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]

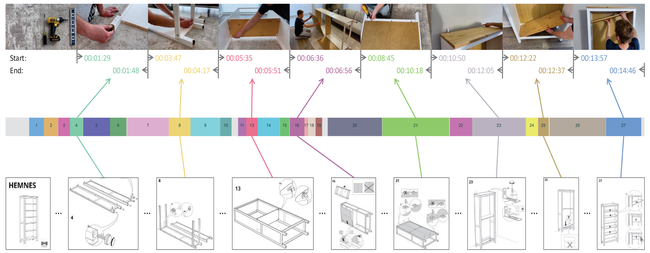
[1]The Australian National University [2]The Australian Institute for Machine Learning [3]Technion Israel Institute of Technology [4]Mitsubishi Electric Research Labs

[1]{first.last}@anu.edu.au [2]{first.last}@adelaide.edu.au [3]sitzikbs@gmail.com [4]cherian@merl.com

Code & Datasets

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

1. Introduction



▶ Problem Statement

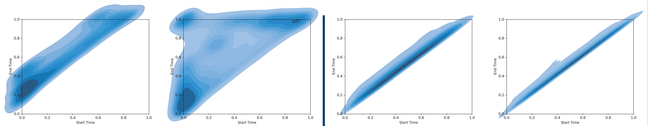
Given an **untrimmed** video that has been evenly divided into N clips for feature extraction, we denote these clips by $\{V_i\}_{i=1}^{N}$, and the set of M diagrams corresponding to the video by $\{I_j\}_{j=1}^{M}$, we aim to develop a model capable of accurately predicting the **timespan** of each diagram $\mathbf{t} = (t_s, t_e)$, where t_s and t_e are the normalized start and end time of a segment.

2. Motivation & Contributions

▶ Motivation

Need to model **multiple unbiased** segments per video.

Dataset	Year	Video Duration	# of Segments per Video
DiDeMo	2017	Max 30s	3.87
Charades-STA	2017	Avg. 30.60s	2.42
ActivityNet Captions	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**

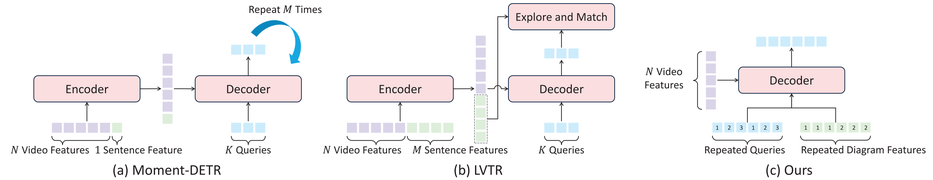


▶ Contributions

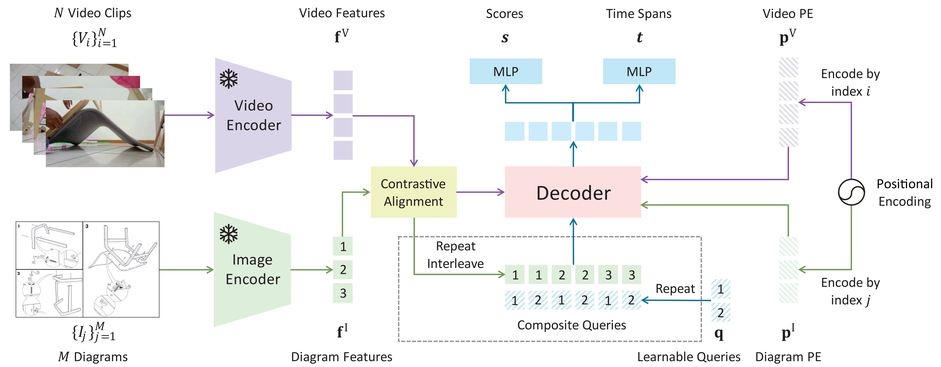
1. **Sequential Grounding with Composite Queries:** Developed a detection transformer that uses composite queries combining both content and positional priors to handle varying query lengths.
2. **Enhanced Attention Mechanisms:** Designed specialized self-attention masks and optimized cross-attention value choices for better grounding.
3. **State-of-the-Art Performance:** Achieved significant improvements on IAW and YouCook2, demonstrating effectiveness for both diagram and sentence sequence grounding.

3. Method - Temporal Diagram Grounding in Videos

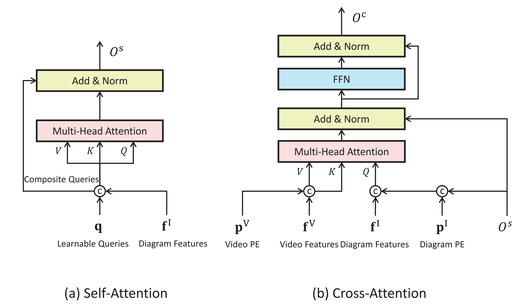
▶ Comparison with Existing Approaches



▶ Overall Pipeline



▶ Decoder Design



▶ Losses

(a) Span L1 Loss:

$$\mathcal{L}_{L_1} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{C_i}\sum_{j=1}^{K}\mathbb{1}_{\{\boldsymbol{s}^{\star}_{i,j}=1\}}\|\hat{\boldsymbol{t}}_{i,j} - \boldsymbol{t}^{\star}_{i,j}\|_1$$

(b) Generalized IoU Loss:

$$\mathcal{L}_{\mathrm{gIoU}} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{C_i}\sum_{j=1}^{K}\mathbb{1}_{\{\boldsymbol{s}^{\star}_{i,j}=1\}} gIoU(\hat{\boldsymbol{t}}_{i,j}, \boldsymbol{t}^{\star}_{i,j})$$

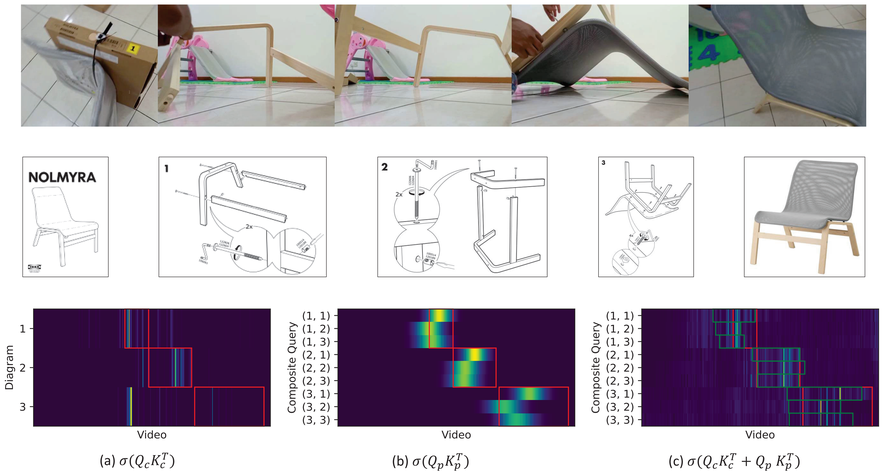
(c) Background Score Loss:

$$\mathcal{L}_{\mathrm{score}} = -\frac{1}{MK}\sum_{i=1}^{M}\sum_{j=1}^{K}\mathbf{s}^{\star}_{i,j}\log(\hat{\mathbf{s}}_{i,j})$$

(d) Total Loss:

$$\mathcal{L} = \lambda_{L_1}\mathcal{L}_{L_1} + \lambda_{\mathrm{gIoU}}\mathcal{L}_{\mathrm{gIoU}} + \lambda_{\mathrm{score}}\mathcal{L}_{\mathrm{score}}$$

4. Content and Position Jointly Guided Cross-Attention Visualization



(a) $\sigma(Q_c K_c^T)$ (b) $\sigma(Q_p K_p^T)$ (c) $\sigma(Q_c K_c^T + Q_p K_p^T)$

5. Final Result

Ikea Assembly in the Wild dataset.

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

YouCook2 dataset.

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