DE²TR: Dual Evidence Detection Transformer for Video Temporal Grounding

Yifan Zhang[1] Chengxu Liu[2] Yujie Dun[1] ✉ Xueming Qian[1]

[1]School of Information and Communication Engineering · [2]School of Software Engineering — Xi'an Jiaotong University, Xi'an, China

ECCV 2026 | MALMÖ | SEPT 8–12

Project Page

Motivation

Video temporal grounding (VTG) aims to localize relevant moments in videos based on language queries. Video moments fade in and out — global semantics alone cannot tell **where they start or end**. DE²TR gives every moment query two evidences semantic (*what is relevant*) ‖ boundary (*where it begins & ends*) — **new SOTA on 4 VTG benchmarks .**

Moments are not objects:

(a) Objects own **clear physical edges** ✓

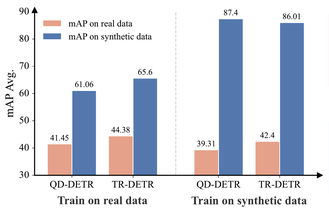
Query: "A woman is making fruit and protein smoothie"

(b) Moments exhibit a **semantic gradient** — ambiguous start / end ✗

DETR-style grounding aggregates only *global semantic evidence* ⇒ severe boundary misalignment.

Boundary cues are the missing evidence:

(c) Splicing GT into irrelevant clips ⇒ **hard synthetic boundaries**

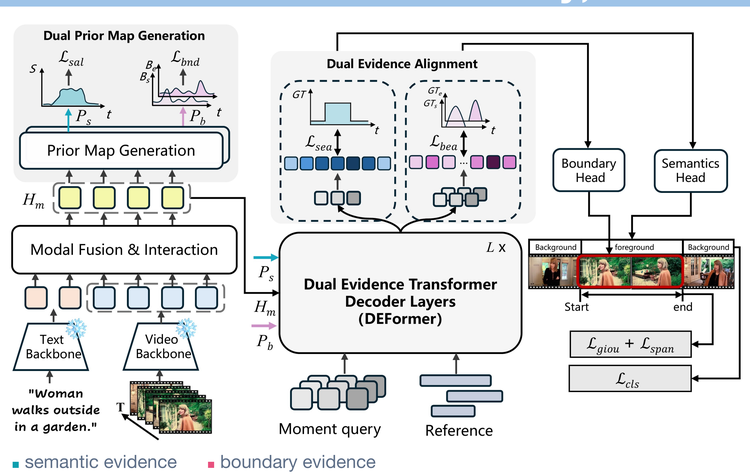


+21.2 mAP
for TR-DETR once boundaries are clear (QD-DETR: **+19.6**)

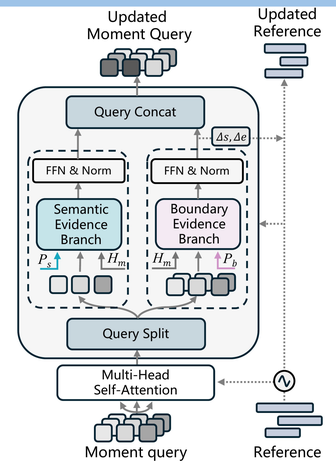
Synthetic-trained ≈ baseline on real videos
more data does *not* help — the boundary *signal* does

Precise grounding needs boundary evidence, not more semantics.

DE²TR — One Query, Two Evidences



(a) Overall framework of DE² TR



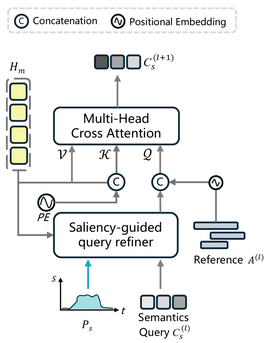
(b) Illustration of DEFormer layer

DUAL PRIOR MAPS
Saliency is dual-natured:
high value ⇒ relevance → P_s;
sharp change ⇒ boundary → P_b via a dedicated projection, soft-label supervised.

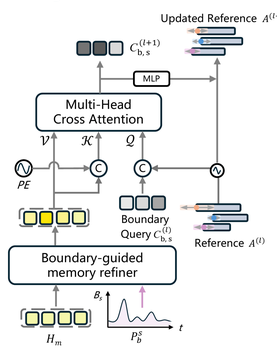
DECOUPLED DECODING
Each query splits channel-wise: C_s scores the foreground ‖ $C_{b,s}$, $C_{b,e}$ refine start / end anchors — evidences are read in parallel branches, layer by layer.

ALIGN & AUGMENT
$\mathcal{L}_{sea}$ / $\mathcal{L}_{bea}$ align each half with its evidence; BAS (motivation.(c)) splices ideal boundaries into training videos.

DEFormer layer



(a) Semantic Evidence Branch



(b) Boundary Evidence Branch

SEMANTIC BRANCH — saliency-guided **query** refiner: P_s-weighted pooling of memory, gated into C_s; $PE(c,w)$ spans the whole moment. $\mathcal{L}_{sea}$ = soft-IoU: similar to in-span clip, silent outside.

BOUNDARY BRANCH — boundary-guided **memory** refiner: $FiLM(\gamma,\beta)$from P_b sharpens boundary responses; $PE(m_s,w)$keeps attention local; offsets $\Delta s, \Delta e$ refine anchors. $\mathcal{L}_{bea}$ = BCE to soft labels.

AMBIGUITY-AWARE SOFT LABELS — $y^s_{k,i} = \exp\left(-\frac{(i-t^s_k)^2}{2\sigma_k^2}\right)$, $\delta_k = \exp(\cos(\mathbf{v}_{in}, \mathbf{v}_{out}))$, sharp boundary ⇒ strict $\delta\downarrow$; ambiguous ⇒ relaxed $\delta\uparrow$.

Where to inject priors?
semantic → query & boundary → memory works best (50.67 Avg mAP).

$$\mathcal{L} = \mathcal{L}_{MR} + \mathcal{L}_{HD} + \lambda_{bnd}\mathcal{L}_{bnd} + \lambda_{sea}\mathcal{L}_{sea} + \lambda_{bea}\mathcal{L}_{bea}$$

Quantitative Results

Tips: MR — Moment Retrieve && HD — Highlight Detection

QVHighlights — MR & HD

Method	R1		mAP		HD (≥VG)	
	@0.5	@0.7	@0.75	Avg.	mAP	HIT@1
M-DETR NeurIPS21	53.94	34.84	—	32.20	35.65	55.55
QD-DETR CVPR23	62.68	46.66	41.82	41.22	39.13	63.03
UniVTG ICCV23	—	—	—	36.13	38.83	—
CG-DETR arXiv23	67.40	52.10	45.70	44.90	40.80	66.70
TR-DETR AAAI24	67.10	51.48	46.42	45.09	40.55	64.77
BAM-DETR ECCV24	65.10	51.61	48.56	47.61	—	—
UVCOM CVPR24	65.10	51.81	—	45.79	40.03	63.29
MS-DETR MM25	66.90	51.68	46.05	46.00	40.57	66.58
TD-DETR‡ ICCV25	65.88	53.67	49.86	49.05	—	—
KDA ICCV25	69.11	53.46	48.04	47.41	—	—
DE²TR (ours)	68.13	53.81	51.61	50.67	40.93	66.65
DE²TR‡ (ours)	**69.81**	**55.29**	**53.43**	**52.12**	**41.20**	**66.97**

‡ = with boundary augmentation · **bold** best · underline second

Charades-STA & TACoS — MR

Method	Charades-STA			TACoS		
	@0.5	@0.7	mIoU	@0.5	@0.7	mIoU
M-DETR NeurIPS21	52.07	30.59	45.54	24.67	11.97	25.49
UniVTG ICCV23	58.01	35.65	50.10	34.97	17.35	33.60
UVCOM CVPR24	59.25	36.64	—	36.39	23.32	—
DualGround NeurIPS25	**61.11**	38.52	—	—	—	—
MS-DETR MM25	59.62	36.48	50.59	39.65	23.42	37.01
FlashVTG WACV25	60.11	38.01	—	41.76	24.74	37.61
KDA ICCV25	60.23	37.63	—	40.13	24.34	—
DE²TR (ours)	60.39	**39.87**	**51.58**	**42.52**	**27.14**	**39.76**

TVSum — HD

Method	VT	VU	GA	MS	PK	PR	FM	BK	BT	DS	Avg.
UMT CVPR22	87.5	81.5	88.2	78.8	81.4	87.0	76.0	86.9	84.4	79.6	83.1
QD-DETR CVPR23	88.2	87.4	85.6	85.0	85.8	86.9	76.4	91.3	89.2	73.7	85.0
TR-DETR AAAI24	89.3	93.0	94.3	85.1	88.0	88.6	80.4	91.3	89.5	81.6	88.1
MS-DETR MM25	89.8	93.3	94.8	88.7	88.5	89.0	80.5	94.0	88.7	76.9	88.4
MQVTG ICCV25	87.7	91.6	92.3	85.2	85.7	**91.3**	78.5	**96.5**	90.6	**82.9**	88.2
DE²TR (ours)	**91.7**	**94.1**	**95.4**	**88.9**	**89.5**	90.4	**81.2**	92.3	**91.6**	80.9	**89.6**

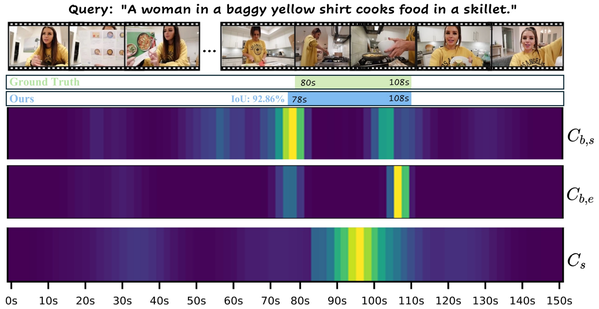
bold best · best on 7/10 categories · +1.2 Avg over prev. best

Contributions

1. **DEFormer** — dual-branch decoder; moment queries decoupled *channel-wise* into semantic ‖ boundary halves, no feature conflict.
2. **Prior-guided refinement** — dual maps P_s/P_b steer each branch toward its regions of interest.
3. **Boundary Augmentation (BAS)** — temporal splicing synthesizes ideal boundary patterns; awareness transfers to real videos.
4. **New SOTA** on QVHighlights, Charades-STA, TACoS & TVSum — e.g. **+3.07 Avg mAP** on QVHighlights.

Qualitative Results

Queries really specialize



Sharper boundaries