Temporally Grounding Instructional Diagrams in Unconstrained Videos


Jiahao Zhang[1]


Frederic Z. Zhang[2]


Cristian Rodriguez[2]


Yizhak Ben-Shabat[1,3]


Anoop Cherian[4]


Stephen Gould[1]

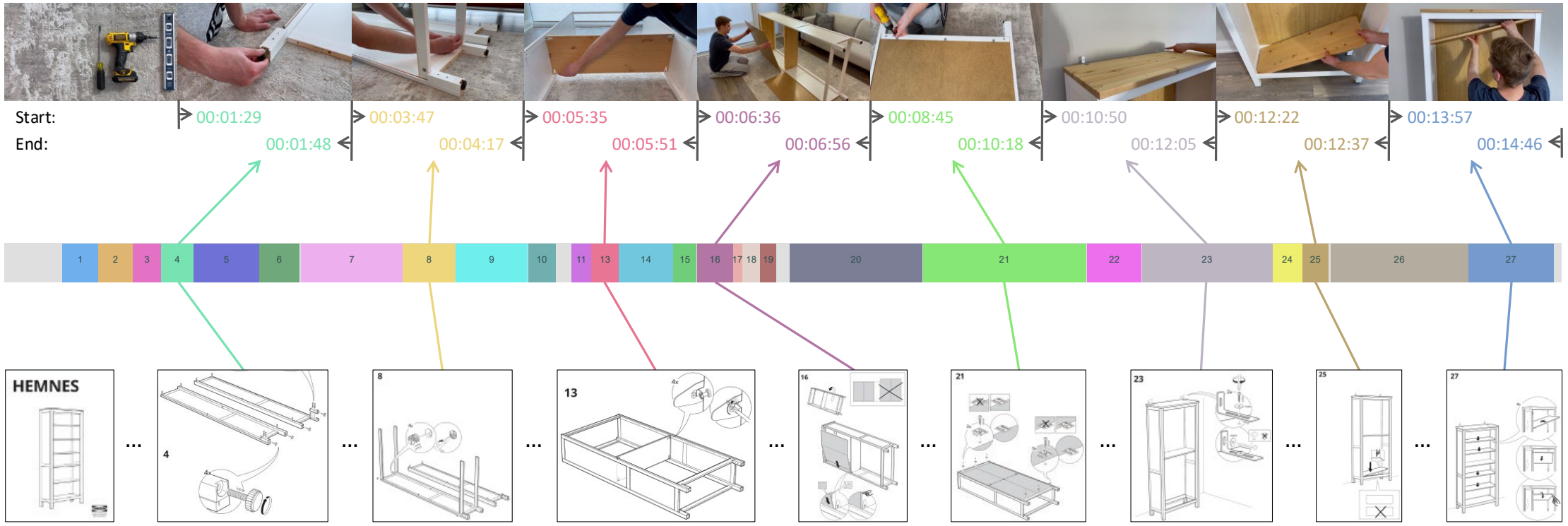
[1]The Australian National University [2]The Australian Institute for Machine Learning

[3]Technion Israel Institute of Technology [4]Mitsubishi Electric Research Labs

Code & Dataset: https://github.com/DavidZhang73/TDGV

Poster: Session 5 (Mon-16:15-18:00)

Problem Statement

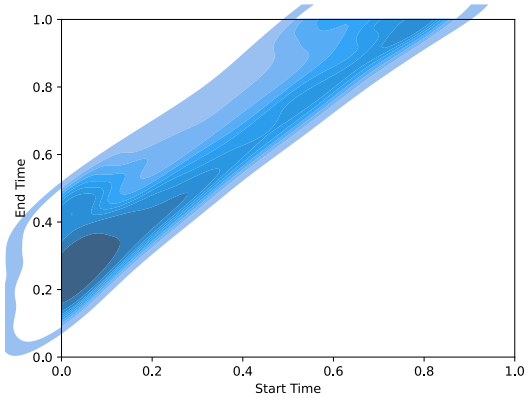


Key Issues

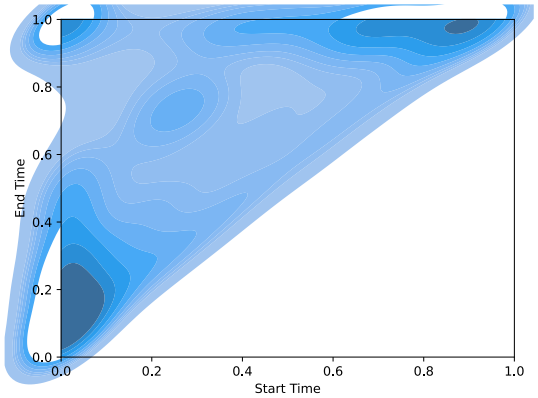
Dataset	Year	Video Length	Avg. #Segments Per video
DiDeMo (YFCC100M)	2017	Max 30s	3.87
Charades-STA (Charades)	2017	Avg. 30.60s	2.42
ActivityNet Captions (ActivityNet)	2017	Avg. 117.60s	4.82
YouCook II	2017	Avg. 5.27 (Max 10) min	7.7
IKEA Assembly in the Wild	2023	**Avg. 11 (Min 1 - Max 79) min**	**15.57**

- More number of segments per video
  - Need to model the relationship among them
- Longer duration

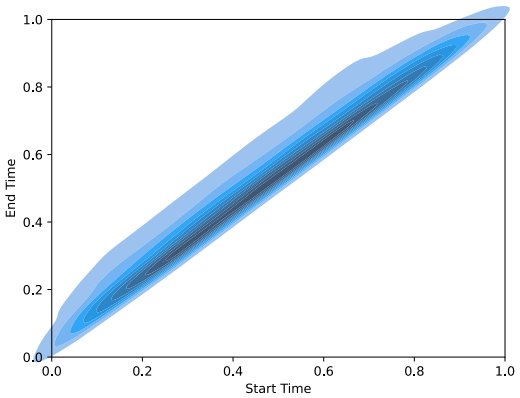
Key Issues



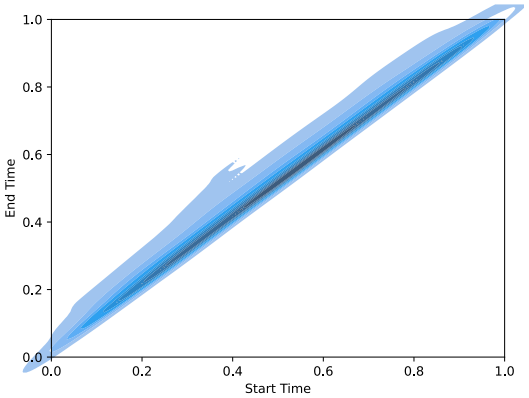
Charades-STA



ActivityNet-Caption



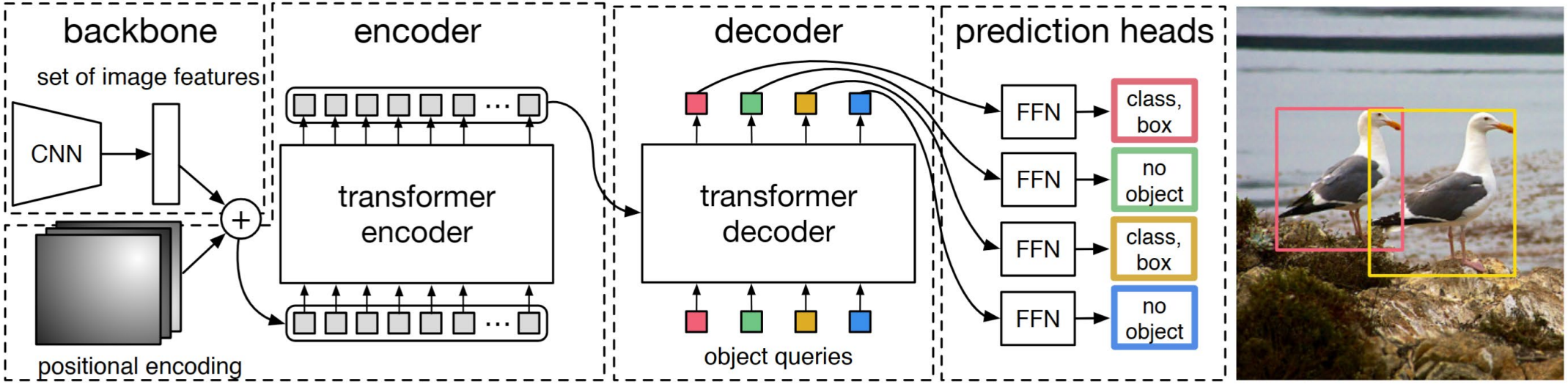
YouCook II



IAW

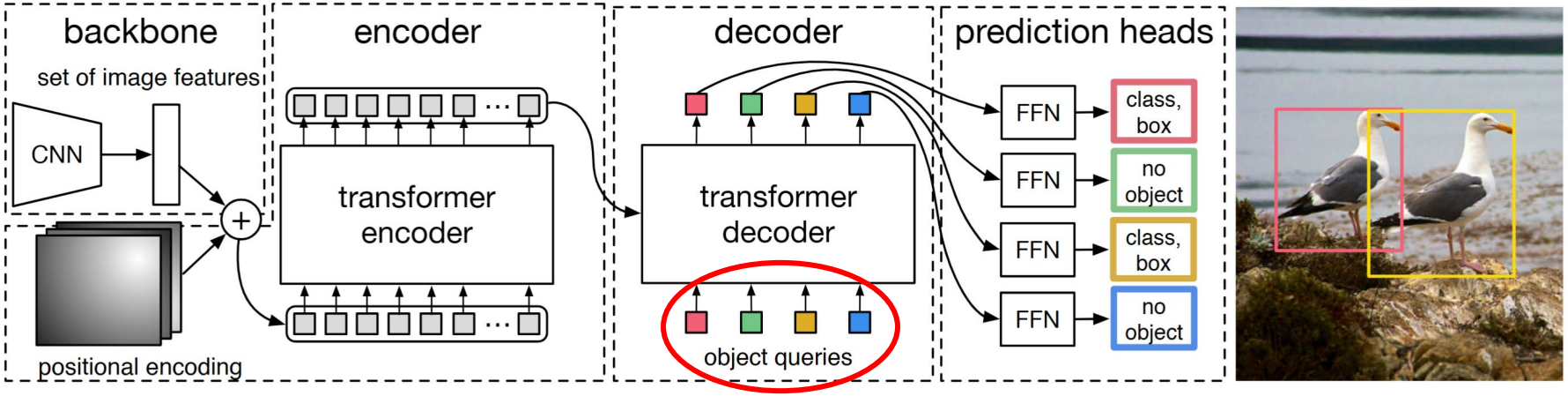
- Longer duration
  - Unbiased segments

Revisit: Detection Transformer (DETR)



Carion et al.: End-to-End Object Detection with Transformers (ECCV'20)

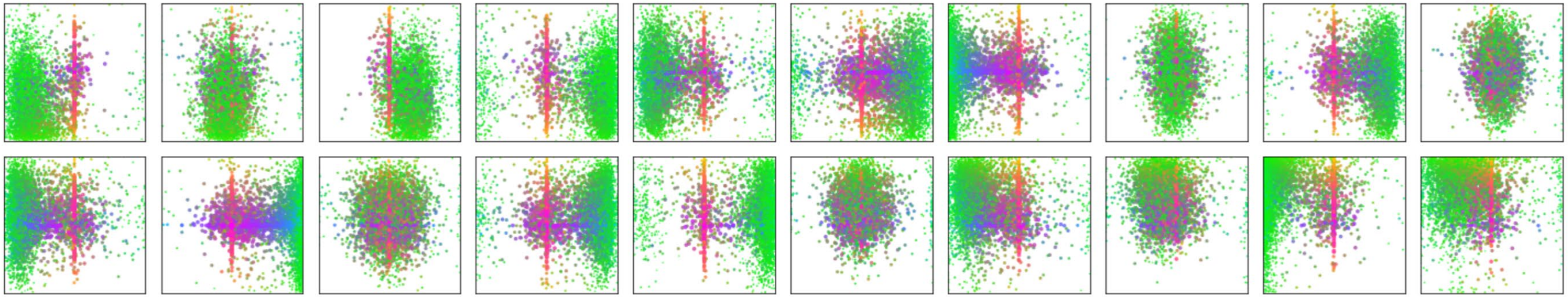
Revisit: Detection Transformer (DETR)



- What is object queries?

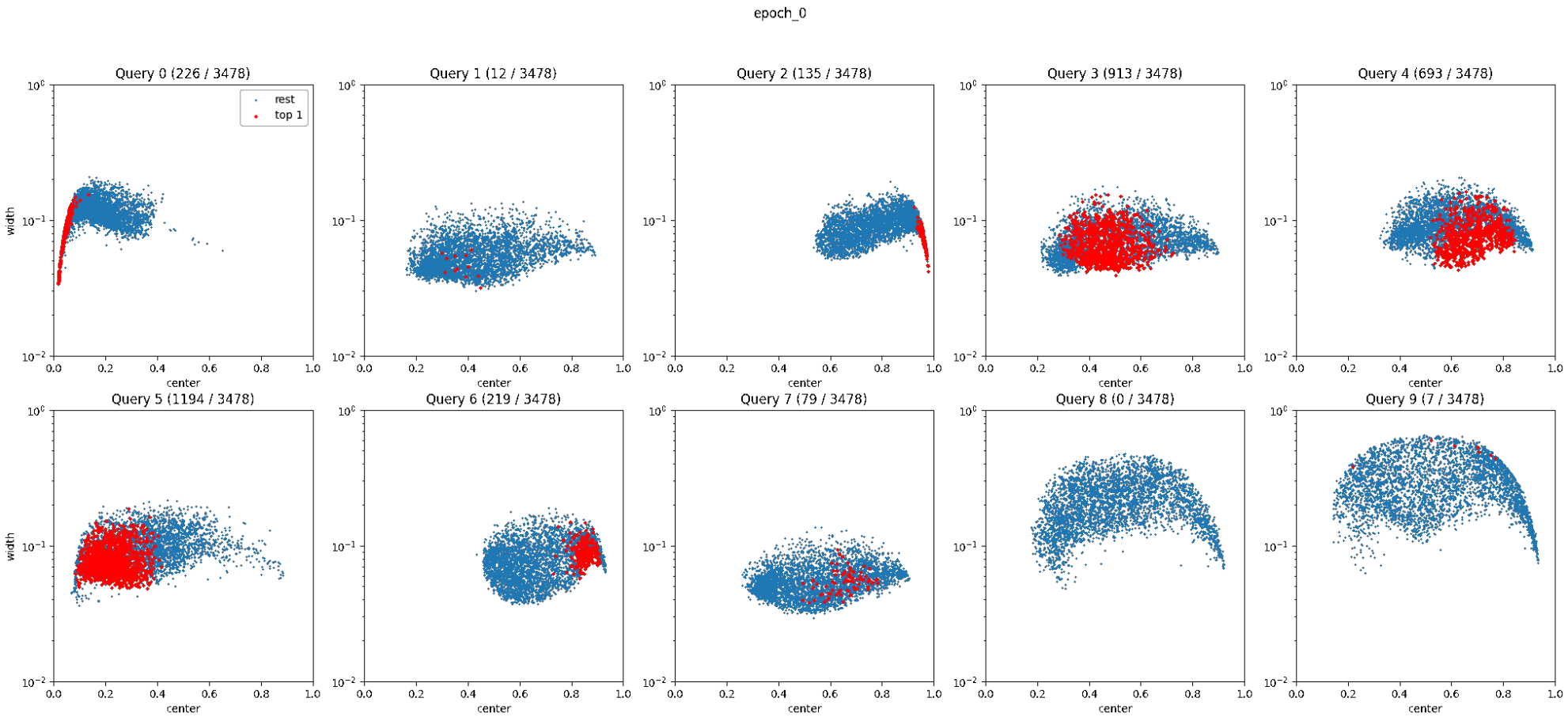
Carion et al.: End-to-End Object Detection with Transformers (ECCV'20)

DETR: Object Queries => Spatial Templates

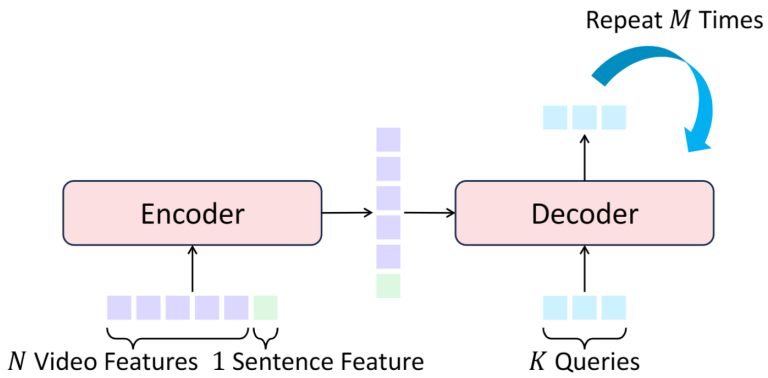


Carion et al.: End-to-End Object Detection with Transformers (ECCV'20)

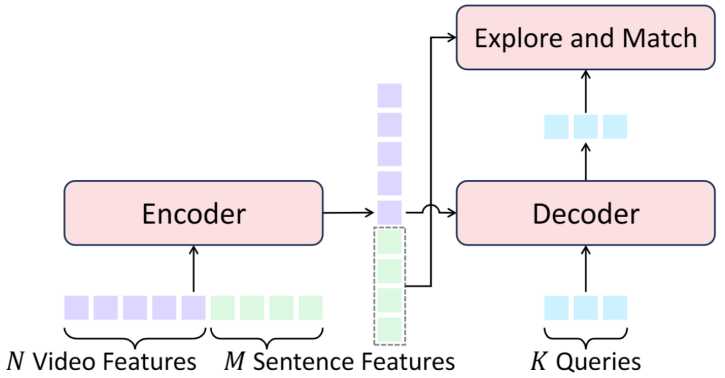
Temporal Templates



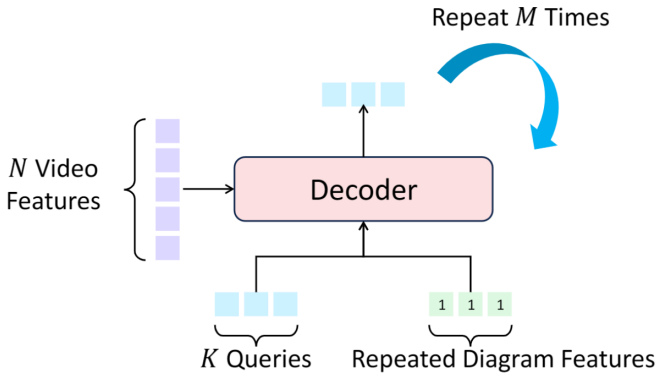
DETR-Based Models



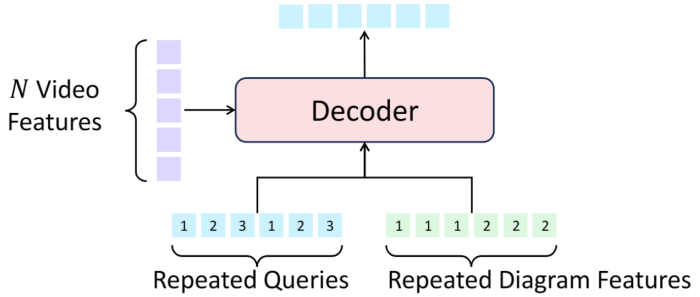
(a) Moment-DETR [21]



(b) LVTR [44]

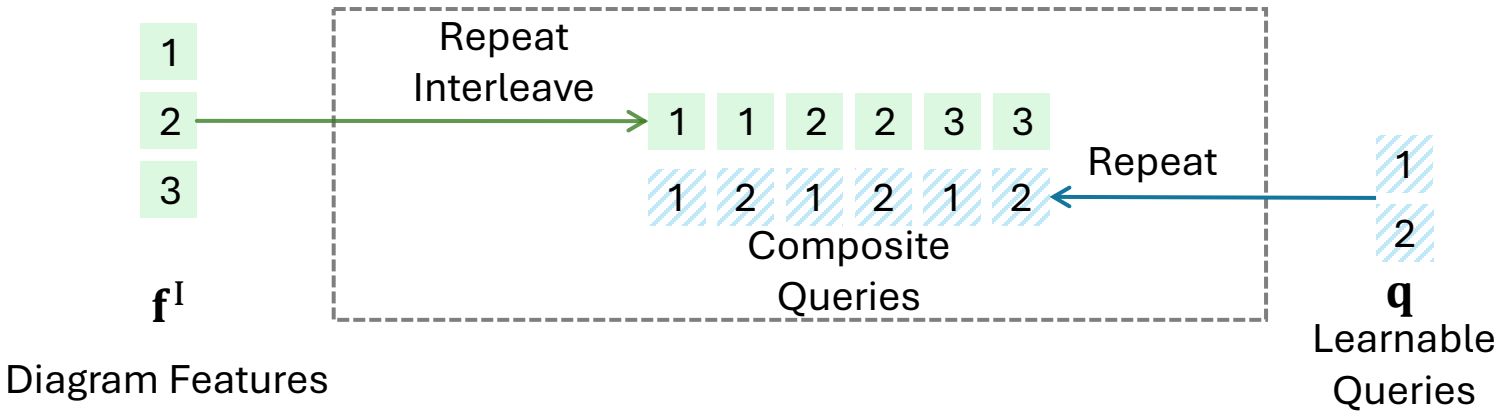


(c) Ours (one diagram at a time)

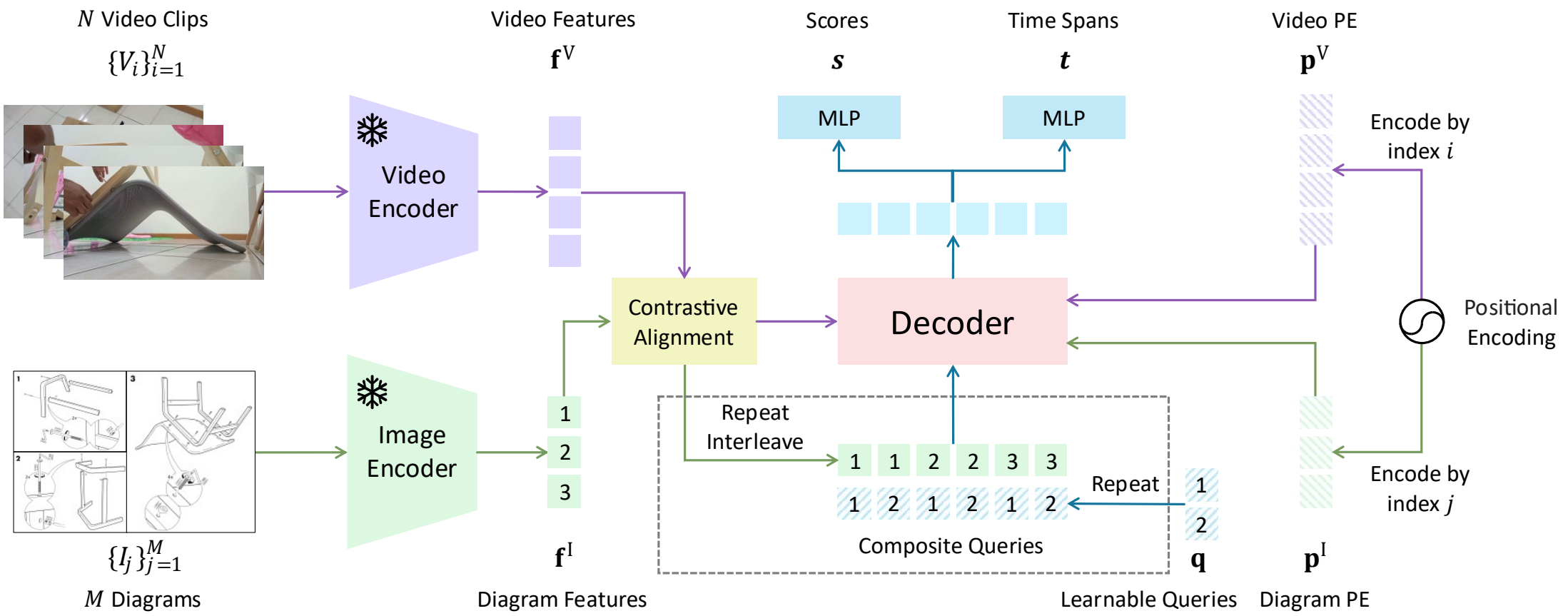


(d) Ours (all diagrams at the same time)

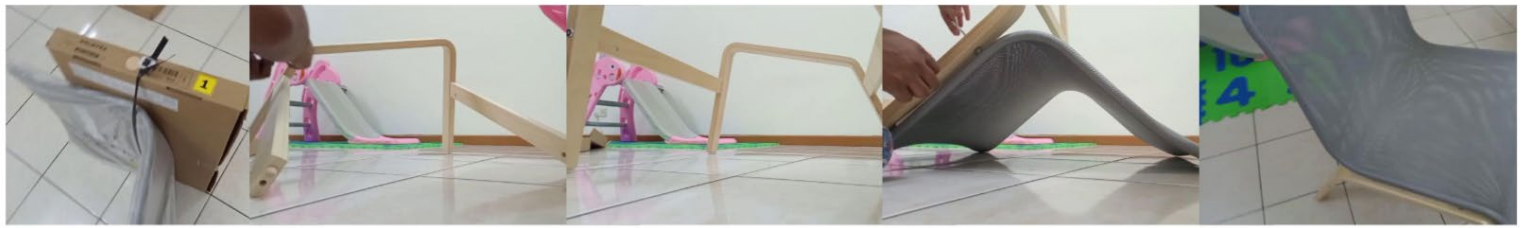
Composite Query via Duplication



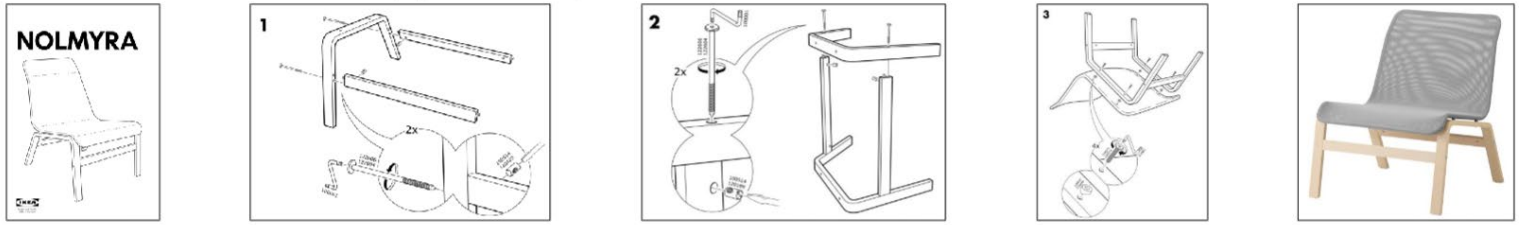
⭐ Temporal Diagram Grounding



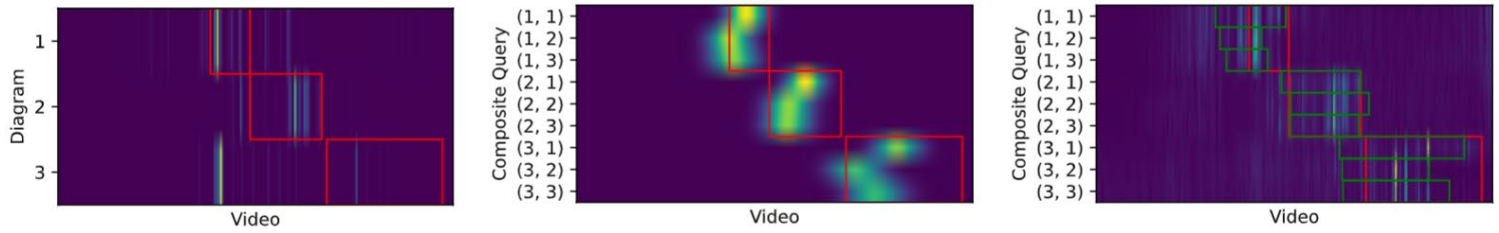
Content and Position Joint Guided Cross-Attention



(a) Assembly video from YouTube 1czDviZ5vG0.



(b) Step diagrams from Ikea furniture 10233607.

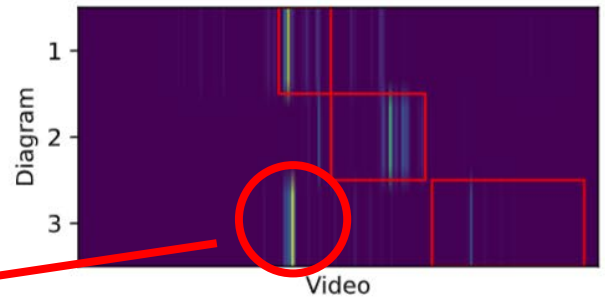


(c) $\sigma(Q_c K_c^\mathsf{T})$ (d) $\sigma(Q_p K_p^\mathsf{T})$ (e) $\sigma(Q_c K_c^\mathsf{T} + Q_p K_p^\mathsf{T})$

⭐ Content and Position Joint Guided Cross-Attention

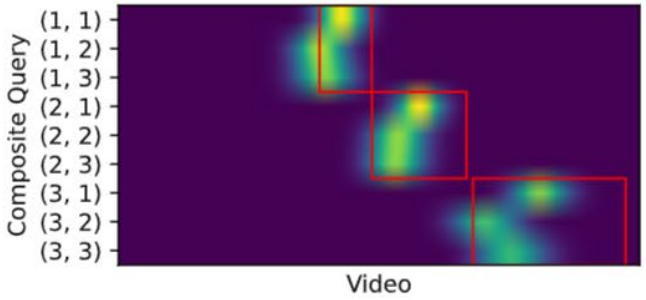
GT Pred

Outlier



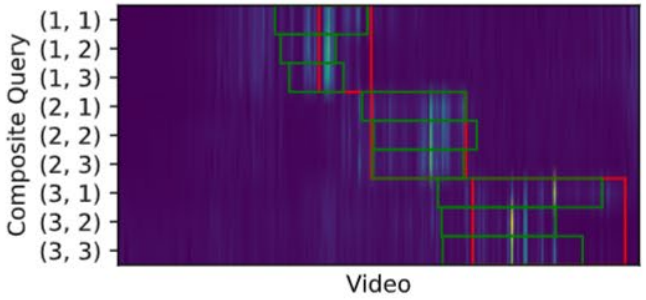
$\mathbf{(c)}\ \sigma(Q_c K_c^\mathsf{T})$

Cross attention between video features and diagram features



$\mathbf{(d)}\ \sigma(Q_p K_p^\mathsf{T})$

Cross attention between video positions and learnable queries



$\mathbf{(e)}\ \sigma(Q_c K_c^\mathsf{T} + Q_p K_p^\mathsf{T})$

Final cross attention

Quantitative Results

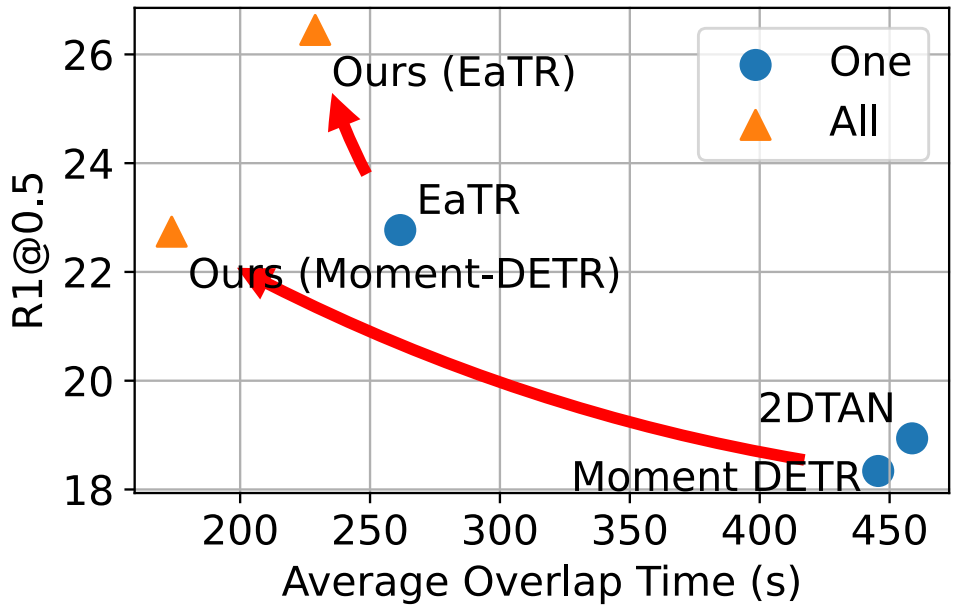
Result on IAW dataset.

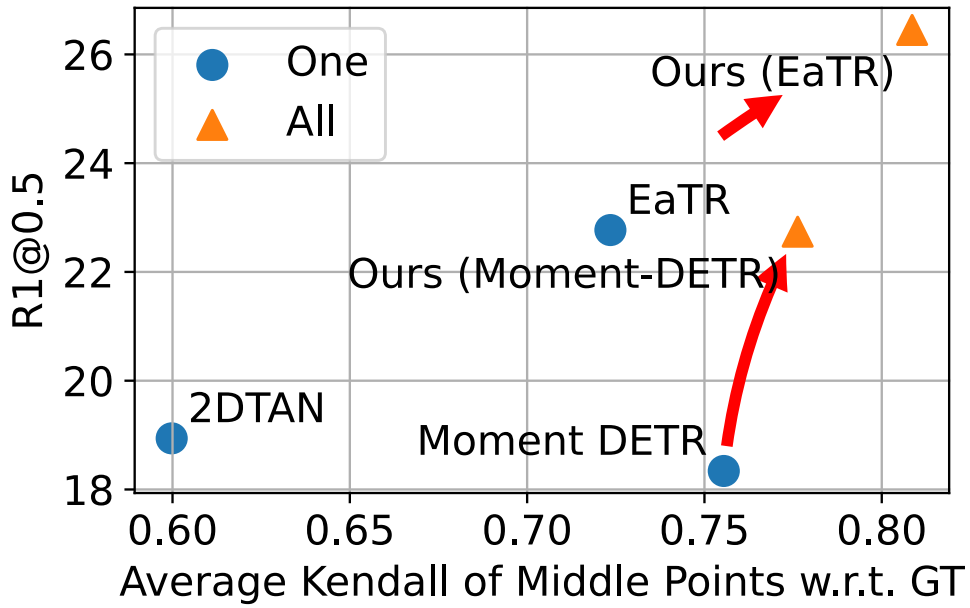
Method	Mode	R@1, IoU=			mIoU
		0.3	0.5	0.7	
Random	-	1.809	0.254	0.057	4.801
LVTR [41]	All	11.26	4.591	1.112	7.515
2D-TAN [53] conv	One	31.24	18.94	8.030	20.51
2D-TAN [53] pool	One	32.94	20.02	8.170	21.21
Moment DETR [18]	One	34.00	18.34	7.290	16.60
Ours w/ Moment DETR	All	37.79	22.74	9.140	23.86
EaTR [11]	One	38.48	22.77	9.540	24.75
Ours w/ EaTR	All	**42.02**	**26.45**	**11.54**	**27.27**

Result on YouCook2 dataset.

Method	Text	R@1, IoU=			mIoU
		0.3	0.5	0.7	
DORi [30]	-	43.36	30.47	18.24	30.46
LocFormer [31]	BERT [8]	46.76	31.33	15.81	30.92
ExCL [10]	BERT ❄ [14]	26.63	16.15	8.51	18.87
TMLGA [29]		34.77	23.05	12.49	24.42
DORi [30]		42.27	29.90	18.38	29.92
Ours w/ EaTR		**52.95**	**36.28**	**18.50**	**35.32**

Quantitative Results





Thanks!

https://github.com/DavidZhang73/TDGV