Decoupling Moment from Event for Video Temporal Grounding

Yuda Zou[1,2], Boxiang Zhou[2], Xin Zhou[2], Yibo Chen[2], Dejia Song[2], Xu Tang[2], Yao Hu[2], and Yongchao Xu[1] ✉

[1] School of Computer Science, Wuhan University
[2] Xiaohongshu Inc.
zouyuda@whu.edu.cn

Abstract. Video temporal grounding (VTG) aims to localize events in untrimmed videos given natural language prompts. Current VTG methods directly inherit the paradigm from 2D object detection, applying rigid IoU-based one-to-one matching between queries and ground truth spans. However, this paradigm fundamentally conflicts with the temporal fluidity of video events: unlike spatial objects, a sub-event (moment) can be semantically aligned with the prompt despite only covering part of the full event span. This matching scheme incorrectly suppresses such moment predictions as negatives, preventing models from learning rich temporal representations. To resolve this, we propose Moment-Event DETR (ME-DETR), a novel framework that embraces the natural part-whole structure of events through dynamic query specialization. During training, after standard Hungarian matching assigns primary event queries to ground truths, we identify other high-confidence predictions within each matched span and designate them as auxiliary moment queries. Through our synergistic supervision strategy, these moment queries are liberated from unreasonable suppression to explore semantically rich sub-events (moments). Extensive experiments demonstrate that ME-DETR establishes new state-of-the-art results, achieving +2.43% mAP improvement on QVHighlights test set without requiring post-processing NMS. Code: https://github.com/zouyuda220/ME-DETR.

Keywords: Video Temporal Grounding · Event Understanding

1 Introduction

Video temporal grounding (VTG) aims to localize the start and end timestamps of an event in an untrimmed video given a natural-language prompt. As a key primitive for video understanding, VTG underpins applications such as content retrieval, video editing, and human-centric interaction. Motivated by the success of DETR-style set prediction in 2D object detection on images [3], recent VTG methods increasingly adopt end-to-end transformer detectors that directly output a set of temporal spans from learnable queries [19–22]. With one-to-one Hungarian matching [9], these approaches avoid hand-crafted post-processing (*i.e.*, NMS), offering a clean and effective paradigm for VTG.

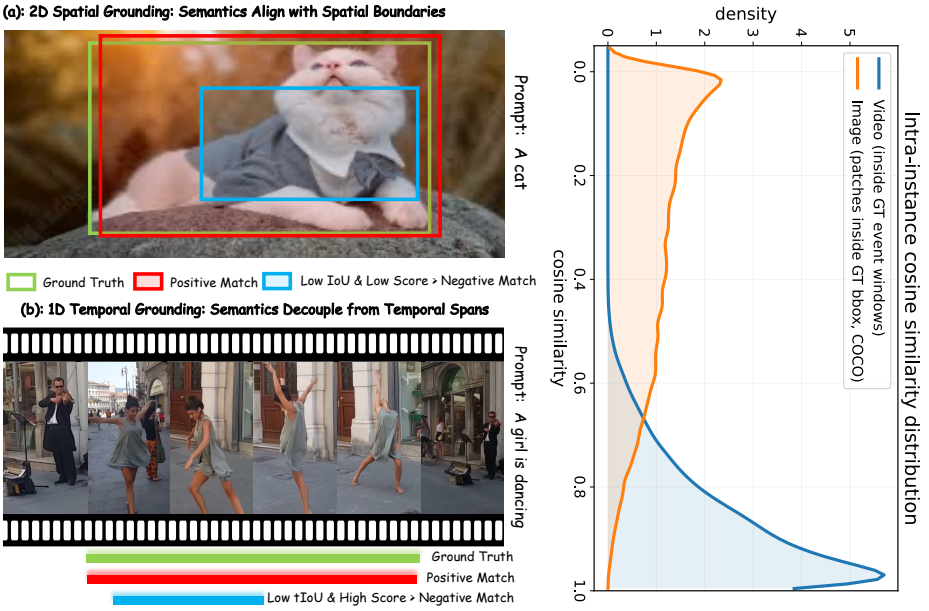

Fig. 1: Paradigm mismatch between 2D spatial grounding and 1D video temporal grounding (VTG). (a) In 2D detection, object semantics are typically coupled with spatial boundaries: a partial box (*e.g.*, *"cat body"*) is not semantically equivalent to the full object (*"a cat"*), hence IoU-based one-to-one matching reasonably suppresses low-IoU predictions. (b) In VTG, events often exhibit *temporal fluidity*: a salient sub-span (a *moment*) can still be strongly *semantically aligned* with the query while covering only part of the annotated event span. Under rigid tIoU-driven one-to-one matching, such semantically aligned but shorter predictions (*blue*) are labeled as negatives, inducing assignment-driven suppression during training. (c) **Empirical evidence of temporal redundancy.** We compare cosine similarity distributions between *(i)* full GT event spans and their randomly cropped sub-spans (QVHighlights, clip embeddings from InternVideo [29]) and *(ii)* full GT bounding boxes and their randomly cropped sub-boxes (COCO, patch embeddings from CLIP ViT-B/32 [23]). All full GTs and sub-regions are encoded independently from raw isolated inputs to ensure a fair comparison. The substantially higher intra-event similarity indicates stronger temporal redundancy, supporting that query-consistent semantics can be preserved by sub-spans even when boundaries do not fully match.

Despite these advances, directly migrating the DETR matching paradigm from 2D images to 1D temporal localization in videos overlooks a fundamental domain-specific discrepancy. As illustrated in Fig. 1, the relationship between *semantic alignment* and *boundary alignment* is fundamentally different in spatial 2D images and temporal 1D videos. In 2D object detection, semantics and spatial extent are usually tightly coupled: a box covering only a "cat body" (low IoU) is not semantically equivalent to the whole "cat", and is therefore a reasonable negative under IoU-based assignment. In contrast, temporal events are often

internally structured and semantically redundant over time. A short sub-span that contains the key action (*e.g.*, a dance move) can be fully consistent with the query (*e.g.*, a girl is dancing) even if it does not cover the entire annotated event. We refer to this property as temporal fluidity, where semantic alignment does not necessarily imply boundary alignment. As further supported in Fig. 1(c), clip representations within GT event windows exhibit markedly higher intra-instance similarity than patch representations within GT bounding boxes in images, indicating stronger temporal redundancy and making query-consistent sub-spans common in VTG.

This discrepancy incurs a fundamental training conflict in DETR-based VTG. Under one-to-one Hungarian matching [9], only a single query with high tIoU (temporal IoU) is assigned as the positive for each ground-truth event span. Other queries that are semantically correct but localize a sub-event (moment) tend to receive low tIoU and are consequently treated as negatives by the classification loss. Over training, these semantically aligned moment predictions are repeatedly penalized, forcing the model to distrust its own semantic evidence. This training scheme not only hinders the model from learning multi-granularity understanding of event content but also fundamentally constrains its representation capacity, which ultimately harms the final grounding accuracy.

To resolve this conflict while preserving end-to-end inference, we propose a novel framework **Moment-Event DETR (ME-DETR)**. ME-DETR follows a DETR-style set prediction paradigm, but introduces a principled query specialization with two parallel heads. First, we initialize queries with a pyramid-shaped span prior (multi-scale reference spans), where different query groups start from different temporal ranges (short-to-long) to cover diverse event durations. These queries are then refined by a standard transformer decoder. Crucially, we perform one-to-one Hungarian matching between predicted spans and GT spans only using the event head. Matched queries are designated as event queries and are optimized as positives in the event head (classification + boundary regression), while all remaining queries are treated as negatives in the event head, ensuring stable de-duplication and precise localization. In parallel, all queries are also fed into a moment head that predicts classification scores only. For each matched event query, we further mine several auxiliary moment queries from other queries whose predicted spans are highly contained within the event query span and have high moment-head confidence. In the moment head, both event queries and mined moment queries are treated as positives, while all other queries remain negatives. This design turns previously suppressed, query-consistent sub-spans into constructive semantic supervision, without interfering with the event head's boundary regression objective.

We evaluate ME-DETR on three standard benchmarks: QVHighlights [11], Charades-STA [5], and TACoS [24]. ME-DETR achieves consistent improvements over prior end-to-end VTG detectors. On the QVHighlights test set, ME-DETR improves Avg. mAP from 52.00% to 53.87% (+1.87%) over the strongest baseline with 1-layer feature. ME-DETR also provides notable gains on

Charades-STA and TACoS, demonstrating that explicitly modeling the part–whole structure of temporal events is broadly beneficial.

Our main contributions can be summarized as three-fold:

- We are the first to identify the forced semantic suppression problem in DETR-based VTG models, revealing its roots in the intrinsic conflict between the temporal fluidity of events and the rigid tIoU-based matching scheme.
- We propose Moment-Event DETR (ME-DETR), a novel framework with dual-head supervision and dynamic query specialization, that decouples moment semantics from event regression while preserving end-to-end inference.
- Extensive experiments show that ME-DETR consistently improves end-to-end VTG performance on QVHighlights, Charades-STA, and TACoS.

2 Related Work

2.1 Video Temporal Grounding

Video Temporal Grounding (VTG) aims to localize a specific event in an untrimmed video given a natural language prompt. Early approaches often followed a two-stage, proposal-based paradigm [1, 5, 6, 13–16, 25, 28, 30, 32, 34–37]. While effective, they may rely on hand-crafted components like Non-Maximum Suppression (NMS). To address these limitations, the field has been revolutionized by the end-to-end, NMS-free paradigm of DETR [3].

Moment-DETR [11] was a pioneering work in this direction, establishing a baseline for using a set of learnable queries and a transformer architecture to directly predict event spans. Subsequent works have built upon this foundation, introducing various refinements. For instance, QD-DETR [21] forces the model to better exploit the query's context by introducing a negative-pair learning strategy, where the model must learn to yield low saliency scores for the same video when paired with an irrelevant text query. Building on this, CG-DETR [20] refines the cross-modal interaction by employing dummy tokens to absorb attention from irrelevant video clips and a clip-word correlation learner to guide attention towards the most salient words within the query. Another direction seeks to enhance the representational power from different perspectives. MS-DETR [19] explicitly disentangles video features into motion and semantic streams within its encoder, and establishes a mutual collaborative mechanism between the moment retrieval and highlight detection tasks in the decoder. In contrast, BAM-DETR [10] tackles the inherent ambiguity of an event's center by proposing a novel boundary-oriented parameterization (anchor, distance-to-start, distance-to-end), which directly predicts boundaries from a flexible anchor point within the moment, making localization more robust.

Despite these architectural advancements, a fundamental challenge persists. Notably, some strong pipelines still report gains with post-processing such as NMS [22], suggesting that multiple temporally-overlapping predictions can be useful in practice. However, DETR-style one-to-one matching assigns supervision

to only one query per ground-truth span and penalizes other query-consistent predictions as negatives. This mismatch becomes particularly problematic under temporal fluidity, where salient sub-spans can preserve strong semantics despite boundary mismatch, giving rise to forced semantic suppression. Our work addresses this training-signal conflict with a principled role-specialized supervision, while keeping end-to-end inference.

2.2 Label Assignment of DETRs

Label assignment in DETR-based models traditionally relies on a one-to-one Hungarian matching [9]. While enabling NMS-free detection, this leads to sparse supervision and slower convergence. To address this, a line of work [4, 8, 38] in object detection has incorporated one-to-many supervision. Methods like Group DETR [4], H-DETR [8], and MS-DETR [38] assign multiple queries to each ground-truth object during training. Critically, these auxiliary positive queries share the same optimization objective as the primary ones: they are all supervised to regress towards the entire ground-truth box. Their purpose is to provide more gradient signals for the same target, thereby accelerating convergence.

While effective for rigid objects, this strategy is fundamentally misaligned with the temporal fluidity of video events. In contrast, our ME-DETR redefines the role of auxiliary supervision. We also identify additional positive samples—our moment queries—but with a crucial distinction: their optimization goal is different. Instead of forcing them to regress to the full event span, we only supervise their classification score in the moment head, explicitly exempting them from the regression loss. Their objective is to correctly identify semantically rich sub-spans (moments) within the ground truth event, while still being treated as negatives in the event head to ensure deduplication. In essence, while previous strategies offer quantitative enrichment (more supervision for the same target), our method provides qualitative enrichment (new supervision for different, valid sub-targets), making our label assignment strategy the first to be explicitly tailored to the part-whole nature of temporal events.

3 Methodology

To address the forced semantic suppression problem induced by rigid one-to-one tIoU matching, we propose the novel **Moment-Event DETR (ME-DETR)**, whose overall architecture is illustrated in Fig. 2. The core idea is to keep the event head as a strict NMS-free localizer (one-to-one positives only), while introducing an additional moment head to reward prompt-consistent sub-spans (moments) through a role-specialized training objective.

3.1 Video and Text Encoder

Following prior work [11,22], we adopt a frozen backbone (*e.g.*, InternVideo2 [29]) to extract per-clip video features and token-level text features, and project them

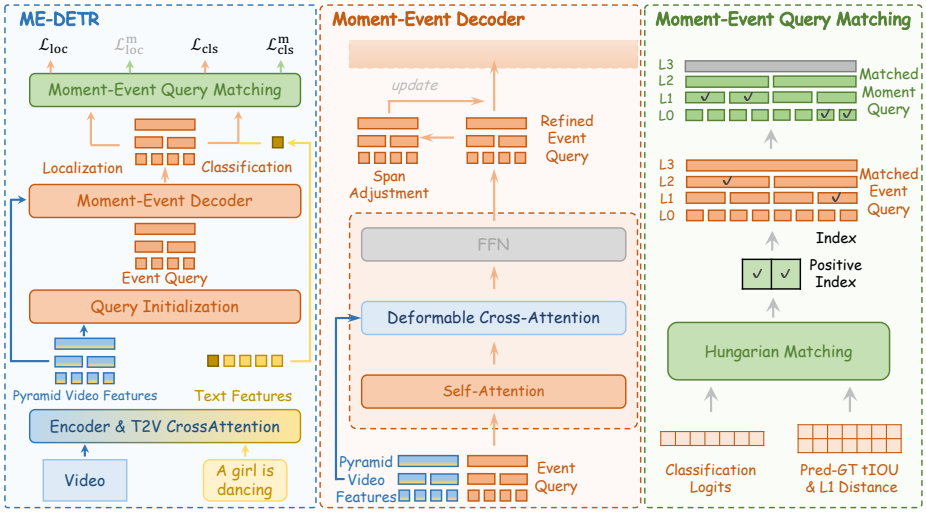

Fig. 2: The overall architecture of our proposed Moment-Event DETR (ME-DETR). We introduce parallel *Event Queries* and *Moment Queries*, which are synergistically supervised to achieve both precise boundary regression and rich semantic understanding.

into a common hidden dimension d with MLPs. We further build a prompt-conditioned video representation and construct a multi-scale temporal feature pyramid $\{M_l\}_{l=0}^{L-1}$ to support events of diverse durations.

3.2 Pyramid Query Initialization

Unlike using a fixed set of learnable query embeddings, we initialize queries with a multi-scale span prior coupled with the temporal pyramid (in the spirit of pyramid-shaped priors used in R^2-Tuning [17] and FlashVTG [2]). Each query q_i consists of a content embedding $z_i \in \mathbb{R}^d$ and a reference span $r_i = (c_i, w_i)$ parameterized by center and width. We generate reference spans densely from the temporal pyramid: each temporal position on feature map M_l serves as a candidate center, and its default width is associated with the feature level l (shorter at high-resolution levels, longer at low-resolution levels). This produces a diverse set of initial spans covering a wide range of locations and durations. For simplicity, we initialize all content embeddings z_i to zeros and let the decoder progressively refine them.

3.3 Moment-Event Decoder with Dual Heads

We employ a standard transformer decoder with N layers. Each layer contains self-attention, deformable cross-attention [40], and a feed-forward network (FFN), iteratively refining query embeddings. After each decoder layer, each

query produces two predictions through two parallel heads: (i) an **event head** that outputs a classification score c_i^e and a span refinement Δs_i^e (used to update the reference span for the next layer); (ii) a **moment head** that outputs a classification score c_i^m only (no regression). Here, $c_i^e \in [0,1]$ denotes the confidence that the predicted span tightly matches a full GT event, while $c_i^m \in [0,1]$ denotes the confidence that the predicted span is semantically aligned with the prompt (regardless of whether it covers the full GT span or only a sub-span). This dual-head design explicitly decouples precise boundary regression (event head) from broad semantic exploration (moment head), and enables deep supervision across decoder layers.

3.4 Moment-Event Query Role Assignment

A key component of ME-DETR is a training-time role assignment strategy that defines which queries are treated as positives for each head.

Event query assignment (event head). We perform one-to-one Hungarian matching [9] between predictions and ground-truth spans using **only the event head outputs**. The matching cost between the i-th query prediction and the j-th ground truth is:

$$\mathcal{C}_{\text{match}}(i,j) = -\lambda_{\text{cls}} c_i^e + \lambda_{\text{L1}} \|s_i - \hat{s}_j\|_1 - \lambda_{\text{tIoU}} \text{tIoU}(s_i, \hat{s}_j), \tag{1}$$

where s_i is the predicted span from the event head, and $\hat{s}_j$ is the GT span. Matched queries form the set $\mathcal{S}_e$ and are designated as **event queries**. In the event head, **only event queries are positives**; all other queries are negatives.

Moment query mining (moment head). In parallel, all queries are fed to the moment head for classification-only prediction. For each event query $q_i \in \mathcal{S}_e$ with predicted span s_i, we mine a small set of **moment queries** from the unmatched queries. Specifically, for any unmatched query $q_k \notin \mathcal{S}_e$ with predicted span s_k, we compute its *containment degree* w.r.t. the event span:

$$\text{Contain}(s_k, s_i) = \frac{|s_k \cap s_i|}{|s_k|}, \tag{2}$$

where $|\cdot|$ denotes temporal length. We collect candidates with high containment, *i.e.*, $\text{Contain}(s_k, s_i) \geq \tau_{\text{in}}$, and then select those with the highest moment-head confidence c_k^m. We further filter candidates by a confidence threshold $c_k^m \geq \tau_{\text{cls}}$ before selecting top-K by confidence. Concretely, we pick at most top-K candidates per event query, where we use $K = 2$ in all experiments. The union of selected candidates over all event queries forms the moment query set $\mathcal{S}_m$. In the moment head, **both event queries $\mathcal{S}_e$ and moment queries $\mathcal{S}_m$ are treated as positives**; all remaining queries are negatives.

3.5 Training Objective

The total loss is a weighted sum of the event-head loss and moment-head loss:

$$\mathcal{L} = \mathcal{L}_{\text{event}} + \lambda_m \mathcal{L}_{\text{moment}}. \tag{3}$$

Table 1: Key configuration of ME-DETR. Most settings follow prior DETR-style VTG models.

Hyperparameter	Value
General	
Hidden dimension (d)	256
Decoder layers (N)	4
Feature pyramid levels (L)	5
Batch size	32
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Epochs	200
Loss and Matching	
Classification loss weight (λ_{cls})	30
L1 loss weight (λ_{L1})	10
tIoU loss weight (λ_{tIoU})	1
ME-DETR Specific	
Moment loss weight (λ_m)	0.2
Containment threshold (τ_{in})	0.7
Moment query confidence threshold (τ_{cls})	0.3
Max moment queries per event (K)	2

Event head supervision. The event head is trained as a strict localizer. We apply a focal classification loss over all queries with label $y_i^e = 1$ if $i \in \mathcal{S}_e$ and 0 otherwise, and apply regression loss only to matched event queries:

$$\begin{aligned}\mathcal{L}_{\text{event}} =& \frac{1}{N_q}\sum_{i=1}^{N_q}\lambda_{\text{cls}}\mathcal{L}_{\text{cls}}(c_i^e, y_i^e) \\ &+ \frac{1}{|\mathcal{S}_e|}\sum_{i\in\mathcal{S}_e}\Big(\lambda_{\text{L1}}\|s_i - \hat{s}_{\sigma(i)}\|_1 + \lambda_{\text{tIoU}}\mathcal{L}_{\text{tIoU}}(s_i, \hat{s}_{\sigma(i)})\Big),\end{aligned} \tag{4}$$

where $\sigma(i)$ denotes the GT index matched to query i.

Moment head supervision. The moment head is trained as a semantic explorer with classification-only supervision. The moment label $y_i^m \in \{0, 1\}$ indicates whether query i is regarded as *semantically aligned* with the prompt: we set $y_i^m = 1$ if $i \in \mathcal{S}_e \cup \mathcal{S}_m$ (*i.e.*, either a matched full-event query or a mined sub-event moment query) and 0 otherwise:

$$\mathcal{L}_{\text{moment}} = \frac{1}{N_q}\sum_{i=1}^{N_q}\mathcal{L}_{\text{cls}}(c_i^m, y_i^m). \tag{5}$$

This role-specialized supervision encourages queries to capture query-consistent sub-spans in the moment head, while preventing them from introducing conflicting regression signals in the event head.

3.6 Inference

All role assignment and moment query mining are used *only during training*. At inference time, ME-DETR is identical to a standard DETR-style VTG model:

Table 2: Results on the QVHighlights Validation Set with Clip & SlowFast as backbone. Bold stands for the best.

Method	Venue	Feature Layers	Moment Retrieval				
			R1		mAP		
			@0.5	@0.7	@0.5	@0.75	Avg.
R^2-Tuning [17]	ECCV'24	4	**68.71**	52.06	–	–	47.59
SDST† [22]	ICCV'25	4	67.42	**54.28**	66.14	51.31	**49.78**
Moment-DETR [11]	NIPS'21	1	53.94	34.84	–	–	32.20
UMT [18]	CVPR'22	1	60.26	44.26	56.70	39.90	38.59
UniVTG [12]	ICCV'23	1	59.74	–	–	–	36.13
MH-DETR [33]	IJCNN'24	1	60.84	44.90	60.76	39.64	39.26
QD-DETR [21]	CVPR'23	1	62.68	46.66	62.23	41.82	41.22
EaTR [7]	ICCV'23	1	61.36	45.79	61.86	41.91	41.74
CG-DETR [20]	PR'26	1	67.35	52.06	65.57	45.73	44.93
TR-DETR [27]	AAAI'24	1	67.10	51.48	66.27	46.42	45.09
BAM-DETR [10]	ECCV'24	1	65.10	51.61	65.41	48.56	47.61
MS-DETR [19]	ACMMM'25	1	66.90	51.68	67.19	46.05	46.00
Flash-VTG [2]	WACV'25	1	**69.03**	**54.06**	**68.44**	52.12	49.85
TD-DETR [39]	ICCV'25	1	64.58	51.87	66.85	49.25	48.48
ME-DETR (Ours)	–	1	65.61	52.58	66.80	**52.99**	**51.62(+1.77)**

we output the final-layer predictions from the event head without any post-processing such as NMS.

4 Experiments

In this section, we conduct comprehensive experiments to validate the effectiveness of our proposed Moment-Event DETR (ME-DETR). We first introduce the experimental settings, including datasets, evaluation metrics, and implementation details. Then, we compare ME-DETR with existing state-of-the-art methods. Finally, we perform extensive ablation studies and qualitative analysis.

4.1 Experimental Settings

Datasets. We evaluate our method on three widely-used VTG datasets.
QVHighlights [11] is the most recently proposed dataset for video temporal grounding. A key feature is that a single query can correspond to multiple disjoint moments, making it a practical and challenging benchmark.
Charades-STA [5] is a widely-used benchmark derived from the Charades [26] dataset. The dataset consists of 9,848 videos, which depict indoor activities on average 30 seconds long.
TACoS [24] is composed of 127 long videos sourced from cooking scenarios.

Evaluation Metrics. Following the standard protocols in VTG, we report Recall@1 (R1) at different temporal Intersection over Union (tIoU) thresholds

Table 3: Results on the QVHighlights Validation Set with InternVideo as backbone. Bold stands for the best. † represents NMS-free version.

Method	Venue	Feature Layers	Moment Retrieval				
			R1		mAP		
			@0.5	@0.7	@0.5	@0.75	Avg.
R^2-Tuning [17]	ECCV'24	4	-	-	**71.40**	53.79	51.49
SDST† [22]	ICCV'25	4	73.68	60.90	70.55	**56.00**	**54.27**
Moment-DETR [11]	NIPS'21	1	60.39	40.58	59.73	35.77	35.71
EaTR [7]	ICCV'23	1	64.9	48.0	64.77	42.97	42.46
QD-DETR [21]	CVPR'23	1	68.71	53.29	68.63	47.58	46.93
CG-DETR [20]	PR'26	1	69.55	54.06	69.02	49.39	48.33
TR-DETR [27]	AAAI'24	1	71.94	56.06	70.23	49.77	48.96
BAM-DETR [10]	ECCV'24	1	68.71	55.42	69.69	53.71	51.76
MS-DETR [19]	ACMMM'25	1	69.86	54.58	68.99	49.75	50.08
Flash-VTG [2]	WACV'25	1	**73.10**	**57.29**	**72.75**	54.33	52.84
TD-DETR [39]	ICCV'25	1	69.68	56.0	72.04	54.51	52.56
ME-DETR (Ours)	–	1	70.71	57.16	71.18	**56.88**	**55.41(+2.57)**

(*e.g.*, 0.5, 0.7). A prediction is considered correct if its tIoU with the ground-truth is above the threshold. We also report the mean Average Precision (mAP) over a range of tIoU thresholds [0.5:0.05:0.95] on QVHighlights [11], and mean IoU (mIoU) on Charades-STA and TACoS. High performance on stricter metrics like R1@0.7 indicates better localization accuracy.

Implementation Details. For fair comparison with recent state-of-the-art methods [19, 22], we adopt a frozen InternVideo [29] (or CLIP+SlowFast) as the video encoder and the corresponding text encoder to extract features. Unless otherwise specified, we use hidden dimension d=256, a L=5-level temporal feature pyramid, and a N=4-layer decoder. Both the event head and the moment head are implemented as lightweight 3-layer MLPs. The model is trained end-to-end using the AdamW optimizer with a fixed learning rate of 1e−4 and weight decay of 1e−4. For the matching and event-head losses, we keep λ_{cls}=30, λ_{L1}=10, and λ_{tIoU}=1. Our role specialization introduces four extra hyperparameters: the moment-head loss weight λ_m, the maximum number of mined moment queries per event K, the containment threshold τ_{in}, and the moment query confidence threshold τ_{cls}. In all experiments we set λ_m=0.2, K=2, τ_{in}=0.7, and τ_{cls}=0.3; a complete list of hyperparameters is summarized in Tab. 1. Importantly, the moment head and moment-query mining are used *only during training*; at inference, ME-DETR is identical to a standard DETR-style VTG model that outputs event-head predictions without NMS, incurring negligible additional inference cost. All experiments are conducted on one NVIDIA H800 GPU.

Table 4: Results on the QVHighlights Test Set. Bold stands for the best.

Method	Venue	Backbone	Feature Layers	Moment Retrieval				
				R1		mAP		
				@0.5	@0.7	@0.5	@0.75	Avg.
R^2-Tuning [17]	ECCV'24	Clip & SlowFast	4	68.03	49.35	69.04	47.56	46.17
Moment-DETR [11]	NIPS'21	Clip & SlowFast	1	52.89	33.02	54.82	29.40	30.73
UMT [18]	CVPR'22	Clip & SlowFast	1	56.23	41.18	53.83	37.01	36.12
UniVTG [12]	ICCV'23	Clip & SlowFast	1	58.86	40.86	57.60	35.59	35.47
MH-DETR [33]	IJCNN'24	Clip & SlowFast	1	60.05	42.48	60.75	38.13	38.38
QD-DETR [21]	CVPR'23	Clip & SlowFast	1	62.40	44.98	62.52	39.88	39.86
CG-DETR [20]	PR'26	Clip & SlowFast	1	65.43	48.38	64.51	42.77	42.86
TR-DETR [27]	AAAI'24	Clip & SlowFast	1	64.66	48.96	63.98	43.73	42.62
UVCOM [31]	CVPR'24	Clip & SlowFast	1	63.55	48.70	64.47	44.01	43.27
BAM-DETR [10]	ECCV'24	Clip & SlowFast	1	62.71	48.64	64.57	46.33	45.36
MS-DETR [19]	ACMMM'25	Clip & SlowFast	1	64.72	48.77	66.41	44.91	44.89
Flash-VTG [2]	WACV'25	Clip & SlowFast	1	66.08	50.00	67.99	48.70	47.59
TD-DETR [39]	ICCV'25	Clip & SlowFast	1	64.53	50.37	66.21	47.32	46.69
ME-DETR (Ours)	–	Clip & SlowFast	1	**66.17**	**51.22**	**68.30**	**52.09**	**50.02(+2.43)**
SDST [22]	ICCV'25	InternVideo	4	70.82	56.23	71.31	54.99	53.31
TD-DETR [39]	ICCV'25	InternVideo	1	69.11	53.98	71.09	52.42	51.43
Flash-VTG [2]	WACV'25	InternVideo	1	**70.69**	53.96	**72.33**	53.85	52.00
ME-DETR (Ours)	–	InternVideo	1	70.10	**55.13**	71.41	**55.52**	**53.87(+1.87)**

4.2 Comparison with State-of-the-Art

Results on QVHighlights [11]. We compare ME-DETR against recent state-of-the-art methods on the QVHighlights dataset, with results presented for the validation set in Tab. 2 and Tab. 3, and the official test set in Tab. 4. For a fair comparison, we explicitly group the methods based on their backbone (CLIP & SlowFast vs. InternVideo) and the number of feature layers utilized (1-layer vs. 4-layer). Our ME-DETR strictly operates on single-layer (1) features, maintaining an efficient, fully end-to-end framework.

The results consistently highlight the superiority of our approach across all settings. On the validation set with CLIP & SlowFast (Tab. 2), ME-DETR achieves an Avg. mAP of 51.62%, outperforming Flash-VTG by +1.77% under the same 1-layer constraint, and surpassing the 4-layer SDST (49.78%). This advantage is further amplified when using the InternVideo backbone (Tab. 3), where our method reaches a remarkable 55.41% Avg. mAP. This translates to a significant gain of +2.57% over Flash-VTG and a clear margin over the advanced SDST [22] with four feature layers (54.27%).

A similar and definitive trend is observed on the unseen official test set (Tab. 4). With the CLIP & SlowFast backbone, ME-DETR achieves 50.02% in Avg. mAP, securing a +2.43% improvement over Flash-VTG. Using the InternVideo backbone, our method establishes a new state-of-the-art of 53.87%, consistently beating both recent 1-layer counterparts like TD-DETR (51.43%) and 4-layer frameworks like SDST (53.31%). This substantial and consistent improvement on Avg. mAP—the strictest localization metric—directly validates our core hypothesis: By allowing the model to correctly learn from semantically rich sub-events via our synergistic supervision, the primary event queries are

Table 5: Results on Charades-STA and TACoS datasets. Bold stands for the best results.

Method	Venue	Backbone	Feature Layers	Charades-STA			TACoS		
				R@0.5	R@0.7	mIoU	R@0.5	R@0.7	mIoU
R^2-Tuning [17]	ECCV'24	Clip & SlowFast	4	59.8	37.0	50.9	38.7	25.1	35.9
Moment-DETR [11]	NIPS'21	Clip & SlowFast	1	52.1	30.6	45.5	24.7	12.0	25.5
UMT [18]	CVPR'22	Clip & SlowFast	1	48.3	29.3	-	-	-	-
UniVTG [12]	ICCV'23	Clip & SlowFast	1	58.0	35.7	50.1	35.0	17.4	33.6
QD-DETR [21]	CVPR'23	Clip & SlowFast	1	57.3	32.6	-	36.8	21.1	35.8
CG-DETR [20]	PR'26	Clip & SlowFast	1	58.4	36.3	50.1	39.6	22.2	36.5
BAM-DETR [10]	ECCV'24	Clip & SlowFast	1	60.0	39.4	52.3	41.5	26.8	39.3
TR-DETR [27]	AAAI'24	Clip & SlowFast	1	57.6	33.5	-	-	-	-
MS-DETR [19]	ACMMM'25	Clip & SlowFast	1	59.6	36.5	50.6	39.7	23.4	37.0
TD-DETR [39]	ICCV'25	Clip & SlowFast	1	60.9	40.4	-	-	-	-
ME-DETR (Ours)	–	Clip & SlowFast	1	**62.6**	**44.5**	**53.2**	**42.3**	**30.7**	**41.6**
R^2-Tuning [17]	ECCV'24	InternVideo	4	68.2	46.3	58.1	38.0	25.3	35.4
SDST [22]	ICCV'25	InternVideo	4	**72.0**	**52.6**	**61.2**	**44.5**	**32.3**	**42.2**
FlashVTG [2]	WACV'25	InternVideo	1	70.3	49.9	–	41.8	24.7	37.6
ME-DETR (Ours)	–	InternVideo	1	**71.8**	**53.2**	**61.9**	**44.9**	**33.2**	**43.0**

freed from conflicting signals, enabling more precise temporal boundary regression.

Table 6: Ablation on the Moment-Query mechanism.

Method	R1		mAP		
	@0.5	@0.7	@0.5	@0.75	Avg.
(a) Standard	70.45	56.22	69.53	53.68	52.83
(b) One-to-Many	**70.91**	56.69	69.92	54.21	53.35
(c) **ME-DETR (Ours)**	70.71	**57.16**	**71.18**	**56.88**	**55.41**

Results on Charades-STA [5] and TACoS [24]. To demonstrate the generalizability of our approach, we further evaluate ME-DETR on Charades-STA and TACoS, with results shown in Tab. 5. Our method consistently establishes new state-of-the-art performance, particularly on the main comparative metric of mean IoU (mIoU). On Charades-STA, we achieve an mIoU of 53.2% under the standard CLIP & SlowFast setting, outperforming the strong baseline BAM-DETR [10] (52.3%). When equipped with the InternVideo backbone, ME-DETR reaches a dominant 61.9% mIoU, outstripping even the 4-layer SDST model (61.2%). On the more challenging TACoS dataset, known for its long, complex cooking videos, ME-DETR also sets a new SOTA. It reaches 41.6% with CLIP & SlowFast and 43.0% with InternVideo, again confirming our superiority over multi-layer architectures like SDST (42.2%). These consistent gains validate that our synergistic supervision mechanism robustly improves localization accuracy across diverse data distributions.

Table 7: Ablation on τ_{cls}.

τ_{cls}	R1		mAP		
	@0.5	@0.7	@0.5	@0.75	Avg.
0.2	70.18	57.12	**71.25**	55.93	55.10
0.3	**70.71**	57.16	71.18	**56.88**	**55.41**
0.4	70.31	**57.25**	71.02	56.21	55.09

Table 8: Ablation on τ_{in}.

τ_{in}	R1		mAP		
	@0.5	@0.7	@0.5	@0.75	Avg.
0.5	70.47	56.81	70.92	56.18	55.09
0.6	70.63	57.03	71.08	56.55	55.28
0.7	**70.71**	**57.16**	**71.18**	**56.88**	**55.41**
0.8	70.52	56.94	70.96	56.34	55.16

Table 9: Ablation on K.

K	R1		mAP		
	@0.5	@0.7	@0.5	@0.75	Avg.
0	70.45	56.22	69.53	53.68	52.83
1	70.63	56.94	70.96	56.07	55.08
2	**70.71**	**57.16**	**71.18**	**56.88**	**55.41**
3	70.59	57.02	71.04	56.39	55.17
5	70.31	56.73	70.77	55.84	54.86

Table 10: Ablation on λ_m.

λ_m	R1		mAP		
	@0.5	@0.7	@0.5	@0.75	Avg.
0.0	70.45	56.22	69.53	53.68	52.83
0.1	70.61	56.98	70.97	56.16	55.14
0.2	**70.71**	**57.16**	**71.18**	**56.88**	**55.41**
0.5	70.47	56.90	70.85	56.02	54.98
1.0	70.18	56.51	70.39	55.34	54.47

4.3 Ablation Studies

We conduct a series of ablation studies on the QVHighlights validation split with InternVideo backbone to thoroughly analyze the effectiveness of our proposed components and design choices.

Effectiveness of Moment-Event Synergistic Supervision. To validate our core contribution, we ablate our Moment-Event synergistic supervision strategy in Tab. 6. We compare three variants: (a) standard one-to-one matching; (b) One-to-Many approach where all positive queries (Event and Moment) are forced to regress the full GT span, mimicking object detection paradigms [8, 38]; and (c) our full ME-DETR with Moment-Event synergistic supervision.

The results clearly show the inadequacy of migrating 2D spatial paradigms directly to the temporal domain. The naive one-to-many strategy (b) yields only minor gains (+0.52% on Avg. mAP), as forcing semantically correct sub-event (moment) predictions to artificially expand and match the full event span introduces conflicting localization signals. In stark contrast, our ME-DETR (c) achieves substantial improvements, boosting the primary mAP Avg. metric by +2.58% (from 52.83% to 55.41%) over the standard. This powerfully confirms our central hypothesis: by decoupling the optimization objectives and allowing Moment Queries to focus strictly on semantic comprehension without strict boundary suppression, our synergistic supervision correctly models the natural composition of events, resolving the underlying training conflict and unlocking superior localization accuracy.

Impact of Hyperparameters. We analyze the four hyperparameters introduced by our role-specialized design: the Moment Query confidence threshold τ_{cls}, the containment threshold τ_{in}, the maximum number of mined Moment

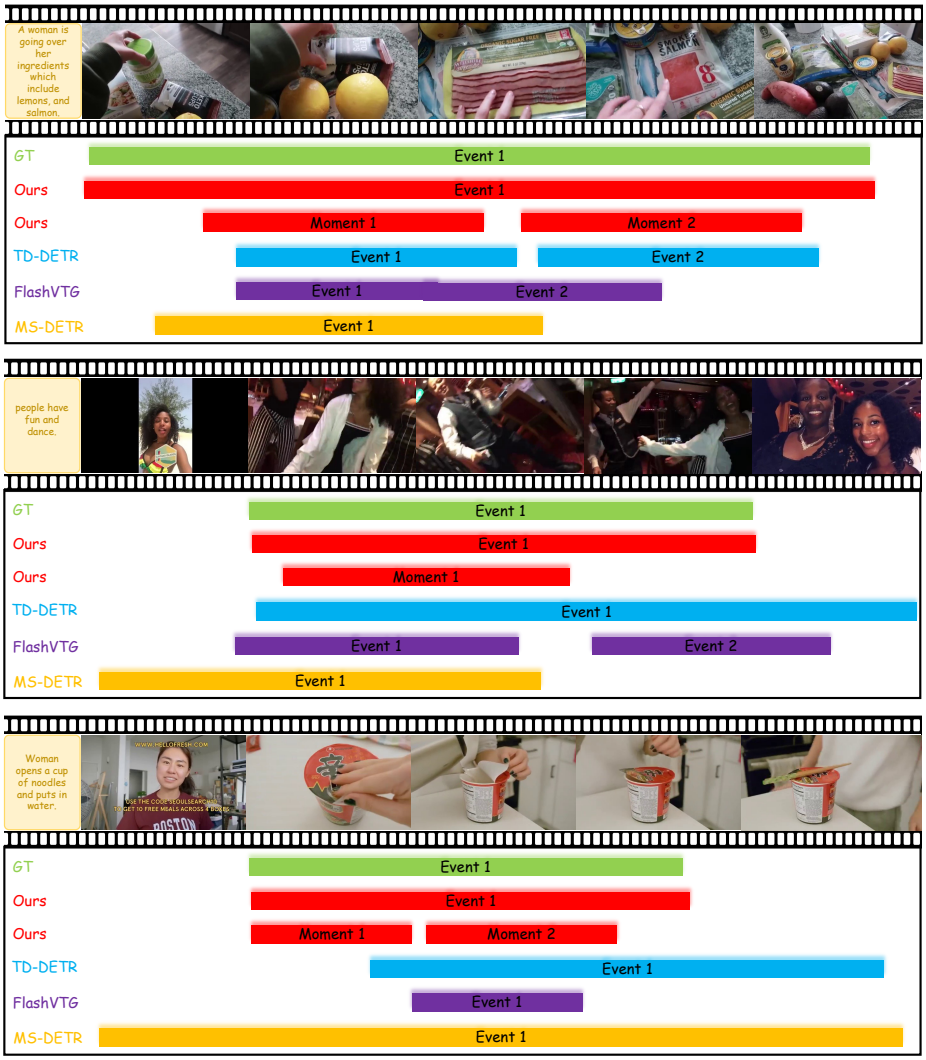

Fig. 3: Qualitative comparison with other advanced methods. Our model successfully generates a precise event query prediction for the full span, while also identifying semantically correct moment queries (*red boxes*) for sub-spans. In contrast, other methods suffering from forced semantic suppression often produce hesitant, shorter predictions despite identifying the correct content.

Queries per event K, and the Moment Head loss weight λ_m. As shown in Tab. 7–Tab. 10, ME-DETR is robust within a reasonable range, and we adopt τ_{cls}=0.3, τ_{in}=0.7, K=2, and λ_m=0.2 as the default.

4.4 Qualitative Analysis

Fig. 3 provides an intuitive visualization of our method's effectiveness. Other methods, suffering from forced semantic suppression, consistently produce short, fragmented, or low-confidence predictions even for semantically correct content. In contrast, our ME-DETR confidently localizes the entire event span (red). This superior performance is achieved by leveraging our auxiliary moment queries (red), which are liberated to identify key semantic sub-events (*e.g.*, the specific "go over lemons" or "go over salmon"). These learned moments act as strong contextual anchors, enabling the primary event query to achieve a more complete and accurate grounding. This vividly demonstrates that our synergistic supervision strategy fundamentally changes the model's behavior to better understand the natural part-whole structure of temporal events.

5 Conclusion

In this paper, we identified and addressed a fundamental, yet overlooked challenge in DETR-based video temporal grounding: the rigid IoU-based matching paradigm inherited from 2D object detection incorrectly suppresses semantically aligned moment predictions as negatives, preventing models from learning rich temporal representations. We propose Moment-Event DETR (ME-DETR), a novel framework that resolves this conflict by embracing the natural part-whole structure of events through dynamic query specialization. During training, our synergistic supervision strategy liberates auxiliary moment queries to explore semantically rich sub-events (moments) without suppression, while primary event queries focus on precise event localization. Extensive experiments demonstrate that ME-DETR establishes new state-of-the-art results across three major benchmarks, achieving up to +2.43% Avg. mAP improvement on QVHighlights test set without complex post-processing. Beyond these results, we hope this work reveals the fundamental paradigm mismatch between spatial and temporal grounding, and inspires more principled strategies for video understanding.

Acknowledgements

This work was partially supported by "Prevention and Control of Emerging and Major Infectious Diseases-National Science and Technology Major Project" (2025ZD01907900), the National Natural Science Foundation of China under Grant 62222112, and Grant 62176186, the Innovative Research Group Project of Hubei Province under Grant 2024AFA017.

References

1. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Int. Conf. Comput. Vis. pp. 5803–5812 (2017)

2. Cao, Z., Zhang, B., Du, H., Yu, X., Li, X., Wang, S.: Flashvtg: Feature layering and adaptive score handling network for video temporal grounding. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 9226–9236. IEEE (2025)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Eur. Conf. Comput. Vis. pp. 213–229. Springer (2020)
4. Chen, Q., Chen, X., Wang, J., Zhang, S., Yao, K., Feng, H., Han, J., Ding, E., Zeng, G., Wang, J.: Group detr: Fast detr training with group-wise one-to-many assignment. In: Int. Conf. Comput. Vis. pp. 6633–6642 (2023)
5. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Int. Conf. Comput. Vis. pp. 5267–5275 (2017)
6. Ge, R., Gao, J., Chen, K., Nevatia, R.: Mac: Mining activity concepts for language-based temporal localization. In: 2019 IEEE winter conference on applications of computer vision (WACV). pp. 245–253. IEEE (2019)
7. Jang, J., Park, J., Kim, J., Kwon, H., Sohn, K.: Knowing where to focus: Event-aware transformer for video grounding. In: Int. Conf. Comput. Vis. (2023)
8. Jia, D., Yuan, Y., He, H., Wu, X., Yu, H., Lin, W., Sun, L., Zhang, C., Hu, H.: Detrs with hybrid matching. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19702–19712 (2023)
9. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly **2**(1-2), 83–97 (1955)
10. Lee, P., Byun, H.: Bam-detr: Boundary-aligned moment detection transformer for temporal sentence grounding in videos. In: Eur. Conf. Comput. Vis. pp. 220–238. Springer (2024)
11. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. In: Adv. Neural Inform. Process. Syst. vol. 34, pp. 11846–11858 (2021)
12. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: Int. Conf. Comput. Vis. pp. 2794–2804 (2023)
13. Liu, D., Qu, X., Dong, J., Zhou, P., Cheng, Y., Wei, W., Xu, Z., Xie, Y.: Context-aware biaffine localizing network for temporal sentence grounding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11235–11244 (2021)
14. Liu, D., Qu, X., Liu, X.Y., Dong, J., Zhou, P., Xu, Z.: Jointly cross-and self-modal graph attention network for query-based moment localization. In: ACM Int. Conf. Multimedia. pp. 4070–4078 (2020)
15. Liu, M., Wang, X., Nie, L., He, X., Chen, B., Chua, T.S.: Attentive moment retrieval in videos. In: The 41st international ACM SIGIR conference on research & development in information retrieval. pp. 15–24 (2018)
16. Liu, M., Wang, X., Nie, L., Tian, Q., Chen, B., Chua, T.S.: Cross-modal moment localization in videos. In: ACM Int. Conf. Multimedia. pp. 843–851 (2018)
17. Liu, Y., He, J., Li, W., Kim, J., Wei, D., Pfister, H., Chen, C.W.: r 2-tuning: Efficient image-to-video transfer learning for video temporal grounding. In: Eur. Conf. Comput. Vis. pp. 421–438. Springer (2024)
18. Liu, Y., Li, S., Wu, Y., Chen, C.W., Shan, Y., Qie, X.: Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3042–3051 (2022)
19. Ma, H., Wang, G., Yu, F., Jia, Q., Ding, S.: Ms-detr: Towards effective video moment retrieval and highlight detection by joint motion-semantic learning. In:

ACM Int. Conf. Multimedia. p. 4514–4523. Association for Computing Machinery (2025)
20. Moon, W., Hyun, S., Lee, S., Heo, J.P.: Correlation-guided query-dependency calibration in video representation learning for temporal grounding. Pattern Recognition p. 112984 (2026)
21. Moon, W., Hyun, S., Park, S., Park, D., Heo, J.P.: Query-dependent video representation for moment retrieval and highlight detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 23023–23033 (2023)
22. Pujol-Perich, D., Escalera, S., Clapés, A.: Sparse-dense side-tuner for efficient video temporal grounding. In: Int. Conf. Comput. Vis. pp. 21515–21524 (2025)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. pp. 8748–8763. PmLR (2021)
24. Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., Pinkal, M.: Grounding action descriptions in videos. Transactions of the Association for Computational Linguistics **1**, 25–36 (2013)
25. Shao, D., Xiong, Y., Zhao, Y., Huang, Q., Qiao, Y., Lin, D.: Find and focus: Retrieve and localize video events with natural language queries. In: Eur. Conf. Comput. Vis. pp. 200–216 (2018)
26. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Eur. Conf. Comput. Vis. pp. 510–526. Springer (2016)
27. Sun, H., Zhou, M., Chen, W., Xie, W.: Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In: AAAI. pp. 4998–5007 (2024)
28. Wang, J., Ma, L., Jiang, W.: Temporally grounding language queries in videos by contextual boundary-aware prediction. In: AAAI. pp. 12168–12175 (2020)
29. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: Eur. Conf. Comput. Vis. pp. 396–416. Springer (2024)
30. Xiao, S., Chen, L., Zhang, S., Ji, W., Shao, J., Ye, L., Xiao, J.: Boundary proposal network for two-stage natural language video localization. In: AAAI. pp. 2986–2994 (2021)
31. Xiao, Y., Luo, Z., Liu, Y., Ma, Y., Bian, H., Ji, Y., Yang, Y., Li, X.: Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18709–18719 (2024)
32. Xu, H., He, K., Plummer, B.A., Sigal, L., Sclaroff, S., Saenko, K.: Multilevel language and vision integration for text-to-clip retrieval. In: AAAI. pp. 9062–9069 (2019)
33. Xu, Y., Sun, Y., Zhai, B., Jia, Y., Du, S.: Mh-detr: Video moment and highlight detection with cross-modal transformer. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2024)
34. Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In: Adv. Neural Inform. Process. Syst. vol. 32 (2019)
35. Zhang, D., Dai, X., Wang, X., Wang, Y.F., Davis, L.S.: Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1247–1257 (2019)
36. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: AAAI. pp. 12870–12877 (2020)

37. Zhang, Z., Lin, Z., Zhao, Z., Xiao, Z.: Cross-modal interaction networks for query-based moment retrieval in videos. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. pp. 655–664 (2019)
38. Zhao, C., Sun, Y., Wang, W., Chen, Q., Ding, E., Yang, Y., Wang, J.: Ms-detr: Efficient detr training with mixed supervision. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 17027–17036 (2024)
39. Zhou, X., Wei, F., Duan, L., Yao, A., Li, W.: The devil is in the spurious correlations: Boosting moment retrieval with dynamic learning. In: Int. Conf. Comput. Vis. pp. 20981–20990 (2025)
40. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: Int. Conf. Learn. Represent. (2021)