Learning Fine-grained View-Invariant Representations from Unpaired Ego-Exo Videos via Temporal Alignment

Zihui (Sherry) Xue[1,2] Kristen Grauman[1,2]

[1]UT Austin [2]FAIR, Meta

TEXAS The University of Texas at Austin | Meta | NEURAL INFORMATION PROCESSING SYSTEMS

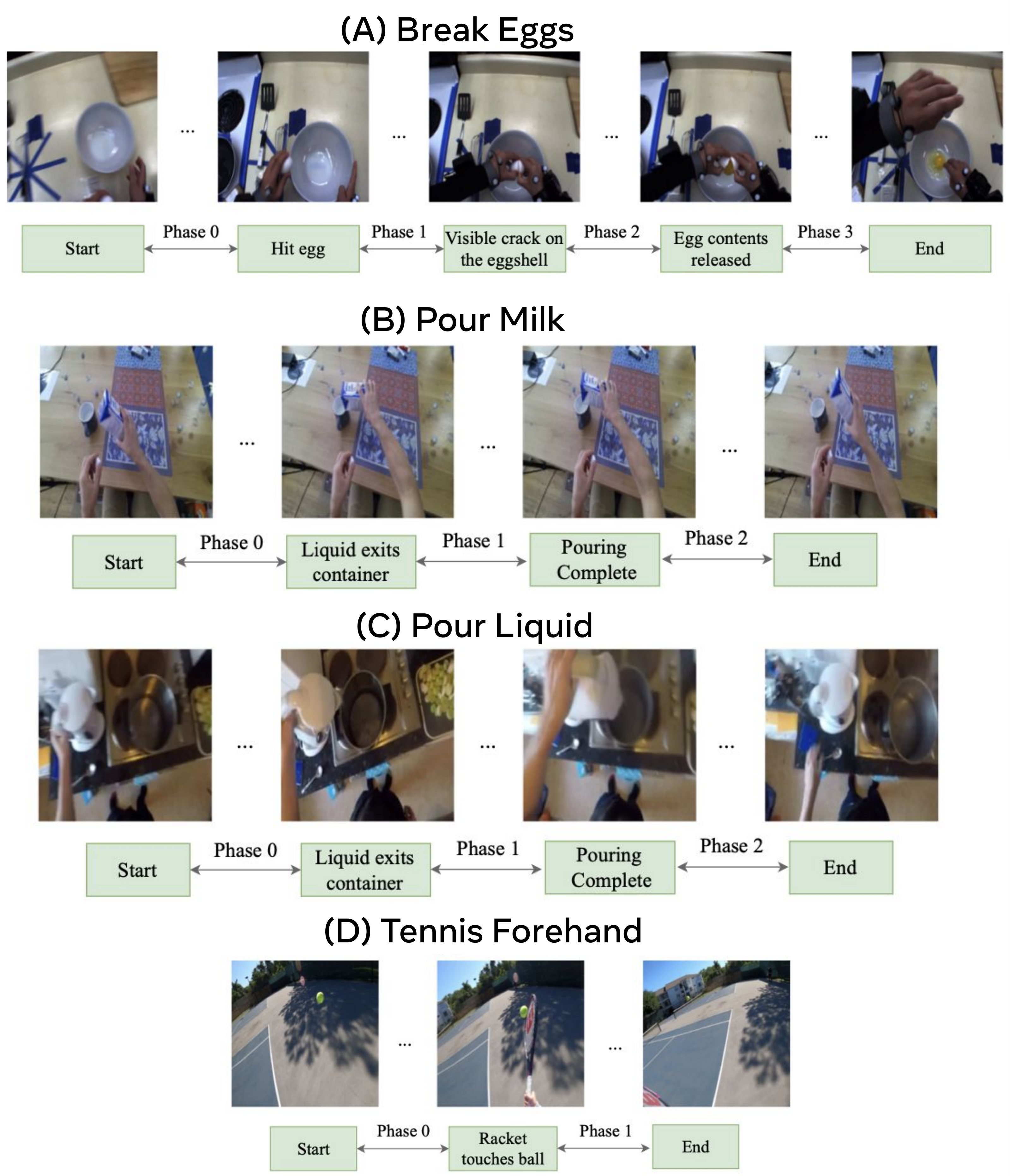
see our website for data, code & qualitative videos →

Motivation

How to bridge the **ego**centric (first-person) and **exo**centric (third-person) viewpoint gap in fine-grained activity understanding?

Ego-Exo Benchmark

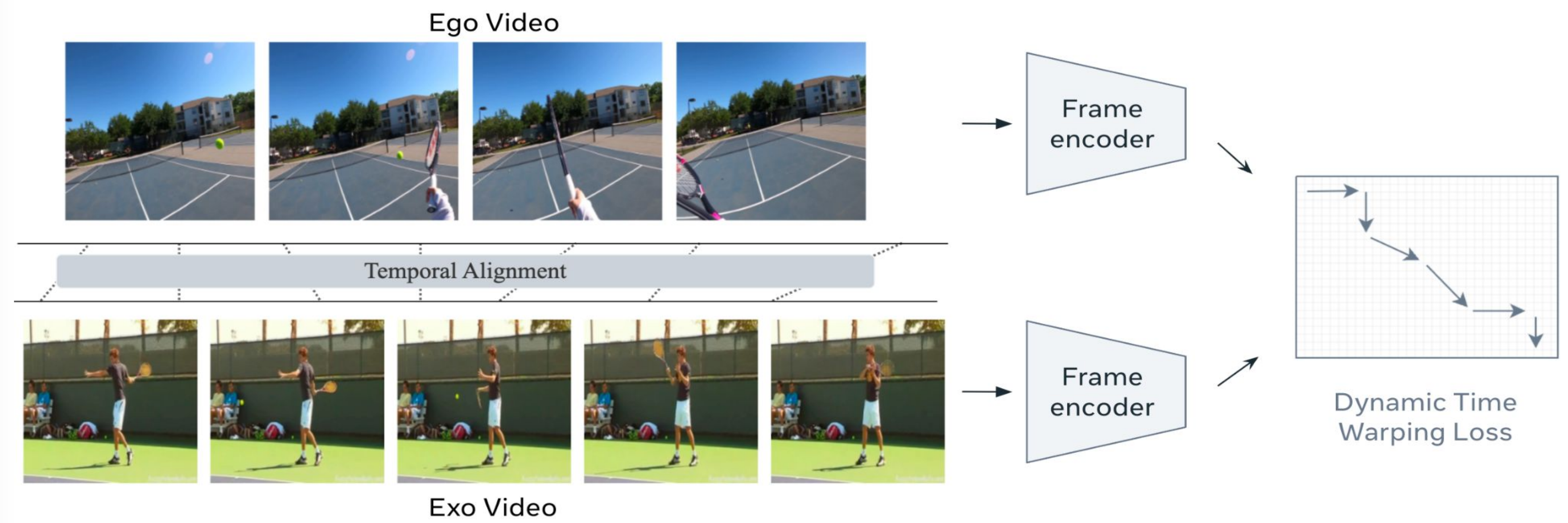
We establish the first ego-exo benchmark for **fine-grained action understanding**.



The benchmark consists of:

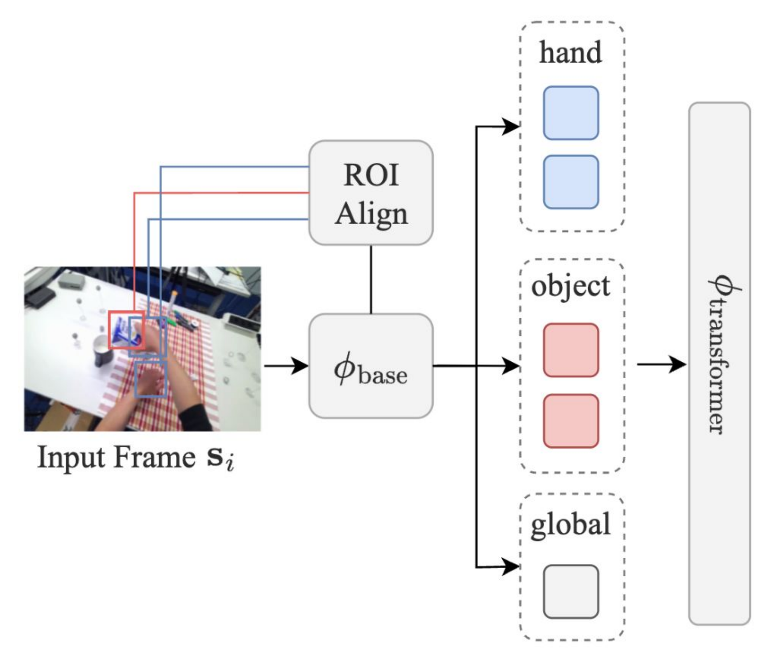
- Four action-specific datasets (videos sourced from five public datasets and an ego tennis dataset we collected)
- Per-frame annotations for every video in the datasets

AE2 overview



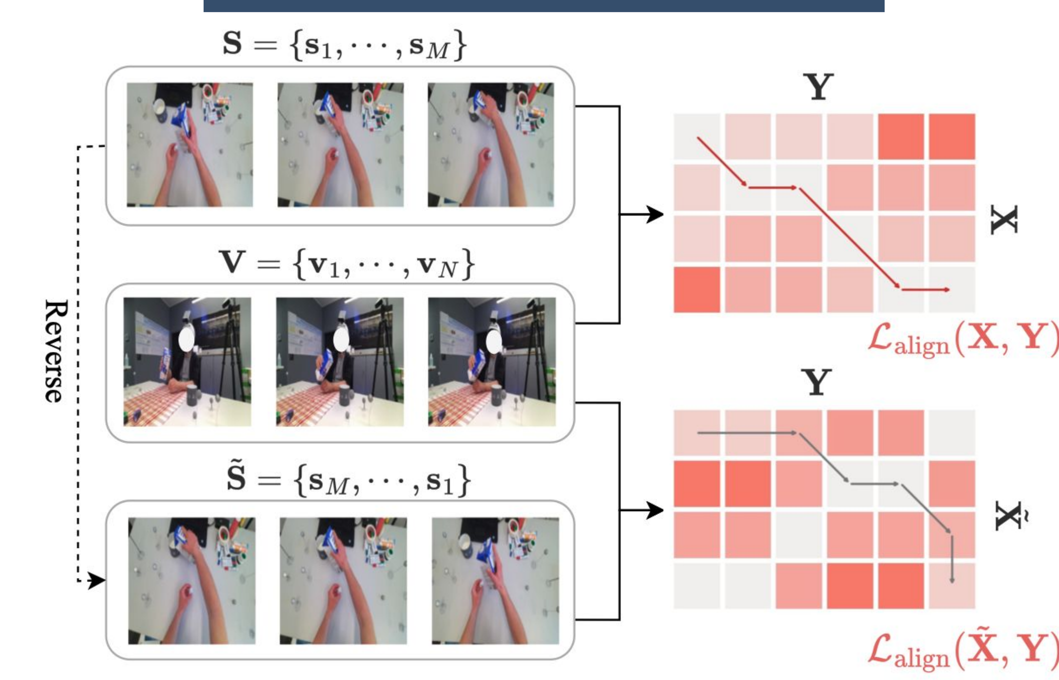
- We propose **AE2 (AlignEgoExo)**, a self-supervised approach for learning fine-grained action features that are invariant to the ego and exo viewpoints, by temporally aligning ego and exo videos of the same action.
- Prior works on view-invariant learning [1,2] rely on **synchronized** multi-view videos for training → AE2 only requires **unpaired** ego and exo videos.

an object-centric encoder



- Integrate regional features on hands and active objects to better bridge the ego-exo gap

a contrastive regularizer



- Enforce the alignment cost of a video pair to be smaller than aligning the same pair when one is played in reverse

Results

AE2 demonstrates superior performance consistently across datasets and downstream tasks, in both regular and cross-view scenarios.

Data set	Method	Classification (F1 score)			Retrieval (mAP@10)			Phase prog.
		regular	ego2exo	exo2ego	regular	ego2exo	exo2ego	
(A)	Prior Best	59.9	54.2	58.4	61.6	61.1	62.0	0.346
	AE2 (ours)	**66.2**	**57.4**	**71.7**	**65.9**	**64.6**	**62.2**	**0.511**
(B)	Prior Best	81.1	74.9	81.5	81.0	75.3	80.3	0.709
	AE2 (ours)	**85.2**	**84.7**	**82.8**	**84.9**	**78.5**	**83.4**	**0.763**
(C)	Prior Best	56.9	47.5	60.0	62.8	58.5	**57.9**	0.116
	AE2 (ours)	**66.6**	**57.2**	**65.6**	**65.5**	**65.8**	57.4	**0.138**
(D)	Prior Best	83.6	82.9	81.8	85.2	78.0	79.1	0.469
	AE2 (ours)	**85.9**	**84.7**	**85.7**	**86.8**	**81.5**	**82.1**	**0.506**

Baselines for comparison:

[1] Sermanet et al.,Time-contrastive networks: Self-supervised learning from video, ICRA 18.

[2] Sigurdsson et al., Actor and observer: Joint modeling of first and third-person videos, CVPR 18.

[3] Dwibedi et al., Temporal cycle-consistency learning, CVPR 19.

[4] Hadji et al., Representation learning via global temporal alignment and cycle-consistency, CVPR 21.

[5] Chen et al., Frame-wise action representations for long videos via sequence contrastive learning, CVPR 22.

Qualitative Results