X-MIC: Cross-Modal Instance Conditioning for Egocentric Action Generalization

Anna Kukleva[1,2], Fadime Sener[1], Edoardo Remelli[1], Bugra Tekin[1], Eric Sauser[1], Bernt Schiele[2], Shugao Ma[1]
[1]Meta Reality Labs; [2]Max Planck Institute for Informatics

CVPR SEATTLE, WA JUNE 17-21, 2024

Summary

◇ Addressing the task of **egocentric generalization**

⇒ real-world applications, e.g. AR adaptation to new environments

◇ **Simple yet effective** framework X-MIC

⇒ adaptation to *each* instance

⇒ encode egocentric and temporal information in text embedding

⇒ adaptation directly in the embedding space

◇ **Ego-Spatio-Temporal** Module

⇒ Combine spatial global context with hand-related information

Motivation

Standard
3rd person view

- Internet data
- Balanced class distribution
- Short clips

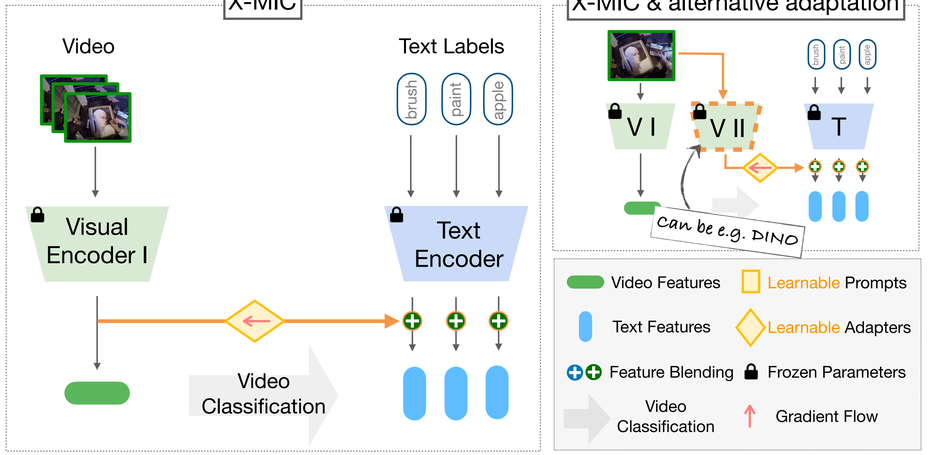
CLIP Zero-Shot is about 60%

Egocentric
1st person view

- Recorded data
- Long-tailed distribution
- Long-term context

CLIP Zero-Shot is less than 10%

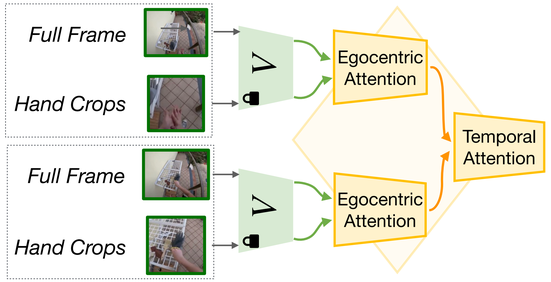
High-level Overview



- **no** changes to original models to keep ***generalization***
- adapt frozen classifier to **each** instance *individually*

• Late Fusion
• Cross-Modal (X-MIC)

Ego-Spatio-Temporal Adapter



egocentric & temporal information encoded **only** in text representations

Results

Within- & Cross- Dataset Evaluation

		Trained on Ego4D (E4D)					
	ta	Nouns			Verbs		
Evaluation dataset		E4D	EK	hm	E4D	EK	hm
ZS CLIP	-	5.89	8.74	7.03	2.18	4.25	2.88
CoOp	-	28.22	10.87	15.70	22.57	20.42	21.44
Co-CoOp	-	30.00	9.51	14.44	21.31	12.99	16.14
CLIP-Adapter	-	30.00	8.95	13.78	22.82	19.94	21.28
CLIP-Adapter*	✓	31.26	10.00	15.16	27.32	22.28	24.54
A5	✓	31.39	7.84	12.55	26.31	22.77	24.41
Vita-CLIP	✓	33.52	10.61	16.11	22.66	25.81	24.13
X-MIC	✓	33.54	15.35	21.06	28.93	26.48	27.65
X-MIC+DINO	✓	35.85	18.96	**24.80**	28.27	29.49	**28.86**

Generalization

Nouns		
shared	novel	hm
10.38	13.58	11.77
16.86	16.02	16.43
16.35	11.51	13.51
12.46	5.99	8.09
16.24	12.22	13.95
15.25	5.24	7.80
15.84	6.15	8.86
20.04	21.51	20.75
25.56	20.52	**22.76**

Influence of ego-spatio-temporal adapter

	Nouns		
	E4D	EK	hm
F	31.68	14.20	19.61
H	31.35	14.02	19.37
F+H	33.54	15.35	**21.06**

F = Full Frame ; H = Hand Crops

Generalization

Verbs		
shared	novel	hm
12.32	4.32	6.40
25.03	5.97	9.64
24.34	0.00	0.00
21.48	3.09	5.40
25.29	1.23	2.35
27.90	3.09	5.56
27.22	4.11	7.14
29.01	7.00	**11.27**
31.92	6.38	10.63

Influence of normalization

norm	Nouns		
	E4D	EK	hm
n1	33.54	15.35	**21.06**
none	32.64	14.34	19.92
n2,n3	32.74	14.59	20.19
n1,n2,n3	31.99	14.49	19.95
n1,n2	15.81	12.3	13.83
n1,n3	12.12	11.34	11.71

[n1] - *l2-norm of features after V encoder and before the adapter*
[n2] - *l2-norm of X- MIC vector before sum*
[n3] - *l2- norm of text features before sum*

	prompts
1	<class>
2	Image of a <class>
3	Video of a <class>
4	Egocentric image a <class>
5	Image of a hand holding a <class>
6	Egocentric image of a hand holding <class>

	Nouns			Verbs			
	ZS		X-MIC	ZS		X-MIC	
	E4D	EK	hm	E4D	EK	hm	
1	5.89	**8.74**	21.06	2.18	4.25	**27.65**	
2	**10.52**	6.75	20.31	3.28	5.40	27.21	
3	10.32	6.80	20.33	2.93	5.97	25.13	
4	9.61	7.11	21.04	2.98	3.83	26.10	
5	10.09	6.32	19.92	**3.29**	**9.87**	22.71	
6	9.23	6.86	**21.45**	2.41	6.24	21.01	