NEURAL INFORMATION PROCESSING SYSTEMS

Egocentric Video-Language Pretraining

Presentor: Kevin Qinghong Lin

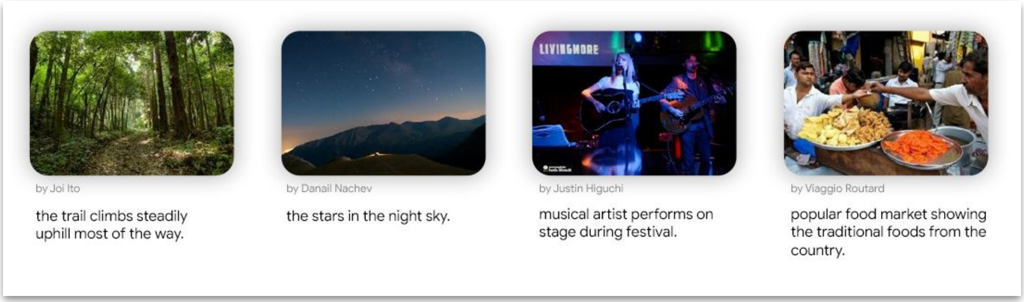
Show Lab @ NUS, U of Bristol, IVUL @ KAUST, Tencent

Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Weijie Kong Wenzhe Zhao, Chengfei Cai, Hongfa Wang, Bernard Ghanem, Dima Damen, Wei Liu, and Mike Zheng Shou. Egocentric Video-Langugae Pretraining. arXiv preprint arXiv:2206.01670, 2022.

Background

- Existing VLP models are pretrained on **Large-scale 3rd-person view datasets**

HowTo100M (ICCV'19)



Conceptual Captions 3M (ACL'18)

Easy to get, Noisy, Edited…

WebVid 2.5M (ICCV'21)

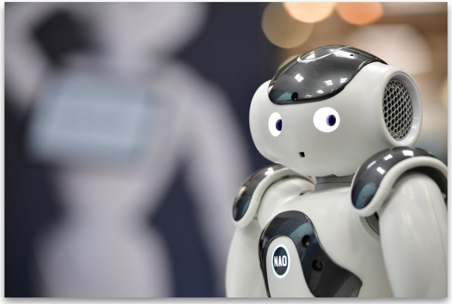
demo video samples from EPIC-Kitchens (PAMI’20)

Background

- In contrast, humans perceive the world in an **egocentric** way.
- Egocentric videos are important in many real-world applications

Egocentric videos

AR/VR
Smart glass



Robot

Motivation

- Would VLP model pretrained on <u>3rd person view videos</u> work well on <u>egocentric videos</u>?

- If not, how can we create an **Egocentric VLP** model?

Motivation

- Previous **Egocentric datasets** are of **small data scale and domain-specific**, making video-language pre-training impossible.

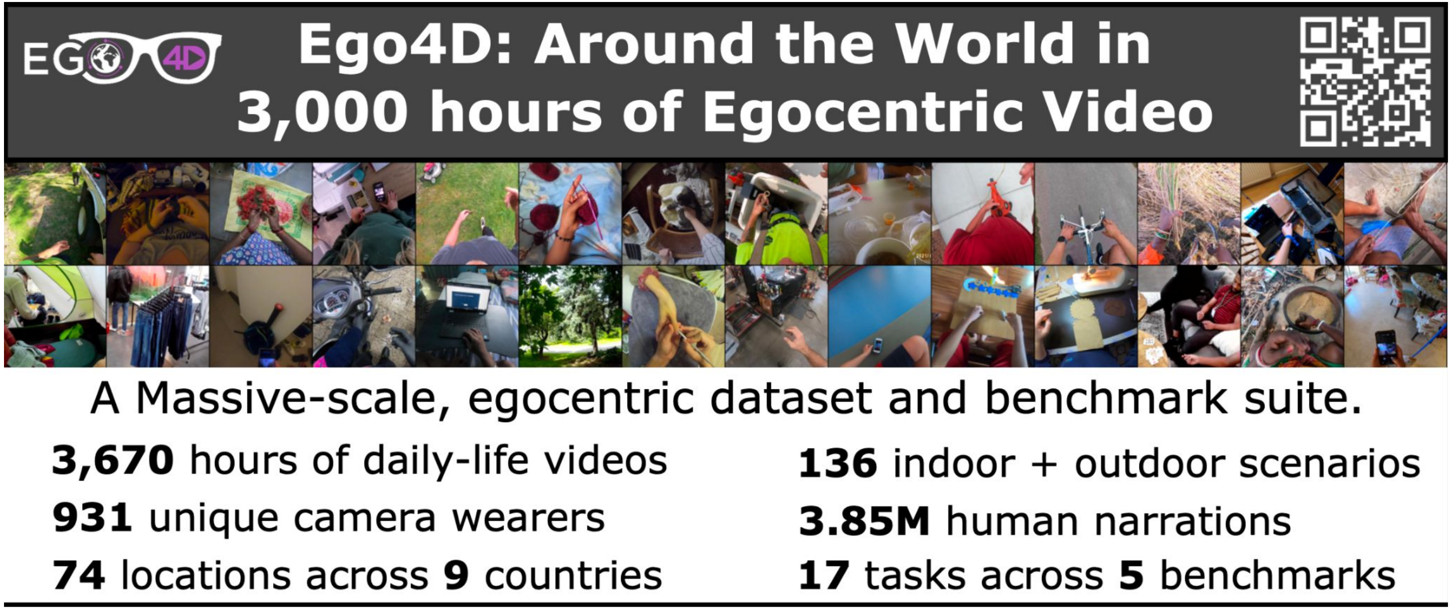
Dataset	Ego?	Domain	Dur (hrs)	# Clips	# Texts	Example
MSR-VTT [17]	✗	diverse	40	10K	200K	
YouCook2 [18]	✗	cooking	176	14K	14K	
ActivityNet Captions [7]	✗	action	849	100K	100K	
WebVid-2M [11]	✗	diverse	13K	2.5M	2.5M	
HowTo100M [10]	✗	instructional	134K	136M	136M	3rd-person view
Charades-Ego [19]	✓	home	34	30K	30K	
UT-Ego [20]	✓	diverse	37	11K	11K	
Disneyworld [21]	✓	disneyland	42	15K	15K	
EPIC-KITCHENS-100 [22]	✓	kitchen	100	90K	90K	
EgoClip	✓	**diverse**	**2.9K**	**3.8M**	**3.8M**	1st-person view

Table 1: Comparison of our proposed EgoClip pretraining dataset against the mainstream video-language datasets (top) and egocentric datasets (bottom).

Egocentric videos are expensive!

Ego4D Recap

- **Ego4D** unlocks the **Egocentric VLP**!



https://ego4d-data.org/

Ego4D for VL Pre-training?

- Research Q1: How to create pre-training **dataset**?

- Research Q2: How to design pre-training **model**?

- Research Q3: What benchmark we shall **evaluate** on?

TL;DR

- Create a Large-scale VL pretraining set of **3.8M video-text pairs** from Ego4D: **EgoClip**

- Propose an Egocentric-friendly video-text pretraining objective: **EgoNCE**

- Construct a development set for evaluating Egocentric VL Pre-training: **EgoMCQ**

- Significant gains on **5 egocentric benchmarks** across **3 datasets:**
  - [Ego4D Challenges] **Object State Change Classifcaition**: Acc from 68.7% to 73.9%. (+5.2%, **1st Place**)
  - [EPIC-KITCHENS Challenges] **Multi-Instance Retrieval**: nDCG (avg) from 53.5% to 59.4%. (+5.9%, **1st Place**)
  - [Ego4D Challenges] **Natural Language Query**: R@1 (IoU=0.3) from 5.45% to 10.84%. (+5.4%, **2nd Place**)
  - [Ego4D Challenges] **Moment Query**: R@1 (IoU=0.3) from 33.45% to 40.43%. (+7.0%)
  - [Charades-Ego] **Action Recognition**: MAP from 30.1% to 32.1%. (+2.0%)

Q1, Egocentric Video-Language Pre-training set 👉 EgoClip

1. Issue of undesired data source and data noise

EGO 4D

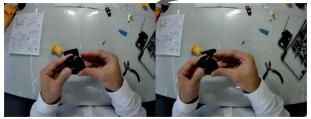
3.6K videos hours &
5M narrations

Data Filtering

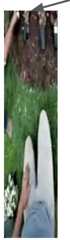
1. Filter unnarrated videos.
2. Do not touch **Ego4D benchmark val/test set.**

Data Cleaning

2.9K videos hours
& 3.8M narrations for VLP



Case 1, Cut half of partial stereo videos



Case 2, Discard videos with ratio of larger/smaller size > 2

#C C writes **unsure** on paper

Case 3, Discard narration with unsure tag

#C C sits

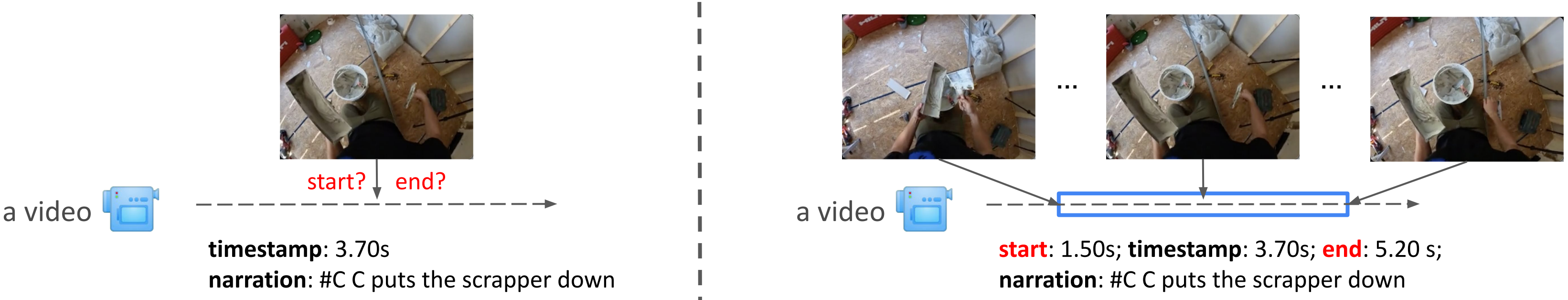
Case 4, narration only one word

Q1, Egocentric Video-Language Pre-training set 👉 EgoClip

2. Issue of no direct <clip, text> pair
    a. **Issue of no direct clip:** Ego4D narration is annotated for a **moment** rather than for an interval
    b. **Our approach**: *a contextual variable-length clip pairing strategy*
        - Measure the clip length β_i according to each video

Watching TV (352.9 sec) v.s.
Cooking in kitchen (0.9 sec)

$$[t_i^{start}, t_i^{end}] = [t_i - \beta_i/2\alpha,\ t_i + \beta_i/2\alpha],$$



a. Current annotation form

b. Our video-text pair definition

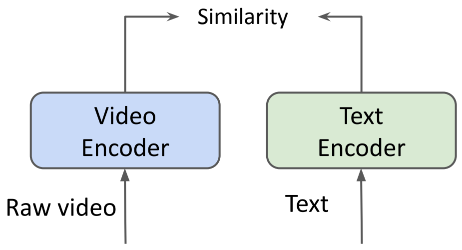
Finally, we create **3.8M** clip-text dataset for video-language pretraining, which we named **EgoClip**.

Bain M, Nagrani A, Varol G, et al. Frozen in time: A joint video and image encoder for end-to-end retrieval[C] ICCV. 2021: 1728-1738.
Lei J, Li L, Zhou L, et al. Less is more: Clipbert for video-and-language learning via sparse sampling[C] CVPR. 2021: 7331-7341.

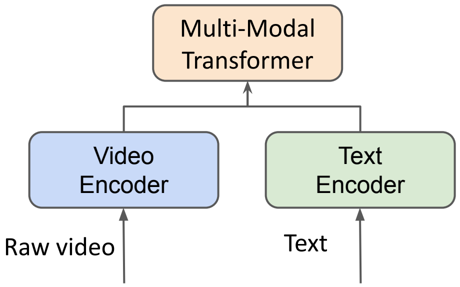
Q2, Egocentric VLP model

- Design of Pretraining Model Framework?
  - What we hope:
    1. Raw videos as input for **end-to-end** training
    2. Efficient for **video-text retrieval** — fundamental task of video-language understanding
    3. Support **video-only tasks** e.g., action recognition

Thus, we go with **Dual-stream transformer** architecture like Frozen (Bain M, ICCV'21) instead of Single-stream, e.g. ClipBert (Lei J, CVPR'21)



a. Dual-stream VLP Model



b. Single-stream VLP Model

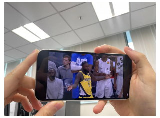
Currently following Frozen, we use **TimeSformer** for video-encoder and **DistillBERT** for text-encoder.

Q2, Egocentric Pretraining Objective 👉 EgoNCE

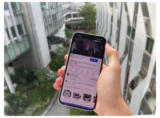
- Design of Pretraining objectives?
    - Universal Video-text contrastive learning: **InfoNCE**

$$\mathcal{L}_{\mathrm{v2t}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \log \frac{\exp(\mathbf{v}_i^T \mathbf{t}_i/\tau)}{\sum_{j\in\mathcal{B}} \exp(\mathbf{v}_i^T \mathbf{t}_j/\tau)},$$

    - InfoNCE regards the sample itself as supervision and others as negatives, which cannot tackle **two challenges in Egocentric pretraining**:
        1. The same action often occurs in different scenarios, e.g.,



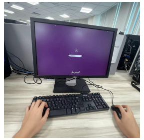
v.s.



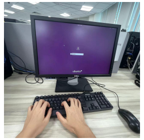
watching the phone when lying in room

watching the phone when walking outdoors

        2. Different actions appearing in the same scenario tend to have minor visual differences, e.g.,



v.s.



moving the mouse when working at a desk

typing on the keyboard when working at a desk

Q2, Egocentric Pretraining Objective 👉 EgoNCE

- Design of Pretraining objectives?
  - Universal Video-text contrastive learning: **InfoNCE**

$$\mathcal{L}_{\mathrm{v2t}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \log \frac{\exp(\mathbf{v}_i^T \mathbf{t}_i/\tau)}{\sum_{j\in\mathcal{B}} \exp(\mathbf{v}_i^T \mathbf{t}_j/\tau)},$$

  - InfoNCE regards the sample itself as supervision and others as negatives, which cannot tackle **two unique challenges in Egocentric pretraining**:
    1. The same action often occurs in different scenarios
    2. Different actions appearing in the same scenario tend to have minor visual differences
  - Novel method **EgoNCE** to leverage positive and negative samples in **egocentric domain**
    1. **Positive Sampling** 👉 based on action = <verb + noun>
    2. **Negative Sampling** 👉 based on temporally adjacent in same video

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

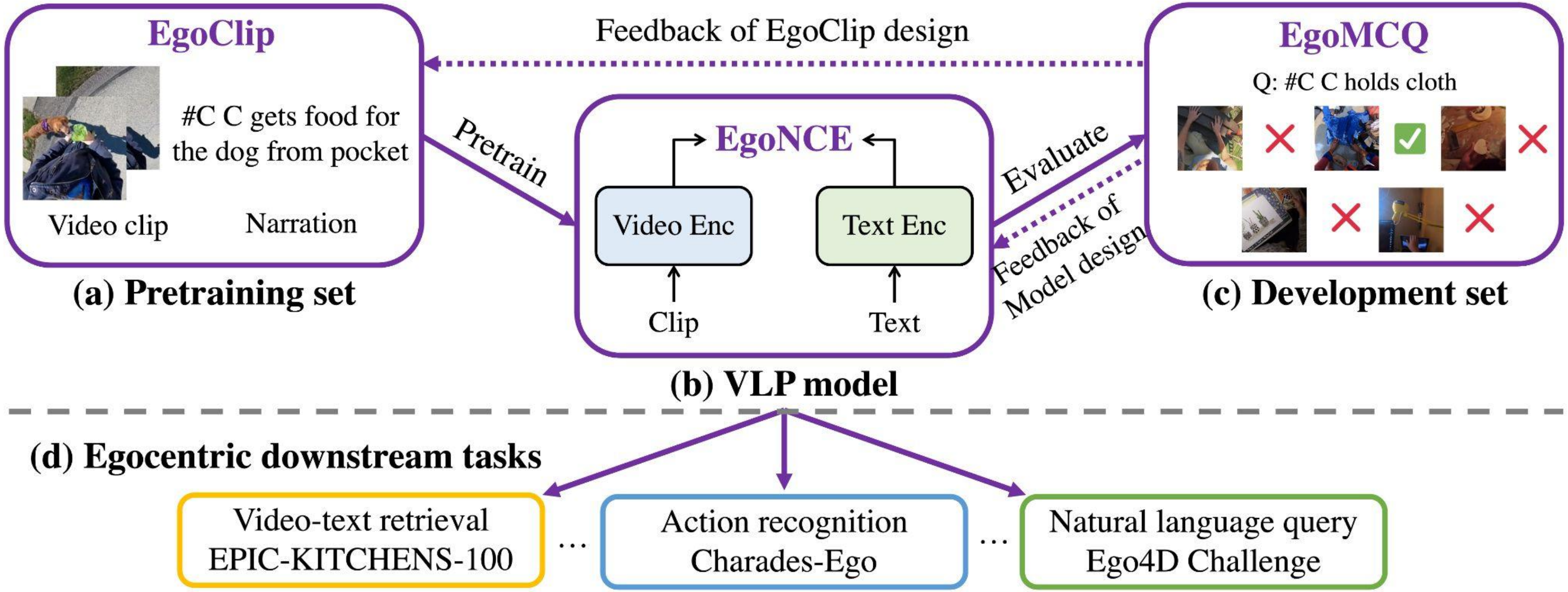
Q3, What benchmark we shall evaluate on?

- Before we transfer our model to other downstream benchmarks…
- We need a dev benchmark that is well aligned with
  - Our pretraining data domain i.e. in-the-wild of Ego4D
  - Our pretraining task i.e. video-text alignment
- This dev benchmark serves as a intermediate step for tuning our pretraining, avoiding the issues may encounter in transfering

Benchmark	**Domain**	**Task**
EPIC-KITCHENs	Cooking ❌	video-text retrieval ✅
Charades-Ego	Indoor ❌	action recognition ❌
Ego4D benchmarks	In-the-wild ✅	moment localization, forecasting ❌
What we'd like to have	In-the-wild ✅	video-text alignment ✅

Q3, What benchmark we shall evaluate on?

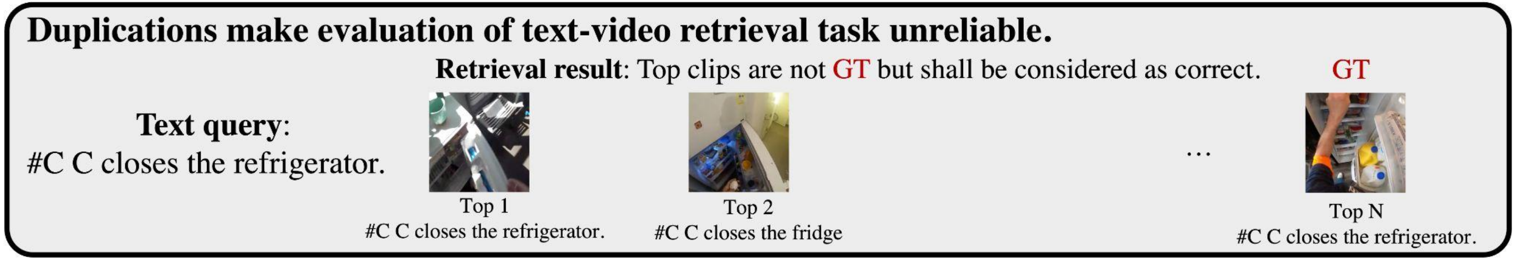
- Data flow of EgoVLP



Damen D, Doughty H, Farinella G M, et al. Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100[J]. IJCV, 2022, 130(1): 33-55.

Q3, A Benchmark for Egocentric VLP Development 👉 EgoMCQ

- **Roadblock 1**: video-text retrieval is not suitable
  - Issue of “one-to-many”
    - Often multiple clips could have similar narration, hard to evaluate for text-video retrieval



- **Solutions:**
  - **nDCG** is designed to tackle this issue (Damen et al. IJCV’21.), but requires additional annotations
  - **Automatic de-dup** methods (e.g. based on Bert text feature similarity thresholding), works not well
  - **Final strategy:**
    - **Repetitions are still allowed** in the pre-training dataset to **ensure diversity**.
    - **Do de-dup** when preparing the dev benchmark for evaluation.

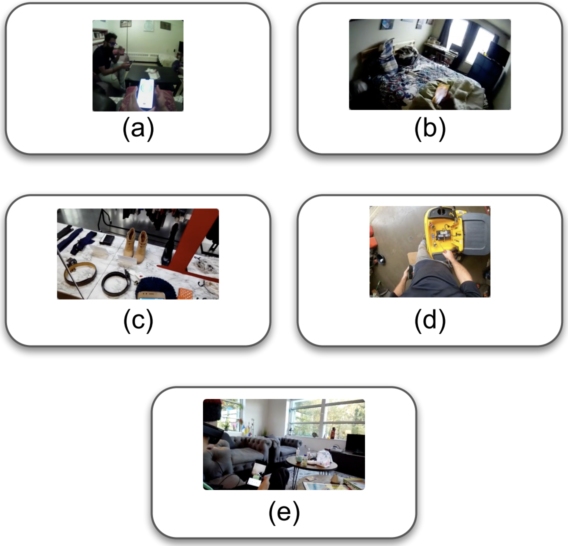
Q3, A Benchmark for Egocentric VLP Development 👉 EgoMCQ

- Our approach: Multi-Choices Question
  - “one-to-many” issue is alleviated **among fewer options (i.e. 5).**
  - A specific form of video-text retrieval and shares the same purpose.
- We propose **EgoMCQ** — what is EgoMCQ?

1. Given a text query:

#C C inserts phone into pocket

2. Choose the correct video clip from following 5 candidates.



3. Ground-Truth

Q3, A Benchmark for Egocentric VLP Development 👉 EgoMCQ

- **Roadblock 2**: How to group question and choices?
- **Our strategy**: De-dup within five options (consider **synonyms**) and propose to evaluate on **2 different modes**
  - **Inter-video**
    - Options from different videos and vary widely in content.
  - **Intra-video**
    - Grouping five continuous clips together, the harder mode.

EgoMCQ	Inter-video					Intra-video				
Text query	#C C carries paint bucket down the ladder					#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 ✅

Ego4D for VL Pre-training?

- Research Q1: How to create pre-training **dataset**? 👉 EgoClip
- Research Q2: How to design pre-training **model**? 👉 EgoNCE
- Research Q3: What benchmark we shall **evaluate** on? 👉 EgoMCQ

Experiments

- How well EgoVLP transfer to egocentric downstream tasks?
- How’s the designs of EgoClip, EgoNCE, and EgoMCQ?

Experiments

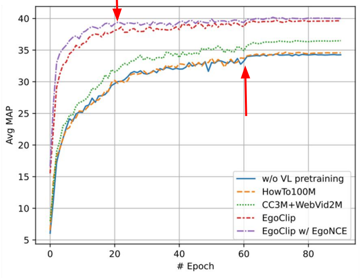
- Transfer EgoVLP to **EPIC-KITCHENS-100**
  - Task of **Multi-Instance Retrieval:** A type of **Text-video retrieval**
  - Metric **mDCG** use label to calculate semantic relevance and is more comprehensive than **mAP**.

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 [39]	32	HowTo100M	34.8	23.6	29.2	47.1	42.4	44.7
MME [40]	TBN † [15]	25	-	43.0	34.0	38.5	50.1	46.9	48.5
JPoSE [40]	TBN † [15]	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+WebVid2M	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+WebVid2M	45.8	36.0	40.9	57.2	54.3	55.8
Frozen+EgoNCE	Raw Videos	16	EgoClip	**49.9**	**40.5**	**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+WebVid2M	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**

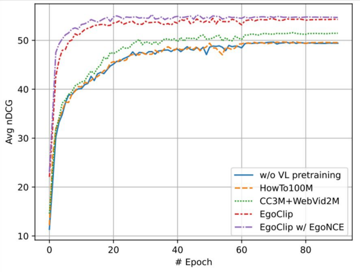
Table 4: Performance of the EPIC-KITCHENS-100 Multi-Instance Retrieval. Note that TBN † feature [15] are a combination of three modalities: RGB, Flow and Audio. Conversely, our approach only relies on RGB input. The grey highlighted rows correspond to **zero-shot evaluation**.

***Same model, different pretraining datasets**: EgoClip outperforms 3rd person view dataset HowTo100M and WebVid*

***EgoClip pretraining convergence faster** (20 v.s. 60 epoches)*



(a) mAP with training epoch



(b) nDCG with training epoch

- Observation
  1. In **zero-shot** settings, **EgoVLP** boosts **Avg mAP and nDCG** of **CC3M+WebVid2M** with **8.6%**
  2. In **fine-tune** settings, **EgoVLP** advances **Avg mDCG** of **JPoSE** with **5.9%** under fewer frames (16 vs 25) and less inputs modality (RGB+Text vs RGB/Flow/Audio+Text)

Experiments

- Transfer EgoVLP to **Ego4D challenge benchmarks**

Task 1: Natural Language Query (Text-video Localization)

Methods	Video-text Pre-extrated Features Vis-text Enc	Vis-text PT	IoU=0.3 R@1	IoU=0.3 R@5	IoU=0.5 R@1	IoU=0.5 R@5
2D-TAN [24]	SlowFast+BERT	-	5.04	12.89	2.02	5.88
VSLNet [44]	SlowFast+BERT	-	5.45	10.74	3.12	6.63
VSLNet [44]	Frozen	HowTo100M	3.95	8.72	2.01	4.62
VSLNet [44]	Frozen	CC3M+WebVid2M	5.06	10.30	2.71	6.69
VSLNet [44]	Frozen	EgoClip	<u>10.53</u>	<u>17.94</u>	<u>5.96</u>	<u>11.85</u>
VSLNet [44]	Frozen+EgoNCE	EgoClip	**10.84**	**18.84**	**6.81**	**13.45**

Table 6: Recall for several IoU on the NLQ task's val. set.

Task 2: Object State Change Classification (Action Recognition)

Methods	Vis-Text PT	Acc. (%)
Always Positive	-	48.1
Bi-d LSTM [46]	ImageNet	65.3
I3D ResNet-50 [47]	-	68.7
Frozen	-	70.3
Frozen	HowTo100M	71.7
Frozen	CC3M+WebVid2M	71.5
Frozen	EgoClip	<u>73.4</u>
Frozen+EgoNCE	EgoClip	**73.9**

Table 8: Accuracy metric on the Object State Change Classification task's val set.

Task 3: Moment Query (Temporal Action Localization)

Methods	Video Pre-extracted Features Vis Enc	Vis-text PT	IoU=0.3 R@1	IoU=0.3 R@5	IoU=0.5 R@1	IoU=0.5 R@5	IoU=0.7 R@1	IoU=0.7 R@5	mAP (%) @ IoU 0.1	0.3	0.5	Avg
VSGN [45]	SlowFast	-	33.45	58.43	25.16	46.18	15.36	25.81	9.10	5.76	3.41	6.03
VSGN [45]	Frozen	HowTo100M	31.40	52.61	22.28	41.29	13.41	23.21	9.83	6.72	3.84	6.72
VSGN [45]	Frozen	CC3M+WebVid2M	32.08	56.40	23.46	43.81	13.73	23.77	9.83	6.40	3.86	6.58
VSGN [45]	Frozen	EgoClip	<u>40.06</u>	<u>63.71</u>	<u>29.59</u>	<u>48.32</u>	<u>17.41</u>	<u>26.33</u>	<u>15.90</u>	<u>10.54</u>	<u>6.19</u>	<u>10.69</u>
VSGN [45]	Frozen+EgoNCE	EgoClip	**40.43**	**65.67**	**30.14**	**51.98**	**19.06**	**29.77**	**16.63**	**11.45**	**6.57**	**11.39**

Table 7: Recall and mAP metrics for several IoU on the Moment Query task's val. set.

Experiments

- Transfer EgoVLP to **Charades-Ego**
  - We prompt video-text knowledge to the task of **Action recognition**

Methods	Vis Enc	# Frames	Vis-Text PT	Train / FT Data	mAP (%)
Actor [41]	ResNet-152	25	-	Charades-Ego (1st + 3rd)	20.0
SSDA [42]	I3D	32	-	Charades-Ego (1st + 3rd)	23.1
I3D [42]	I3D	32	-	Charades-Ego (1st).	25.8
Ego-Exo [43]	SlowFast (Res-101)	32	-	Charades-Ego (1st)	30.1
Frozen	TimeSformer	16	-	Charades-Ego (1st)	28.8
Frozen	TimeSformer	16	HowTo100M	Charades-Ego (1st)	28.3
Frozen	TimeSformer	16	CC3M+WebVid2M	Charades-Ego (1st)	30.9
Frozen	TimeSformer	16	EgoClip	Charades-Ego (1st)	31.2
Frozen+EgoNCE	TimeSformer	16	EgoClip	Charades-Ego (1st)	**32.1**
Frozen	TimeSformer	16	HowTo100M	-	9.2
Frozen	TimeSformer	16	CC3M+WebVid2M	-	20.9
Frozen	TimeSformer	16	EgoClip	-	23.6
Frozen+EgoNCE	TimeSformer	16	EgoClip	-	$\boxed{\textbf{25.0}}$

Table 5: Performance of the action recognition on Charades-Ego dataset (First-person test set). The grey highlighted rows correspond to **zero-shot evaluation**.

Experiments

- Ablation Studies on EgoMCQ
  - EgoClip (Clip creation strategy)

Clip creation strategy	Clip's length (s) Avg ± Std	EgoMCQ Acc (%) Inter-video	EgoMCQ Acc (%) Intra-video	Zero-shot T↔V Retrieval [22] mAP (avg)	Zero-shot T↔V Retrieval [22] nDCG (avg)
(a) $[t_i, t_i{+}\alpha]$	5.0 ± 0.0	87.66	39.72	19.6	12.3
(b) $[t_i{-}\alpha/2, t_i{+}\alpha/2]$	5.0 ± 0.0	89.23	41.68	20.6	13.7
(c) $[t_{i-1}, t_{i+1}]$	10.0 ± 38.2	88.13	40.62	20.6	13.7
(d) $[t_i{-}\beta_i/2, t_i{+}\beta_i/2]$	4.9 ± 4.7	89.74	44.82	21.1	14.5
(e) $[t_i{-}\beta_i/4, t_i{+}\beta_i/4]$	2.4 ± 2.4	**90.23**	49.67	21.9	15.3
(f) $[t_i{-}\beta_i/2\alpha, t_i{+}\beta_i/2\alpha]$	1.0 ± 0.9	89.36	**51.51**	**22.1**	**15.5**

Table 2: Results on our development set EgoMCQ and video-text retrieval on EPIC-KITCHENS-100 when using different strategies in the creation of EgoClip, where t_i, α, β_i are defined in Eq. 1. In all experiments, we bold the **best results** and underlined the second best results.

Under same average length, our ***varied-length*** *(d) outperform fixed-length (b)*

  - EgoNCE (positive & negative sampling)

Variants	Accuracy (%) Intra-video	Accuracy (%) Inter-video
InfoNCE	89.4	51.5
(a) w/ Pos, noun	82.9 (6.5 ↓)	42.3 (9.2 ↓)
(b) w/ Pos, verb	86.9 (2.5 ↓)	50.5 (1.0 ↓)
(c) w/ Pos, noun & verb	89.7 (0.4 ↑)	53.6 (2.1 ↑)
(d) w/ Neg, random	88.3 (1.1 ↓)	49.9 (1.6 ↓)
(e) w/ Neg, within video	89.7 (0.3 ↑)	53.0 (1.5 ↑)
(f) w/ Neg, within 1 min	89.5 (0.2 ↑)	54.5 (3.0 ↑)
(g) w/ Pos & Neg, **EgoNCE**	**90.6** (1.3 ↑)	**57.2** (5.7 ↑)

Table 3: Pretraining sampling strategy ablation. We evaluate accuracy performance on our development benchmark EgoMCQ.

Only noun / verb decrease performance, our ***<noun, verb>*** *brings gains*

Negative from same video is help, while ***close-in-time*** *further boost performance*

Potential directions

- Egocentric-Exocentric Domain adaptation

Pretraining set	**EPIC-Kitchens (1st)**		**MSR-VTT (3rd)**	
	nDCG	mAP	T2V R@1	V2T R@1
EgoClip (1st)	**23.4**	**16.5**	4.2	4.8
WebVid (3rd)	14.2	7.8	**18.1**	**15.9**
EgoClip + WebVid (1st+3rd)	22.0 ↓	15.0 ↓	17.5 ↓	15.3 ↓

- Egocentric foundation model like human
    - Generation task
    - Other modality e.g., audio

Want to know more?

- Preprint: https://arxiv.org/abs/2206.01670
- Github: https://github.com/showlab/EgoVLP
- Contact: kevin.qh.lin@gmail.com & mike.zheng.shou@gmail.com

Thank you!

Appreciate any questions and comments