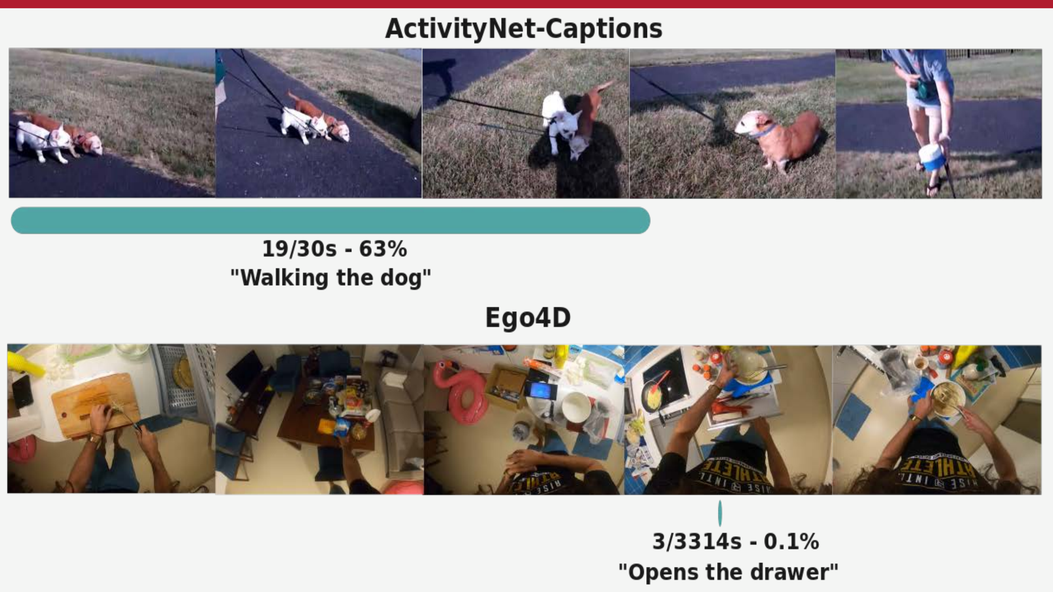
Learning Temporal Sentence Grounding in Narrated EgoVideos

Kevin Flanagan Dima Damen Michael Wray

University of BRISTOL

BMVC 2023

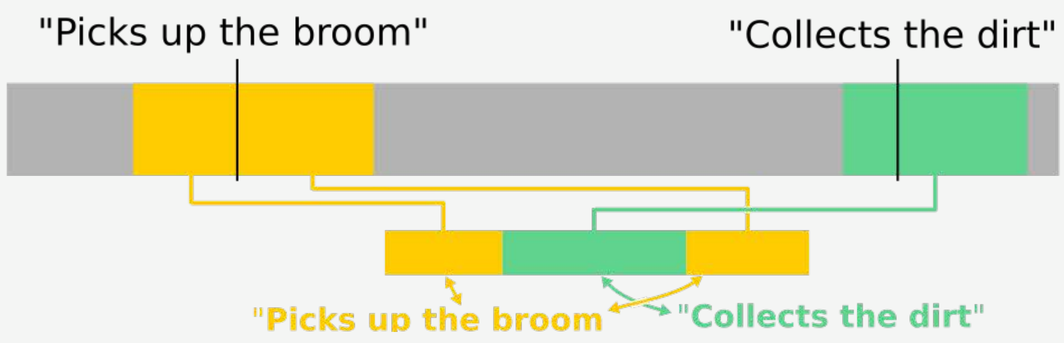
Motivation



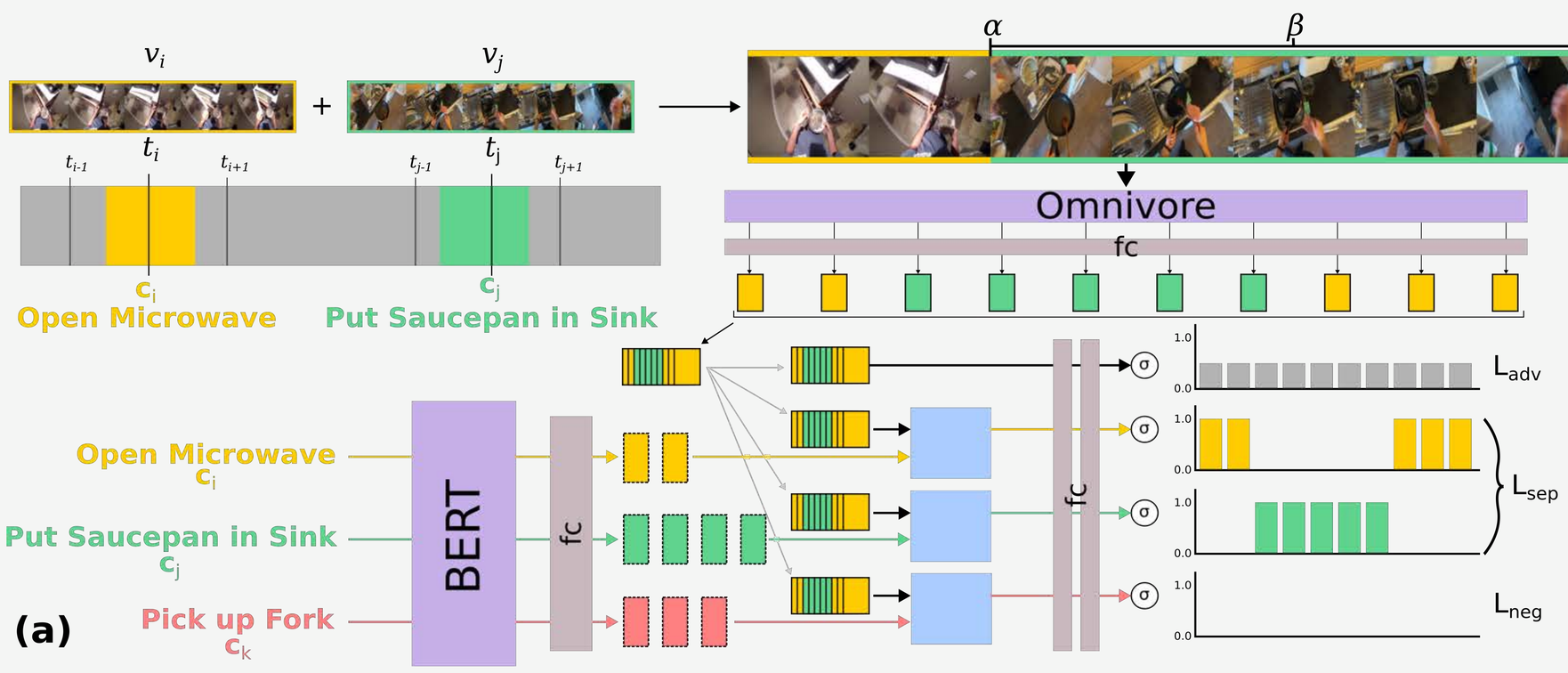
Comparison of Temporal Sentence Grounding Datasets

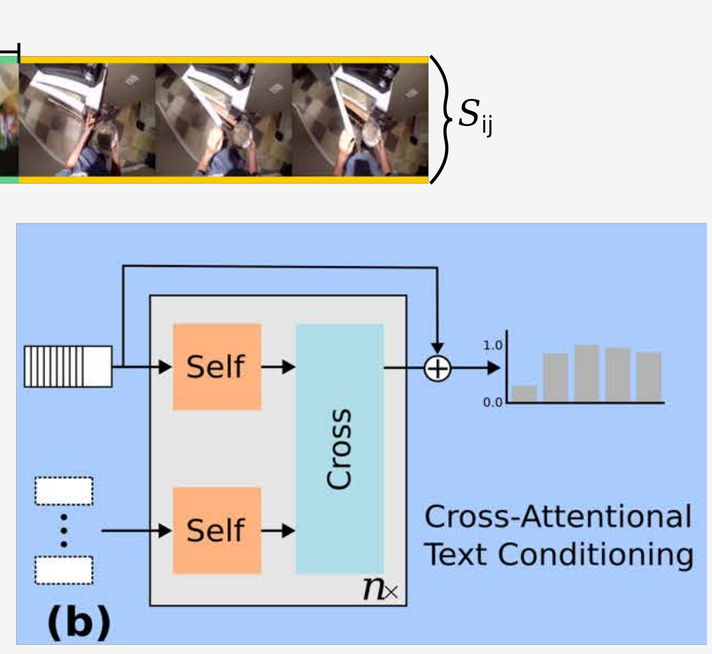
	Total Vid Duration	Avg. Vid Duration	Avg. Mom. Duration	Annotations / Video	Total Annotations	Avg. Coverage ↓
ANet-Captions [21]	487.6h	2.0min	37.1s	4.9	72k	30.90%
Charades-STA [14]	57.1h	0.5min	8.1s	2.3	16k	27.00%
DiDeMo [1]	88.7h	0.5min	6.5s	3.9	41k	21.70%
TACoS [40]	10.1h	4.8min	27.9s	143.6	18k	9.70%
Ego4D [17]	234.9h	17.3min	2.0s	214.1	223k	**0.19%**
EPIC-Kitchens [12]	73.4h	8.9min	3.1s	134.1	67k	0.58%

Clip Merging & Rough Timestamps

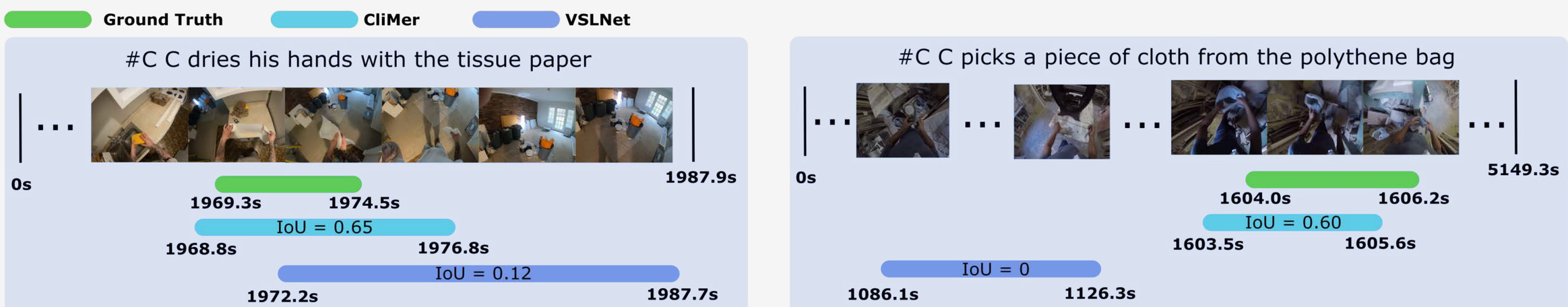


CliMer





Results - Qualitative



Results - Quantitative

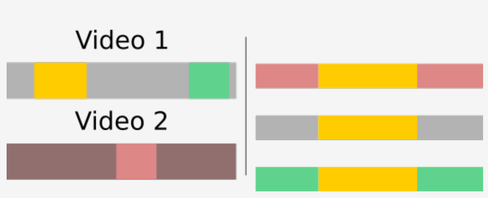
Ego4d

	IoU=0.1	IoU=0.3	IoU=0.5	mR
Random Baseline	0.47	0.09	0.02	0.19
VSLNet [50]	7.32	3.16	1.35	3.94
CliMer	**9.68**	**5.03**	**2.24**	**5.65**

EPIC-Kitchens

	IoU=0.1	IoU=0.3	IoU=0.5	mR
Random Baseline	0.78	0.13	0.03	0.31
VSLNet [50]	19.32	8.76	3.90	10.66
CliMer	**22.20**	**11.57**	**5.25**	**13.01**

Ablation - Clip Merging & Hard Negatives



Merged Segment	Hard Negatives	IoU=0.1	IoU=0.3	IoU=0.5	mR
✓	-	7.28	3.09	1.08	3.82
-	✓	7.23	4.42	2.20	4.62
✓	✓	**9.68**	**5.03**	**2.24**	**5.65**

Conclusion

- Clip merging proves to be an effective method of training using clips segmented from rough timestamps.

- We explore Ego4D and EPIC-Kitchens for Temporal Sentence Grounding showcasing their difficulty.

- Qualitatively CliMer shows a better ability to pick out precise moments for given sentence queries.