Egocentric Video-Language Pretraining

Kevin Qinghong Lin[1], Alex Jinpeng Wang[1], Mattia Soldan[3], Michael Wray[2], Rui Yan[1], Eric Zhongcong Xu[1], Difei Gao[1], Rongcheng Tu[4], Wenzhe Zhao[4], Weijie Kong[4], Chengfei Cai[4], Hongfa Wang[4], Dima Damen[2], Bernard Ghanem[3], Wei Liu[4], Mike Zheng Shou[1]

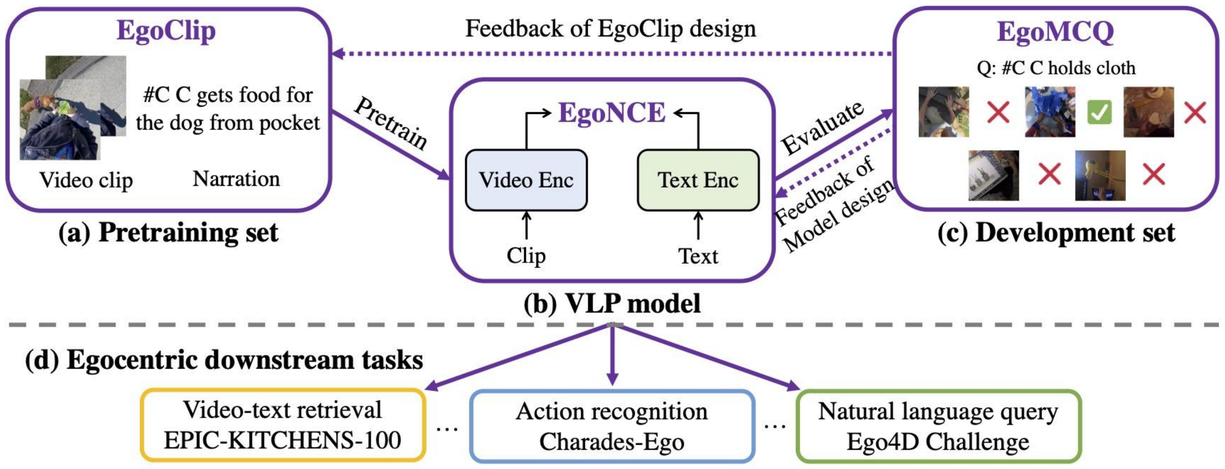
[1]Show Lab @ NUS [2]University of Bristol [3]IVUL @ KAUST [4]Tencent Data Platform

Contact: kevin.qh.lin@gmail.com
Code and model are available at github.com/showlab/EgoVLP

Highlight: the first egocentric vision-language pretrained model

Motivations: Existing VLP models are pretrained on Large-scale 3rd-person view datasets. In contrast, humans perceive the world in an egocentric way. How can we create an Egocentric VLP model?

Contributions: We pioneer Egocentric Video-Language Pretraining from pretraining dataset, model and development benchmark; the resulted pretrained model exhibits strong performance on five downstream tasks across three egocentric datasets.



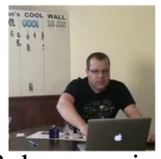
Experimental Results

Methods	Vis Enc Input	# Frames	Vis-text PT	mAP (%) V→T	mAP (%) T→V	mAP (%) Avg	nDCG (%) V→T	nDCG (%) T→V	nDCG (%) Avg
Random	-	-	-	5.7	5.6	5.7	10.8	10.9	10.9.
MI-MM	S3D [42]	32	HowTo100M	34.8	23.6	29.2	47.1	42.4	44.7
MME [43]	TBN † [14]	25	-	43.0	34.0	38.5	50.1	46.9	48.5
JPoSE [43]	TBN † [14]	25	-	49.9	38.1	44.0	55.5	51.6	53.5
Frozen	Raw Videos	4	-	38.8	29.7	34.2	50.5	48.3	49.4
Frozen	Raw Videos	4	HowTo100M	39.2	30.1	34.7	50.7	48.7	49.7
Frozen	Raw Videos	4	CC3M+WebVid-2M	41.2	31.6	36.4	52.7	50.2	51.4
Frozen	Raw Videos	4	EgoClip	44.5	34.7	39.6	55.7	52.9	54.3
Frozen+EgoNCE	Raw Videos	4	EgoClip	**45.1**	**35.3**	**40.2**	**56.2**	**53.5**	**54.8**
Frozen	Raw Videos	16	CC3M+WebVid-2M	45.8	36.0	40.9	57.2	54.3	55.8
Frozen+EgoNCE	Raw Videos	16	EgoClip	**49.9**	**40.1**	**45.0**	**60.9**	**57.9**	**59.4**
Frozen	Raw Videos	4	HowTo100M	6.8	6.3	6.5	11.6	12.8	12.2
Frozen	Raw Videos	4	CC3M+WebVid-2M	8.6	7.4	8.0	14.5	14.6	14.5
Frozen	Raw Videos	4	EgoClip	17.9	13.1	15.5	23.0	21.2	22.1
Frozen+EgoNCE	Raw Videos	4	EgoClip	**19.4**	**13.9**	**16.6**	**24.1**	**22.0**	**23.1**

Results on EPIC-Kitchens-100 text-video retrieval, the grey color rows are zero-shot evaluation.

Key observations: In the Egocentric domain, the same VLP model, different pretraining datasets; EgoClip (3.8M) significantly outperforms 3rd person view datasets HowTo100M and CC3M + WebVid2M in both zero-shot and fine-tune settings. EgoNCE further boosts the performance.

An Egocentric Video-Langauge Pretraining dataset 👉 EgoClip

Dataset	Ego?	Domain	Dur (hrs)	# Clips	# Texts	Example
MSR-VTT [1]	✗	diverse	40	10K	200K	
YouCook2 [16]	✗	cooking	176	14K	14K	
ActivityNet Captions [7]	✗	action	849	100K	100K	
WebVid-2M [3]	✗	diverse	13K	2.5M	2.5M	
HowTo100M [10]	✗	instructional	134K	136M	136M	3rd-person view
Charades-Ego [17]	✓	home	34	30K	30K	
UT-Ego [18]	✓	diverse	37	11K	11K	
Disneyworld [19]	✓	disneyland	42	15K	15K	
EPIC-KITCHENS-100 [20]	✓	kitchen	100	90K	90K	
EgoClip	✓	**diverse**	**2.9K**	**3.8M**	**3.8M**	1st-person view

EgoClip, a 1st-person video-text pretraining dataset comprising 3.8M clip-text pairs well-chosen from Ego4D, covering a large variety of human daily activities.

An Egocentric-friendly Pretraining Objective 👉 EgoNCE

$$\mathcal{L}_{\text{v2t}}^{\text{ego}} = -\frac{1}{|\widetilde{\mathcal{B}}|}\sum_{i\in\widetilde{\mathcal{B}}}\log\frac{\sum_{k\in\mathcal{P}_i}\exp(\mathbf{v}_i^T\mathbf{t}_k/\tau)}{\sum_{j\in\mathcal{B}}\left(\exp(\mathbf{v}_i^T\mathbf{t}_j/\tau)+\exp(\mathbf{v}_i^T\mathbf{t}_{j'}/\tau)\right)}.$$

Postive sample: share at least one noun and one verb.
Negative sample: close in time within the same video.

EgoNCE, a novel pretraining objective, which adapts video-text contrastive learning to the egocentric domain by mining egocentric-aware positive and negative samples.

A Benchmark for Egocentric VLP Development 👉 EgoMCQ

EgoMCQ	Inter-video					Intra-video				
Text query	#C C picks the silicone sealant					#C C carries paint bucket down the ladder				
Select the correct video clip from 5 candidates	 (a)	 (b)	 (c)	 (d)	 (e)	 (a)	 (b)	 (c)	 (d)	 (e)
Answer with GT	#C C places the camping seat down ✗	#C C holds the power drill with both hands. ✗	#C C picks the silicone sealant ✓	#C C takes a stone ✗	#C C cuts the green bean into pieces ✗	#C C holds paintbrush with both hands ✗	#C C turns paintbrush in his left hand ✗	#C C shifts paintbrush to right hand ✗	#C C drops paintbrush on paint bucket ✗	#C C carries paint bucket down the ladder ✓

EgoMCQ, a development benchmark that is close to EgoClip and hence can support effective validation and fast exploration of design decisions in pretraining dataset and model. EgoMCQ includes two settings: ``Inter-video'' (left) and ``Intra-video'' (right). The latter is more challenging.