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

Supplementary Material

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, Saarland Informatics Campus
{annakukleva, famesener}@meta.com

In this supplementary, we provide additional implementation details in Sec. 6, extend the discussion on generalization performance to shared and novel action classes in Sec. 7, explore the synergies between various adaptation techniques and X-MIC framework in Sec. 8, and present additional ablation experiments of our X-MIC framework in Sec. 9.

6. Implementation Details

Transformer Block. In our Ego-spatio-temporal attention module, we utilize a sequence of transformer blocks b_S and b_T to capture spatial and temporal dependencies, respectively. The general structure of the transformer block given input tensors x of dimensionality D is depicted in Alg. 1. LN denotes LayerNorm [2], MHA denotes multi-head attention with 8 heads [35], and MLP denotes 2-layer multi-layer perception with the bottleneck $D/4$ and QuickGELU [11] as an activation function.

Algorithm 1 Transformer Block

```
Require: $x \in \mathcal{R}^D$
  $x \leftarrow x + MHA(LN(x))$
  $x \leftarrow x + MLP(LN(x))$
```

We note that b_S consist of one transformer block and b_T includes two sequential transformer blocks.

Data. For augmentations, we exclusively employ frame flipping during training. Furthermore, for the CLIP backbone, we resize frames so that the shortest side is 224, followed by a center crop of 224x224 and normalization. For the Lavila backbone, frames are directly resized to 224x224.

For Epic-Kitchens, we use the provided annotation boundaries to define action clips. Conversely, in the Ego4D FHO challenge, the boundaries for each clip are initially set as 8 seconds. However, upon observation, we note that actions typically span a duration shorter than 8 seconds, prompting us to uniformly shorten all clips to 4 seconds.

7. Zero-Shot Generalization Discussion

In Table 2, we present comprehensive cross-dataset results showcasing the generalization performance on both the Epic-Kitchens and Ego4D datasets. Notably, the Epic-Kitchens dataset is exclusive to kitchen-related scenes and actions, while the Ego4D dataset encompasses a diverse range of daily activities. We distinguish between subsets of shared and novel classes and provide additional details in the following paragraph.

Shared-Novel Classes. We categorize classes as ”shared” when there is an exact match in their names across datasets. For noun classes in Ego4D and Epic Kitchens, there exist 163 such shared classes, including examples like “apple”, “toaster” or “washing machine”. Consequently, the set of ”novel” noun classes for Ego4D comprises 358 classes, encompassing items such as “transistor”, “ambulance” and “stroller”. In contrast, the ”novel” noun classes for Epic-Kitchens total 137 and predominantly represent more detailed kitchen-related categories such as “mint”, “onion ring”, “scale”. In the domain of verb classes for both Ego4D and Epic Kitchens, we identify 51 *shared* classes, including actions such as “hold”, “hang” or “attach”. This results in 66 *novel* verb classes for Ego4D with examples like “park”, “repair” and “wave”. On the other hand, *novel* verb classes for Epic-Kitchens amount to 46, primarily encompassing more detailed kitchen-related actions like “slide”, “stab”, “unfreeze”. For a comprehensive list of classes, refer to the detailed class separation in Sec. 10.

8. Complementary of X-MIC

In Table 12, we illustrate the compatibility of our framework with other adaptation methods. Early fusion methods, where the uni-modal (U) method corresponds to CoOp [44] and the cross-modal (X) method corresponds to CoCoOp [43], demonstrate enhanced performance in both within- and cross-dataset evaluations. However, the integration of our framework with late fusion uni-modal adapters (Tt and Vv) does not further enhance the generalization while maintaining the overall high performance.

	Nouns			Verbs		
α	E4D	EK	hm	E4D	EK	hm
0.1	31.23	14.46	19.77	22.85	24.91	23.83
0.5	32.43	14.40	19.94	26.81	23.80	25.21
1.0	33.54	15.35	**21.06**	28.93	26.48	**27.65**
2.0	33.20	14.73	20.40	28.14	26.39	27.24
5.0	32.29	14.24	19.77	27.50	27.00	27.25

Table 8. **Influence of scale of X-MIC vector.** Trained on Ego4D.

	Nouns			Verbs		
	EK	E4D	hm	EK	E4D	hm
F	24.89	12.48	16.62	42.10	18.86	26.05
H	31.13	11.58	16.88	44.20	18.49	26.07
F+H	30.64	12.32	**17.57**	50.01	18.10	**26.58**

Table 9. **Influence of Ego-Spatial-Temporal attention.** F denotes full frames, H denotes hand crops. F+H correspond to our proposed attention module. All models share the same architecture of the temporal attention module. Trained on Epic-Kitchens.

			Nouns			Verbs		
Evaluation dataset			EK	E4D	hm	EK	E4D	hm
CLIP	ViT-L/16	Zero-Shot	8.74	5.89	7.03	4.25	2.18	2.88
		X-MIC	30.64	12.32	17.57	50.01	18.10	26.58
	ViT-L/14	Zero-Shot	13.88	8.40	10.46	9.70	8.57	9.10
		X-MIC	39.02	14.24	20.86	48.12	18.83	27.07
Lavila	ViT-L/14	Zero-Shot	31.06	24.99	27.69	15.74	6.19	8.88
		X-MIC	41.78	29.62	34.67	46.14	9.35	15.54

Table 10. **Influence of different backbones.** We compare the performance of CLIP ViT-L/14 with ViT-L/16. Additionally, we provide a comparison of CLIP backbone, pretrainied on text-image pairs, to Lavila backbone, pretrained on pairs of egocentric videos and narrations from full Ego4D. Trained on Epic-Kitchens (EK).

norm	Trained on Ego4D (E4D)						Trained on Epic-Kitchens (EK)					
	Nouns			Verbs			Nouns			Verbs		
	E4D	EK	hm	E4D	EK	hm	EK	E4D	hm	EK	E4D	hm
n1	33.54	15.35	**21.06**	28.93	26.48	**27.65**	30.64	12.32	**17.57**	50.01	18.10	26.58
none	32.64	14.34	19.92	27.79	25.76	26.74	32.54	10.34	15.69	43.25	19.53	**26.91**
n2,n3	32.74	14.59	20.19	27.55	24.49	25.93	34.08	10.48	16.03	49.32	16.49	24.71
n1,n2,n3	31.99	14.49	19.95	24.30	22.69	23.47	32.46	10.21	15.54	49.02	18.05	26.39
n1,n2	15.81	12.3	13.83	24.3	22.69	23.47	19.88	8.14	11.55	19.72	8.41	11.79
n1,n3	12.12	11.34	11.71	22.88	18.49	20.46	16.16	8.72	11.33	23.68	17.68	20.24

Table 11. **Influence of feature normalization**. Extended Table 6(main). [n1] corresponds to the normalization of visual features after the V_{II} encoder and before the adapter and demonstrates an optimal balance between normalization and no normalization. [n2] corresponds to the normalization of the X-MIC vector before summation with text representation. [n3] corresponds to the normalization of text representation before summation with the X-MIC vector.

	U/X	Nouns			Verbs		
	-Modal	E4D	EK	hm	E4D	EK	hm
Early Fusion	U	28.22	10.87	15.70	22.57	20.42	21.44
+X-MIC	U+X	29.83	11.94	17.05	24.99	22.72	23.80
Early Fusion	X	30.00	9.51	14.44	21.31	12.99	16.14
+X-MIC	X+X	30.33	10.34	15.42	25.53	25.06	25.29
X-MIC	X	33.54	15.35	21.06	28.93	26.48	27.65
+ Tt	X+U	33.66	14.82	20.58	28.41	26.69	27.52
+ Vv	X+U	32.75	15.20	20.77	28.36	25.85	27.05
+ Tt + Vv	X+U	33.23	15.13	20.79	28.20	26.29	27.21

Table 12. **X-MIC framework with other adaptation methods.** U denotes uni-modal methods, X denotes cross-modal methods. Tt denotes text uni-modal adapter for late fusion, Vv similarly denotes video uni-modal adapter for late fusion. Trained on Ego4D (E4D).

# frames	Nouns			Verbs		
	E4D	EK	hm	E4D	EK	hm
32	33.02	14.77	20.41	28.56	25.34	26.85
16	33.54	15.35	**21.06**	28.93	26.48	**27.65**
8	32.06	14.82	20.27	27.30	26.90	27.10
4	31.74	14.99	20.36	26.41	26.17	26.29
2	29.95	12.77	17.91	23.85	25.73	24.75

Table 13. **Influence of number of frames**. Trained on Ego4D.

	Nouns			Verbs		
	E4D	EK	hm	E4D	EK	hm
w/o	31.41	13.31	18.69	22.65	20.19	21.34
w/	33.54	15.35	**21.06**	28.93	26.48	**27.65**

Table 14. **Influence of temporal attention.** Replacing the temporal module with a simple average decreases verb and noun recognition. The models share the same architecture for Ego-Spatial attention module.

9. Ablations of X-MIC

In this section, we conduct additional ablation experiments to validate the efficacy of our design choices.

Temporal Attention. In Table 14, we present an ablation where we keep our ego-spatial attention but replace the temporal module with a simple average over frames. The results demonstrate significant improvements in verb recognition on Ego4D by 6.28%, and in verb generalization to Epic by 6.29% when employing temporal modeling. This aligns with expectations, as the temporal component encodes movements. We also observe improvements in noun recognition and generalization, indicating that recognizing nouns in an egocentric dataset requires more than just appearance; motion encoding is also beneficial *e.g.* for recognizing ”slicing an apple” action.

Number of frames. In Table 13, we showcase the influence of the number of frames sampled during both training and evaluation. Notably, the performance peaks with 16 and 8 sampled frames. Conversely, sampling only two frames per clip significantly diminishes performance across all classes.

Scale of X-MIC-vector. In Eq. 1 in our main paper, we extend our analysis to validate the importance of the scale of the X-MIC-vector. To be specific, according to Eq. 1 in our main paper is the following:

$$c = \mathrm{argmax}_t < \overline{e_t + \alpha A(\overline{V_{II}(x_v))}}, \bar{e}_v >, \tag{4}$$

where α is a scale factor. In Table 8, we vary the scale factor α from 0.1 to 5 and observe that higher values result in improved performance, particularly in the evaluation of verbs both within and across datasets.

Ego-Spatial-Temporal Attention. In Table 9, we provide supplementary results to highlight the impact of our ego-spatial-temporal attention module. We note consistent performance across models trained on Epic-Kitchens and Ego4D, see Table 3 in our main paper.

Different backbones. Table 10 presents a comparison between CLIP ViT-B/16 and ViT-L/14 models when trained on Epic-Kitchens. Furthermore, we evaluate the performance of image-text pre-training (CLIP model) and video egocentric pre-training (Lavila). Our findings f those outlined in Table 5 in our main paper.

Importance of normalization. Table 11 offers supplementary results to complement those in Table 6 our main paper. The outcomes closely align with the findings presented our main paper.

10. List of Classes

Shared Nouns:

spoon; plate; knife; pan; lid; bowl; drawer; sponge; glass; hand; fridge; cup; fork; bottle; onion; cloth; chopping board; bag; spatula; container; dough; water; meat; pot; potato; oil; cheese; bread; food; tray; pepper; colander; carrot; tomato; kettle; pasta; oven; sauce; paper; garlic; towel; egg; rice; mushroom; chicken; coffee; glove; leaf; sink; milk; jug; salad; dishwasher; cucumber; peach; flour; courgette; filter; butter; scissors; chopstick; blender; mat; spice; sausage; napkin; microwave; pizza; button; stock; grater; ladle; yoghurt; cereal; broccoli; brush; lemon; juicer; light; squash; leek; fish; lettuce; seed; foil; washing machine; corn; soup; clip; lighter; ginger; tea; nut; vinegar; rolling pin; pie; burger; book; tongs; cream; banana; paste; plug; teapot; floor; lime; bacon; sandwich; phone; thermometer; orange; basket; tablet; cake; avocado; chair; pancake; toaster; apple; chocolate; ice; handle; pea; yeast; coconut; spinach; apron; grape; kale; wire; asparagus; mango; kiwi; bean; whisk; remote control; label; celery; cabbage; ladder; battery; pear; funnel; wall; strawberry; shelf; straw; cork; window; bar; heater; watch; melon; popcorn; candle; balloon; computer; key; pillow; pen; plum; tape; camera;

Novel Nouns Ego4D:

arm; artwork; awl; axe; baby; baking soda; ball; ball bearing; baseboard; bat; bat; bathtub; batter; bead; beaker; bed; belt; bench; berry; beverage; bicycle; blanket; block; blower; bolt extractor; bookcase; bracelet; brake; brake pad; branch; brick; broom; bubble gum; bucket; buckle; butterfly; cabinet; calculator; caliper; can opener; canvas; car; card; cardboard; carpet; cart; cat; ceiling; cello; cement; chaff; chain; chalk; chip; chip; chip;

chisel; cigarette; circuit; clamp; clay; clock; coaster; coffee machine; comb; cooker; cookie; corner; countertop; crab; cracker; crayon; crochet; crowbar; curtain; cushion; cutter; decoration; derailleur; detergent; dice; dog; door; doorbell; dough mixer; doughnut; dress; drill; drill bit; drum; dumbbell; dust; duster; dustpan; eggplant; engine; envelope; eraser; facemask; fan; faucet; fence; file; filler; fishing rod; flash drive; flower; foam; foot; fries; fuel; game controller; garbage can; gasket; gate; gauge; gauze; gear; generator; glasses; glue; glue gun; golf club; gourd; grain; grapefruit; grass; grill; grinder; guava; guitar; hair; hammer; hanger; hat; hay; haystack; head; headphones; helmet; hinge; hole; horse; hose; house; ice cream; ink; iron; jack; jacket; ketchup; keyboard; leash; leg; lever; lock; lubricant; magnet; manure; mask; matchstick; medicine; metal; microscope; mirror; mixer; mold; money; mop; motorcycle; mouse; mouthmower; multimeter; nail cutter; nail gun; nail polish; necklace; needle; net; nozzle; nut; okra; paddle; paint; paint roller; paintbrush; palette; panel; pantspapaya; pastry; peanut; pedal; peel; peeler; peg; pencil; photo; piano; pickle; picture; pilot jet; pin; pipe; planer; plant; playing cards; plier; pole; pot; pump; pumpkin; purse; puzzle or game piece; rack; radio; rail; rake; razor blade; ring; rod; root; rope; router; rubber band; ruler; sand; sander; sandpaper; saw; scarf; scoopscraper; screw; screwdriver; sculpture; seasoning; set square; sewing machine; sharpener; shears; sheet; shell; shirt; shoe; shovel; shower head; sickle; sieve; sketch pad; skirt; slab; snorkel; soap; sock; socket; sofa; soil; solder iron; spacer; speaker; sphygmomanometer; spirit level; spray; spring; squeezer; stairs; stamp; stapler; steamer; steering wheel; stick; sticker; stone; stool; stove; strap; string; stroller; switch; syringe; table; taco; tape measure; television; tent; test tube; tie; tile; timer; toilet; toilet paper; toolbox; toothbrush; toothpick; torch; toy; tractor; trash; treadmill; tree; trimmer; trowel; truck; tweezer; umbrella; undergarment; vacuum; vacuum cleaner; valve; vase; video game; violin; wallet; wallpaper; watermelon; weighing scale; welding torch; wheat; wheel; wheelbarrow; windshield; wiper; wood; worm; wrapper; wrench; yam; zipper; zucchini; ambulance; back; bamboo; bandage; baton; bird; brownie; cash register; cassava; cocoa; cow; cupcake; drone; earplug; hotdog; marble; person; pipette; plunger; printer; putty; racket; ratchet; road; scaffold; stereo; transistor;

Novel Nouns Epic-Kitchens:

tap; cupboard; washing liquid; box; hob; package; bin; salt; jar; top; skin; coffee maker; rubbish; cutlery; can; heat; aubergine; chilli; mixture; clothes; tofu; olive; potato peeler; cover; kitchen towel; vegetable; plastic wrap; sugar; biscuit; wrap; scale; rest; drying rack; alarm; salmon; freezer; spreads; cap; curry; oatmeal; spring onion; holder; powder; egg shell; pork; oregano; food processor; recipe; liquid; pak choi; slow cooker; utensil; noodle; salami; kitchen; tuna; omelette; parsley; salad spinner; presser; coriander; bottle opener; lentil; blueberry; extractor fan; salt cellar; hummus; juice; green bean; knob; wine; pith; fishcakes; raisin; basil; paprika; caper; drink; stalk; turmeric; whetstone; thyme; lady finger; beef; blackberry; slicer; hoover; breadstick; roll; cocktail; crisp; beer; dust pan; washing powder; backpack; cumin; pizza cutter; air; quorn; almond; tv; egg scotch; stand; vide sous machine; masher; hand guard; shrimp; fruit; artichoke; cherry; sprout; sushi mat; crab stick; onion ring; pestle; gin; mint; lemon grass; rubber; gherkin; breadcrumb; cinnamon; dumpling; rosemary; power; syrup; pineapple; sheets; soda; raspberry; airer; turkey; face; whiskey; kitchen door; cd; vanilla extract;

Shared Verbs:

take; put; wash; open; close; insert; turn on; cut; turn off; pour; mix; move; remove; throw; shake; scoop; adjust; squeeze; peel; press; turn; scrape; fill;apply; fold; break; pull; lift; hold; unroll; hang; sprinkle; spray; roll; search; stretch; knead; divide; sharpen; water; attach; wear; measure; unscrew; grate; screw; serve; uncover; lock; carry; mark;

Novel Verbs Ego4D:

arrange; blow; catch; clap; clean; climb; consume; count; cover; crochet; detach; dig; dip; draw; drill; drive; enter; feed; file; fry; give; grind; hit; inspect; iron; kick; knit; loosen; mold; operate; pack; paint; park; pet; plant; play; point; pump; push; read; repair; sand; scroll; sew; shuffle; sieve; sit; smooth; stand; step; stick; swing; talk; tie; tighten; tilt; touch; unfold; untie; walk; weld; wipe; write; zip; watch; wave;

Novel Verbs Epic-Kitchens:

dry; empty; flip; check; scrub; pat; eat; wrap; filter; look; sort; rip; cook; add; crush; set; feel; rub; soak; brush; drop; drink; slide; gather; turn down; coat; transition; increase; wait; lower; form; smell; use; let go; finish; stab; unwrap; choose; flatten; switch; season; unlock; prepare; bake; bend; unfreeze;