Zero-shot Video Moment Retrieval With Off-the-Shelf Models

Anuj Diwan*, Puyuan Peng*, Raymond Mooney
Department of Computer Science, The University of Texas at Austin
Austin, Texas, USA (* Equal contribution)

Transfer Learning for NLP Workshop

1. Introduction

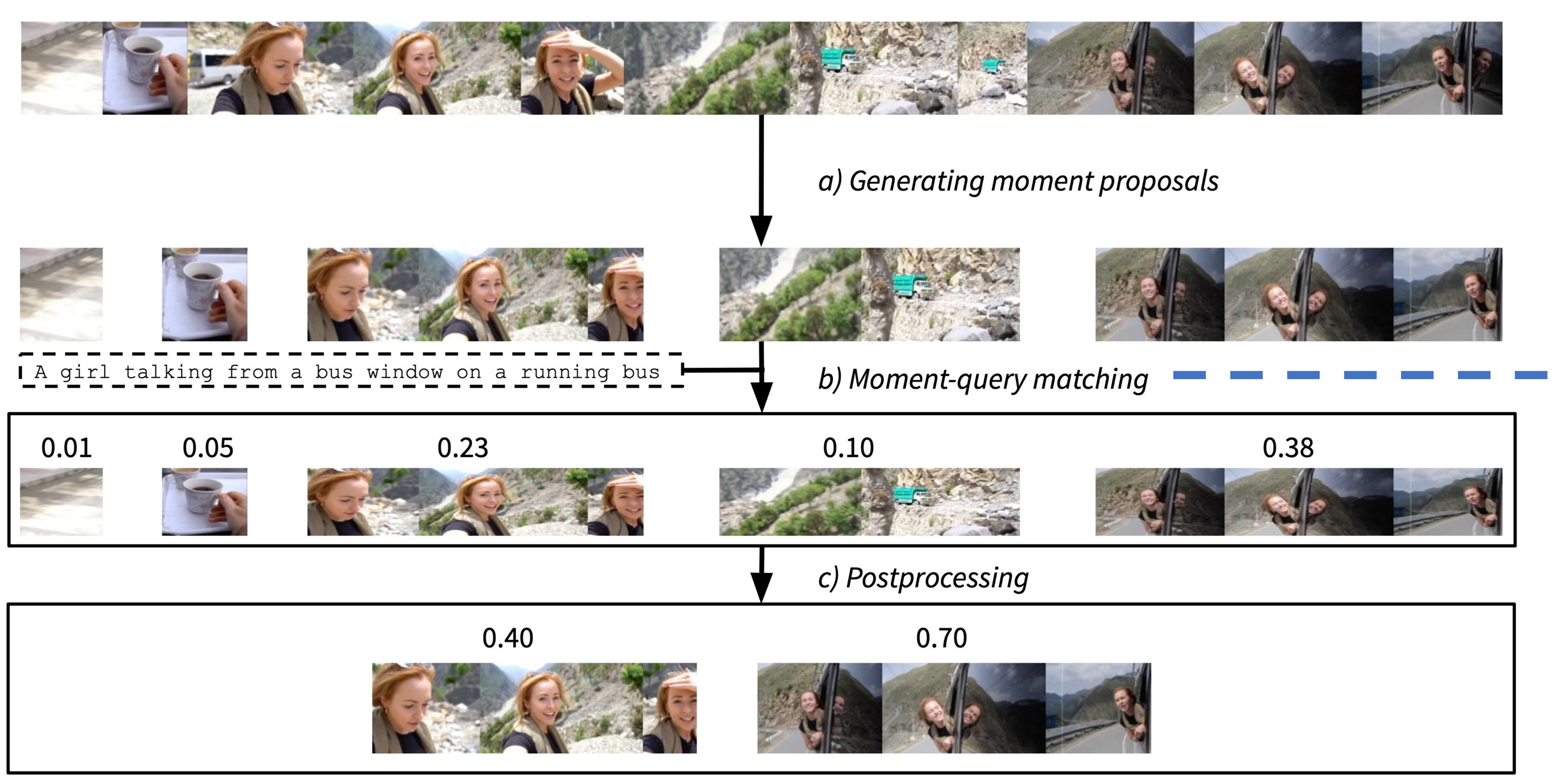
Given a video and a natural language query, the task of **Video Moment Retrieval** (VMR) involves temporally localizing moments (video segments) within the given video that are relevant to the query

Query: a girl talking from a bus window on a running bus

2. Our Approach

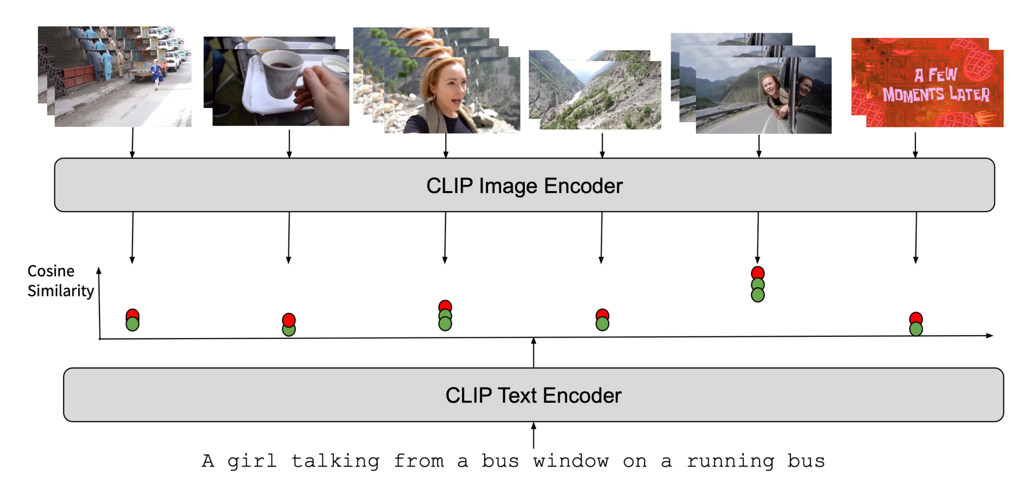
***No newly trained models**; Zero-shot transfer from off-the-shelf models (CLIP/UniVL) to VMR* **!**

2.1 Overall Procedure

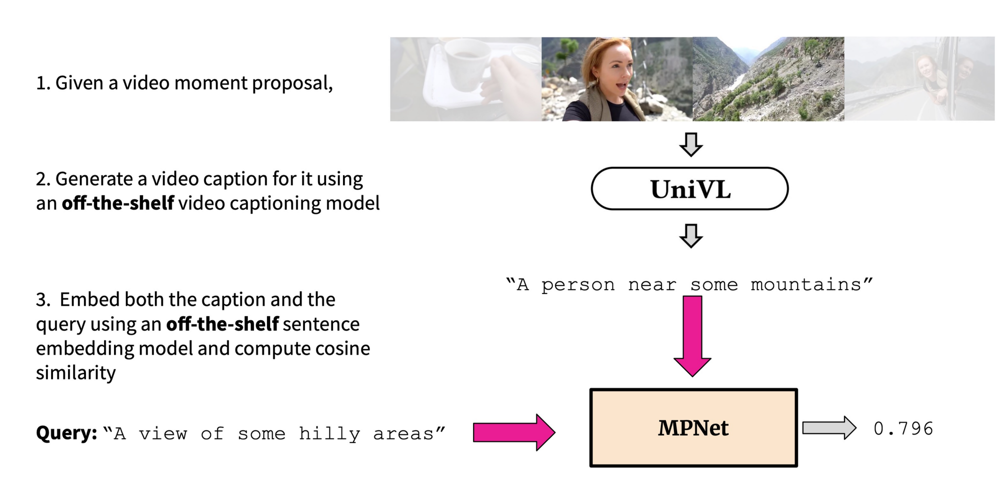


2.2 Moment-Query Matching

CLIP



UniVL Captions



3. Key Results

1. Best Zero-shot Performance

2. Better Performance than some Supervised approaches!
(that do not use labor-intensive frame-wise saliency loss or compute-intensive pretraining)

3. Great Performance for **Short** Segment Detection

Method	Long	Medium	Short
M-DETR w/ PT	**45.18**	**37.53**	3.50
ShotDetect+CLIP	27.49	26.15	**7.08**

4. But there's room for improvement

4.1 Our query-moment matchers are still far away from the best possible 'oracle' query-moment matchers

Method	R1		mAP
	@0.5	@0.7	avg
ShotDetect + CLIP	40.24	25.94	24.82
ShotDetect + Fine-tuned-CLIP	42.12	27.89	25.50
ShotDetect + Oracle Matcher	63.94	41.49	30.98
Supervised UMT (SoTA)	60.83	43.26	38.08

4.2 We do not beat pretrained + supervised VMR approaches, despite using a pretrained V+L model (CLIP)

Category	Method	R1		mAP		
		@0.5	@0.7	@0.5	@0.75	avg
Zero Shot	CLIP+Watershed((Lei et al., 2021))	16.88	5.19	18.11	7.00	7.67
	SlidingWindow + VideoCaptioning **(Ours)**	19.60	6.00	25.94	6.00	9.58
	ShotDetect+VideoCaptioning **(Ours)**	22.25	14.71	28.90	17.30	18.06
	SlidingWindow + CLIP **(Ours)**	29.71	8.86	35.26	8.31	13.42
	ShotDetect+CLIP **(Ours)**	40.24	25.94	41.74	24.11	24.82
	ShotDetect+CLIP+SimpleWatershed **(Ours)**	**48.33**	**30.96**	**46.94**	**25.75**	**27.96**
VMR-Sup	MCN*(Hendricks et al. (2017))	11.41	2.72	24.94	8.22	10.67
	CAL*(Escorcia et al. (2019))	25.49	11.54	23.40	7.65	9.89
	XML*(Lei et al. (2020))	41.83	**30.35**	44.63	**31.73**	**32.14**
	M-DETR w/o saliency loss(Lei et al. (2021))	**45.03**	25.81	**48.42**	21.91	24.68
VMR-Sup+Saliency	XML+*(Lei et al. (2021))	46.69	33.46	47.89	34.67	34.90
	M-DETR*(Lei et al. (2021))	52.89	33.02	54.82	29.40	30.73
	M-DETR w/ PT(Lei et al. (2021))	59.74	41.10	59.90	35.42	36.19
	M-DETR w/ PT*(Lei et al. (2021))	59.78	40.33	**60.51**	35.36	36.14
	UMT*(Liu et al. (2022))	56.23	41.18	53.83	37.01	36.12
	UMT w/ PT*(Liu et al. (2022))	**60.83**	**43.26**	57.33	**39.12**	**38.08**

4. Conclusion

1. A simple shot detector reveals CLIP's query-moment matching power and leads to performance **close to some supervised approaches on VMR**.
2. The shot detection method is espeically good at **detecting short segments**, outperforming strong supervised and pretrained models.
3. Despite the simplicity of the shot detector, CLIP is not the best query-moment matcher for it; there's quite some room for improved matchers.