.1109/ICCV51070.2023.00609.

Guiqin Wang, Peng Zhao, Cong Zhao, Shusen Yang, Jie Cheng, Luziwei Leng, Jianxing Liao, and Qinghai Guo. Weakly-supervised action localization by hierarchically-structured latent attention modeling. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10169–10179, 2023a. doi: 10.1109/ICCV51070.2023.00936.

Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6402–6411. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.678. URL https://doi.org/10.1109/CVPR.2017.678.

Yu Wang, Yadong Li, and Hongbin Wang. Two-stream networks for weakly-supervised temporal action localization with semantic-aware mechanisms. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18878–18887, 2023b. doi: 10.1109/CVPR52729.2023.01810.

Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Qi Tian, Yongdong Zhang, and Feng Wu. Uncertainty guided collaborative training for weakly supervised temporal action detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 53–63. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00012. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yang_Uncertainty_Guided_Collaborative_Training_for_Weakly_Supervised_Temporal_Action_Detection_CVPR_2021_paper.html.

Wulian Yun, Mengshi Qi, Chuanming Wang, and Huadong Ma. Weakly-supervised temporal action localization by inferring salient snippet-feature. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 6908–6916. AAAI Press, 2024. doi: 10.1609/AAAI.V38I7.28516. URL https://doi.org/10.1609/aaai.v38i7.28516.

Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 16010–16019. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01575. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_CoLA_Weakly-Supervised_Temporal_Action_Localization_With_Snippet_Contrastive_Learning_CVPR_2021_paper.html.

Quan Zhang, Jinwei Fang, Rui Yuan, Xi Tang, Yuxin Qi, Ke Zhang, and Chun Yuan. Weakly supervised temporal action localization via dual-prior collaborative learning guided by multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24139–24148, 2025a. doi: 10.1109/CVPR52734.2025.02248.

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 Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 10085–10093. AAAI Press, 2025b. doi: 10.1609/AAAI.V39I10.33094. URL https://doi.org/10.1609/aaai.v39i10.33094.

Jingqiu Zhou, Linjiang Huang, Liang Wang, Si Liu, and Hongsheng Li. Improving weakly supervised temporal action localization by bridging train-test gap in pseudo labels. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23003–23012, 2023. doi: 10.1109/CVPR52729.2023.02203.

Wei Zhou, Kang Lin, Weipeng Hu, Chao Xie, Tao Su, Haifeng Hu, and Yap-Peng Tan. Snippet-inter difference attention network for weakly-supervised temporal action localization. *IEEE Transactions on Multimedia*, 27:3610–3624, 2025. doi: 10.1109/TMM.2025.3535336.

Yuanbing Zou, Qingjie Zhao, Prodip Kumar Sarker, Le Yang, and Binglu Wang. Action-to-action diffusion network for weakly supervised temporal action localization. *IEEE Transactions on Multimedia*, 27:9371–9384, 2025. doi: 10.1109/TMM.2025.3613174.