SEMANTIC FREQUENCY BASED QUERY RELATIONS FOR TEMPORAL ACTION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurately identifying the boundaries of actions is crucial for temporal action detection (TAD), but this task remains challenging as sudden visual changes do not always indicate action boundaries. Existing frequency-domain TAD methods can capture the global temporal changes in videos, but they ignore that some of these changes do not necessarily correspond to action boundaries but are caused by background changes, thereby affecting the accuracy of action boundary localization. During the decoding stage, existing query-based detectors update queries based on the temporal features of the video, but a large number of queries corresponding to different action boundaries may temporally overlap or cover different parts of the same action instance, resulting in redundancy or fragmentation between the predicted action segments. To solve this problem, we propose Semantic Frequency based Query Relations Model for Temporal Action Detection (SQRAct). Specifically, we first design a global frequency-domain semantic filter in the encoder, which dynamically generates a filter bank based on the action and background semantic information of the video, enhancing the ability to distinguish action boundaries through semantically guided frequency filtering. Moreover, considering the redundancy and fragmentation of different queries for the same action instance, we present an inter-layer query relations refinement (QRR) module in the decoder. The QRR module analyzes the containment or adjacency relationships between queries with similar features and further refines the features and temporal boundaries of the queries. SQRAct simultaneously models both the global frequency-domain semantic features and the explicit query relationships to enhance the ability to distinguish action boundaries and refine redundant and fragmented action predictions, thereby significantly improving the accuracy of action boundary localization. Extensive experiments on THUMOS14, HACS-Segment, and ActivityNet v1.3 show that the proposed SQRAct model outperforms the existing state-of-the-art methods, which validates that enhancing boundary distinguishability and explicitly modeling instance-level query relationships is crucial for TAD.

1 INTRODUCTION

Temporal Action Detection (TAD) aims to identify action instances in untrimmed videos and locate their temporal boundaries. By associating action labels with specific time intervals, TAD can support various applications such as video indexing, event retrieval, and highlight generation. However, accurately localizing temporal boundaries is challenging due to significant variations in action durations and the fact that similar local motions may belong to different actions. For example, a running clip could represent an independent running action, or it could merely be the starting phase before the long jump. Therefore, recognizing action categories and their complete temporal spans requires broader temporal context. Existing TAD methods can be broadly categorized into proposal-based approaches (Zhao et al., 2017; Lin et al., 2019; Qing et al., 2021), dense prediction-based methods (Zhang et al., 2022; Shi et al., 2023; Yang et al., 2024b), and query-based methods (Liu et al., 2022b; Kim et al., 2024; 2025a). The proposal and dense prediction based approaches identify action instances by scoring candidate segments or temporal locations and estimating their boundaries. In contrast, the query-based detectors directly model TAD as set prediction using an encoder-decoder

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.

architecture. Through attention mechanism, each query can flexibly focus on relevant video features and capture their dependencies over long temporal spans.

Frequency-domain modeling can capture the global temporal dynamic changes in the video that span the entire time process. TFFormer (Yang et al., 2024a) employed discrete wavelet transform to extract multi-scale frequency-domain features and used cross-domain attention to exchange information between the time and frequency domains. FDDet (Zhu et al., 2025) separated and reweighted low-frequency and high-frequency components to suppress background redundancy in pretrained features while preserving key details related to the action. However, during the movement, both human poses and scene backgrounds may undergo abrupt changes. These changes do not necessarily indicate the start or end of an action but can interfere with boundary prediction. Global filters used in existing models have fixed cutoff frequencies and scaling parameters, and cannot dynamically adjust their frequency responses according to the motion dynamics and background changes of different videos. It results in weakening genuine boundary cues while amplifying irrelevant variations. For instance, in a diving videos, the athlete rapidly transition from a tucked position to an extended position during the same action, and the camera' s panning might cause the swimming pool and the audience to move within the frame. These changes in scenarios will lead to obvious feature changes, yet they do not signify the termination of the current action or the initiation of a new action. Nevertheless, the detectors may misidentify internal changes within an action or background changes as action boundariea, thereby leading to incorrect boundary predictions.

To accommodate large variations in action duration, the query-based TAD methods typically construct multi-scale temporal features, adaptively selecting initial queries from encoded features (Kim et al., 2024), and iteratively updating action queries via self-attention and cross-attention mechanisms in the decoder. These mechanisms facilitate interactions among queries and between queries and video features. ReAct (Shi et al., 2022) leveraged relational attention, overlap regularization, and quality estimation to distinguish duplicate queries from those associated with different action instances. However, existing methods rely solely on feature similarity and metrics such as temporal intersection over union (tIoU) to filter duplicate queries. Different queries in TAD may cover distinct time segments of the same action, but the filtering process cannot correct their boundaries, thus leading to redundant or fragmented predictions. We observe that query pairs can be categorized based on both feature similarity and temporal relationships, as illustrated in Figure 1. When two highly similar queries form a containment relationship, they may represent redundant predictions of the same action instance, as show in Figure 1(a). The more confident query provides a reference for refining another query's features and boundaries. Conversely, low feature similarity suggests that temporally containment predictions may correspond to different actions and should remain separate, as illustrated in Figure 1(b). For adjacent predictions, high similarity supports boundary expansion toward the neighboring segment to alleviate potential fragmentation, whereas low similarity argues against such interaction, as shown in Figures 1(c) and 1(d), respectively. These observations motivate jointly considering feature similarity and temporal relationships when determining whether and how to refine queries.

Based on the above analysis, we propose a semantic frequency based query relation model for temporal action detection (SQRAct). To enhance the boundary discriminability of action features, we design a multi-granularity temporal aggregation module (MTAM). This module's global frequency-domain branch dynamically generates a filter bank guided by action and background semantics, enhancing the frequency responses related to the action while suppressing background interference. Subsequently, we build a spatial feature enhancement module to facilitate information transmission between multiscale features, integrating fine-grained boundary details with high-level semantics to alleviate boundary information loss during progressive downsampling. To address the redundancy and fragmentation issues arising from multiple queries covering partial temporal regions of the same action instance, we design a layer-wise query relation refinement module (QRR). QRR explicitly models the containment and adjacency relationships between query prediction segments, and leverages these relationships to refine query features and boundaries. By jointly performing semantics-aware frequency-domain feature enhancement, cross-scale feature interaction, and explicit query relation refinement, SQRAct significantly improves performance of TAD. Our main contributions are as follows:

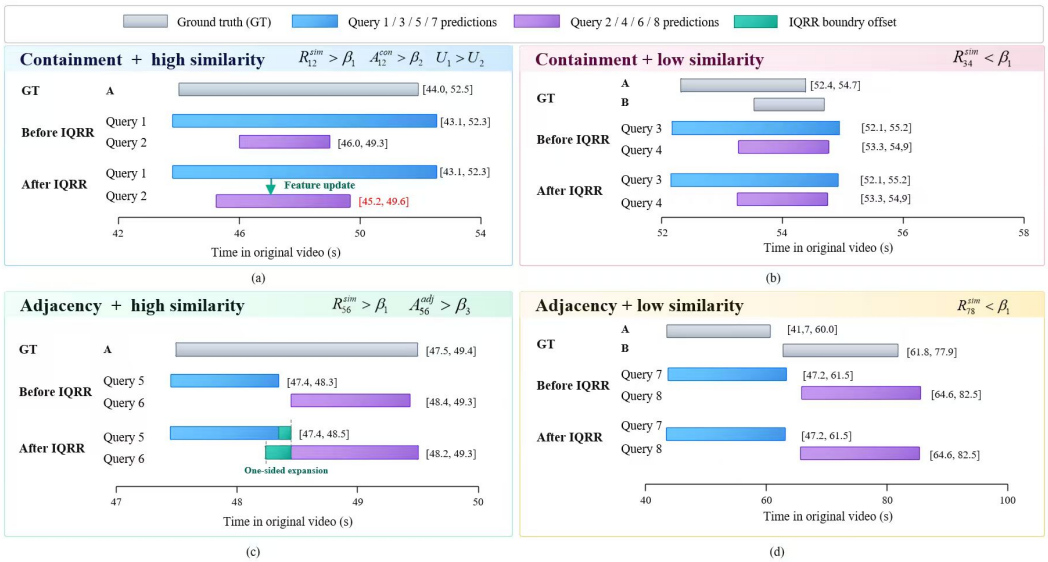

Figure 1: Illustration of query refinement based on feature similarity and temporal relationships. Gray bars indicate two different ground-truth (GT) A and B.

- We propose SQRAct, which improves the accuracy of temporal action boundary localization by integrating semantic frequency-domain feature enhancement, cross-level pyramid feature interaction, and explicit query relationship modeling.
- We design the MTAM module to suppress non-boundary temporal variations and strengthen boundary discriminability. This module consists of three branches: global semantic frequency-domain filtering, local temporal offset correction, and temporal feature enhancement, which respectively capture global temporal patterns through frequency-aware perception, adaptively refine local boundary details, and generate discriminative channel-temporal representations.
- We introduce the QRR module to address the issue of instance-level prediction repetition in queries. This module models the containment and adjacency relationships among queries and optimizes their features and boundaries based on the types of relationships.
- The experiment results of SQRAct in THUMOS14, ActivityNet v1.3, and HACS-Segments datasets show that SQRAct surpasses most advanced models.

2 Related Work

2.1 Query-based End-to-End Action Detection

DETR (Carion et al., 2020) first formulated object detection as a set prediction problem, enabling end-to-end detection via learnable queries and Hungarian matching. Subsequently, Deformable DETR (Zhu et al., 2021) introduced a sparse deformable attention mechanism, significantly improving model convergence speed and computational efficiency. Inspired by these advances, query-based end-to-end detection methods have been extended to TAD. TadTR (Liu et al., 2022b) achieved end-to-end action detection using temporal deformable attention. Self-DETR (Kim et al., 2023) identified the collapse issue in decoder self-attention during temporal detection and restores attention diversity using encoder cross-attention maps. Pred-DETR (Kim et al., 2025b) further alleviated collapse and shortcut behaviors in cross-attention by aligning attention with prediction results via intersection-over-union (IoU) relations. TE-TAD (Kim et al., 2024) proposed a time-aligned coordinate representation and dynamically adjusts the number of queries according to video duration. DiGIT (Kim et al., 2025a) designed a multi-dilated gated encoder and a center-neighborhood region fusion decoder. However, significant internal action variations and background motion in complex and long videos still affect the boundary prediction of action.

2.2 Frequency-Domain Modeling

According to the convolution theorem, point-wise multiplication in the frequency domain corresponds to circular convolution in the time domain. Therefore, the frequency-domain filtering enables global interactions across entire video sequences and has been widely applied to efficient feature aggregation and long-range dependency modeling (Rao et al., 2021; Guibas et al., 2022; Huang et al., 2023; Yi et al., 2024). In TAD, LPF (Jin et al., 2021) applied low-pass filters before temporal downsampling and dynamically adjusts the cutoff frequency for each input to reduce aliasing while preserving discriminative high-frequency information. TFFormer (Yang et al., 2024a) employed discrete wavelet transform to build a dual-branch network that jointly modeled multi-scale temporal and frequency information. FDDet (Zhu et al., 2025) leverageed discrete Fourier transform to separate high-frequency and low-frequency features, suppressing background redundancy while enhancing action-related details. Recently, W-StrataFormer (Chen et al., 2026) combined wavelet decomposition with hierarchical feature enhancement to improve action representations at different temporal scales. These studies demonstrated that the frequency information is of great significance in the action representation and boundary localization. However, most existing methods rely on predefined wavelet bases or frequency band divisions. Even adaptive filtering only globally adjusts the frequency content of the video representation, without considering the scene variations and action changes within individual videos.

2.3 Query Relationship Modeling

The query-based TAD methods predict actions through a large number of queries. Through information exchange and competition during prediction, multiple queries may predict different temporal segments of the same action instance, thereby affecting the final detection result. Therefore, modeling relationships between queries is crucial for reducing duplicate predictions and improving action detection accuracy. In the image object detection task, relation networks (Hu et al., 2018) improved proposal representations by incorporating appearance and geometric relations between candidate instances. DAB-DETR (Liu et al., 2022a) represented queries as iteratively updated anchor boxes and introduceed explicit spatial position priors during decoding. EASE-DETR (Gao et al., 2024) identified leading-lagging relationships between queries based on intermediate predictions to alleviate competition over the same target. In TAD, ReAct (Shi et al., 2022) explicitly distinguished between similar instance queries, dissimilar instance queries, and duplicate queries, and addressed the issue of query redundancy through relational attention, IoU decay, and segment quality prediction. Self-Feedback DETR (Kim et al., 2023) supervised self-attention using a relationship graph generated via cross-attention to mitigate attention collapse. Pred-DETR (Kim et al., 2025b) constructed a relationship graph based on IoU between predicted segments and aligns self-attention and cross-attention with predicted relationships via auxiliary objectives. Although these methods improve query prediction performance through query competition, attention calibration, or quality estimation, most of them achieve this by suppressing or selecting queries predicting the same instance, or by regularizing attention using the predicted relationships. This selection mechanism treats related queries as mutually exclusive candidates, potentially discarding complementary action information and boundary cues. For the longer actions, different queries may locate distinct parts of the same instance. In such cases, overlapping or adjacent local segments may fail to form consistent boundaries or complete instances, resulting in redundant or fragmented predictions that degrade the accuracy of final action boundary localization.

3 Method

3.1 Overview

Figure 2 presents the framework of SQRAct, a query-based encoder–decoder model. Given an untrimmed video, we first use a frozen pretrained video backbone to extract a sequence of snippet-level feature $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T} \in \mathbb{R}^{T \times C}$, where T is the temporal length, $\mathbf{x}_t \in \mathbb{R}^{C}$ is the feature vector of snippet t, and C is the feature dimension. The sequence is processed by N encoder layers, each of which contains an MTAM and gradually downsamples to generate a multiscale feature pyramid. To alleviate information loss caused by the commonly used downsampling operations during pyramid construction, we propose a spatial feature enhancement module (SFE), composed of cross-scale fea-

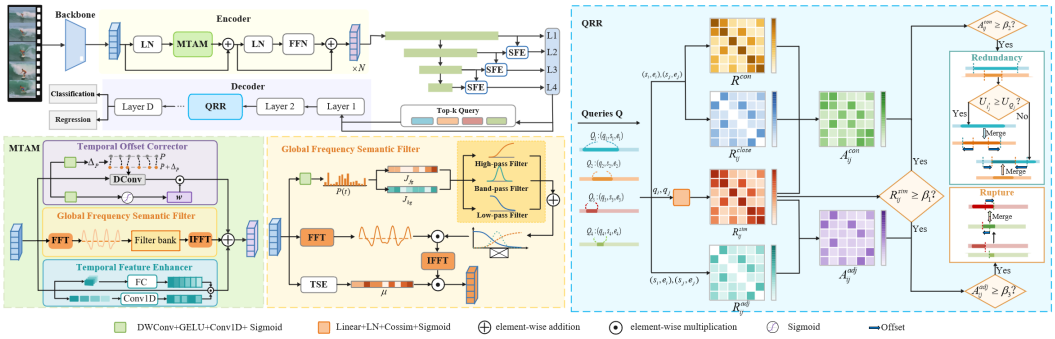

Figure 2: Framework of SQRAct.

ture alignment, adaptive weight estimation, and residual feature fusion. This module propagates the high-resolution feature information layer by layer into the low-resolution features, effectively mitigating feature degradation due to chained propagation and enhancing the discriminative capability of pyramid features. Subsequently, the enhanced pyramid features are filtered through an adaptive Top-K selection to generate initial queries, with each query consisting of a feature vector and temporal coordinates. In the decoder, SQRAct adopts the Center-Adjacent Region Integration Decoder (CAID) proposed by DiGIT (Kim et al., 2025a) as the basic decoding layer. Furthermore, we introduce an inter-layer query relation refinement module (QRR) inserted between decoding layers, which leverages current queries to guide the subsequent updates of query features and temporal coordinates. Finally, the classification and boundary regression heads are used to output the category of each action as well as the start and end times, respectively.

3.2 Multigranularity Temporal Aggregation Module

The goal of the encoder is to effectively model long-range temporal context while preserving fine-grained action boundary details, thereby providing discriminative temporal representations for subsequent prediction. However, relying solely on local convolutions or global modeling mechanisms is insufficient to capture temporal information at different granularities. Moreover, the importance of action features varies significantly across different channels and temporal positions, making it difficult for a single feature representation to adequately highlight the key action regions. To address these challenges, we propose Multi-Granularity Temporal Aggregation Module (MTAM), illustrated in Figure 2. MTAM consists of three parallel branches: temporal offset corrector (TOC), global frequency-spatial filter (GFSF), and temporal feature enhancer (TFE).

Inspired by deformable convolution (Zhu et al., 2019), TOC learns content-adaptive sampling offsets and modulation weights along the temporal dimension to aggregate local temporal context.. Specifically, TOC employs two lightweight convolutional branches to predict sampling offsets $\Delta p_{t,k}$ and modulation weights $w_{t,k}$ for each time position. Then, linear interpolation sampling is performed at the offset positions $p_{t,k} + \Delta p_{t,k}$ of the continuous time, and the sampled features are aggregated using learnable depthwise kernels (DWConv) to obtain local temporal features. GFSF applies semantic filtering over the entire sequence of video segment features to derive global action features. TFE computes channel-wise and temporal-wise attention weights separately from the input features, element-wise multiplies them with the input features to emphasize important channels and critical moments, and produces enhanced temporal features. Subsequently, the output features obtained by the TOC, GFSF, and TFE branches are concatenated along the channel dimension, then the concatenated feature sequentially passed through one-dimensional convolution with kernel size 1, a GELU activation function, and another one-dimensional convolution with kernel size 1, followed by a Sigmoid function to generate the per-timestep weights α_r shared across all three branch channels. Finally, the features of each branch are weighted and fused according to the corresponding weights to generate the output feature of MTAM.

3.3 GLOBAL FREQUENCY-DOMAIN SEMANTIC FILTER

Different videos vary in action duration, boundary patterns, and background complexity, making it difficult for the fixed frequency-domain filters to cover such diverse scenarios.To mitigate the impact of complex environments on action-related feature extraction, we propose a global frequency semantic filter (GFSF), illustrated in Figure 2. GFSF dynamically generates global frequency-domain filters based on the foreground and background semantic information of video, enabling each video to adopt a content-adaptive frequency enhancement strategy. First, the input temporal features are transformed into the frequency domain using the fast Fourier transform (FFT) to model global temporal patterns $\hat{X} = \mathcal{F}_t(X)$. In parallel, an action prediction head in a separate branch is utilized to estimate the foreground and background semantic information of video. These semantic representations are then used to dynamically generate three types of global frequency-domain filters: low-pass, high-pass, and band-pass filters. The generated filters perform semantic filtering on the frequency-domain features, which are subsequently transformed back into the time domain to obtain an enhanced representation containing global semantic information. Meanwhile, a Temporal Semantic Enhancement (TSE) module (Hu et al., 2018) generates the channel-wise weights. Finally, these weights are multiplied element-wise with the enhanced temporal features to further strengthen the representation of action segments.

Action Semantic Modeling The action head comprises a depthwise separable 1-D convolution, GELU, a 1-D convolution, and a sigmoid function. For $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$, it predicts an action score $P(t) \in (0, 1)$ at each temporal position. The foreground and background semantic descriptions are obtained by weighting the action scores of the input features:

$$\begin{aligned} J_{\text{fg}} &= \text{FC}_{\text{fg}}\left(\frac{\sum_{t=1}^{T} P(t)x_t}{\sum_{t=1}^{T} P(t)}\right), \\ J_{\text{bg}} &= \text{FC}_{\text{bg}}\left(\frac{\sum_{t=1}^{T}(1-P(t))x_t}{\sum_{t=1}^{T}(1-P(t))}\right). \end{aligned} \tag{1}$$

The foreground descriptor $\mathbf{J}_{\text{fg}}$ summarizes action regions, whereas the background descriptor $\mathbf{J}_{\text{bg}}$ captures non-action regions. They jointly provide semantic guidance for the generation of subsequent dynamic filters.

Filter Generation We use the background descriptor $\mathbf{J}_{\text{bg}}$ and the foreground descriptor $\mathbf{J}_{\text{fg}}$ to generate the low-frequency response and the high-frequency response, respectiely. At the same time, the both descriptors are utilized to obtain the response centered around the mid-frequency band:

$$\begin{aligned} H_l(f) &= \sigma\Big(\gamma_l\left(\sigma(\text{MLP}_l([J_{\text{bg}}; E_{\text{mid}}])) - \widehat{f}\right)\Big), \\ H_h(f) &= 1 + 2\sigma\Big(\gamma_h\left(\widehat{f} - \sigma(\text{MLP}_h([J_{\text{fg}}; E_{\text{mid}}]))\right)\Big), \\ H_b(f) &= 1 + 2\sigma\left(\gamma_b \exp\left(-\frac{\left(\widehat{f} - \sigma(\text{MLP}_b([J_{\text{fg}}; J_{\text{bg}}; E_{\text{mid}}]))\right)^2}{2\rho^2}\right)\right). \end{aligned} \tag{2}$$

Where $\widehat{f} = f/F$ is the normalized frequency, f and F are the current and maximum frequency indices, respectively. σ is the sigmoid function, ρ determines the band width, γ_l, γ_h, and γ_b control the response shapes, and $\mathbf{E}_{\text{mid}}$ represents the midpoint of the frequency axis. The complete parameterization is provided in the Appendix.

Since the relative contributions of the low-pass, high-pass, and band-pass responses vary with video content,, GFSF further employs an MLP to predict the importance weights of each filter, combining the three responses into a final global frequency-domain filter $H(f)$, and then transforming the filtered spectrum back to the time domain to obtain feature Y. Feature Y is subsequently weighted by channel weights generated from the Temporal Semantic Enhancement (TSE) module, producing the output features of the GFSF branch. The GFSF dynamically adjusts the response strength of different frequency bands according to the specific content of each video, reducing the interference caused by the complex backgrounds and enhancing the discriminative ability of action representations.

3.4 INTER-LAYER QUERY RELATION REFINEMENT

Most existing methods implicitly model the interactions among queries through self-attention, without exploring the temporal relations between query prediction segments. Some queries may detect the same action instance and yield redundant predictions. Two temporally adjacent queries might also cover the beginning and end segments of a long-duration action separately, resulting in fragmented action predictions. Therefore, we introduce the plug-in QRR module. First, for any pair of queries Q_i and Q_j, their semantic similarity R_{ij}^{sim}, containment R_{ij}^{con}, center proximity R_{ij}^{close}, and adjacency R_{ij}^{adj} are computed resepectively. Subsequently, the containment scores A_{ij}^{con} and the adjacency scores A_{ij}^{adj} are calculated based on the semantic information and the temporal relations, and different refinement strategies are adopted for the two cases.

Containment refinement We compute the containment score of Q_i and Q_j, and select the best query with inclusion relationships for each query:

$$\begin{aligned} A_{ij}^{\mathrm{con}} &= R_{ij}^{\mathrm{con}} R_{ij}^{\mathrm{close}} R_{ij}^{\mathrm{sim}} \, \sigma\left(\frac{R_{ij}^{\mathrm{con}} - \tau_{\mathrm{con}}}{0.05}\right), \\ i_j^* &= \arg\max_i A_{ij}^{\mathrm{con}}. \end{aligned} \tag{3}$$

Where τ_{con} is a predefined threshold to reduce the contribution of weaker inclusion relationships. For each Q_j, we select the candidate Q_i with the highest containment score A_{ij}^{con}, denote its index by i_j^*. When its classification confidence is higher $U_{i_j^*} \geq U_{Q_j}$, $Q_{i_j^*}$ guides the feature and boundary updates of Q_j. An MLP receives $\left[\mathbf{q}_{i_j^*}; \mathbf{q}_j\right]$ and predicts a channel-wise fusion weight $\mathbf{G}_j$. Subsequently, feature transformation is performed through linear mapping, Dropout, residual connections, and LayerNorm (LN) to obtain the refined query representation $\widehat{q_j}$:

$$\Delta q_j^{\mathrm{con}} = A_{i_j^*,j}^{\mathrm{con}} G_j \odot (q_{i_j^*} - q_j), \tag{4}$$

$$\widehat{q_j} = \mathrm{LN}\big(q_j + \mathrm{Dropout}(\mathrm{Linear}(\Delta q_j^{\mathrm{con}}))\big). \tag{5}$$

The reference $Q_{i_j^*}$ also guides containment-based boundary refinement of Q_j. We concatenate their features and use a shared MLP and Sigmoid function to predict boundary offset weights δ_j^s and δ_j^e. The start and end boundaries are updated separately as follows::

$$\begin{aligned} \Delta s_j^{\mathrm{con}} &= A_{i_j^*,j}^{\mathrm{con}} \delta_j^s \, \mathrm{clip}\Big(s_{i_j^*} - s_j, -r_{\mathrm{con}} w_j, r_{\mathrm{con}} w_j\Big), \\ \Delta e_j^{\mathrm{con}} &= A_{i_j^*,j}^{\mathrm{con}} \delta_j^e \, \mathrm{clip}\Big(e_{i_j^*} - e_j, -r_{\mathrm{con}} w_j, r_{\mathrm{con}} w_j\Big). \end{aligned} \tag{6}$$

Where r_{con} is used to constrain the offset of the boundary Q_j, allowing the two endpoints to move toward the corresponding endpoints of the reference fragment with controllable magnitude. The center and width are updated accordingly:

$$\begin{aligned} c_j' &= c_j + \frac{\Delta s_j^{\mathrm{con}} + \Delta e_j^{\mathrm{con}}}{2}, \\ w_j' &= w_j + \Delta e_j^{\mathrm{con}} - \Delta s_j^{\mathrm{con}}. \end{aligned} \tag{7}$$

If Q_j has the higher classification confidence($U_{i_j^*} < U_{Q_j}$), the refinement direction is reversed. The refined queries are passed to the subsequent decoder layer. The feature fusion among potentially redundant queries can make their representations more consistent, allowing subsequent layers to produce better-aligned predictions.

Adjacency refinement We multiply the pairwise adjacency coefficient R_{ij}^{adj} by semantic similarity R_{ij}^{sim} to obtain the adjacency score:

$$
\begin{aligned}
A_{ij}^{\mathrm{adj}} &= R_{ij}^{\mathrm{adj}} R_{ij}^{\mathrm{sim}}, \\
j^{+} &= \arg\max_{i} A_{ij}^{\mathrm{adj}}, \qquad j^{-} = \arg\max_{i} A_{ji}^{\mathrm{adj}}, \\
a_j^{-} &= A_{j^{+},j}^{\mathrm{adj}}, \qquad a_j^{+} = A_{j,j^{-}}^{\mathrm{adj}}.
\end{aligned} \tag{8}
$$

For Q_j, the selected predecessor j^- and successor j^+ are the highest-scoring eligible neighbors ($A_{ij}^{adj}>\beta_3$) in their respective temporal directions, with scores a_j^- and a_j^+. A shared MLP and sigmoid produce adaptive weights λ_j^- and λ_j^+ from the concatenated features of the current query and its selected neighbors. The query aggregates context from both directions and is updated using the same transformation way as the containment refinement. The aggregated feature is given by:

$$
\Delta q_j^{\mathrm{adj}} = a_j^{-} \lambda_j^{-} \odot q_{j^{+}} + a_j^{+} \lambda_j^{+} \odot q_{j^{-}}. \tag{9}
$$

The nonnegative expansion ratios δ_j^- and δ_j^+ from the current query and its selected neighbors are predicted by MLP and Sigmoid function. The start boundary expands toward the predecessor and the end boundary towards the successor. The corresponding center and width updates are:

$$
\begin{aligned}
c_j' &= c_j + \frac{r_{\max} w_j}{2} \left(a_j^{+} \delta_j^{+} - a_j^{-} \delta_j^{-} \right), \\
w_j' &= w_j \left[1 + r_{\max} \left(a_j^{-} \delta_j^{-} + a_j^{+} \delta_j^{+} \right) \right].
\end{aligned} \tag{10}
$$

Where $r_{\max}$ controls the expansion scale. Expanding either endpoint leaves the opposite endpoint unchanged. The refined query feature and coordinates $\left(c_j', log w_j'\right)$ are passed to the next decoder layer for further classification and boundary regression.

Temporally nested or consecutive segments can belong to different actions when their query features are semantically dissimilar. Therefore, we exclude pairs of queries with the low-similarity from the feature fusion and adjacency-based boundary refinement to avoid confusing different actions or blurring their boundaries.

4 Experiment

4.1 Comparison with State-of-the-Art Methods

To evaluate the overall performance of SQRAct, we compare it on THUMOS14(Idrees et al., 2017), ActivityNet v1.3(Heilbron et al., 2015), and HACS-Segments(Zhao et al., 2019) with three representative categories of temporal action detection (TAD) methods: anchor-based or anchor-free dense detectors, boundary-based methods, and query-based end-to-end methods. The results are summarized in Tables 1. The results of the HACS-Segments dataset are shown in the appendix.

THUMOS14 As shown in Table 1,SQRAct achieves an average mAP of 74.7% on the THUMOS14 dataset, outperforming all compared methods. Compared to dense prediction approaches such as ActionMamba, ActionFormer, and ViT-TAD, SQRAct demonstrates higher overall detection accuracy, surpassing the second-best method DiGIT by 0.9 percentage points. THUMOS14 contains action instances with highly variable durations and dense distributions, along with pronounced background variation and complex action transitions. SQRAct jointly models global frequency patterns, cross-scale boundary information, and temporal relationships among queries, contributing to reduce interference from background variations on boundary localization and alleviate redundant or fragmented predictions caused by different queries for the same action instance. These results demonstrate that our model achieves favorable action detection performance under complex temporal scenarios.

ActivityNet v1.3 As shown in Table 1, SQRAct achieves an average mAP of 41.5% on ActivityNet v1.3, improving by 0.3 percentage points over our re-implementation of DiGIT with the same features. Among all compared methods in Table 1, it is second only to ActionMamba (42.0%), and is superior to the remaining dense and query-based detectors. In this dataset, each video contains on average only about 1.5 annotated action instances, and actions typically span a large fraction of

Method	Feature	THUMOS14						ActivityNet-v1.3			
		0.3	0.4	0.5	0.6	0.7	Avg.	0.5	0.75	0.95	Avg.
AFSD Lin et al. (2021)	I3D	67.3	62.4	55.5	43.7	31.1	52.0	52.4	35.3	6.5	34.4
TALLFormer Cheng & Bertasius (2022)	Swin-B	76.0	–	63.2	–	34.5	59.2	54.1	36.2	7.9	35.6
ViT-TAD Yang et al. (2024c)	ViT-B	85.1	80.9	74.2	61.8	45.4	69.5	55.8	38.5	8.8	37.4
TriDet Shi et al. (2023)	I3D/R(2+1)D	83.6	80.1	72.9	62.4	47.4	69.3	54.7	38.0	8.4	36.8
DyFADet Yang et al. (2024b)	I3D/R(2+1)D	84.0	80.1	72.7	61.1	47.9	69.2	58.1	39.6	8.4	38.5
ActionFormer Zhang et al. (2022)	I3D/R(2+1)D	82.1	77.8	71.0	59.4	43.9	66.8	54.7	37.8	8.4	36.6
ActionFormer Zhang et al. (2022)	InternVideo2	82.3	81.9	75.1	65.8	50.3	71.9	61.5	44.6	12.7	41.2
ActionMamba Chen et al. (2024)	InternVideo2	86.9	83.1	76.9	65.9	50.8	72.7	62.4	43.5	10.2	**42.0**
RTD-Net Tan et al. (2021)	I3D/TSN	68.3	62.3	51.9	38.8	23.7	49.0	47.2	30.7	8.6	30.8
ReAct Shi et al. (2022)	I3D/TSN	69.2	65.0	57.1	47.8	35.6	55.0	49.6	33.0	8.6	32.6
Self-DETR Kim et al. (2023)	I3D	74.6	69.5	60.0	47.6	31.8	56.7	52.3	33.7	8.4	33.8
TadTR Liu et al. (2022b)	I3D/R(2+1)D	74.8	69.1	60.1	46.6	32.8	56.7	53.6	37.5	10.6	36.8
DualDETR Zhang et al. (2024)	I3D	82.9	78.0	70.4	58.5	44.4	66.8	52.6	35.0	7.8	34.3
TE-TAD Kim et al. (2024)	I3D/R(2+1)D	83.3	78.4	71.3	60.7	45.6	67.9	54.2	38.1	10.6	37.1
DiGIT Kim et al. (2025a)	I3D/R(2+1)D	83.6	79.6	71.9	61.5	48.6	69.0	54.3	38.4	10.6	37.2
TadTR Liu et al. (2022b)	InternVideo2	84.8	79.3	70.4	58.2	43.8	67.3	57.1	38.8	11.0	38.2
TE-TAD Kim et al. (2024)	InternVideo2	84.3	81.1	73.7	62.6	49.5	70.3	61.3	41.8	10.9	41.1
DiGIT* Kim et al. (2025a)	InternVideo2	87.6	84.2	77.6	67.3	52.5	73.8	54.8	39.0	10.2	37.8
FRQAct	InternVideo2	88.7	84.7	79.0	67.8	53.7	**74.7**	56.6	40.5	10.9	39.2

Table 1: Comparison with state-of-the-art methods on THUMOS14 and ActivityNet-v1.3. For THUMOS14 we report mAP at tIoU thresholds $\{0.3, 0.4, 0.5, 0.6, 0.7\}$ and the average mAP. For ActivityNet-v1.3 we report mAP at $\{0.5, 0.75, 0.95\}$ and the average mAP over $\{0.5{:}0.05{:}0.95\}$. - indicates that the result is not available. * denotes results reproduced by us under the same environment as our method.

the video duration. Small boundary adjustments induce only minor changes in temporal IoU, so the boundary refinement fails to demonstrate its advantages. At the same time, ActionMamba' s long-sequence state-space modeling is better suited to capturing long-range dependencies in sparse, long actions, so SQRAct achieves a slightly lower average mAP on the ActivityNet v1.3 dataset compared to ActionMamba.

Overall, SQRAct achieves strong performance across all three datasets and remains highly competitive against representative methods from all three detection paradigms. These experiments demonstrate that jointly modeling frequency-domain temporal features, cross-scale boundary information, and explicit inter-query relations effectively improves temporal action detection across diverse video scenarios. In addition, the gains of SQRAct over baselines are more pronounced on THUMOS14 and HACS-Segments than on ActivityNet v1.3, suggesting that our query-relation refinement is particularly effective for denser action instances.

5 Conclusion

In this paper, we propose SQRAct, a TAD framework centered on action boundary modeling, establishing a collaborative optimization pipeline from feature encoding to query decoding. In the encoder, MTAM jointly models local temporal, global frequency-domain, and channel–temporal dual-dimensional representations to capture multi-level temporal features. SFE enhances boundary details while preserving semantic information through cross-layer interactions between shallow spatial features and deep semantic features. In the decoder, QRR explicitly models containment and adjacency temporal relations among query-predicted segments, and performs feature fusion and boundary refinement according to the relationship type. The three modules respectively optimize action boundary modeling from the perspectives of global semantics, cross-layer interaction, and instance-level query relations, thereby jointly improving boundary localization accuracy and consistency in instance-level predictions.

REFERENCES

N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, 2020.

G. Chen, Y. Huang, J. Xu, B. Pei, Z. Chen, Z. Li, J. Wang, K. Li, T. Lu, and L. Wang. Video Mamba Suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*, 2024.

Y. Chen, Y. Li, X. Cui, F. Lin, and S. Lv. W-StrataFormer: A temporal action localization model based on wavelet transform and strata feature enhancement. *Applied Soft Computing*, 190:114613, 2026. doi: 10.1016/j.asoc.2026.114613.

F. Cheng and G. Bertasius. TALLFormer: Temporal action localization with a long-memory transformer. In *Proceedings of the European Conference on Computer Vision*, 2022.

Gao et al. EASE-DETR: Easing the competition among object queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17282–17291, 2024.

J. Guibas, M. Mardani, Z. Li, A. Tao, A. Anandkumar, and B. Catanzaro. Efficient token mixing for transformers via adaptive fourier neural operators. In *International Conference on Learning Representations*, 2022.

F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2015.

H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei. Relation networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Z. Huang, Z. Zhang, C. Lan, Z.-J. Zha, Y. Lu, and B. Guo. Adaptive frequency filters as efficient global token mixers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. The THUMOS challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017.

C. Jin, Y. Chen, G. Li, T. Zhang, and T. H. Li. Low pass filter for anti-aliasing in temporal action localization. *arXiv preprint arXiv:2104.11403*, 2021.

H.-J. Kim, J.-H. Hong, H. Kong, and S.-W. Lee. TE-TAD: Towards full end-to-end temporal action detection via time-aligned coordinate expression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

H.-J. Kim, Y. Lee, J.-H. Hong, and S.-W. Lee. DiGIT: Multi-dilated gated encoder and central-adjacent region integrated decoder for temporal action detection transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

J. Kim, M. Lee, and J.-P. Heo. Self-feedback DETR for temporal action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10286–10296, 2023.

J. Kim, M. Lee, C.-H. Cho, J. Lee, and J.-P. Heo. Prediction-feedback DETR for temporal action detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025b.

C. Lin, C. Xu, D. Luo, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu. Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

T. Lin, X. Liu, X. Li, E. Ding, and S. Wen. BMN: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022a.

X. Liu, Q. Wang, Y. Hu, X. Tang, S. Zhang, S. Bai, and X. Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31:5427–5441, 2022b.

Z. Qing, H. Su, W. Gan, D. Wang, W. Wu, X. Wang, Y. Qiao, J. Yan, C. Gao, and N. Sang. Temporal context aggregation network for temporal action proposal refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 485–494, 2021.

Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou. Global filter networks for image classification. In *Advances in Neural Information Processing Systems*, 2021.

D. Shi, Y. Zhong, Q. Cao, J. Zhang, L. Ma, J. Li, and D. Tao. ReAct: Temporal action detection with relational queries. In *Proceedings of the European Conference on Computer Vision*, 2022.

D. Shi, Y. Zhong, Q. Cao, L. Ma, J. Li, and D. Tao. TriDet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

J. Tan, J. Tang, L. Wang, and G. Wu. Relaxed transformer decoders for direct action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Y. Wang, K. Li, X. Li, J. Yu, Y. He, G. Chen, B. Pei, R. Zheng, Z. Wang, Y. Shi, et al. InternVideo2: Scaling foundation models for multimodal video understanding. In *Proceedings of the European Conference on Computer Vision*, 2024.

M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem. G-TAD: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

J. Yang, P. Wei, and N. Zheng. Cross time-frequency transformer for temporal action localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):4625–4638, 2024a.

L. Yang, Z. Zheng, Y. Han, H. Cheng, S. Song, G. Huang, and F. Li. DyFADet: Dynamic feature aggregation for temporal action detection. In *Proceedings of the European Conference on Computer Vision*, pp. 305–322, 2024b.

M. Yang, H. Gao, P. Guo, and L. Wang. Adapting short-term transformers for action detection in untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024c.

K. Yi, J. Fei, Q. Zhang, H. He, S. Hao, D. Lian, and W. Fan. FilterNet: Harnessing frequency filters for time series forecasting. In *Advances in Neural Information Processing Systems*, 2024.

C.-L. Zhang, J. Wu, and Y. Li. ActionFormer: Localizing moments of actions with transformers. In *Proceedings of the European Conference on Computer Vision*, pp. 492–510, 2022.

Y. Zhang, X. Zhang, and L. Wang. DualDETR: Dual-level query DETR for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

H. Zhao, A. Torralba, L. Torresani, and Z. Yan. HACS: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8731–8740, 2019.

Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017.

X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

X. Zhu, Y. Zhu, T. Chen, W. Wu, and Y. Dang. FDDet: Frequency-decoupling for boundary refinement in temporal action detection. In *Advanced Intelligent Computing Technology and Applications*, volume 15847 of *Lecture Notes in Computer Science*, pp. 205–217. Springer, 2025.

Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable ConvNets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9308–9316, 2019. doi: 10.1109/CVPR.2019.00953.