HyrCap: Hybrid Rank-Calibration of Action Proposals for Temporal Event Understanding

Rui Chen[1] **Xingyu Chen**[2,✉*] **Pengxin Xu**[4,2] **Kai Chen**[3,2] **Junzhi Yu**[1,✉]
[1]Peking University [2]Zhongguancun Academy
[3] Zhongguancun Institute of Artificial Intelligence
[4] Harbin Institute of Technology, Shenzhen
chenruiii@stu.pku.edu.cn, chenxingyu@bza.edu.cn

Abstract

Temporal action detection in long untrimmed videos still suffers from inaccurate temporal localization and low-quality proposals, especially under sparse temporal annotations and ambiguous event boundaries. In the context of open-ended video understanding and LLM-driven event analysis, reliable temporal action proposals have emerged as a critical interface bridging temporal localization, event annotation, and semantic understanding. However, conventional temporal action detection (TAD) pipelines lack explicit modeling of proposal localization quality, cross-representation support, and inter-proposal relations, making it difficult to consistently provide high-quality proposals. We propose Hybrid Rank-Calibration of Action Proposals (HyrCap), a proposal-centric query-based TAD framework for temporal event understanding. In HyrCap, Hybrid denotes the use of complementary video features that provide different temporal localization cues, while Rank-Calibration refers to calibrating the ranking confidence of dense action proposals so that it better reflects proposal localization reliability and candidate competition. HyrCap treats dense query-based predictions as an over-complete temporal event hypothesis space and improves the ranking reliability, localization quality, and candidate discrimination of action proposals through modules such as Cross-Representation Consensus Calibration, Localization Quality Estimation, and Proposal Redundancy Suppression. Furthermore, inspired by the strong capability of LLMs in fine-grained semantic recognition, event description, and reasoning, we introduce HyrCap for Semantic Action Understanding: calibrated proposals can be used for standard TAD detection and as inputs for LLM-based semantic understanding. We further define a staged evaluation protocol to separately measure event localization ability, segment-level semantic understanding, and their combined event understanding performance. Experiments on public benchmarks show that HyrCap achieves state-of-the-art performance across multiple metrics and provides a more fine-grained paradigm and evaluation protocol for query-based event segmentation and annotation. The project page is available at `https://zgca-hmi-lab.github.io/HyrCap/`.

1 Introduction

Existing video data typically consist of long untrimmed videos with sparse temporal annotations and imprecise event boundaries, making accurate temporal localization difficult and often leading to low-quality proposals [10, 3, 32]. Localizing reliable event segments from such videos is a fundamental step for video classification, event description, retrieval, and high-level semantic understanding.

*Project leader. ✉ Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

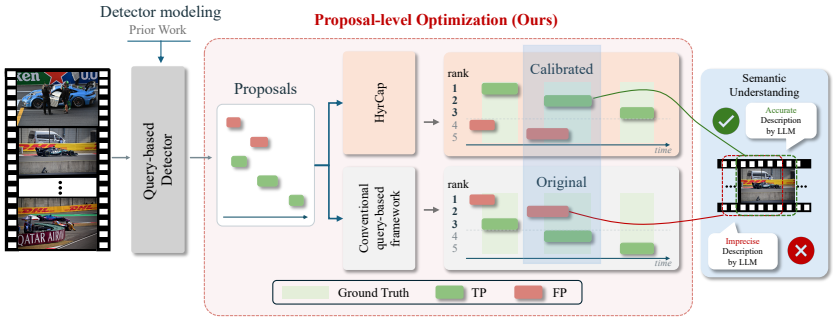

Figure 1: Comparison of HyrCap and conventional query-based TAD: HyrCap produces higher-quality temporal proposals that better support subsequent semantic understanding.

Temporal action detection (TAD) aims to localize action instances in untrimmed videos and recognize their semantic categories [10, 3]. Conventional methods usually formulate TAD as a closed-set detection problem, where a model predicts temporal boundaries, assigns action labels, and ranks candidate detections by confidence scores [14, 31, 19]. However, in the context of open-ended video understanding and LLM-driven event analysis, reliable temporal action proposals have become a key interface between temporal localization, event annotation, and semantic understanding [22, 17, 29]. Therefore, TAD is expected not only to produce final detections, but also to provide high-quality proposals.

Recent query-based TAD methods generate dense temporal candidates with learnable queries and have shown strong potential for localizing events in long videos [4, 15, 12, 13]. However, detection performance depends not only on whether the model generates proposals that cover ground-truth events, but also on how these proposals are ranked, suppressed, and matched during evaluation. Existing pipelines lack explicit modeling of proposal localization quality, cross-representation support, and inter-proposal relations, making dense query predictions with localization potential difficult to preserve, rank, and exploit. This suggests that improving TAD should not only rely on generating more candidates, but should also understand and optimize how proposals behave throughout the final detection pipeline.

Motivated by this observation, we first revisit the role of proposals in query-based TAD and analyze different video features for this task. We study how different types of false positives (FPs), proposal ranking, evaluation matching, and video feature representations affect proposal quality and final detection performance [5, 27, 1]. This analysis clarifies the relationship among proposal behavior, feature representation, and final mAP, and indicates that the localization potential of query-based detectors should be released through more reliable proposal-level ranking calibration.

Based on the above analysis, we propose Hybrid Rank-Calibration of Action Proposals (HyrCap), as shown in Figure 1, a proposal-centric query-based TAD framework for temporal event understanding. In HyrCap, Hybrid denotes the use of complementary video feature representations that provide different temporal localization cues, while Rank-Calibration refers to calibrating the ranking confidence of dense action proposals so that it better reflects proposal localization reliability and candidate competition. HyrCap does not directly treat dense query predictions as final detections. Instead, it regards them as an over-complete temporal event hypothesis space and improves the ranking reliability, localization quality, and candidate discrimination of action proposals through modules such as Cross-Representation Consensus Calibration, Localization Quality Estimation, and Proposal Duplication Suppression.

Furthermore, inspired by the growing capability of LLMs in fine-grained event description, semantic recognition, and spatio-temporal reasoning [22, 17, 29], we introduce HyrCap for Semantic Action Understanding. This paradigm argues that query-based TAD should not be assessed only by overall mAP under conventional one-stage loc-cls evaluation, but further decomposed into proposal localization ability and segment-level semantic understanding ability. Specifically, in the first stage, action categories are collapsed into foreground event and background, and the category-agnostic localization quality, ranking reliability, and event coverage of proposals are evaluated independently. In the second stage, an LLM or semantic classifier is applied to localized segments to evaluate classification, description, or reasoning performance. This two-stage protocol decouples whether an

event is accurately localized from whether its semantics are correctly understood, making reliable generation, ranking, and selection of temporal action proposals the central capability of query-based detectors.

Our contributions are summarized as follows:

- We provide a systematic analysis of proposals in query-based TAD, studying how FPs, proposal ranking, evaluation matching, and different video features affect proposal quality and final detection performance, thereby clarifying the relationship among proposal behavior, feature representation, and final mAP.
- We propose HyrCap, a proposal-centric rank-calibration framework that improves proposal localization reliability, ranking quality, and candidate discrimination through modules such as Cross-Representation Consensus Calibration, Localization Quality Estimation, and Proposal Duplication Suppression.
- We introduce a two-stage paradigm for semantic action understanding, where proposal localization ability and segment-level semantic understanding ability are evaluated separately. HyrCap achieves state-of-the-art performance across multiple metrics on public benchmarks and provides a more fine-grained paradigm and evaluation protocol for query-based event segmentation and annotation.

2 Related Work

Video representations for action understanding. Action recognition is a fundamental task in video analysis, aiming to classify video sequences into specific action categories. I3D [5] extends the Inception network to the video domain by introducing 3D convolutions; R(2+1)D [24] improves computational efficiency by decomposing 3D convolutions into 2D spatial operations and 1D temporal operations. TSN [26] divides a video into multiple temporal segments and aggregates segment-level predictions through a consensus function, enabling effective long-range temporal modeling for action recognition. VideoMAEv2 [25] adopts a Transformer-based masked reconstruction pretraining strategy to learn more robust video representations. InternVideo2 [27] further enhances video representation learning with large-scale training data and a highly scalable model. V-JEPA2 [1] learns high-level video representations through self-supervised joint embedding prediction, improving spatiotemporal modeling capability. These models are commonly used as feature extractors for downstream tasks such as TAD.

Temporal action detection. Anchor-free and query-based detectors are two mainstream paradigms in current TAD. Anchor-free methods such as ActionFormer [31], TriDet [19], and ActionMamba [6] directly predict action segments, avoiding predefined anchors and improving temporal modeling with Transformer architectures, trident prediction schemes, or Mamba-based structures, but still typically rely on NMS to remove redundant proposals. Query-based methods, inspired by DETR [4], formulate TAD as a set prediction problem and generate candidate events with learnable queries, thereby reducing reliance on hand-crafted components. RTD-Net [23], ReAct [20], TadTR [15], and DualDETR [34] have advanced query-based TAD, but some still depend on sliding windows or NMS. In contrast, TE-TAD [12] achieves a complete end-to-end detection framework by redesigning coordinate representation; building on it, DiGIT [13] further introduces MDGE, a multi-scale adapter, and CAID to improve encoder representation capability and decoder temporal context modeling.

Localization quality, temporal refinement, and video-language understanding. Localization-aware ranking and post-processing have been studied in the 2D object detection domain, including Soft-NMS [2], IoU-Net [11], and Weighted Boxes Fusion [21]. These methods improve detection from the perspectives of score suppression, localization quality estimation, and box-level fusion. However, TAD focuses on temporal action segments, where boundaries are more ambiguous and proposal ranking is affected by redundancy, one-to-one matching, and global score ordering across videos. Therefore, post-processing strategies from 2D detection cannot be directly regarded as equivalent solutions for TAD. Meanwhile, video-language models such as VideoBERT [22] show the potential of event-level semantic understanding. For query-based TAD, reliable temporal localization is the foundation for subsequent semantic understanding, making high-quality temporal proposals essential for connecting detection with video-language understanding.

3 Analysis of Proposal Behavior

FP Categorization. For dense query predictions, final detection quality is determined not only by the generated candidate segments, but also by temporal boundary accuracy, confidence ranking,

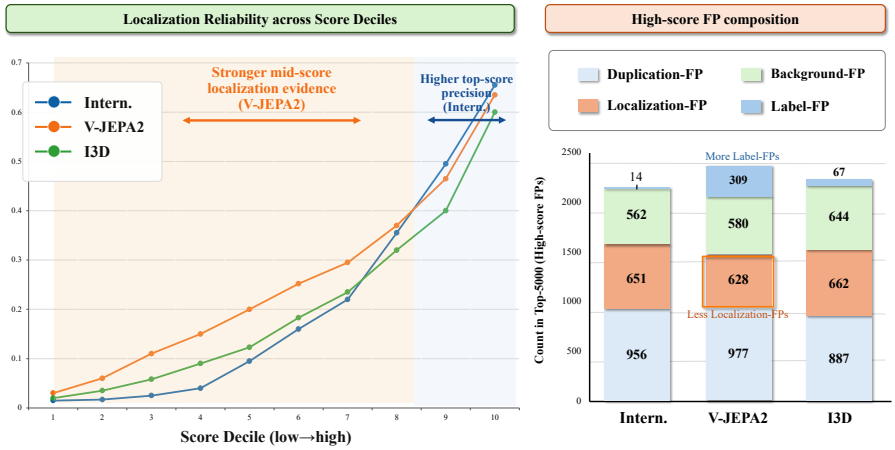

Figure 2: Visualization of different features under TAD. Intern. denotes InternVideo2 features.

proposal redundancy, and one-to-one evaluation matching. Therefore, how proposals are preserved, ranked, and matched in the final detection pipeline is critical for understanding the performance bottleneck of TAD. To analyze the sources of false positives (FPs), we instantiate the analysis with DiGIT and compare the proposal behaviors of I3D [5], InternVideo2 [27], and V-JEPA2 [1]. We categorize FPs into four types: duplication FP denotes redundant proposals caused by one-to-one matching when multiple candidates correspond to the same ground-truth event; localization FP denotes proposals that overlap with a ground-truth event but have inaccurate temporal boundaries; background FP denotes proposals with low overlap to all ground-truth events; and label FP denotes proposals with correct temporal coverage but incorrect class prediction.

Feature-dependent Proposal Behavior. As shown in Fig. 2, different feature inputs lead to distinct proposal behaviors. InternVideo2 shows the strongest proposal quality: it achieves the highest TP rate in the top deciles and produces fewer duplication, background, and label FPs among high-score predictions, indicating cleaner high-confidence candidates and more stable category prediction. In contrast, although V-JEPA2 is weaker than InternVideo2 as a detection representation, it achieves higher localization reliability in the middle score range and produces fewer localization FPs in the top-5000 predictions, suggesting complementary localization evidence. I3D exhibits weaker localization reliability and produces more background FPs. These observations indicate that InternVideo2 is suitable as the primary proposal source, while V-JEPA2 provides complementary localization reliability, motivating the hybrid rank-calibration design of HyrCap.

4 Method

The preceding analysis indicates that InternVideo2 and V-JEPA2 exhibit complementary proposal behaviors in TAD. Based on this observation, we propose HyrCap, a hybrid rank-calibration framework for query-based temporal action detection. As illustrated in Fig. 3(a), dense query-based proposals are first generated from the primary InternVideo2 representation and the auxiliary V-JEPA2 representation. HyrCap then operates in the proposal space, as shown in Fig. 3(b), and recalibrates existing proposal scores by combining learned cross-representation temporal consensus, proposal-level localization reliability, and local proposal competition:

$$s_i^{cal} = s_i(1 + \beta c_i m_i)(q_i^{loc})^{\gamma_q}(1 - d_i)^{\gamma_d}, \tag{1}$$

where s_i is the detector score, m_i denotes cross-representation temporal support, c_i is the learned consensus gate, q_i^{loc} estimates localization reliability, and d_i is the relation-suppression score. After calibration, HyrCap refines proposal boundaries by temporal voting and applies multi-class NMS to obtain final detections.

4.1 Cross-Representation Consensus Calibration

As shown in Fig. 3(b)-1, CPCC uses InternVideo2 proposals as the primary candidates and V-JEPA2 proposals as class-agnostic auxiliary temporal evidence. The primary proposal set is denoted as $\mathcal{P}^I = \{p_i = (b_i, y_i, s_i)\}_{i=1}^N$, where b_i is the temporal segment, y_i is the predicted category, and s_i is the detector confidence. The auxiliary proposal set is denoted as $\mathcal{P}^A = \{\bar{p}_j = (\bar{b}_j)\}_{j=1}^M$.

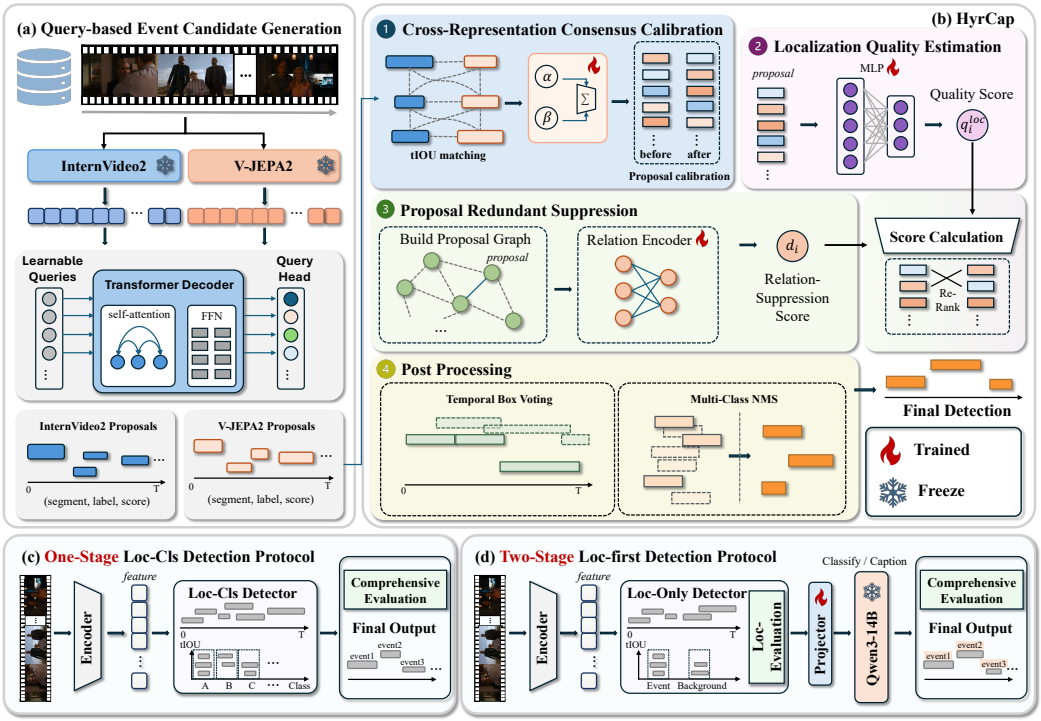

Figure 3: Overview of HyrCap. (a) Query-based event candidate generation with InternVideo2 and V-JEPA2 representations. (b) HyrCap calibrates and refines dense proposals through cross-representation consensus calibration, localization quality estimation, proposal duplication suppression, and post-processing. (c) Conventional one-stage loc-cls detection protocol. (d) Two-stage loc-first protocol for staged semantic understanding. Loc denotes localization.

For each primary proposal, we compute its auxiliary temporal support by the maximum tIoU:

$$m_i = \max_{\bar{p}_j \in \mathcal{P}^A} \mathrm{tIoU}(b_i, \bar{b}_j). \tag{2}$$

Instead of using m_i as a fixed heuristic, CPCC trains a lightweight consensus head to predict a gate $c_i = f_\theta^{con}(r_i)$ from proposal-level consensus cues. The gate is supervised by the localization quality of the primary proposal:

$$\hat{c}_i = \begin{cases} 1, & \max_{g \in \mathcal{G}_v} \mathrm{tIoU}(b_i, b_g) \geq \mu_c, \\ 0, & \text{otherwise}, \end{cases} \qquad \mathcal{L}_{con} = \mathrm{BCE}(c_i, \hat{c}_i). \tag{3}$$

The consensus-calibrated score is defined as $s_i^{con} = s_i(1 + \beta c_i m_i)$. This module keeps the original segment and category unchanged, and only adjusts proposal ranking through learned cross-representation temporal consensus.

4.2 Localization Quality Estimation

Fig. 3(b)-2 shows the localization quality estimation module. Detector confidence does not always reflect temporal localization quality. HyrCap therefore estimates a proposal-level localization score q_i^{loc} from a compact descriptor x_i^{loc}, which includes confidence statistics, temporal geometry, consensus evidence, and local ranking cues:

$$q_i^{loc} = f_\theta^{loc}(x_i^{loc}), \quad q_i^{loc} \in [0, 1]. \tag{4}$$

The supervision is constructed offline from ground-truth annotations. A proposal is localization-positive when it sufficiently overlaps with any ground-truth event in the same video:

$$\mathcal{L}_{loc} = \mathrm{BCE}(q_i^{loc}, \hat{q}_i), \quad \hat{q}_i = \begin{cases} 1, & \max_{g \in \mathcal{G}_v} \mathrm{tIoU}(b_i, b_g) \geq \mu_q, \\ 0, & \text{otherwise}. \end{cases} \tag{5}$$

This target is category-agnostic and measures temporal alignment rather than semantic correctness. During inference, q_i^{loc} is used as a multiplicative factor in the calibrated ranking score.

4.3 Proposal Duplication Suppression

As illustrated in Fig. 3(b)-3, dense query predictions often contain multiple overlapping proposals around the same event. To model local proposal competition, HyrCap builds same-video and same-class neighborhoods according to temporal overlap. For each proposal, pairwise cues such as temporal IoU, center distance, length ratio, and score gap are aggregated by a lightweight relation encoder:

$$d_i = f_\phi^{rel}(h_i'), \quad d_i \in [0,1]. \tag{6}$$

Here, d_i estimates whether proposal p_i should be suppressed under its local proposal context. The suppression target is constructed from offline proposal matching diagnosis, where duplicate proposals and locally inferior candidates in overlapping groups are treated as suppression-positive samples:

$$\mathcal{L}_{cal} = \mathcal{L}_{loc} + \lambda_{rel}\mathcal{L}_{rel}, \quad \mathcal{L}_{rel} = \mathrm{BCE}(d_i, \hat{d}_i). \tag{7}$$

The relation factor $(1-d_i)^{\gamma_d}$ reduces the ranking priority of redundant or inferior candidates while preserving proposals with stronger local evidence.

After reranking, HyrCap further refines proposal boundaries by temporal voting, as shown in Fig. 3(b)-4. For each proposal p_i, we collect its same-video and same-class neighbors $\mathcal{B}_i$ whose tIoU with p_i exceeds a threshold c. The refined boundaries are computed as

$$\hat{t}_i^s = \frac{\sum_{p_j\in\mathcal{B}_i} w_j t_j^s}{\sum_{p_j\in\mathcal{B}_i} w_j + \epsilon}, \quad \hat{t}_i^e = \frac{\sum_{p_j\in\mathcal{B}_i} w_j t_j^e}{\sum_{p_j\in\mathcal{B}_i} w_j + \epsilon}, \quad w_j = \max(s_j^{cal} - \tau, 0). \tag{8}$$

The final refined proposal is $\hat{p}_i = (\hat{b}_i, y_i, s_i^{cal})$, where $\hat{b}_i = (\hat{t}_i^s, \hat{t}_i^e)$. Multi-class NMS is then applied to produce the final detection results.

4.4 HyrCap for Staged Semantic Understanding

Beyond closed-set TAD evaluation, HyrCap supports a staged semantic understanding protocol. As shown in Fig. 3(c), the conventional one-stage loc-cls protocol jointly performs temporal localization and closed-set classification, where proposal ranking, boundary quality, and semantic prediction are coupled in the final class-specific detections.

In contrast, Fig. 3(d) shows our two-stage loc-first protocol. In the first stage, HyrCap focuses on category-agnostic event localization and converts dense query predictions into a compact set of event proposals through rank calibration, boundary refinement, and class-agnostic NMS:

$$\mathcal{E}(V) = \mathrm{TopK}\left(\mathrm{NMS}_{ca}\left(\{(\hat{b}_i, s_i^{cal})\}_{i=1}^N\right)\right), \tag{9}$$

where $\mathcal{E}(V)$ denotes the selected event units and NMS_{ca} denotes class-agnostic NMS. The selection only depends on the calibrated localization score s_i^{cal}, without using ground-truth annotations or semantic model confidence.

In the second stage, each selected event segment is passed to a semantic model through a trained projector for segment-level classification, captioning, or reasoning:

$$(\tilde{y}_i, \tilde{r}_i) = \mathcal{M}_{sem}\left(g_\omega\left(\psi(V[\hat{b}_i])\right), \mathcal{C}, \pi\right), \tag{10}$$

where $\psi(V[\hat{b}_i])$ denotes the segment-level visual representation, $g_\omega(\cdot)$ is the trained projector, $\mathcal{C}$ is the candidate action vocabulary, and π is the task prompt. The final staged output is

$$\mathcal{D}_{stage}(V) = \{(\hat{b}_i, \tilde{y}_i, s_i^{cal}, \tilde{r}_i)\}_{i=1}^K. \tag{11}$$

The ranking score remains s_i^{cal} rather than the semantic model confidence, keeping temporal proposal ranking and semantic interpretation decoupled. For joint evaluation, a prediction is considered correct only when both its temporal segment and semantic prediction are correct.

5 Experiment

5.1 Setup

Datasets. We conduct experiments on three public TAD datasets: THUMOS14 [10], ActivityNet v1.3 [3], and HACS-Segment [32]. THUMOS14 contains 20 action classes with 200/213 validation/test videos and 3,007/3,358 action instances. ActivityNet v1.3 and HACS-Segment both cover 200 classes, with 10,024/4,926 and 37,613/5,981 training/validation videos, respectively.

Table 1: Comparison with state-of-the-art methods on THUMOS14 [10] and ActivityNet v1.3 [3]. For TE -TAD + Ours, we additionally apply our methods on TE-TAD.

Training Type	Head Type	Method	Feature	THUMOS14						ActivityNet v1.3			
				0.3	0.4	0.5	0.6	0.7	Avg.	0.5	0.75	0.95	Avg.
Full	-	AFSD	I3D [5]	67.3	62.4	55.5	43.7	31.1	52.0	52.4	35.3	6.5	34.4
		TALLFormer	Swin-B [16]	76.0	-	63.2	-	34.5	59.2	54.1	36.2	7.9	35.6
		ViT-TAD	ViT-B [8]	85.1	80.9	74.2	61.8	45.4	69.5	55.8	38.5	8.8	37.4
		AdaTAD	VideoMAEv2-g [25]	**89.7**	**86.7**	**80.9**	**71.0**	**56.1**	**76.9**	**61.7**	**43.4**	**10.9**	**41.9**
Head only	Anchor -free	TriDet	I3D [5] / R(2+1)D [24]	83.6	80.1	72.9	62.4	47.4	69.3	54.7	38.0	8.4	36.8
		DyFADet	I3D [5] / R(2+1)D [24]	84.0	80.1	72.7	61.1	47.9	69.2	58.1	39.6	8.4	38.5
		ActionFormer	I3D [5] / R(2+1)D [24]	82.1	77.8	71.0	59.4	43.9	66.8	54.7	37.8	8.4	36.6
		ActionFormer	InternVideo2 [27]	82.3	81.9	75.1	65.8	50.3	71.1	61.5	**44.6**	**12.7**	41.2
		ActionMamba	InternVideo2 [27]	**86.9**	**83.1**	**76.9**	**65.9**	**50.8**	**72.7**	**62.4**	43.5	10.2	**42.0**
	Query -based	RTD-Net	I3D [5] / TSN [26]	68.3	62.3	51.9	38.8	23.7	49.0	47.2	30.7	8.6	30.8
		ReAct	I3D [5] / TSN [26]	69.2	65.0	57.1	47.8	35.6	55.0	49.6	33.0	8.6	32.6
		Self-DETR	I3D [5]	74.6	69.5	60.0	47.6	31.8	56.7	52.3	33.7	8.4	33.8
		TadTR	I3D [5] / R(2+1)D [24]	74.8	69.1	60.1	46.6	32.8	56.7	53.6	37.5	10.6	36.8
		DualDETR	I3D [5]	82.9	78.0	70.4	58.5	44.4	66.8	52.6	35.0	7.8	34.3
		TE-TAD	I3D [5] / R(2+1)D [24]	83.3	78.4	71.3	60.7	45.6	67.9	54.2	38.1	10.6	37.1
		DiGIT	I3D [5] / R(2+1)D [24]	83.6	79.6	71.9	61.5	48.6	69.0	54.4	38.2	10.7	37.3
		TadTR	InternVideo2 [27]	84.8	79.3	70.4	58.2	43.8	67.3	57.1	38.8	11.0	38.2
		TE-TAD	InternVideo2 [27]	84.3	81.1	73.7	62.6	49.5	70.3	61.3	41.8	10.9	41.1
		DiGIT	InternVideo2 [27]	87.6	84.2	77.6	67.3	52.5	73.8	62.0	43.1	11.3	42.0
		TE-TAD + Ours	InternVideo2 [27] + V-JEPA2 [1]	85.9	81.4	74.1	66.9	48.6	71.4	62.1	42.5	11.0	41.9
		HyrCap	InternVideo2 [27] + V-JEPA2 [1]	**88.1**	**85.2**	**79.1**	**70.8**	**54.6**	**75.6**	**64.4**	**44.0**	**11.4**	**43.3**

Evaluation Metrics. We follow the standard evaluation protocols of each dataset. For THUMOS14, we report mAP at tIoU thresholds of 0.3, 0.4, 0.5, 0.6, and 0.7, together with the average mAP. For ActivityNet v1.3 and HACS-Segment, we report mAP@0.5, mAP@0.75, mAP@0.95, and the average mAP over 0.5:0.95. In addition to standard class-specific detection evaluation, we report category-agnostic localization results. For staged semantic evaluation, we further measure event localization, segment-level semantic understanding, and their combined performance.

Implementation Details. For better visualization and analysis, we use DiGIT as the baseline and deploy our method within its framework. More implementation details are provided in the appendix.

5.2 Comparison with State-of-the-Art Methods

We first compare HyrCap with existing state-of-the-art temporal action detection methods. The compared methods include anchor-based, anchor-free, and query-based detectors. The main results are reported on multiple public benchmarks to evaluate the effectiveness of HyrCap under different dataset scales, action category distributions, and video characteristics.

Table 1 reports the comparison on THUMOS14 and ActivityNet v1.3. On THUMOS14, HyrCap improves over strong query-based baselines in terms of average mAP and remains effective under stricter tIoU thresholds. This indicates that hybrid rank-calibration not only improves proposal ranking, but also preserves proposals with more accurate temporal boundaries. On ActivityNet v1.3, HyrCap also achieves consistent improvement, showing that the proposal-level calibration strategy is not limited to short action instances or a single benchmark.

Table 2 reports the results on HACS-Segment. The performance on this large-scale benchmark shows that HyrCap generalizes to diverse action categories and video distributions. These results validate HyrCap as a general proposal-centric framework for improving query-based TAD.

5.3 Ablation Study

To validate the effectiveness of each component, we conduct ablation experiments on each module, as shown in Table 3. CPCC improves proposal ranking by exploiting cross-representation temporal consistency, LQE explicitly estimates proposal localization reliability and prioritizes candidates with more accurate temporal boundaries, and PRS suppresses redundant candidates to reduce high-score false positives. When combined, these modules achieve the best performance on different baselines.

Table 2: Performance comparison with state-of-the-art methods on HACS-Segment.

Head Type	Method	Feature	mAP			
			0.5	0.75	0.95	Avg.
Anchor-based	SSN [33]	I3D [5]	28.8	18.8	5.3	19.0
	G-TAD [28]	I3D [5]	41.1	27.6	8.3	27.5
	BMN [14]	SlowFast [9]	52.5	36.4	10.4	35.8
	TCANet [18]	SlowFast [9]	54.1	37.2	11.3	36.8
Anchor-free	TALLFormer [7]	Swin-B [16]	55.0	36.1	11.8	36.5
	TriDet [19]	I3D [5]	54.5	36.8	11.5	36.8
	TriDet [19]	SlowFast [9]	56.7	39.3	11.7	38.6
	TriDet [19]	VideoMAEv2-g [25]	62.4	44.1	13.1	43.1
	DyFADet [30]	SlowFast [9]	57.8	39.8	11.8	39.2
	DyFADet [30]	VideoMAEv2-g [25]	64.0	44.8	14.1	44.3
	ActionFormer [31]	InternVideo2 [27]	62.6	44.6	12.7	43.3
	ActionMamba [6]	InternVideo2 [27]	**64.0**	**45.7**	**13.3**	**44.6**
Query-based	TadTR [15]	I3D [5]	47.1	32.1	10.9	32.1
	TadTR [15]	InternVideo2 [27]	54.2	38.8	12.8	37.8
	TE-TAD [12]	InternVideo2 [27]	60.4	45.6	16.5	44.1
	DiGIT [13]	InternVideo2 [27]	61.1	47.5	17.8	45.5
	HyrCap	InternVideo2 [27] + V-JEPA2 [1]	**62.7**	**48.8**	**17.9**	**47.1**

Table 3: Component Analysis. CPCC denotes Cross-Representation Consensus Calibration, LQE denotes Localization Quality Estimation, and PRS denotes Proposal Redundancy Suppression. Dup. represents Duplication FPs, Bg. Background FPs, Loc. Localization FPs, and Lab. Label FPs.

Baseline	Input Feature		CPCC	LQE	PRS	mAP				FPs in Top-100			
	Intern.	V-JEPA2				0.3	0.5	0.7	Avg.	Dup.	Bg.	Loc.	Lab.
TE-TAD [12]	✓					85.0	73.0	46.5	70.3	2	2	1	1
		✓				67.5	56.7	38.4	55.2	0	1	1	0
	✓	✓	✓			85.2	73.8	48.1	71.1	0	1	1	0
	✓			✓		85.0	73.0	47.2	70.9	2	2	0	0
	✓				✓	85.2	73.1	46.6	70.7	0	4	1	1
	✓	✓	✓	✓		85.0	73.0	47.2	70.8	0	2	0	1
	✓	✓	✓		✓	85.6	73.9	48.3	70.4	0	3	0	2
	✓			✓	✓	85.0	73.4	48.0	71.0	0	4	1	1
	✓	✓	✓	✓	✓	**85.9**	**74.1**	**48.6**	**71.4**	**0**	**1**	**0**	**0**
DiGIT [13]	✓					86.6	77.3	52.8	73.8	2	7	6	0
		✓				73.9	64.3	40.9	60.8	1	1	0	2
	✓	✓	✓			86.9	77.7	54.4	75.0	1	4	1	0
	✓			✓		86.8	77.4	53.7	74.5	2	6	2	0
	✓				✓	87.7	77.8	52.8	74.7	1	5	5	0
	✓	✓	✓	✓		87.3	77.8	54.5	75.1	2	5	1	0
	✓	✓	✓		✓	87.7	78.5	54.2	75.4	0	2	4	0
	✓			✓	✓	87.6	77.7	52.2	74.8	0	5	6	0
	✓	✓	✓	✓	✓	**88.1**	**79.1**	**54.6**	**75.6**	**0**	**1**	**1**	**0**

The Top-100 FP statistics further show that HyrCap effectively reduces duplicate, background, and localization FPs, leading to more reliable proposal ranking and improved detection performance.

Controlled Comparison of Fusion and Calibration Strategies. Because CPCC introduces proposals from an auxiliary representation, we further conduct controlled comparisons to distinguish the benefit of additional proposals from that of the proposed calibration mechanism. All methods use exactly the same InternVideo2 primary proposals and V-JEPA2 auxiliary proposals, together with the same proposal budget, score normalization, Top-K setting, NMS threshold, and evaluation protocol. Direct Union + NMS directly merges proposals from both branches before NMS; Score Average and Max-Score Fusion combine the scores of matched proposals; Temporal WBF fuses their temporal boundaries using proposal-score-based weights; and Fixed Consensus uniformly sets the consensus gate to $c_i = 1$. Therefore, the performance differences reflect only the fusion, calibration, and suppression strategies.

As shown in Table 4, directly combining the two proposal sources or applying simple score-level fusion remains substantially inferior to HyrCap, showing that the improvement cannot be attributed

Table 4: Controlled comparison of fusion and calibration strategies on THUMOS14 using identical dual-branch proposal inputs.

Method	AP@0.3	AP@0.5	AP@0.7	Avg. mAP
Direct Union + NMS	80.9	72.2	48.2	68.2
Score Average	81.8	72.6	48.4	68.6
Max-Score Fusion	80.5	71.1	45.7	66.8
Temporal WBF	83.8	75.2	50.8	71.1
Fixed Consensus	87.3	78.0	53.8	74.1
HyrCap	**88.1**	**79.1**	**54.6**	**75.6**

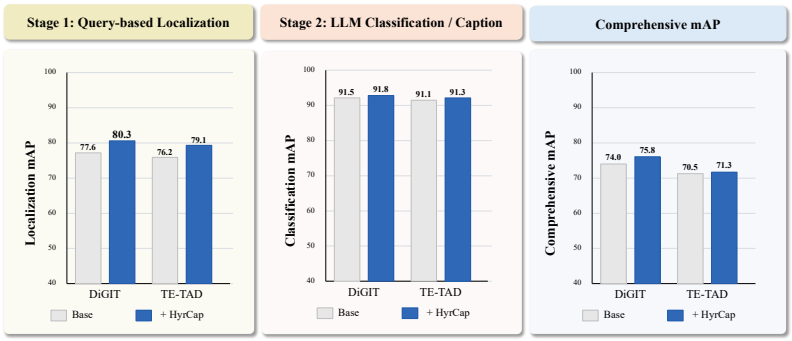

Figure 4: Two-Stage Evaluation.

merely to the availability of additional proposals. Temporal WBF improves over simple fusion, while Fixed Consensus further confirms that cross-branch temporal agreement provides useful reliability evidence. HyrCap surpasses Fixed Consensus by 1.5 points in average mAP, demonstrating that learned proposal-level calibration can selectively exploit beneficial auxiliary evidence while limiting the influence of unreliable cross-branch agreement.

5.4 Staged Semantic Evaluation

Beyond closed-set TAD evaluation, we further evaluate HyrCap under a staged semantic understanding protocol on THUMOS14 with InternVideo2 features. This protocol first evaluates whether a detector can produce category-agnostic event proposals, and then sends the selected event segments to Qwen3-14B [29] for segment-level classification or captioning.

Specifically, for localization evaluation, all action categories are collapsed into foreground event and background, and proposals are ranked by localization scores and matched to ground-truth segments using standard tIoU-based AP computation. For semantic evaluation, we select the top-ranked proposals after class-agnostic NMS, extract their segment features, map them with the trained projector, and feed them into Qwen3-14B for segment-level semantic prediction. For joint event evaluation, proposals are ranked by the calibrated localization score, and a prediction is counted as a true positive only when its temporal segment matches an unmatched ground-truth instance under the required tIoU threshold and its semantic prediction is correct. Thus, semantic confidence is not used for proposal ranking, keeping temporal localization and semantic interpretation decoupled.

As shown in Fig. 4, HyrCap consistently improves localization mAP on different query-based detectors. For DiGIT and TE-TAD, the localization mAP increases from 77.6 and 76.2 to 80.3, 79.1, respectively. Since the semantic model operates on localized event segments, better localization provides more reliable inputs for segment-level semantic prediction. Under joint event evaluation, HyrCap also improves final event understanding performance, showing that calibrated proposals benefit both closed-set TAD and LLM-based semantic understanding.

5.5 Qualitative Results

We conduct a qualitative comparison between HyrCap and the baseline DiGIT detector on representative videos. As shown in Figure 5, HyrCap generates proposals that more accurately cover the ground-truth action segments, while simultaneously reducing duplicate, background, and localization-error false positives. In contrast, DiGIT produces more redundant or misaligned proposals.

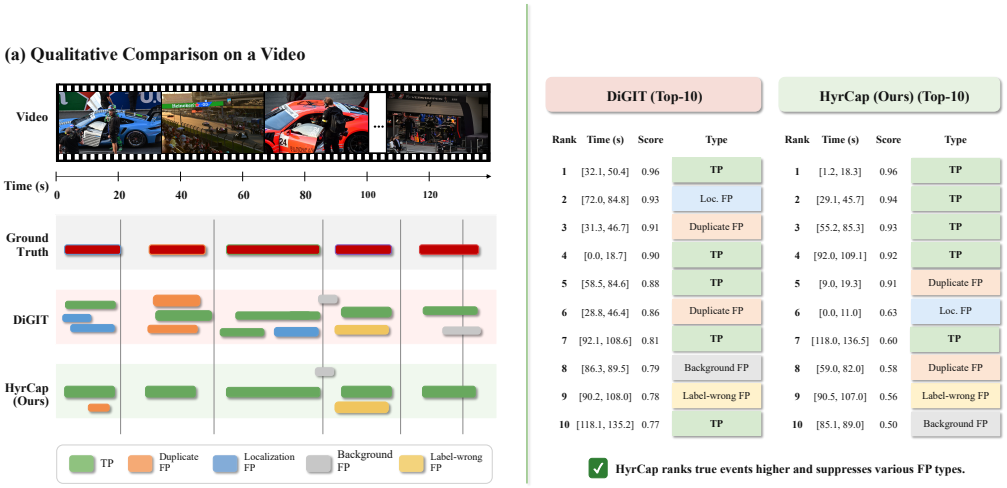

Figure 5: Qualitative visualization of proposal quality and ranking behavior.

On the top-ranked proposals (Figure 5), HyrCap consistently ranks true positives higher and suppresses various types of false positives, demonstrating that its cross-representation consensus and proposal calibration effectively improve both proposal quality and ranking reliability. This visual inspection highlights HyrCap's ability to produce temporally precise and stable event segments suitable for downstream semantic understanding.

6 Conclusion

In this paper, we proposed HyrCap, a proposal-centric hybrid rank-calibration framework for query-based temporal action detection and temporal event understanding. HyrCap treats dense query predictions as an over-complete temporal event hypothesis space and improves proposal quality through cross-representation consensus calibration, localization quality estimation, Proposal Redundancy Suppression, and temporal boundary refinement. Experiments on public TAD benchmarks show that HyrCap consistently improves strong query-based baselines, while ablation and qualitative results verify each component. Beyond closed-set detection, HyrCap further supports a staged semantic understanding protocol, where calibrated proposals serve as reliable event-level inputs for LLM-based classification, captioning, and reasoning. These results demonstrate the importance of high-quality temporal proposals for connecting TAD with higher-level video understanding.

Limitation and future work Although HyrCap consistently improves query-based TAD baselines, it still has several limitations. First, HyrCap is a proposal calibration framework and therefore depends on the quality of the base detector. If the base detector fails to generate candidates that cover true action instances, the missing events cannot be recovered by subsequent rank calibration or boundary refinement. Second, the cross-representation consensus module relies on auxiliary video representations to provide complementary temporal evidence. While this design improves proposal reliability, it may introduce additional feature extraction, storage, and inference cost. Reducing the dependence on auxiliary representations or learning more efficient consensus cues is an important direction for future work.

Third, the staged semantic understanding protocol depends on the selected event proposals, the trained projector, and the semantic model. Errors in temporal localization or visual-semantic projection may affect segment-level classification, captioning, and reasoning. Moreover, our current semantic evaluation is mainly built on predefined action categories and public TAD benchmarks, which does not fully cover open-vocabulary event discovery or complex long-form video reasoning. In the future, we will explore more adaptive proposal selection, more efficient proposal calibration, and stronger integration with open-vocabulary semantic models for temporal event understanding in more diverse real-world videos.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China under Grant 62403012, Grant 62673073 and Grant T2121002, and Zhongguancun Academy under Grant C20250503.

References

[1] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. 2025. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985* (2025).

[2] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. 2017. Soft-NMS–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*. 5561–5569.

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

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*. Springer, 213–229.

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

[6] Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Jiahao Wang, Zhe Chen, Zhiqi Li, Tong Lu, and Limin Wang. 2026. Video mamba suite: State space model as a versatile alternative for video understanding. *International Journal of Computer Vision* 134, 1 (2026), 20.

[7] Feng Cheng and Gedas Bertasius. 2022. Tallformer: Temporal action localization with a long-memory transformer. In *European Conference on Computer Vision*. Springer, 503–521.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6202–6211.

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

[11] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. 2018. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European conference on computer vision (ECCV)*. 784–799.

[12] Ho-Joong Kim, Jung-Ho Hong, Heejo Kong, and Seong-Whan Lee. 2024. 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*. 18837–18846.

[13] Ho-Joong Kim, Yearang Lee, Jung-Ho Hong, and Seong-Whan Lee. 2025. Digit: Multi-dilated gated encoder and central-adjacent region integrated decoder for temporal action detection transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 24286–24296.

[14] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. 2019. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3889–3898.

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

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.

[17] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 12585–12602.

[18] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junjie Yan, Changxin Gao, and Nong Sang. 2021. Temporal context aggregation network for temporal action proposal refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 485–494.

[19] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. 2023. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18857–18866.

[20] Dingfeng Shi, Yujie Zhong, Qiong Cao, Jing Zhang, Lin Ma, Jia Li, and Dacheng Tao. 2022. React: Temporal action detection with relational queries. In *European conference on computer vision*. Springer, 105–121.

[21] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. 2021. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing* 107 (2021), 104117.

[22] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7464–7473.

[23] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. 2021. Relaxed transformer decoders for direct action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 13526–13535.

[24] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6450–6459.

[25] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14549–14560.

[26] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*. Springer, 20–36.

[27] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. 2024. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*. Springer, 396–416.

[28] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. 2020. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10156–10165.

[29] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[30] Le Yang, Ziwei Zheng, Yizeng Han, Hao Cheng, Shiji Song, Gao Huang, and Fan Li. 2024. Dyfadet: Dynamic feature aggregation for temporal action detection. In *European conference on computer vision*. Springer, 305–322.

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

[32] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. 2019. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8668–8678.

[33] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. 2017. Temporal action detection with structured segment networks. In *Proceedings of the IEEE international conference on computer vision*. 2914–2923.

[34] Yuhan Zhu, Guozhen Zhang, Jing Tan, Gangshan Wu, and Limin Wang. 2024. Dual detrs for multi-label temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18559–18569.

Table 5: Benchmark statistics.

Dataset	Classes	Video splits	Action instances
THUMOS14	20	200 validation 213 test	3,007 validation 3,358 test
ActivityNet v1.3	200	10,024 training 4,926 validation	15,410 training 7,654 validation
HACS-Segment	200	37,613 training 5,981 validation	105,868 training 16,436 validation

A Implementation Details

A.1 Dataset Characteristics

We evaluate HyrCap on THUMOS14 [10], ActivityNet v1.3 [3], and HACS-Segment [32], following the standard data splits used in prior temporal action detection works. These benchmarks provide complementary evaluation settings in terms of dataset scale, action duration, category coverage, and instance density.

THUMOS14 contains relatively short and densely occurring actions in untrimmed videos. Repeated actions and substantial background make precise boundary localization, proposal ranking, and redundancy suppression challenging. ActivityNet v1.3 covers a broader category space, with actions generally longer and more sparsely distributed than those in THUMOS14, emphasizing semantic diversity and long-duration action modeling. HACS-Segment has the largest video count among the three benchmarks, supporting evaluation on larger-scale and more diverse video distributions.

Together, these datasets evaluate the effectiveness of proposal calibration across different temporal structures, category spaces, and dataset scales. Their statistics are summarized in Table 5.

A.2 Training Configuration

The base detectors are trained or evaluated following their official configurations, and HyrCap is applied to their generated proposal predictions.

During proposal calibration training, we freeze the video encoders and the base detectors, and train the Cross-Representation Consensus Calibration (CPCC), Localization Quality Estimation (LQE), and Proposal Redundancy Suppression (PRS) modules. For semantic projector training, we freeze the video encoders, the base detectors, and Qwen3-14B, and train only the projector.

Training is conducted on four NVIDIA A100-SXM4-80GB GPUs, each with 80 GB of memory. The calibration modules are trained for 50 epochs using the AdamW optimizer with a batch size of 4. The NMS threshold is set to 0.70, and Top-K is set to 900.

A.3 Measurement Settings

We use DiGIT as the baseline to measure the computational requirements of HyrCap. Measurements are performed on a single NVIDIA RTX 4090 GPU using pre-extracted video features, FP32 precision, and a batch size of one. Inference time for the detection and calibration stages is averaged over all THUMOS14 test videos, whose mean duration is 217.27 seconds.

B Robustness Analysis

B.1 Temporal-Structure Perturbations

To evaluate robustness to temporal-structure changes, we construct a temporally perturbed evaluation set by moving a segment from the latter part of each video to the beginning. This operation changes the ordering of the video content and its temporal context. We evaluate HyrCap on the clean and

Table 6: Performance under temporal-structure perturbations on THUMOS14.

Evaluation setting	**AP@0.3**	**AP@0.5**	**AP@0.7**	**Avg. mAP**
Clean set	88.1	79.1	54.6	75.6
Temporally perturbed set	85.9	77.0	53.1	73.3
Change	-2.2	-2.1	-1.5	-2.3

Table 7: Performance under proposal-level perturbations on THUMOS14.

Proposal setting	**Avg. mAP**	**Change**
Clean proposals	75.6	—
$\pm 5\%$ boundary jitter	73.9	-1.7
Random removal of 5% of proposals	74.0	-1.6

perturbed sets using the THUMOS14 detection metrics. Average mAP is computed over five tIoU thresholds from 0.3 to 0.7 with a step of 0.1.

As shown in Table 6, HyrCap achieves an average mAP of 73.3 on the perturbed set, a decrease of 2.3 points from the clean set. AP decreases by 2.2, 2.1, and 1.5 points at tIoU thresholds of 0.3, 0.5, and 0.7, respectively. These results show that HyrCap maintains stable detection performance under temporal rearrangement.

B.2 Proposal-Level Perturbations

We further evaluate robustness to imperfect proposal inputs using two perturbations: $\pm 5\%$ boundary jitter and random removal of 5% of proposals. Boundary jitter introduces temporal localization noise, while proposal removal reduces the available candidate evidence.

As shown in Table 7, HyrCap achieves average mAP values of 73.9 and 74.0 under boundary jitter and proposal removal, respectively, corresponding to decreases of 1.7 and 1.6 points. These results demonstrate robustness to proposal-boundary noise and missing proposals.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions of HyrCap, including hybrid rank calibration, proposal refinement, and staged semantic understanding. These claims are supported by the method design, ablation studies, standard TAD evaluation, and staged semantic evaluation.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the scope and limitations of HyrCap, including its reliance on pretrained video features, base query-based detectors, and the quality of generated proposals for downstream semantic understanding.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present theoretical theorems or formal proofs. The equations are used to define the proposed empirical framework and evaluation protocol.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper specifies the datasets, feature representations, base detectors, evaluation metrics, module designs, training settings, and post-processing procedures needed to reproduce the main results. Additional implementation details and configurations are provided in the experimental section and supplemental material.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments are conducted on public datasets and public feature/model resources. We will release anonymized code, configuration files, and reproduction scripts for the proposed modules and evaluation protocols.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper reports the datasets, splits, feature representations, base detectors, evaluation metrics, module settings, and ablation configurations used in the experiments. Training and inference details are included in the implementation details and supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: The paper reports standard TAD benchmark results following common practice, but does not include repeated-run error bars because training and evaluating multiple large video detection models is computationally expensive. The main claims are supported by consistent improvements across datasets, base detectors, and ablation settings.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides the main compute settings used for training and evaluation, including GPU-based training, pretrained feature extraction, and lightweight module training. Additional resource details are included in the implementation section or supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public video understanding benchmarks and publicly available pretrained models or features. The research is conducted for temporal action detection and event understanding, and does not involve private data collection or human-subject experiments.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential benefits for video retrieval, event annotation, and assistive video understanding. It also acknowledges that improved video understanding systems may be misused in surveillance or privacy-sensitive applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not introduce a new high-risk pretrained foundation model, image generator, or scraped dataset. It uses existing public datasets and pretrained resources for research evaluation.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original works for the datasets, pretrained video features, base detectors, and language model resources used in the experiments. We follow the corresponding licenses and terms of use for public assets such as THUMOS14, ActivityNet v1.3, HACS-Segment, I3D, InternVideo2, V-JEPA2, DiGIT, TE-TAD, TadTR, and Qwen3-14B.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces code, configuration files, calibration scripts, and evaluation scripts for HyrCap rather than a new dataset. These assets will be documented with usage instructions, environment requirements, and reproduction commands.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human-subject experiments, user studies, or newly collected human data, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper explicitly describes the use of Qwen3-14B in the staged semantic understanding protocol. The LLM is used for segment-level semantic prediction, captioning, or reasoning after proposal localization, and its confidence is not used for proposal ranking.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.